UNIVERSITÉ DE GRENOBLE

VISUAL GRAPH MODELING AND RETRIEVAL
A LANGUAGE MODEL APPROACH
FOR SCENE RECOGNITION

PHAM TRONG-TÔN

MRIM TEAM & IPAL LAB

02 December 2010

Jury :

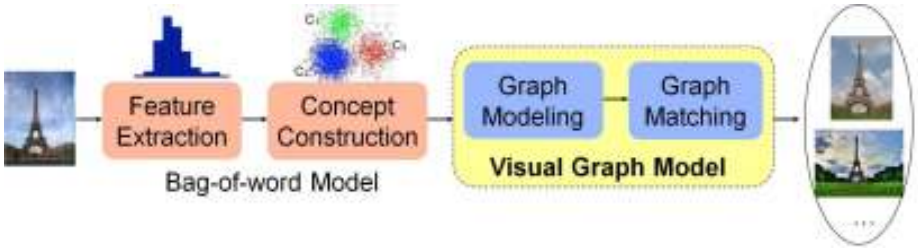
Augustin LUX (Président)
Philippe MULHEM (Directeur)
Joo-Hwee LIM (Co-directeur)

Mohand BOUGHANEM (Rapporteur)
Salvatore-Atoine TABBONE (Rapporteur)
Florent PERRONNIN (Examinateur)

Context: content-based image retrieval

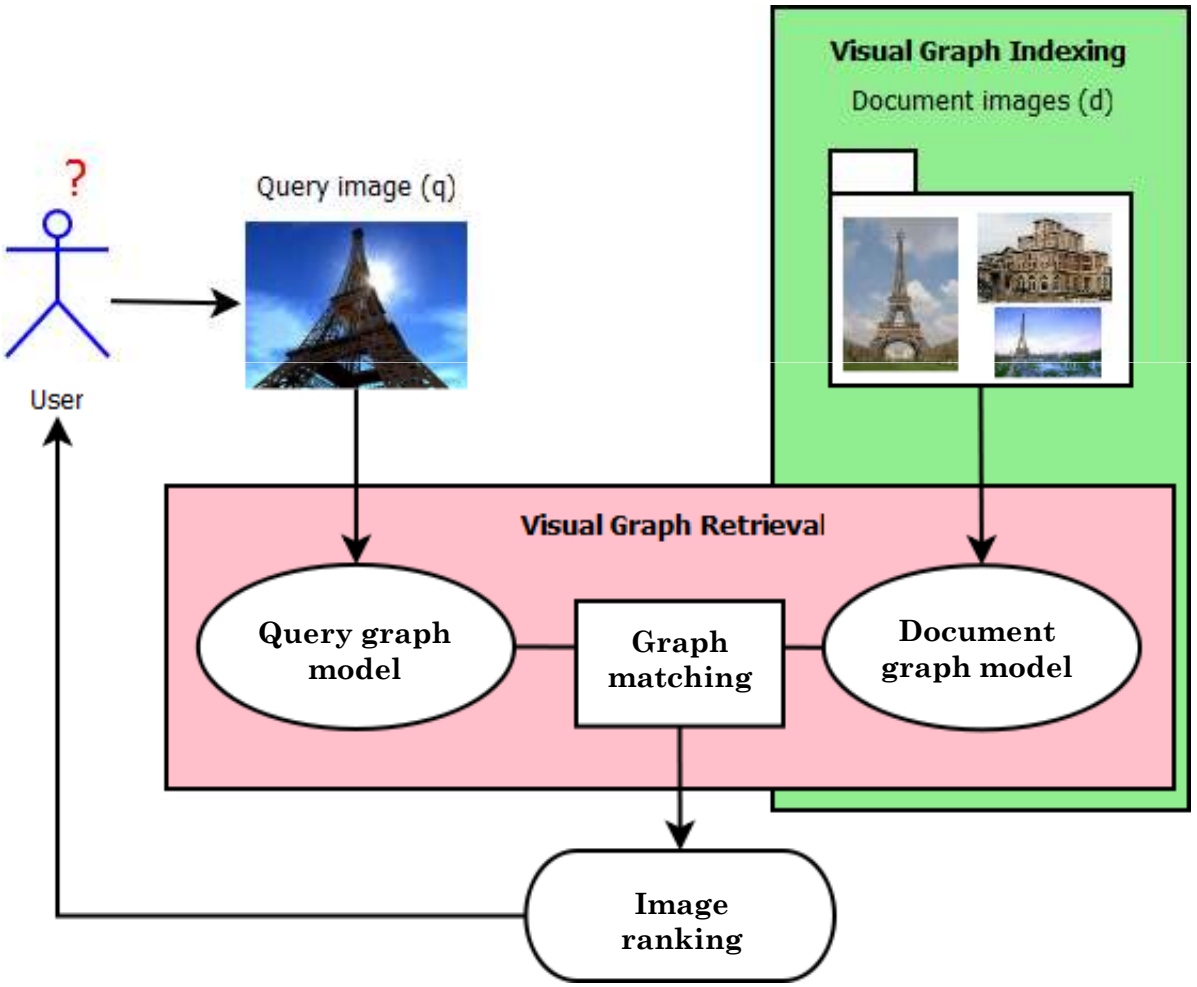
- **Goal**
  - Multiple image representations
  - Integration of spatial information
  - Fast and reliable image matching algorithm

