Learning from Large-Scale Online Images

Jianping Fan
Department of Computer Science
UNC-Charlotte

Course Website:
http://webpages.uncc.edu/jfan/itcs5152.html

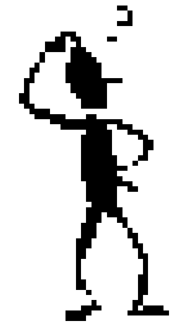
Two Huge Image Sources

- Social images such as Flickr images

- Web images such as Google images

How to harvest both social images and web images for computer vision tasks?

1. Research Motivation



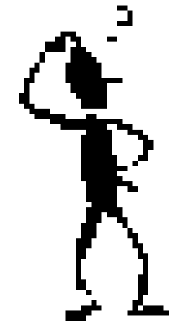
- **Computer Vision Tasks: Why we need large-scale labeled training images?**

  » **Number of objects and concepts could be large;**

  » **Learning complexity for some objects and concepts could be very high!**

Labeling large-scale training images is label-intensive!

1. Research Motivation



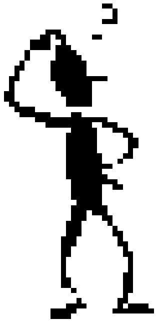
- **What Collaboratively-Tagged Images can do for us?**

  - **They are sufficient to characterize the diverse visual properties of large amounts of objects and concepts;**

  - **They can obtained easily by leveraging the collaborative efforts of Internet users.**

Why not using collaboratively-tagged images for classifier training?

1. Research Motivation

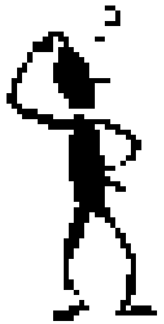


- **What are the problems of collaboratively-tagged images?**

    - **Spam tags & junk images;**
    - **Synonymous & Ambiguous tags;**
    - **Loose tags;**

We call such collaboratively-tagged images as ***weakly-tagged images***!

1. Research Motivation



- **What is the problem of classifier training algorithms?**

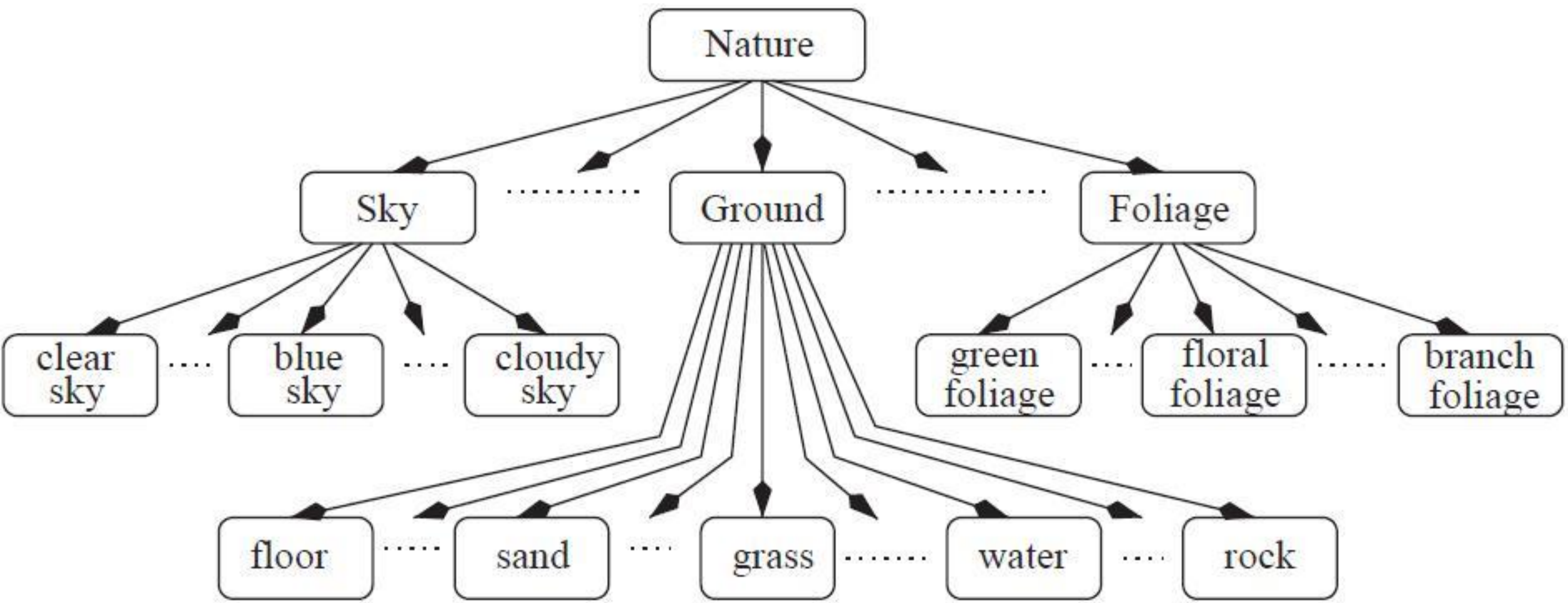
  - **Inter-Task Correlation Exploration;**

  - **Scalability with the number of objects and image concepts;**

  - **Discrimination power for visually-similar objects and image concepts.**

2. Image Crawling

- **Flickr Images & Others**
  - **Keywords for image crawling**



2. Image Crawling

- **Flickr Images & Others**

  - **Images for each keyword**

    ---5000 images and their tags and comments;

    - **Some of these 5000 images are junks;**
    - **Some of these 1000 keywords are synonymous or ambiguous**

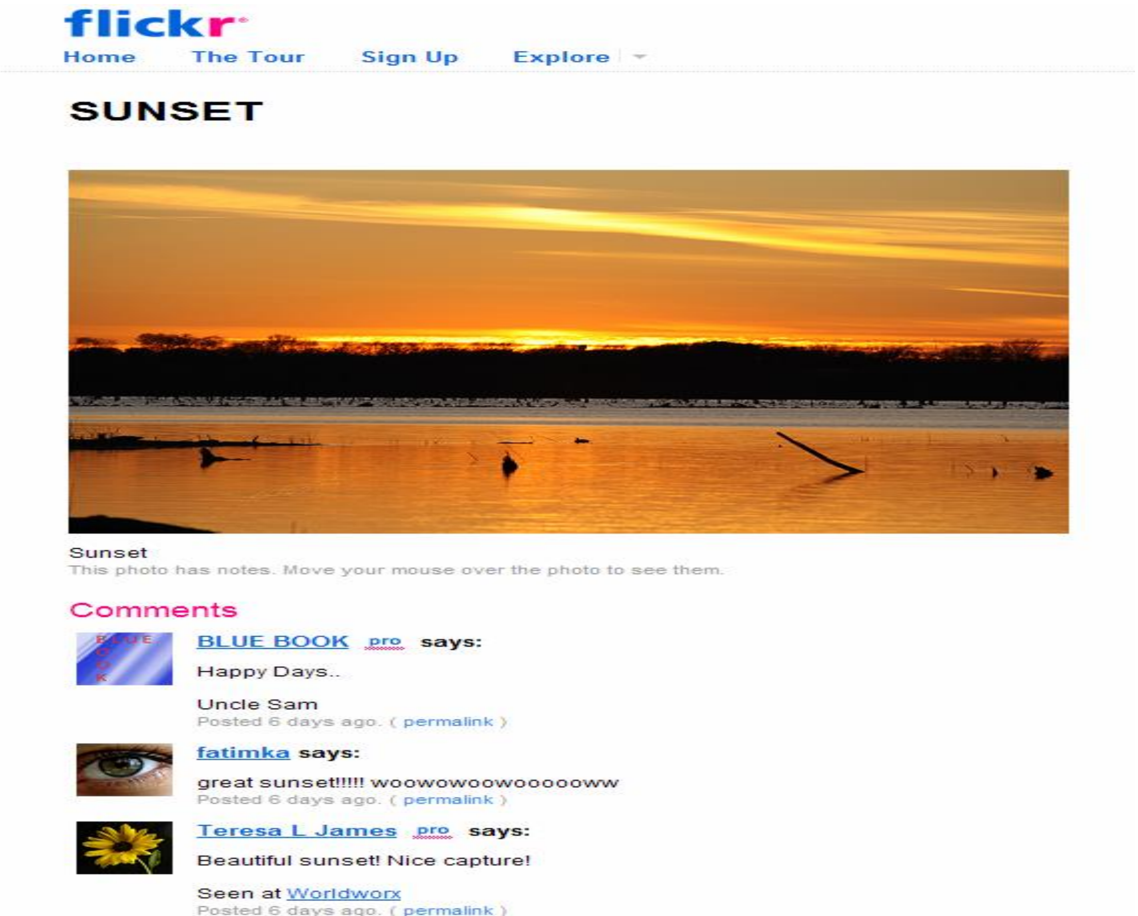
These weakly-tagged images cannot directly be used for classifier training!

- **Image & Tag Cleansing;**

- **Classifier Training with noisy images**

2. Image Crawling

- **Multiple Information Sources**



3. Image Content Representation

- **Multi-Resolution Image Grids**

- **Computational cost for feature extraction**
- **Discrimination power of visual features for classifier training**