- **Thesis focus**
  - Graph-based image modeling
  - Graph matching method

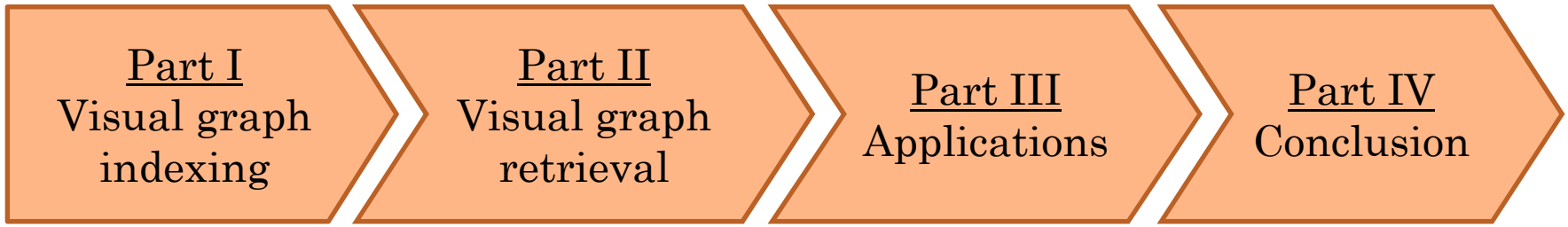


- **Applications**
  - Outdoor scene recognition
  - Indoor robot localization

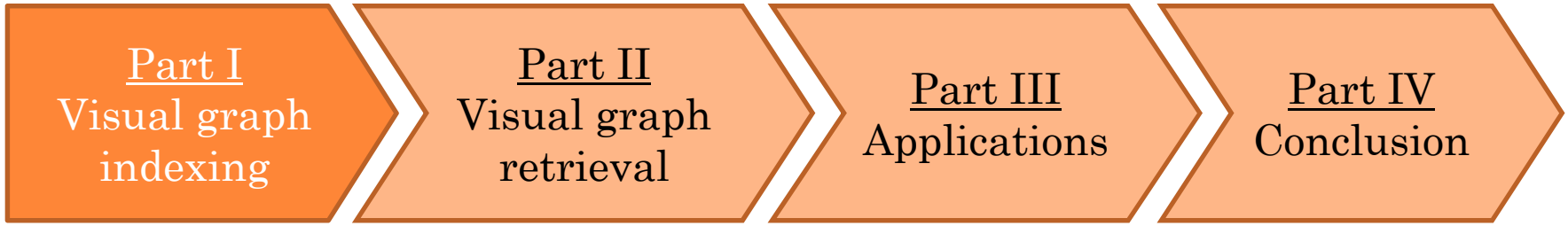
APPROACH OVERVIEW



OUTLINE


Part I
Visual graph indexing
Part II
Visual graph retrieval
Part III
Applications
Part IV
Conclusion

PART I: VISUAL GRAPH INDEXING

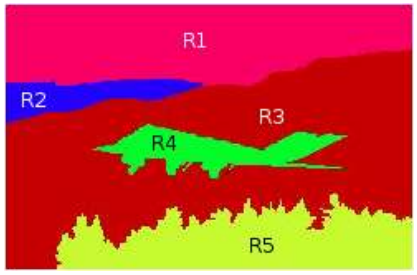


1. State of the art
2. Visual graph indexing

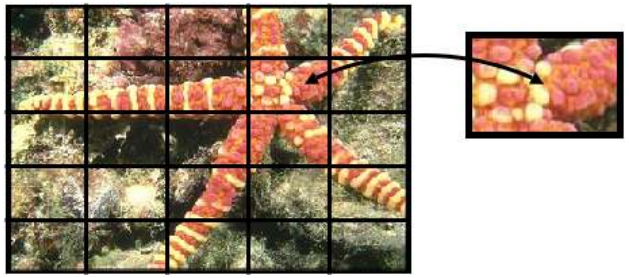
MULTIPLE IMAGE POINT OF VIEWS

- Different ways of image decompositions (region segmentation, grid partitioning, interest point detection)
- Different visual features extracted (color, edge, local invariant features)

[Mikolajczyk & Schmid 2002]