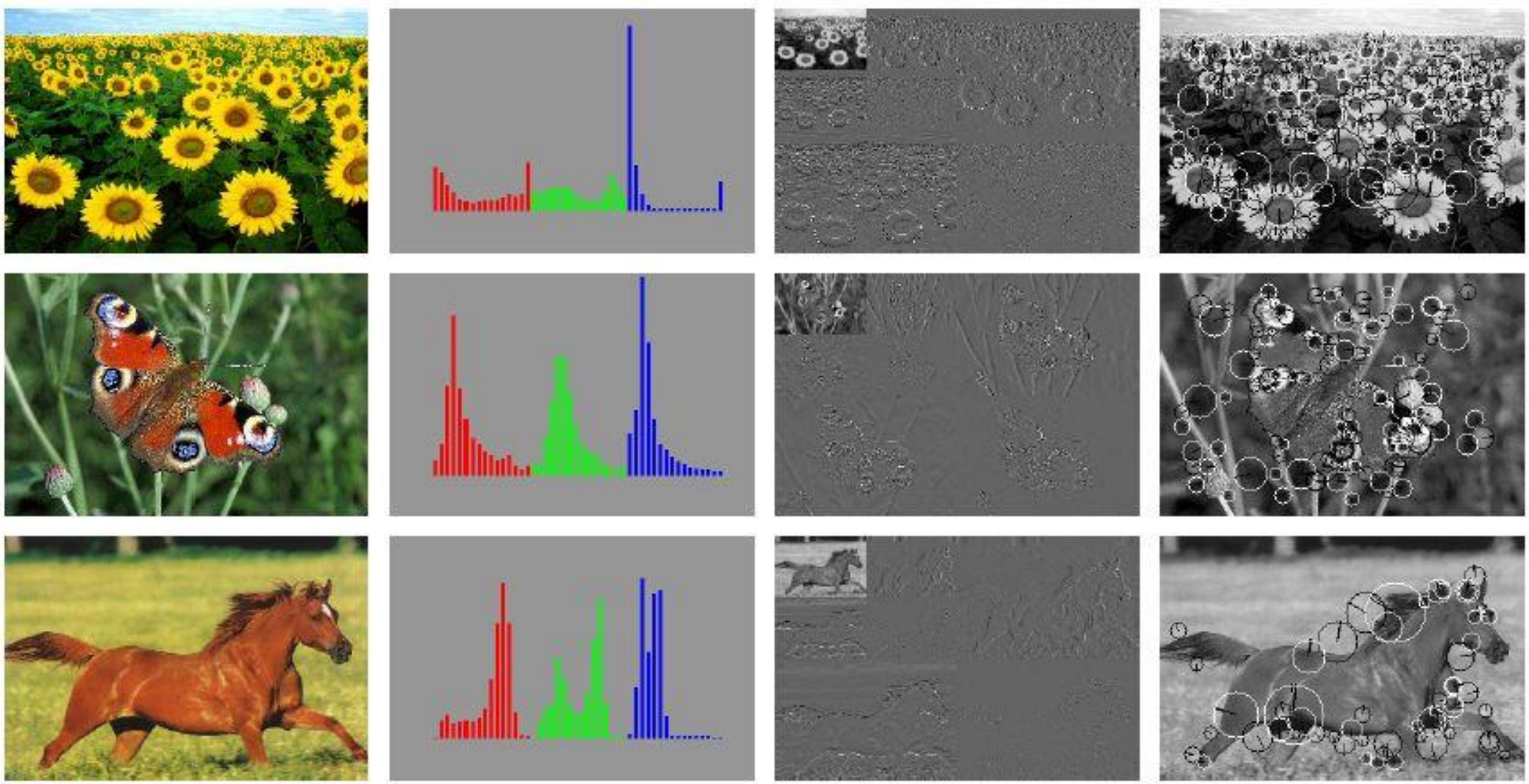
3. Image Content Representation

- **Multi-Resolution Image Grids**

- **Object information characterization at certain accuracy;**
- **Good trade-off between computational cost and accuracy.**

4. Image Similarity Characterization

- **Multi-Modal Visual Features & Mixture-of-Kernels**



4. Image Similarity Characterization

- **Mixture-of-Kernels**

$$\kappa(x, y) = \sum_{l=1}^{\tau} \alpha_l \kappa_l(x, y), \qquad \sum_{l=1}^{\tau} \alpha_l = 1$$

- **Image distributions under different feature subsets may have different statistical properties**

- **One kernel cannot handle such diversity!**

5. Junk Image Filtering

- **Positive Comments *vs.* Negative Comments**

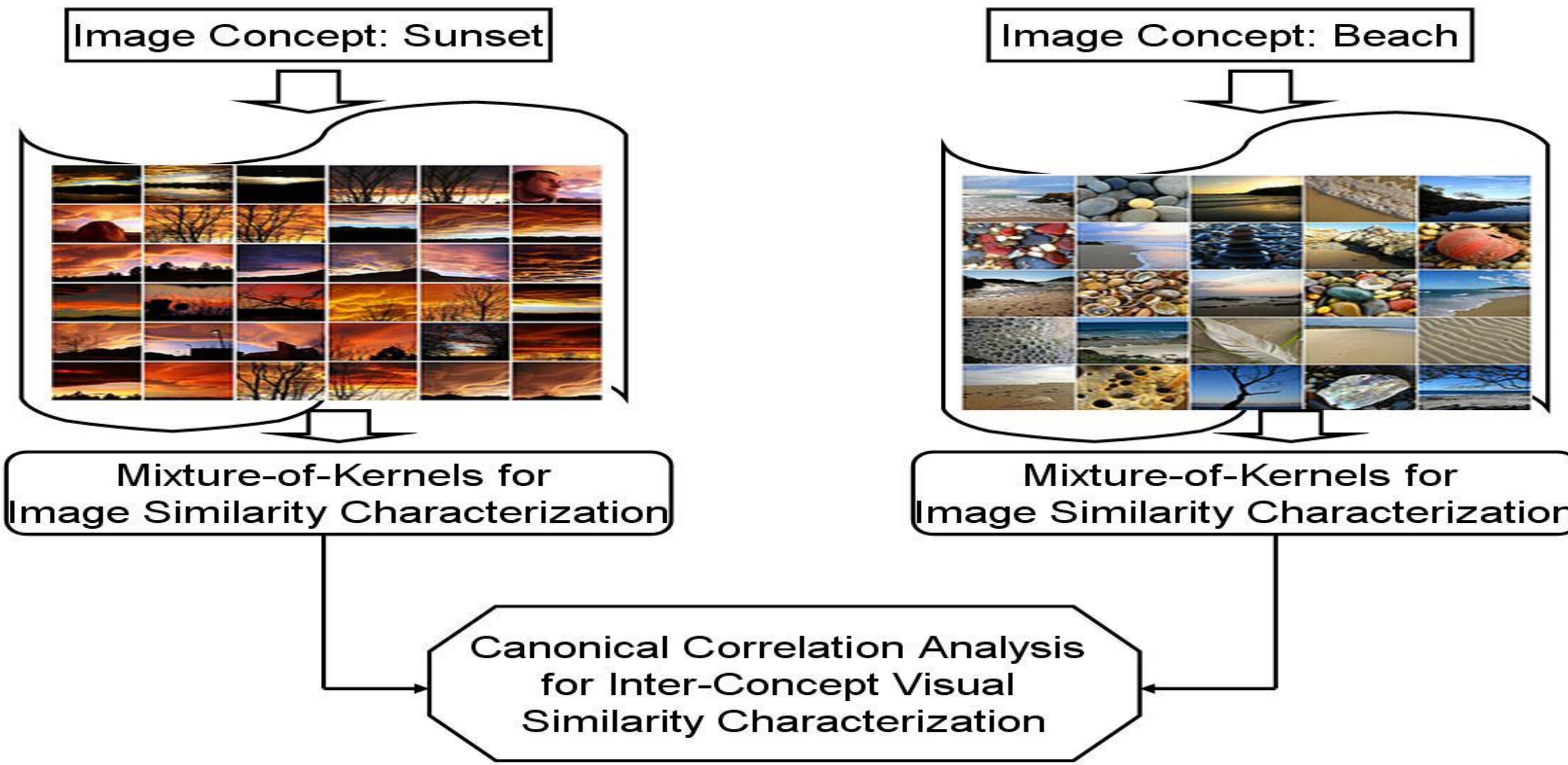
$$PMI(C,\Phi) = \sum_{i=1}^{N} \log \frac{P(C \cap P\_word_i)}{P(C)P(P\_word_i)}$$

$$PMI(C,\Psi) = \sum_{i=1}^{N} \log \frac{P(C \cap N\_word_i)}{P(C)P(N\_word_i)}$$

$$PMI(C) = PMI(C,\Phi) - PMI(C,\Psi)$$

6. Visual Concept Network

- **Kernel Canonical Correlation Analysis**



6. Visual Concept Network

- **Kernel Canonical Correlation Analysis**

$$\gamma(C_i, C_j) = \max_{\theta, \vartheta} \frac{\theta^T \kappa(S_i)\kappa(S_j)\vartheta}{\sqrt{\theta^T \kappa^2(S_i)\theta \cdot \vartheta^T \kappa^2(S_j)\vartheta}}$$

$$\kappa(S_i) = \sum_{x_l, x_m \in S_i} \kappa(x_l, x_m) \quad \kappa(S_j) = \sum_{x_h, x_k \in S_j} \kappa(x_h, x_k)$$

$$\kappa(S_i)\kappa(S_i)\theta - \lambda_\theta^2 \kappa(S_i)\kappa(S_i)\theta = 0$$

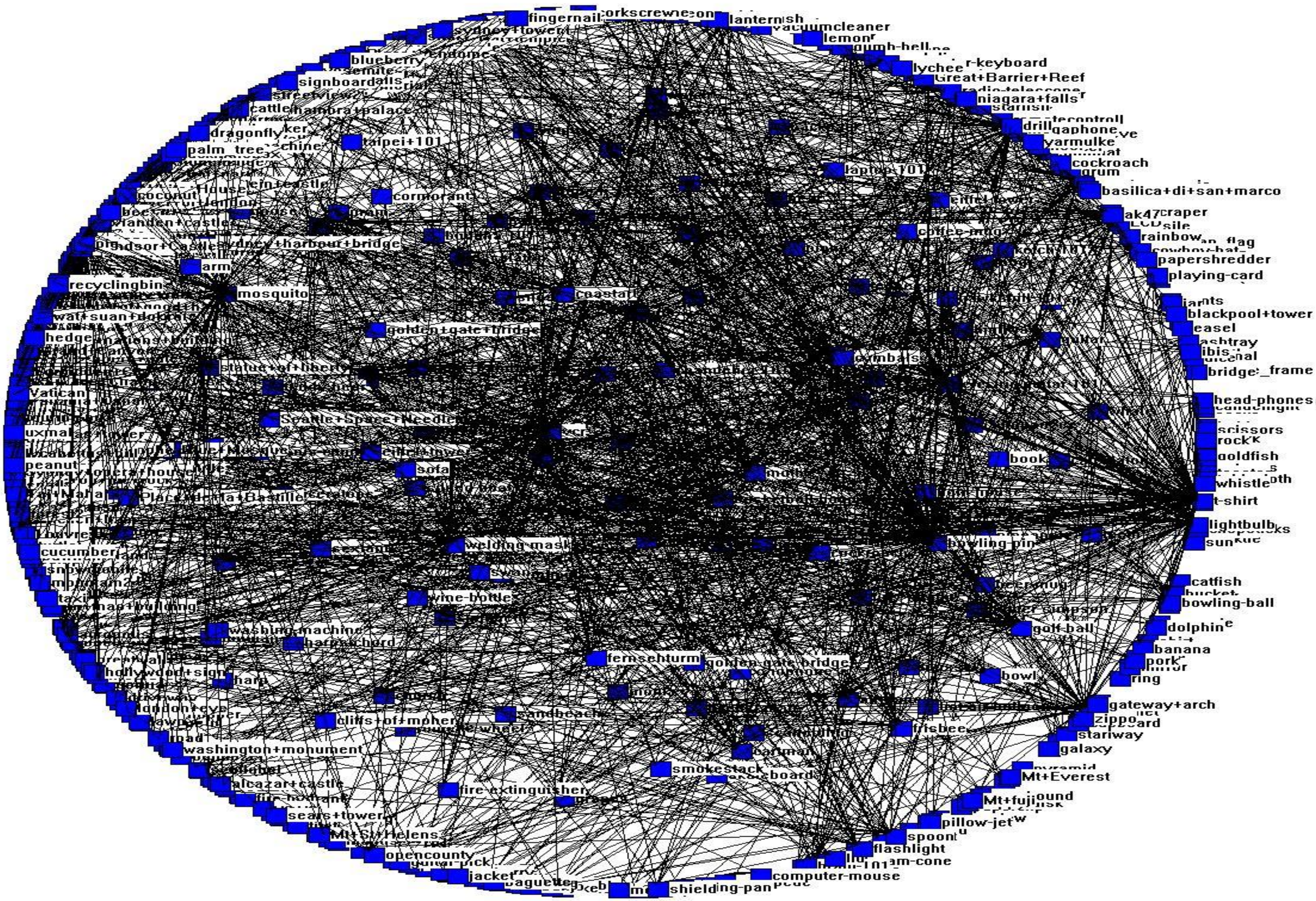
$$\kappa(S_j)\kappa(S_j)\vartheta - \lambda_\vartheta^2 \kappa(S_j)\kappa(S_j)\vartheta = 0$$