Region segmentation

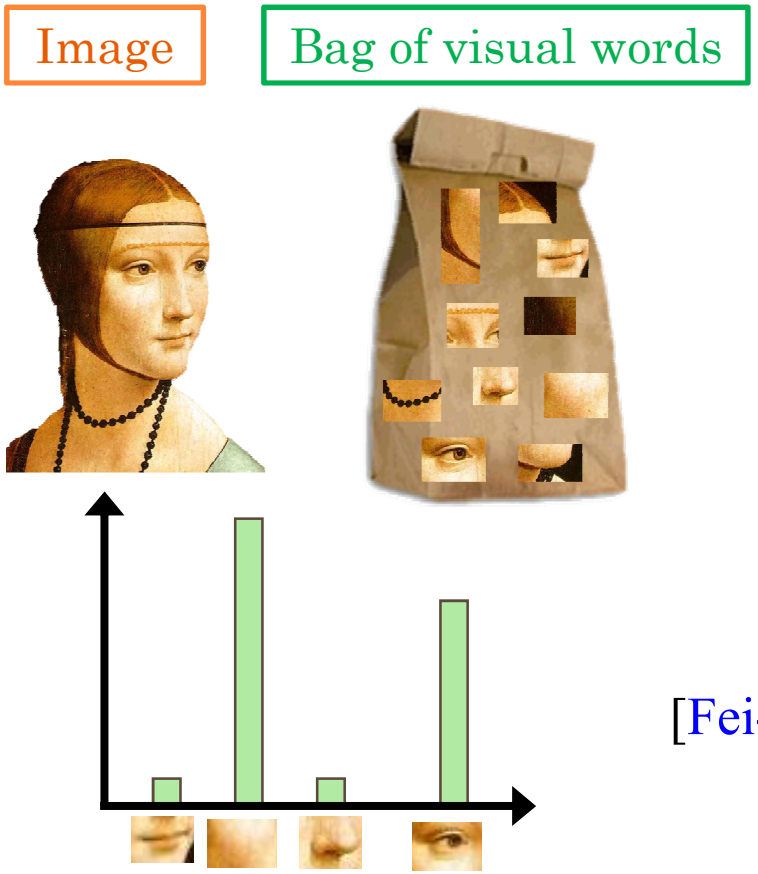


Grid partitioning

Interest point

VECTOR-BASED IMAGE REPRESENTATION

- Inspired from text retrieval domain



+ simple
+ easy to implement
+ memory efficiency

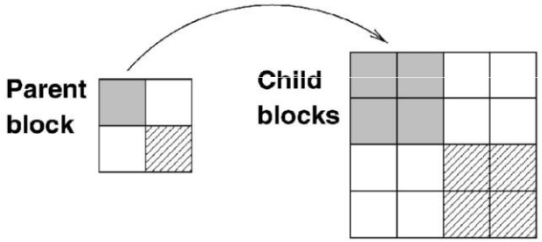
- flat representation
- sparse vectors
- lack of spatial information

[Fei-Fei & Perona 2005]

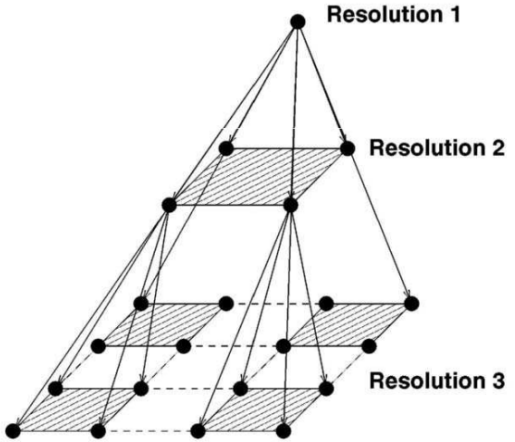
STRUCTURED IMAGE REPRESENTATION (CONT.)

- Multi-resolution hierarchical image structure

Generic photo



Pyramid structure



[Li & Wang 2003]

+ automatic block partitioning
+ require less computation

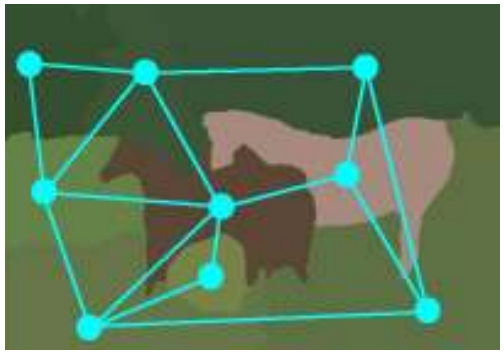
- non-weighted nodes and links

STRUCTURED IMAGE REPRESENTATION (CONT.)

- Forming planar graph from the connected regions

Natural scene

Planar graph



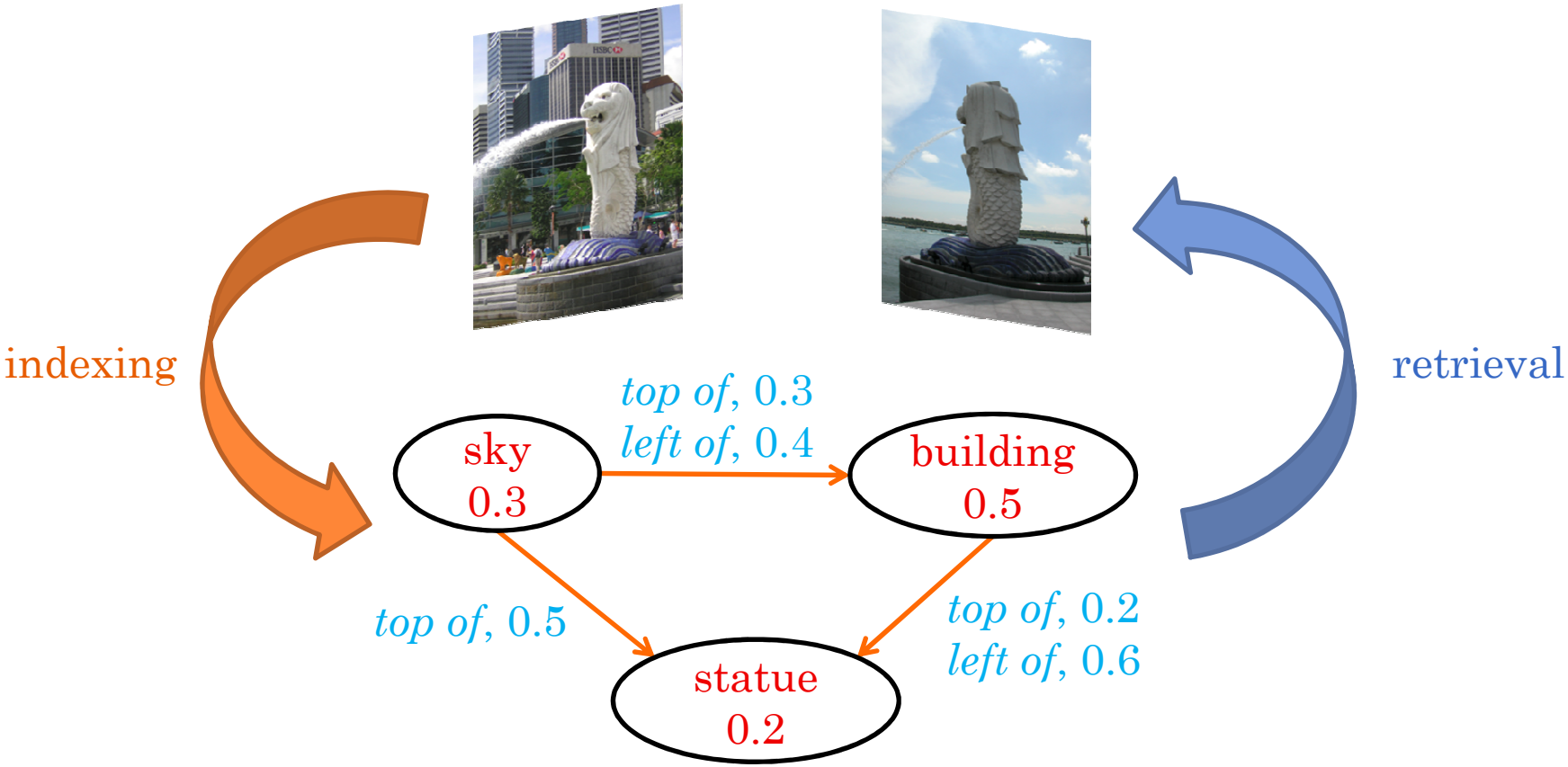
[Harchaoui & Bach 2007]