6. Visual Concept Network

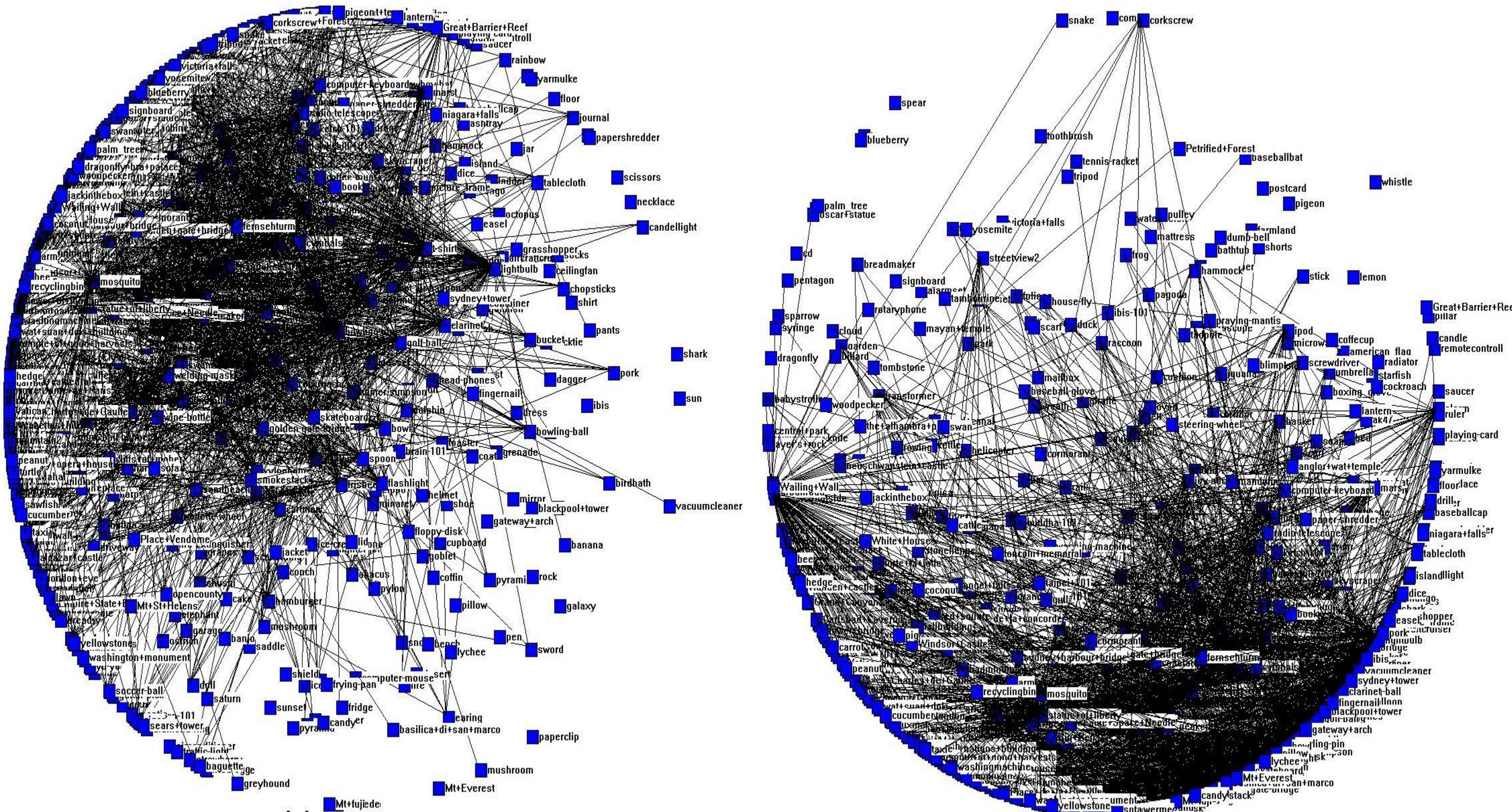
- **Kernel Canonical Correlation Analysis**

concept pair	γ	concept pair	γ	concept pair	γ	concept pair	γ
urbanroad-streetview	0.99	cat-dog	0.81	kerb-saucer	0.28	tweezer-corn	0.19
frisbee-pizza	0.80	dolphin-cruiser	0.73	fridge-vest	0.29	journal-grape	0.19
moped-bus	0.75	habor-outview	0.71	stick-cupboard	0.29	sheep-greatwall	0.26
monkey-humanface	0.71	guitar-violin	0.71	mushroom-moon	0.32	whistle-watermelon	0.28
lightbulb-firework	0.69	mango-broccoli	0.69	cannon-ruler	0.41	snake-ipod	0.31
porcupine-lion	0.68	bridge-warship	0.68	tombstone-crab	0.42	helicopter-city	0.63
doorway-street	0.65	statue-building	0.68	pylon-highway	0.61	LCD-container	0.65
windmill-bigben	0.63	cat-lion	0.66	beermug-bar	0.62	sailboat-cruiser	0.66

6. Visual Concept Network



6. Visual Concept Network

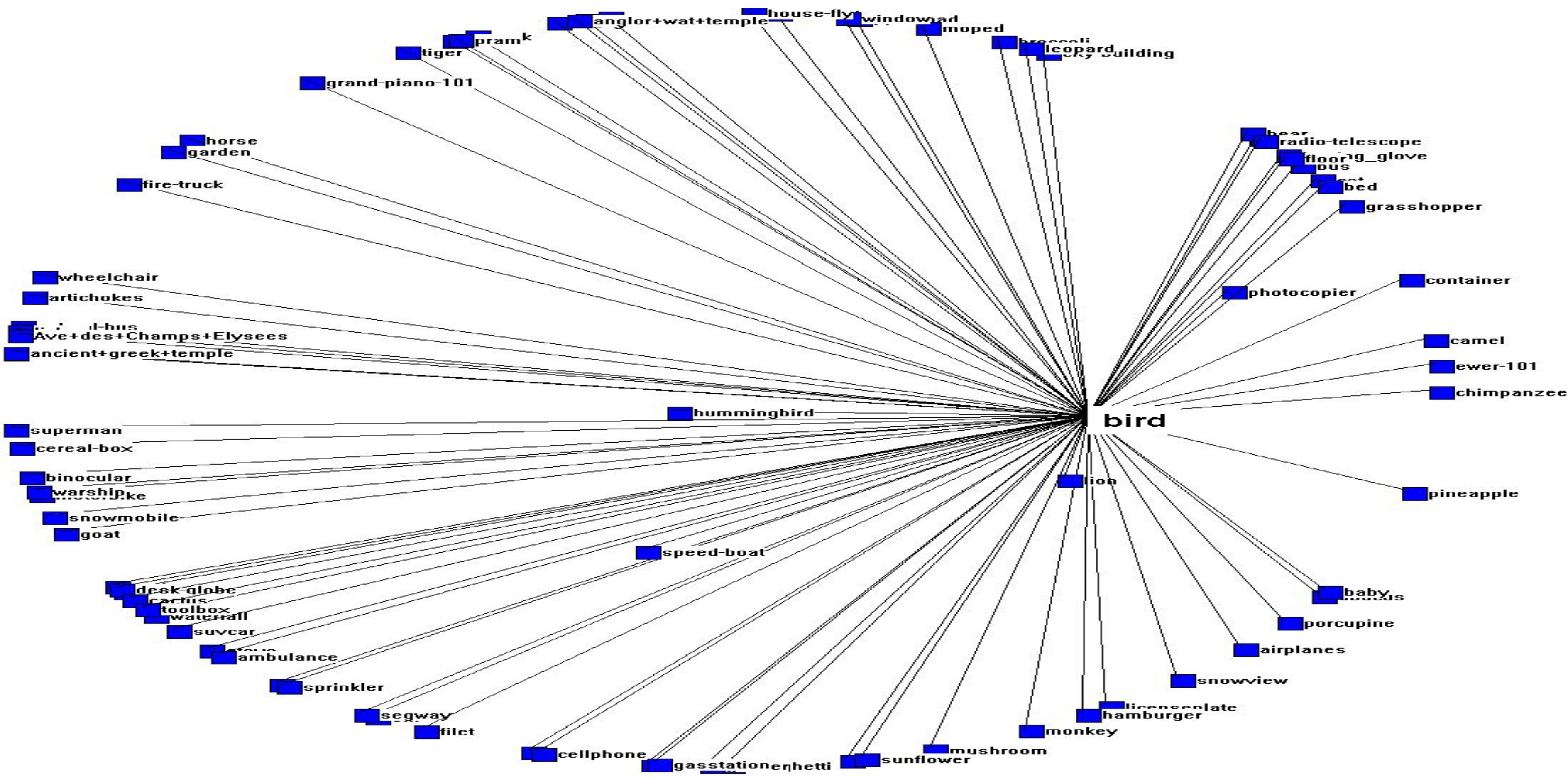


(a)