+ automatic image segmentation

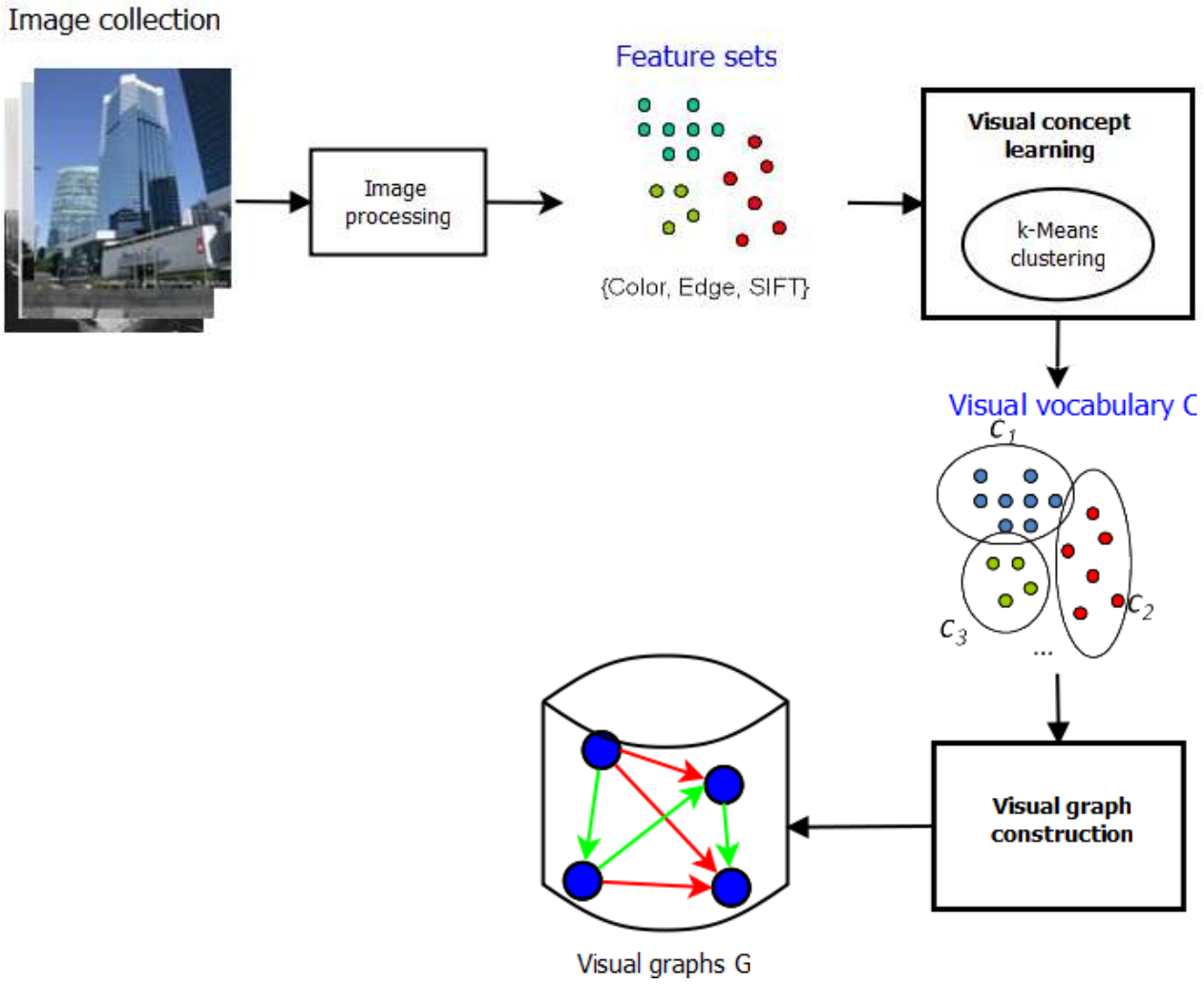
- non-weighted nodes and links

OUR PROPOSAL: VISUAL GRAPH MODELING

- A **visual graph** that combines both *visual concepts*, *spatial relations* and their *weights/probabilities*