(b)

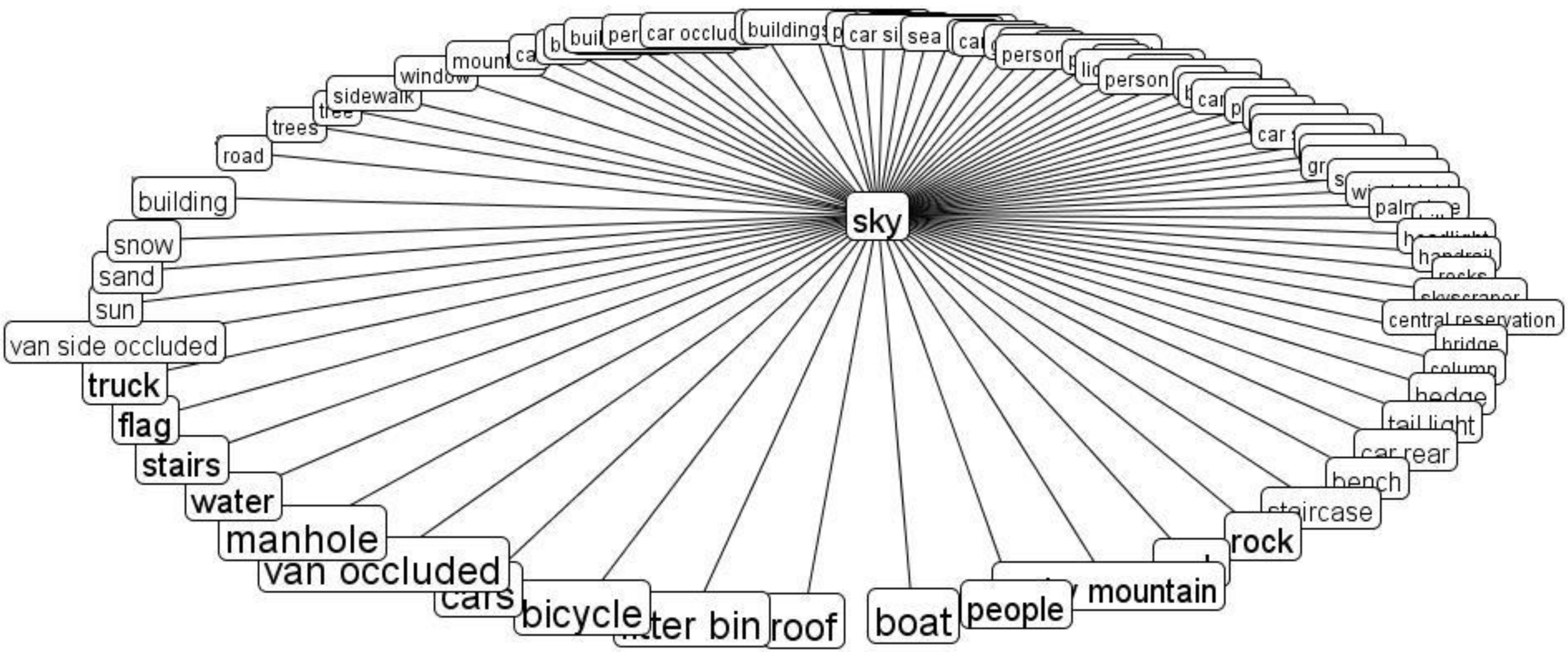
6. Visual Concept Network

- **First-Order Nearest Neighbors**



6. Visual Concept Network

- **First-Order Nearest Neighbors**



6. Visual Concept Network

- **Why we need a visual concept network?**
  - **Inter-Related Learning Tasks, e.g., inter-related objects and concepts;**
  - **Discrimination power of classifiers, e.g., if our classifiers can identify the visually-similar objects and concepts, they will have better discrimination power.**

7. Cross-Modal Tag Cleansing

- **Synonymous Tags: Visual Similarity**

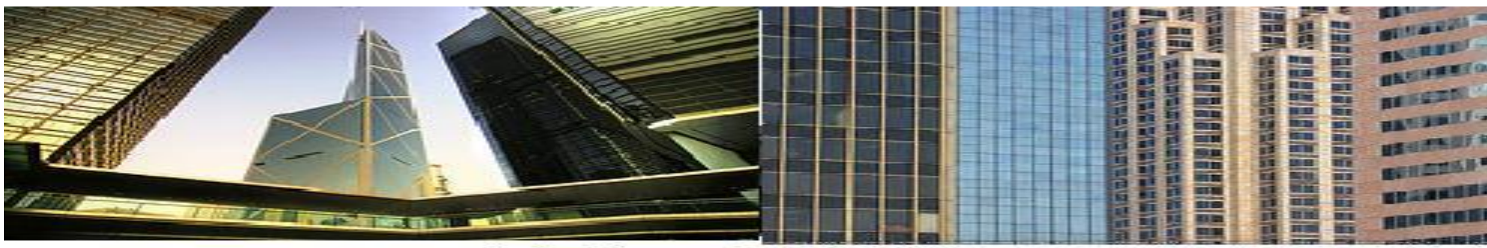
(a) Auto

(b) Automobile

(c) Car

7. Cross-Modal Tag Cleansing

- **Ambiguous Tags: Visual Diversity**



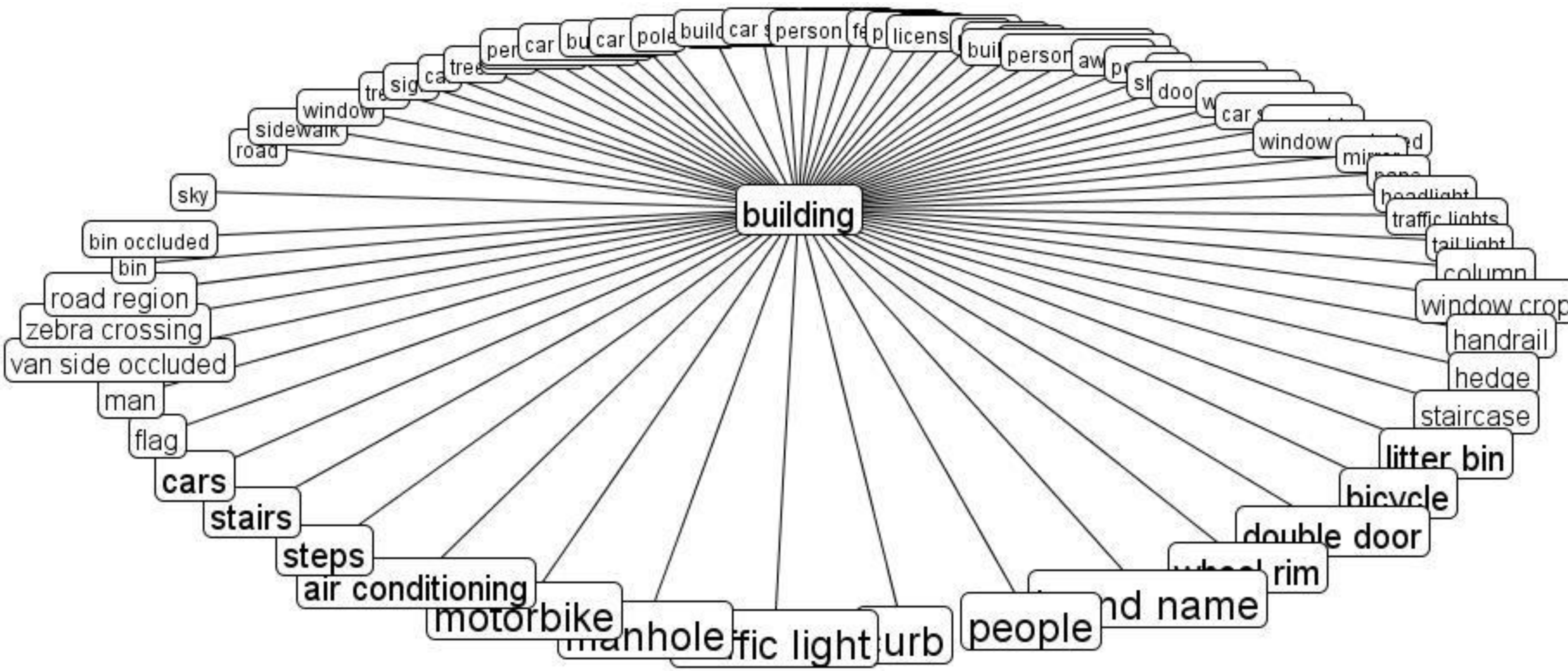
(a) Bank Office

(b) River Bank

(c) Cloud Bank

8. Inter-Related Classifier Training

- **Which Object and Concepts are correlated?**



Our visual concept network can provide a good environment for this task!

8. Inter-Related Classifier Training

- **How to model such inter-concept correlation?**

  ----Structured Max-Margin Networking

  - **Support Vector Machine (SVM)**

    ----It is able to handle high-dimensional issue effectively, but it cannot model the inter-related structure!

  - **Graphical Models such as CRF**

    ----It is able to model the inter-related structure effectively, but it cannot handle high-dimensional issue!

Our learning situation is both high-dimension and correlation structure!

8. Inter-Related Classifier Training

- **How to model such inter-concept correlation?**

$$P(C_j, X) = \frac{1}{Z} exp \left(\sum_{C_j \in \Xi_j} f(C_j, X) + \sum_{C_j \in \Xi_j} \sum_{C_i \in \Xi_i} f(C_j, C_i, X) \right)$$

$$Z = \sum_{j=1}^{T} exp \left(\sum_{C_j \in \Xi_j} f(C_j, X) + \sum_{C_j \in \Xi_j} \sum_{C_i \in \Xi_i} f(C_j, C_i, X) \right)$$

8. Inter-Related Classifier Training

- **How to model such inter-concept correlation?**

$$P(C_j|X) \propto P(C_j, X) \propto exp\left(\sum_{C_j \in \Xi_j} f(C_j, X) + \sum_{C_j \in \Xi_j} \sum_{C_i \in \Xi_i} f(C_j, C_i, X)\right)$$

$$H_{C_j}(X) = argmax\left(\sum_{C_j \in \Xi_j} f(C_j, X) + \sum_{C_j \in \Xi_j} \sum_{C_i \in \Xi_i} f(C_j, C_i, X)\right)$$

8. Inter-Related Classifier Training

- **How to model such inter-concept correlation?**

$$f(C_j, X) = sign\left(\sum_{l=1}^{N}\sum_{m=1}^{\tau} \beta_{lj} Y_{lj} \alpha_m \kappa_m(X_{lj}, X) + b\right)$$

$$f(C_j, C_i, X) = sign\left(\sum_{j=1}^{M}\sum_{l=1}^{N}\sum_{m=1}^{\tau} \hat{\beta}_{lj} Y_{lj} \hat{\alpha}_m \kappa_m(X_{lj}, X) + b\right)$$

8. Inter-Related Classifier Training

- **How to model such inter-concept correlation?**

$$\min_{\beta} \max_{\alpha} \sum_{r=1}^{\tau} \alpha_l \Psi(r) + \min_{\hat{\beta}} \max_{\hat{\alpha}} \sum_{r=1}^{\tau} \hat{\alpha}_l \Phi(r)$$

Subject to:

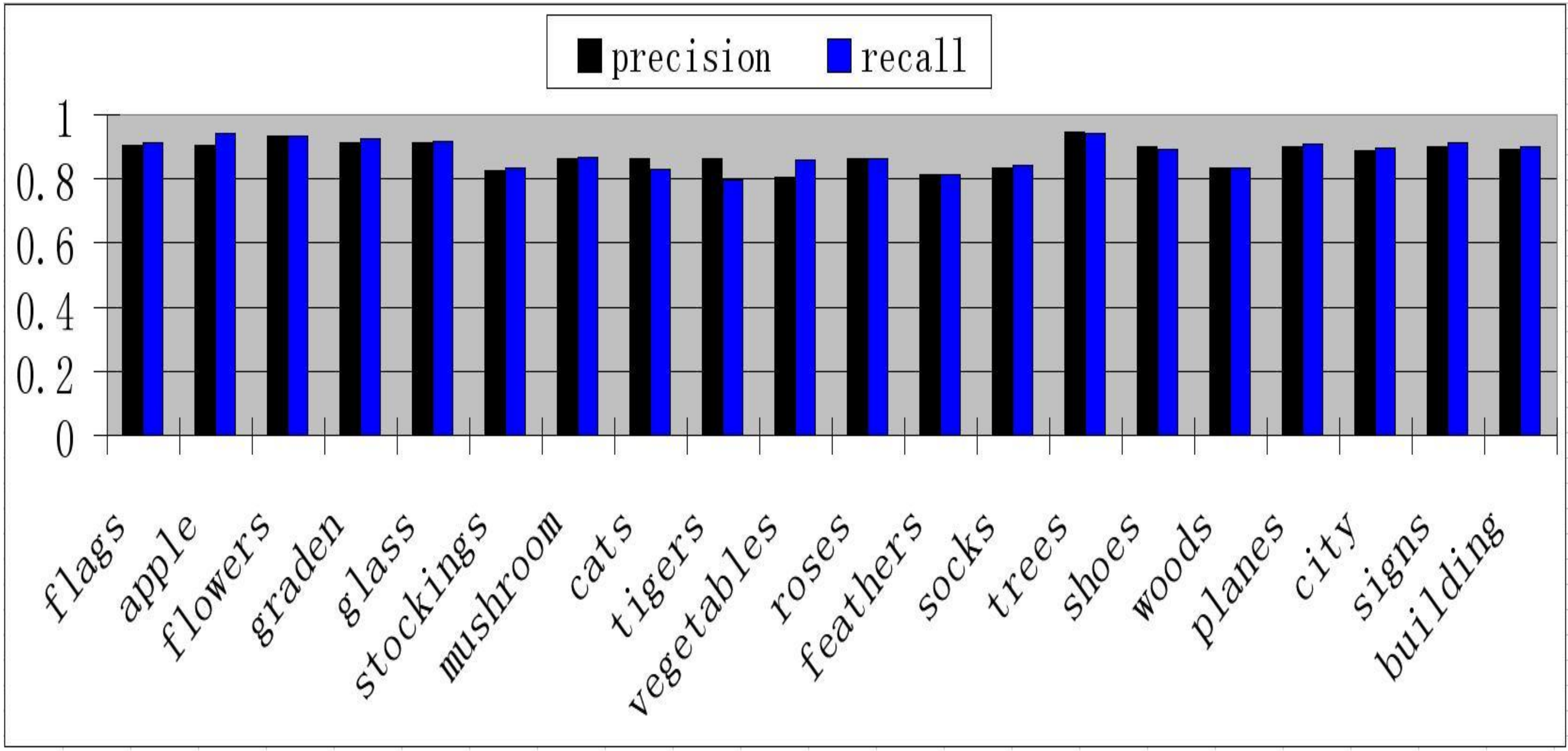
$$\forall_{l=1}^{N}: \quad 0 \le \beta_l \le \lambda, \quad \sum_{l=1}^{N} \beta_l Y_l = 0; \quad \forall_{r=1}^{\tau}: \quad \alpha_r \ge 0, \quad \sum_{r=1}^{\tau} \alpha_r = 1$$

$$\forall_{i=1}^{N} \ \forall_{j=1}^{M}: \quad 0 \le \hat{\beta}_{ij} \le \frac{M}{2\lambda}, \quad \sum_{j=1}^{M} \sum_{i=1}^{N} \hat{\beta}_{ij} Y_{ij} = 0; \quad \forall_{r=1}^{\tau}: \quad \hat{\alpha}_r \ge 0, \quad \sum_{r=1}^{\tau} \hat{\alpha}_r = 1$$

$$\Psi(r) = \sum_{l,m=1}^{N} \beta_l \beta_m Y_l Y_m \kappa_r(X_l, X_m) - \sum_{l=1}^{N} \beta_l \qquad \Phi(r) = \sum_{j=1}^{M} \sum_{i=1}^{N} \sum_{h=1}^{M} \sum_{l=1}^{N} \hat{\beta}_{ih} Y_{ih} \hat{\beta}_{jl} Y_{jl} \kappa_r(X_{ih}, X_{jl}) - \sum_{j=1}^{M} \sum_{i=1}^{N} \hat{\beta}_{ij}$$

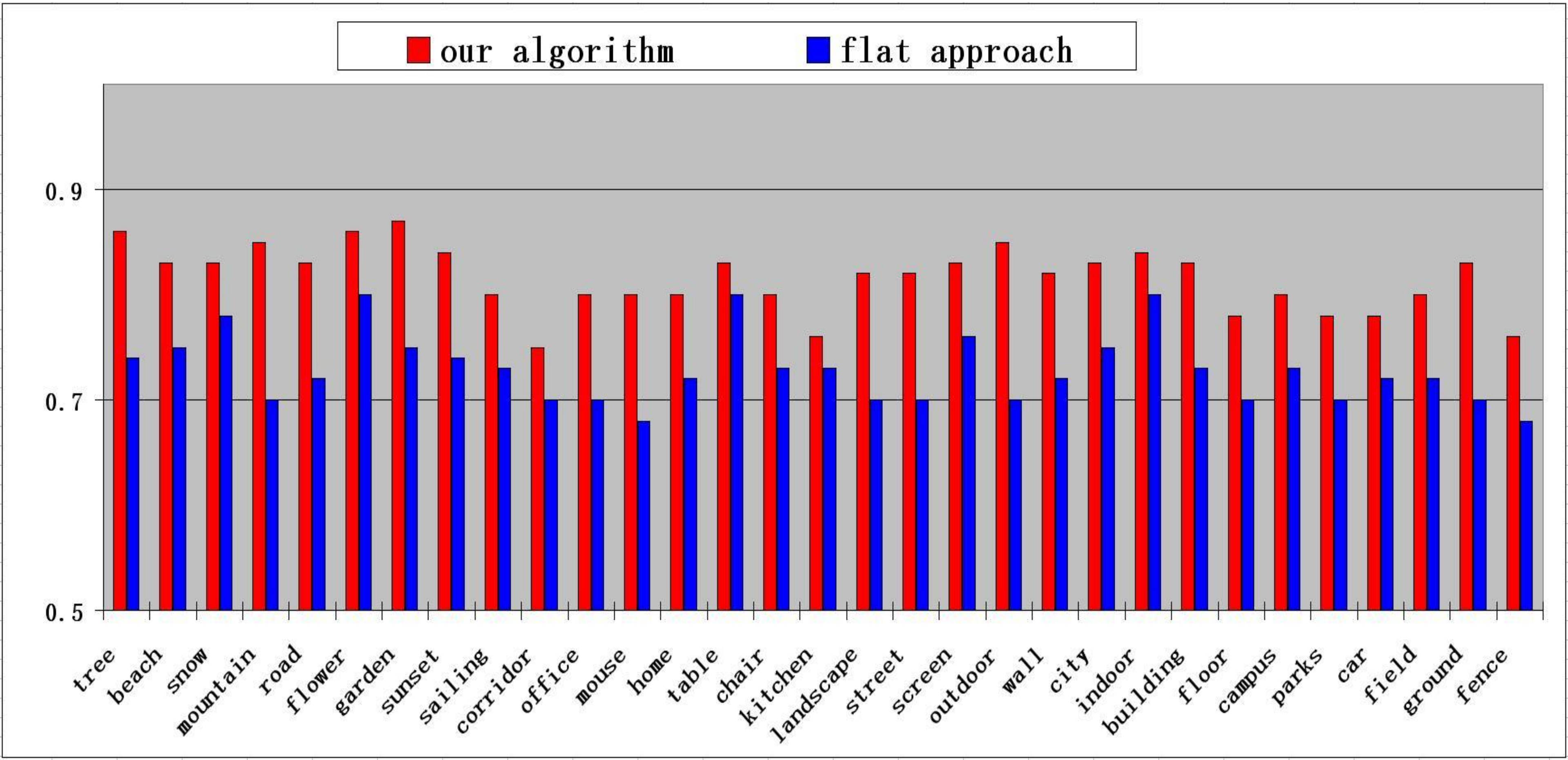
9. Algorithm Evaluation

- **Junk Image Filtering**



9. Algorithm Evaluation

- **Inter-related Classifier Training**



9. Algorithm Evaluation

- **Computational Cost for classifier training**
  - **Our Algorithm** $O(\hat{M} \times T) \cdot O(\tau N^3)$
  - **GentleBoosting** $O(T^2) \cdot O(\widetilde{N}^3)$
- **Computational Cost for image classification**
  - **Our Algorithm** $O(\hat{M} + T)$
  - **GentleBoosting** $O(T^2)$

Web Image Indexing

- Research Motivation
- Image and Auxiliary Text Extraction
  - Image-Block Generation
  - Image Clustering
- Automatic Image-Text Alignment
  - Term-Image Relevance Estimation
  - Term Correlation Network
  - Relevance Refinement
- Evaluation