VISUAL GRAPH INDEXING SCHEME



VISUAL GRAPH INDEXING SCHEME

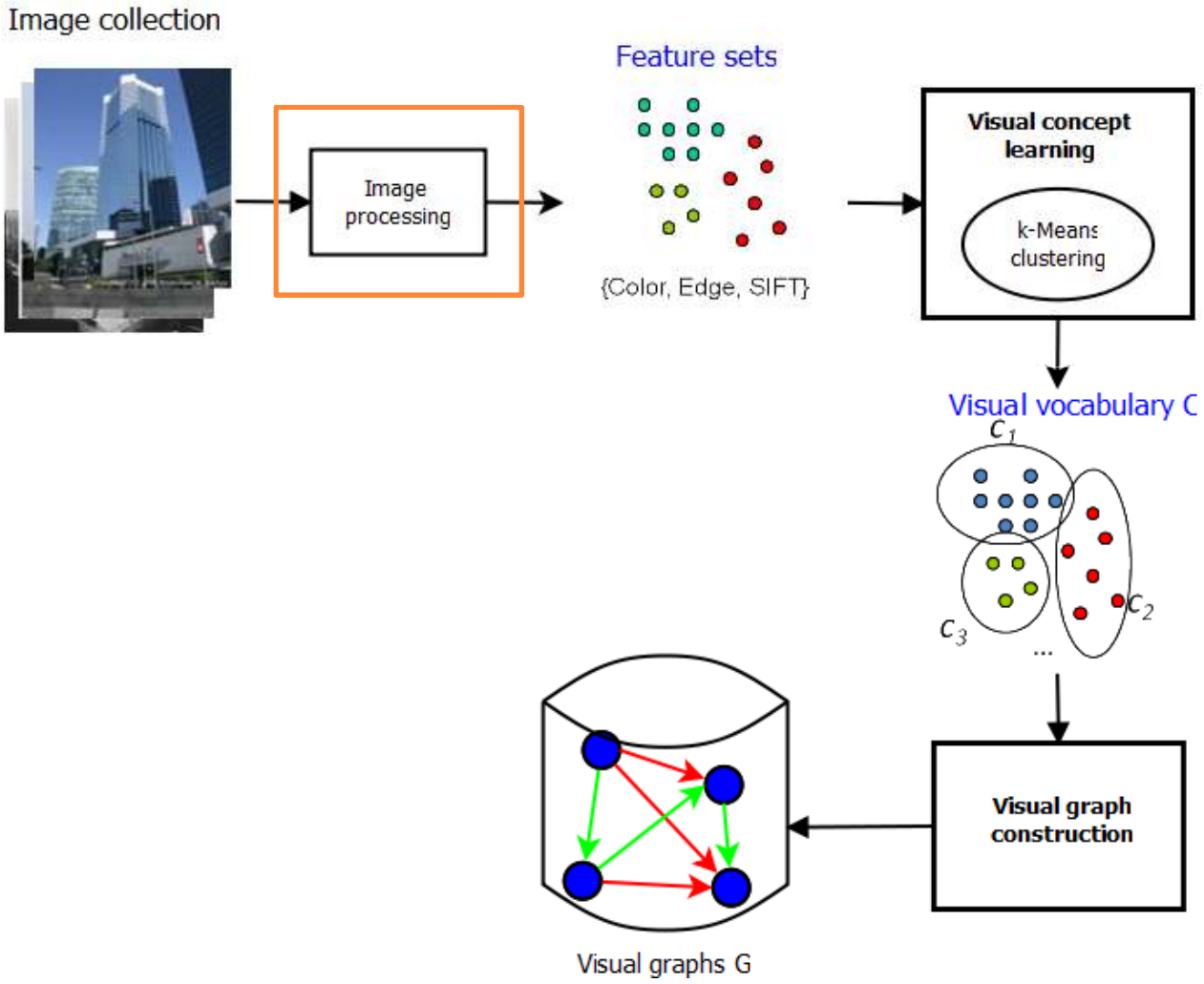


IMAGE PROCESSING

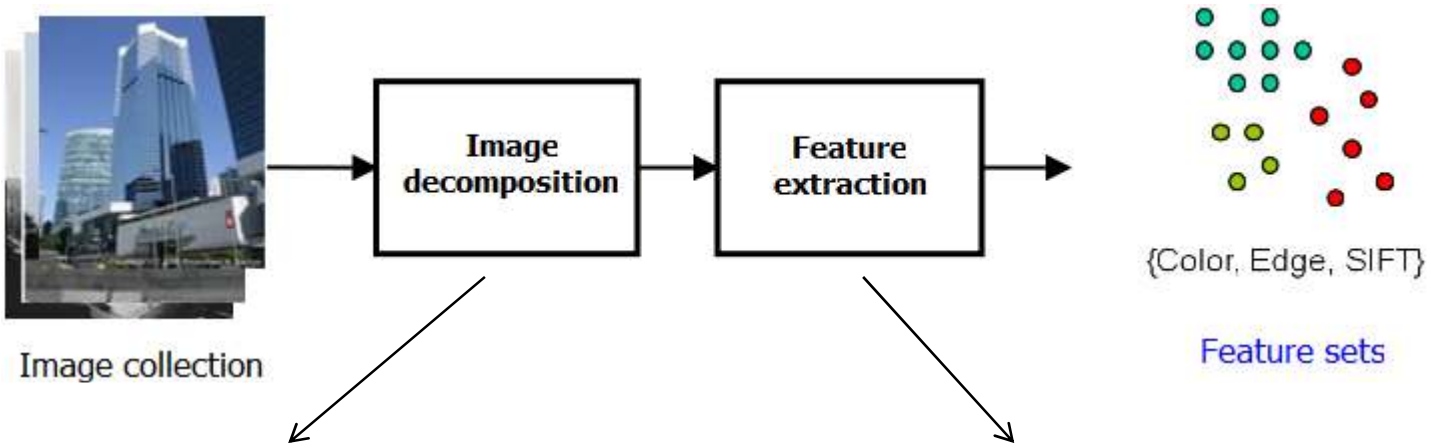
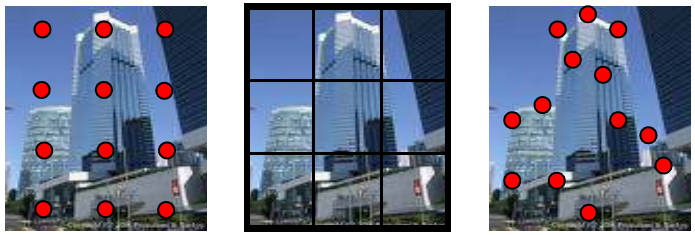


Image decomposition



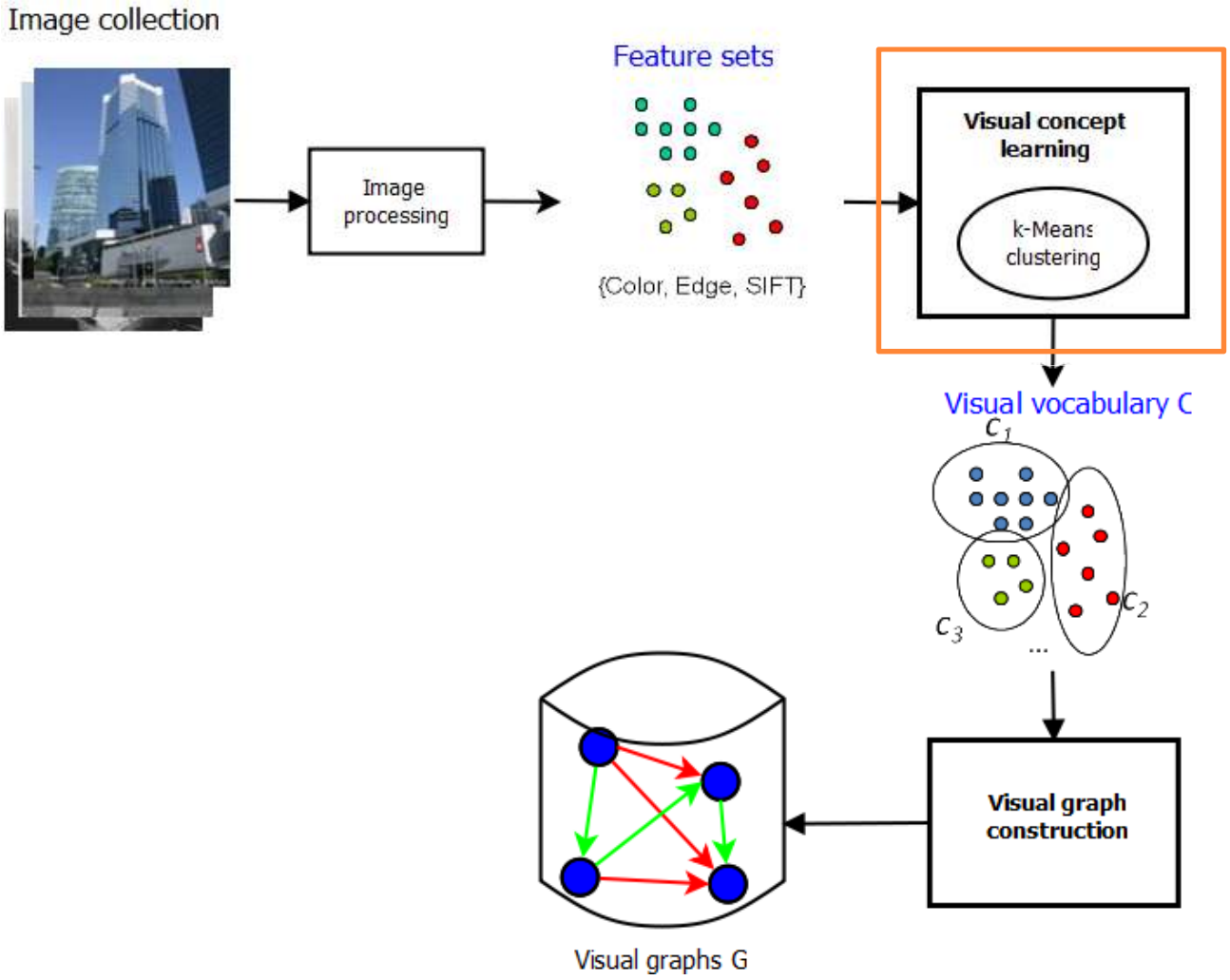
Pixel sampling | Grid partitioning | Interest point

Feature extraction

- Color histogram
- Edge histogram [Won *et al.* 2002]
- SIFT* descriptors [Lowe 1999]

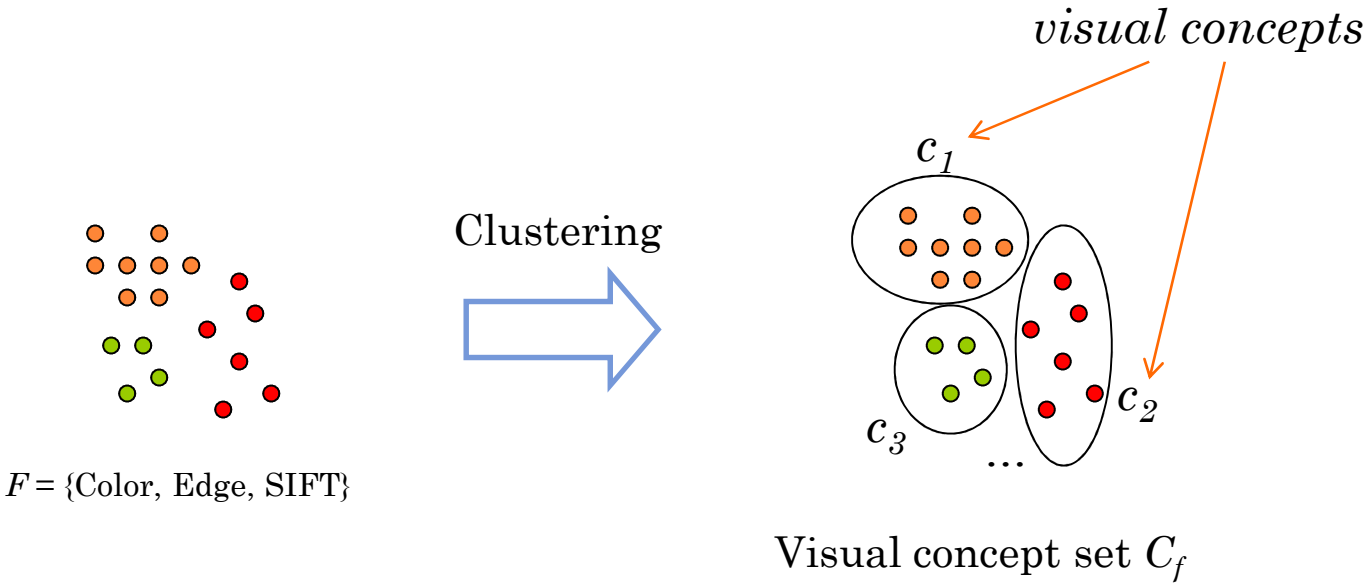
*SIFT: Scale Invariant Feature Transform