Research Motivation

- Leveraging large-scale web images with reliable labels for vision tasks
  - Most modern web-pages are composed by Images and auxiliary texts
  - Image labels can be learned from the auxiliary texts

- Challenges
  - Most of text terms are weakly related or even irrelevant to the semantics of the web images in the same hosted webpage

Image-Text Alignment Framework

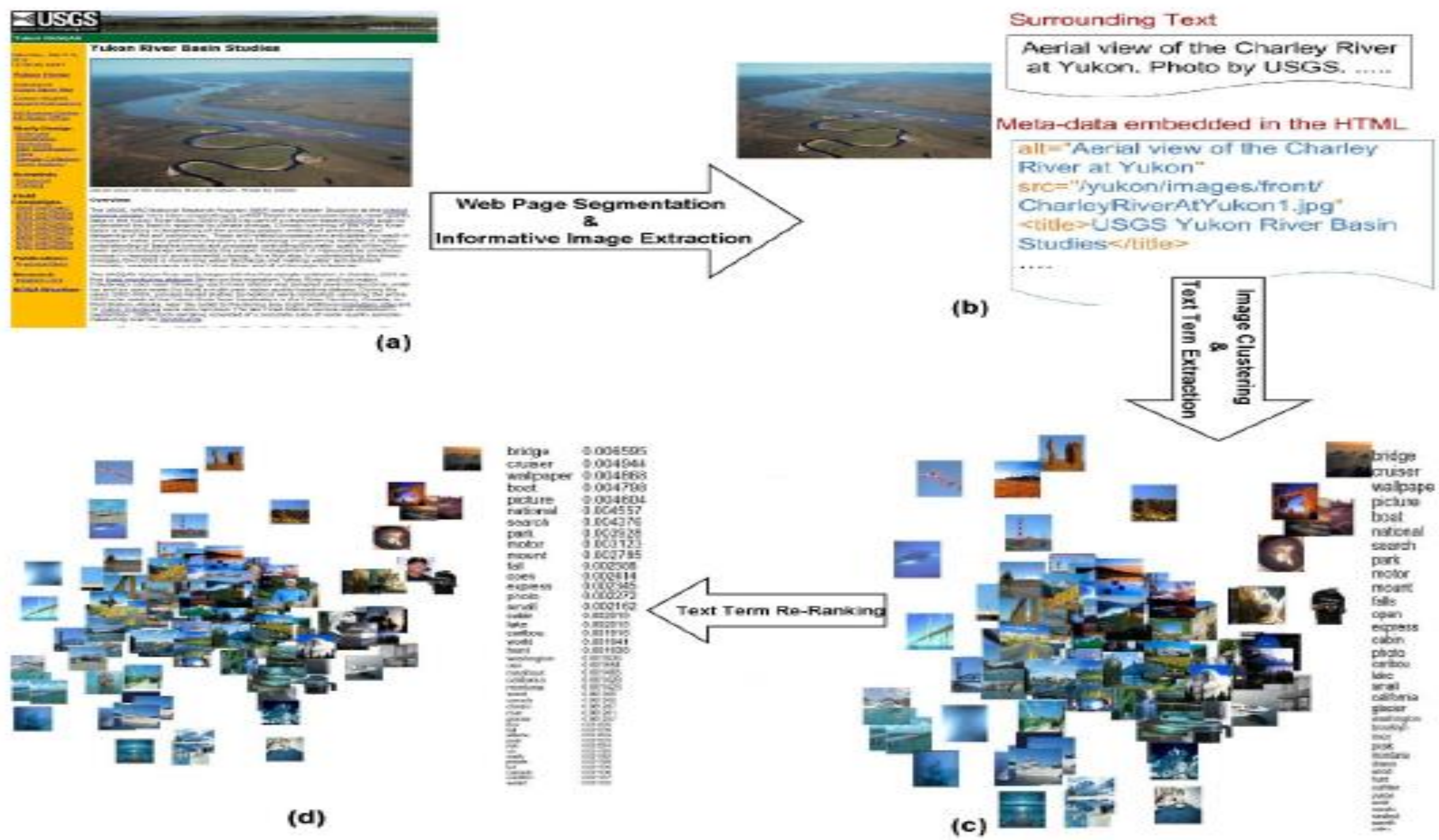
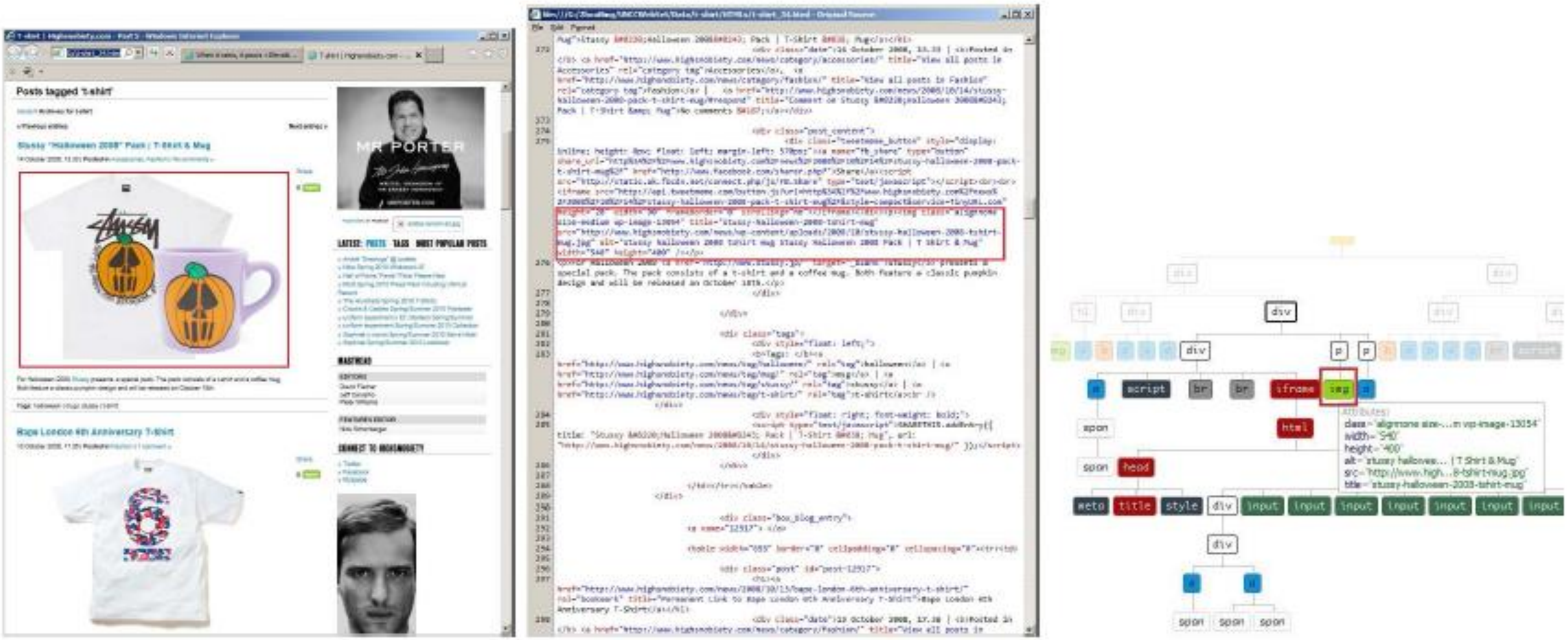


Figure 1: The illustration of the key components of our image-text alignment scheme: (a) web page; (b) image-block pair; (c) image cluster and ranked auxiliary text terms; (d) image cluster and re-ranked auxiliary text terms.

- **Text-Image Alignment for Web Image Indexing**



(a) A web page rendered by IE (b) The html document (c) The DOM-Tree

WWW2010, TPAMI2012

Image and Auxiliary Text Extraction

- Informative Image Extraction
  - Plenty of “noise” images: navigation menus, advertisement images, snippet previews,…
  - Still a open problem in the research community
- Method
  - Aspect ratio (>0.2 or <5)
  - Image size (min(width, height) > 60 pixel)
  - Not perfect but can produce satisfied results
  - Unsupervised and computationally efficient

Image and Auxiliary Text Extraction

- Auxiliary text extraction
  - The text content in a webpage is diverse and most of them are irrelevant to the images in the webpage
- Assumption: texts which are visually close to the web image are more likely to be related to the semantics of the image
- Webpage segmentation
  - Visually-based: precise but computationally expensive
  - DOM(Document Object Model) based: computationally efficient

Image and Auxiliary Text Extraction

- DOM-based region growing for most relevant text block(s) extraction
  - the corresponding image node in the DOM-tree is set as the start point
  - a upward growing search is performed until it reaches any text node
  - the inner texts embedded in the text node(s)are extracted as the text block(s)

Image and Auxiliary Text Extraction

- Meta data embedded i
  - Alternate text
  - Image titles
  - Image filename
  - Webpage title

Surrounding Text

Aerial view of the Charley River at Yukon. Photo by USGS. .....

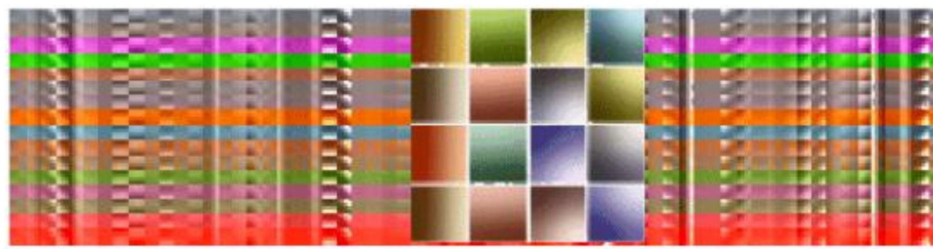
Meta-data embedded in the HTML source

alt="Aerial view of the Charley River at Yukon"
src="/yukon/images/front/CharleyRiverAtYukon1.jpg"
<title>USGS Yukon River Basin Studies</title>
....

Image Clustering

- Image as a bag of visual words

- Codebook



- Dista

$$d(\mathbf{x}_m, \mathbf{x}_n) = \sum_{\forall i} \frac{|\mathrm{ASPH}_m(i) - \mathrm{ASPH}_n(i)|}{1 + \mathrm{ASPH}_m(i) + \mathrm{ASPH}_n(i)} + \sum_{\forall j} \frac{|\mathrm{CSPH}_m(j) - \mathrm{CSPH}_n(j)|}{1 + \mathrm{CSPH}_m(j) + \mathrm{CSPH}_n(j)}. \quad (1)$$

Image Clustering

- Clustering method: Affinity propagation
- Image pair wise similarity is taken as the negative distance between these two images
- Text blocks belong to the same image cluster are merged as a single joint text document
- Text terms are extracted from this document using NLTK tool kit

Automatic Image-Text Alignment

- Term-Image relevance estimation

$$\rho(C,t) = \frac{\sum_{x\in\Theta(t)} P(x,t)}{\sum_{y\in\Theta}\sum_{r\in\mathcal{W}} P(y,r)}, \qquad (2)$$

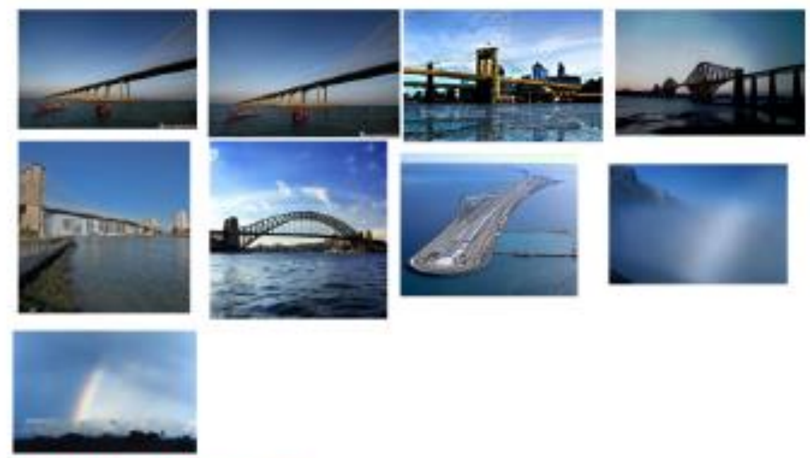
- Image clusters with ranked terms by relevance score

Auxiliary Terms: **portrait, mount rushmore, national monument, south dakota, arch, nature, canvas, joshua tree, california, mountain, mist, glacier, national park, montana, brooklyn bridge**

Automatic Image-Text Alignment

- Image clusters with ranked terms by relevance score

Auxiliary Terms: river, national park, zambia, pool, africa, age, news, tiger, brain, pond, united states, popularity, uniform, green, pan, vegetable, source, vitamin a, potassium, peak season, ...



Auxiliary Terms: brooklyn bridge, bridge, rainbow, island, completion, peak, sunset, traffic, gallery, photo, dawn, front, devil , tree, order, transfer, water, causeway, history, harbour, prince edward island, ...

Term Correlation Network

- Terms are not alone but inter-related
  - Multiple terms can have similar meaning
  - Some terms can have multiple senses under different context
  -
- Inter-term correlation characterization
  - Term co-occurrences
  - Semantic similarity from WordNet

Term Correlation Network

- Term co-occurrences
- Semantic

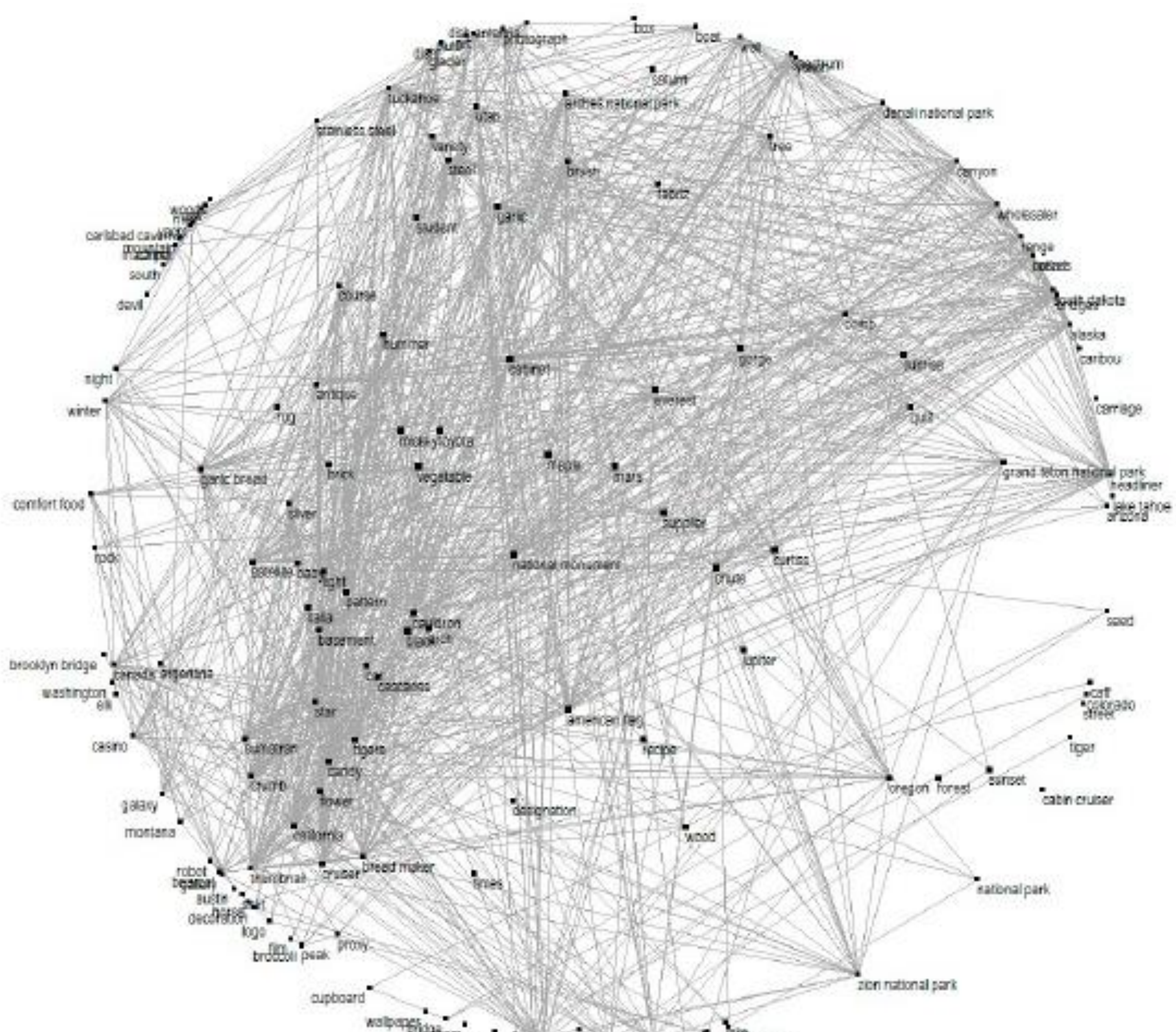
$$\beta(t_i, t_j) = -P(t_i, t_j) log \frac{P(t_i, t_j)}{P(t_i) + P(t_j)},$$

- Integrati

$$\gamma(t_i, t_j) = P(t_i, t_j) \cdot \log \frac{L(t_i, t_j)}{2 \cdot D}$$

$$\phi(t_i, t_j) = \alpha \cdot \gamma(t_i, t_j) + (1 - \alpha) \cdot \beta(t_i, t_j),$$

Visualization of the term correlation network



Relevance Refinement

- Random walk over term correlation network

- Transmission $$\phi_{ij} = \frac{\phi(i,j)}{\sum_k \phi(i,k)},$$

- Random walk process

$$\rho_k(t) = \theta \sum_{j \in \Omega_j} \rho_{k-1}(j)\phi_{tj} + (1-\theta)\rho(C,t),$$

Refinement example

(a)

(b) **Auxiliary Terms:** photo, tiger, friendship, oct, animal, photograph, tag, carrot, caribou, cat, meow, chimpanzee, wallpaper, format, wife, stab, source, bridge, inspiration, turnip, balloon, calla lily, kid, golf, ...

(c) **Re-ranked Terms:** friendship, chimpanzee, tag, tiger, animal, oct, photo , photograph, cat, carrot, caribou, meow, stab, yukon, shark, usmc, balloon, source, calla lily, achievement, launch, stride, ...

- **Text-Image Alignment for Web Image Indexing**

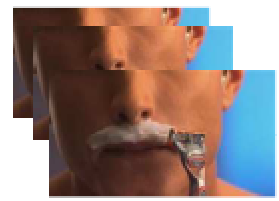
Cluster No.: 3598, 10 duplicates

Phrase list 1: sterilization equipment, water, sterilizer, china mainland
Phrase list 2: autoclave, sterilizer, water, china mainland, manufacturer
Phrase list 3: retort, heating, sterilizer, water, china mainland, manufacturer
Phrase list 4: sterilizer, water, china mainland, manufacturer
Phrase list 5: sterilization equipment, water, sterilizer, china mainland, manufacturer
....
Aggregation: sterilizer, sterilization equipment, water, retort, manufacturer, ….

Cluster No.: 6244, 13 duplicates

Phrase list 1: cimarron, roper, saddle, roper saddle, horse, ...
Phrase list 2: cimarron, roper, saddle, roper saddle,...
Phrase list 3: saddle, roper, roper saddle, horse, sale
Phrase list 4: roper saddle, saddle, cimarron, horse
....
Aggregation: saddle, roper, roper saddle, cimarron, ….

Cluster No.: 16263, 33 duplicates



Phrase list 1: face, area, drive stick, rule safety
Phrase list 2: face, grip, play tennis, tennis racket
Phrase list 3: face, , tennis racket, maintenance
Phrase list 4: face, shaver, tennis preparation tip,.
....
Aggregation: face, shaver, gillete, ….

Cluster No.: 29906, 8 duplicates

Phrase list 1: pisa feb, pisa, leaning tower, location, photo
Phrase list 2: pisa, leaning tower, location, photo
Phrase list 3: pisa, location, leaning tower, photo
Phrase list 4: pisa, leaning tower, photo….
Aggregation: pisa, learning tower, pisa feb, location, ….

Cluster No.: 35950, 27 duplicates

Phrase list 1: venture snowmobile, indonesia
Phrase list 2: venture snowmobile, arctic, snowmobile, ...
Phrase list 3: venture snowmobile, snowmobile
Phrase list 4: venture snowmobile, snowmobile manufacture
Aggregation: venture snowmobile, snowmobile, ….

Near-duplicates share similar semantics!

WWW2010, TPAMI2012

Evaluation

- Data set
  - 500, 000 web pages crawled from the Internet
  - 5,000,000 informative image have been extracted
  - Randomly select 5,000 images for evaluation because of the computational cost consideration

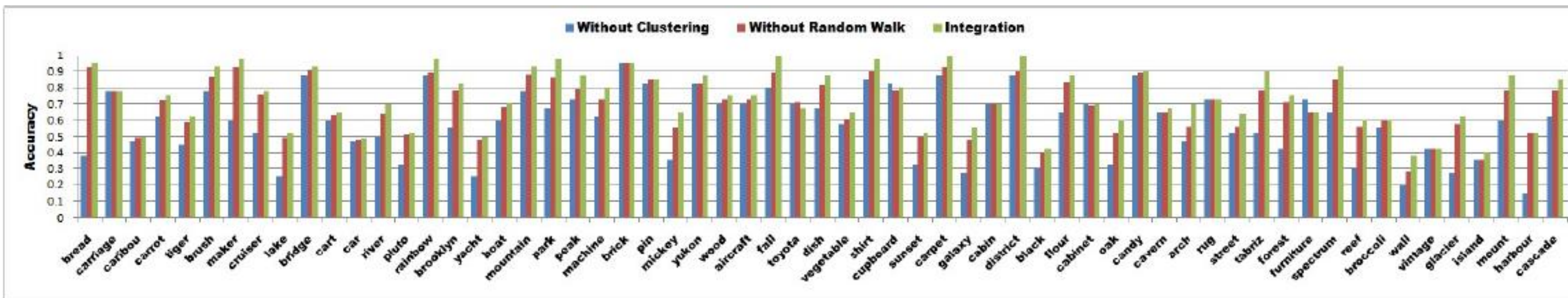
- Evaluation metrics
  - Accuracy rate

$$\varrho = \frac{\sum_{i=1}^{N} \delta(L_i, R_i)}{N},$$

$$\delta(x,y) = \begin{cases} 1, & x = y, \\ 0, & otherwise \end{cases}$$

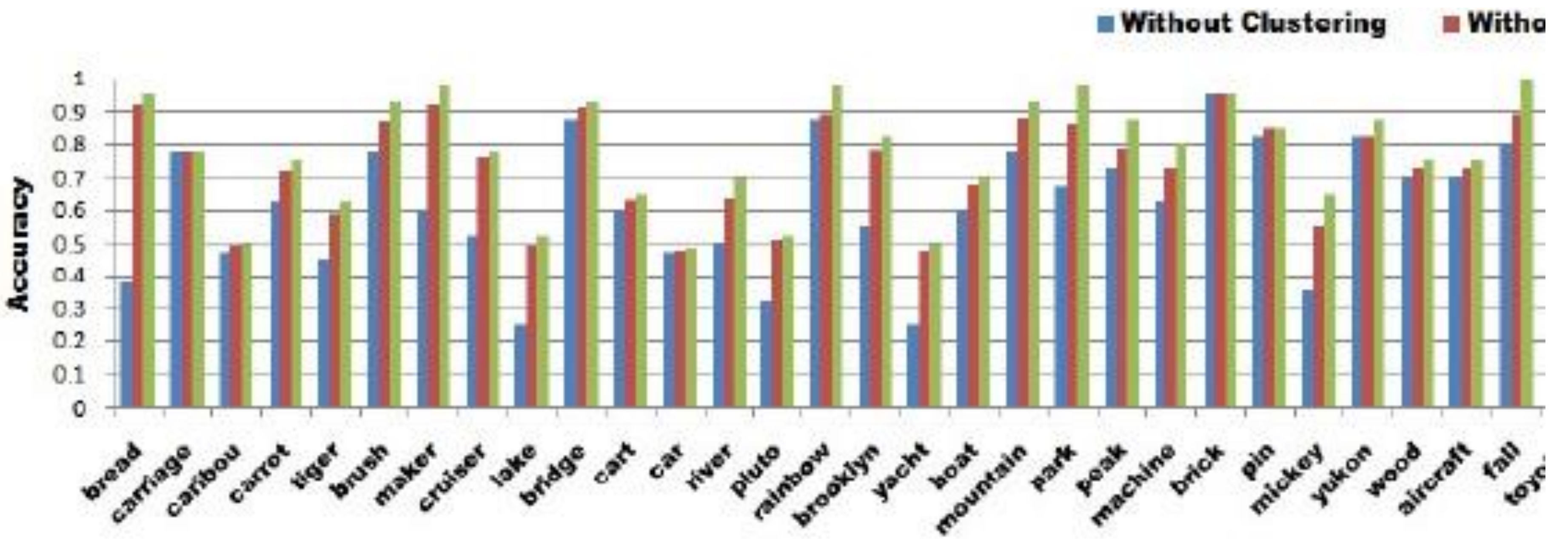
Results

- Effectiveness of image clustering and random walk for refinement



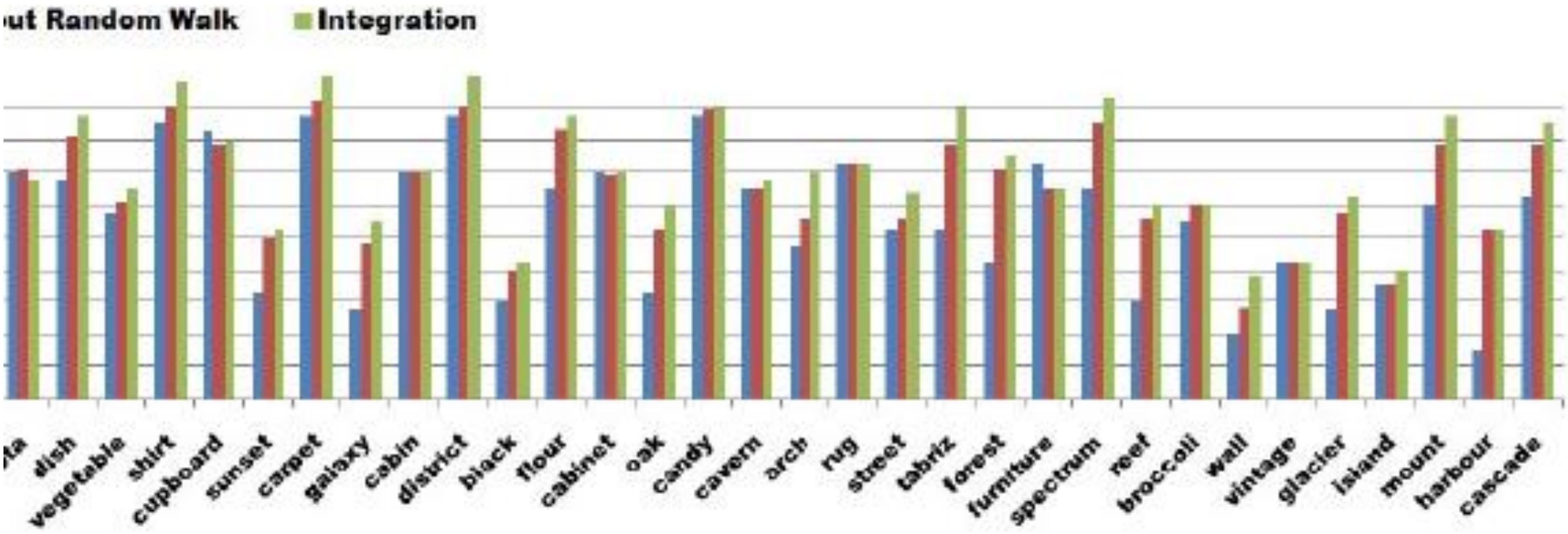
- Average accuracy: without clustering = 0.5828; Without random walk = 0.6939; Integration = 0.7373

Results



Without Clustering
Witho
Accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
bread
carriage
caribou
carrot
tiger
brush
maker
cruiser
lake
bridge
cart
car
river
pluto
rainbow
brooklyn
yacht
boat
mountain
park
peak
machine
brick
pin
mickey
yukon
wood
aircraft
fall

Results

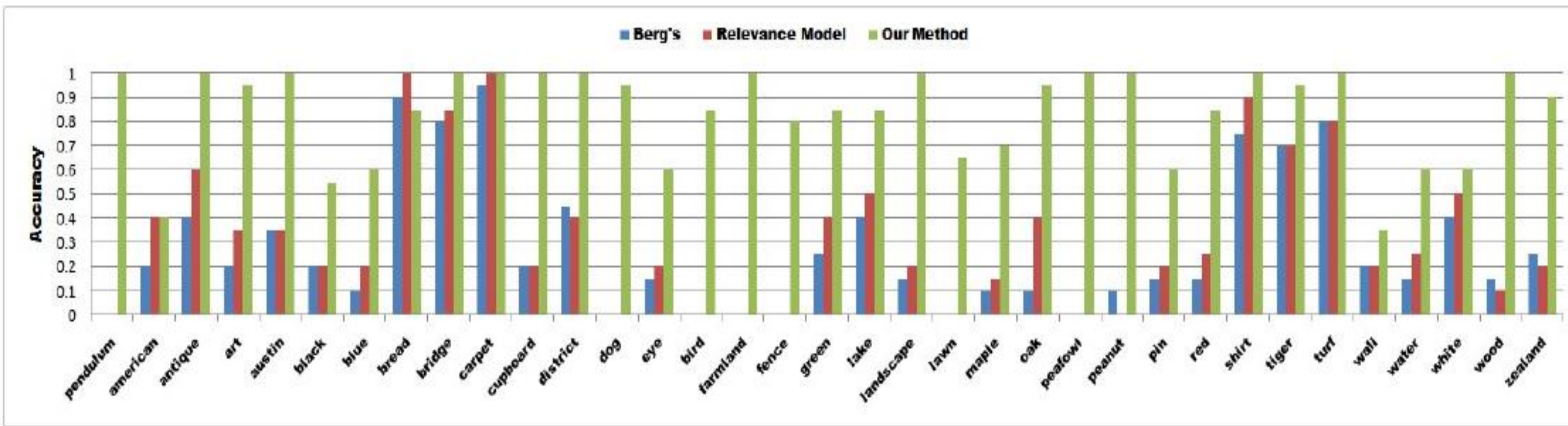


ut Random Walk
Integration
ta
dish
vegetable
shirt
cupboard
sunset
carpet
galaxy
cabin
district
black
flour
cabinet
oak
candy
cavern
arch
rug
street
tabriz
forest
furniture
spectrum
reef
broccoli
wall
vintage
glacier
island
mount
harbour
cascade

Results

- Compare with other image-text alignment model
- Models (both are supervised ones)
  - Berg’s
  - Cross-media relevance model
- Each concept we randomly select 60% samples as training samples and the other as test
- Our method was compared to the two methods on the test partion

Results



- Average accuracy: Berg's = 0.2771; Relevance Model = 0.3286; Our method = 0.8400