VISUAL GRAPH INDEXING SCHEME

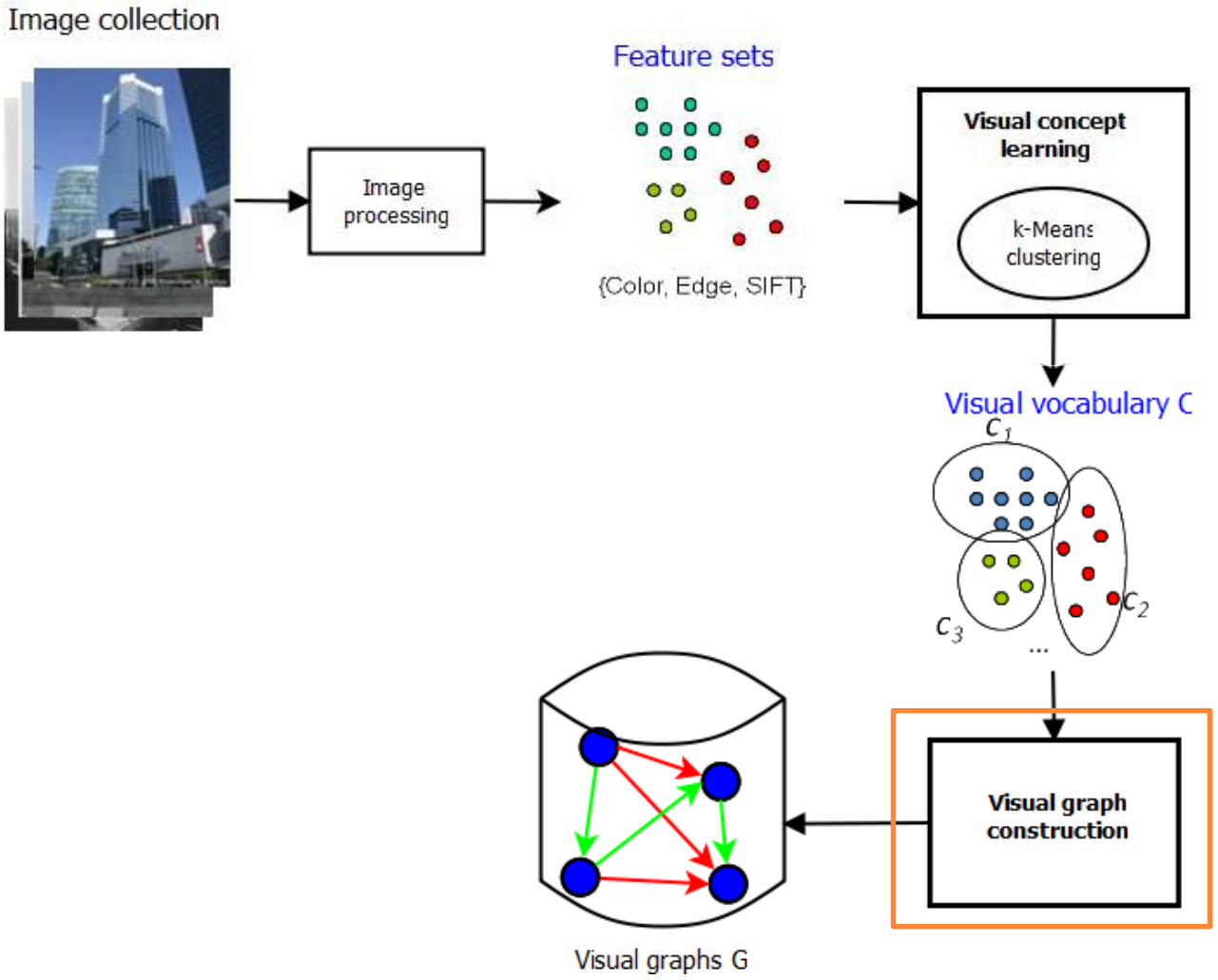


VISUAL CONCEPT LEARNING

- Unsupervised learning with *k-means clustering*
  - *k*: number of clusters (visual concepts)
  - One visual concept set C_f for each feature $f \in F$

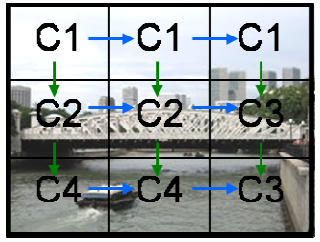


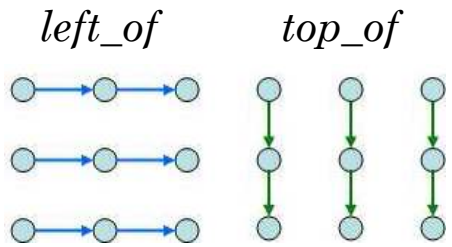
VISUAL GRAPH INDEXING SCHEME



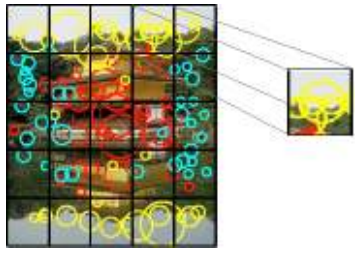
VISUAL GRAPH CONSTRUCTION

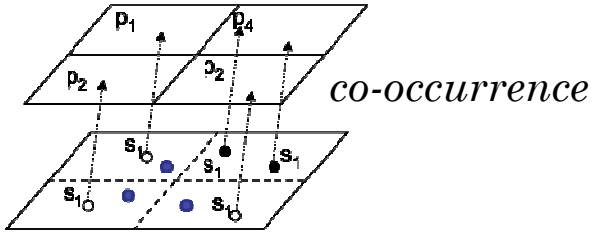
- For each image *I*:
  - Extraction of **weighted concept set** $\boldsymbol{WC^I_f}$ for each visual feature *f*
  - Extraction of **weighted relation set** $\boldsymbol{WE^I_l}$ between two concept sets WC^I_f and $WC^I_{f'}$ for each labeled relation *l*
    - intra-relation: $WC^I_f = WC^I_{f'}$
    - inter-relation: $WC^I_f \neq WC^I_{f'}$





Intra-relation





Inter-relation

VISUAL GRAPH CONSTRUCTION (CONT.)

- For an image collection C
  - Set of weighted concept sets

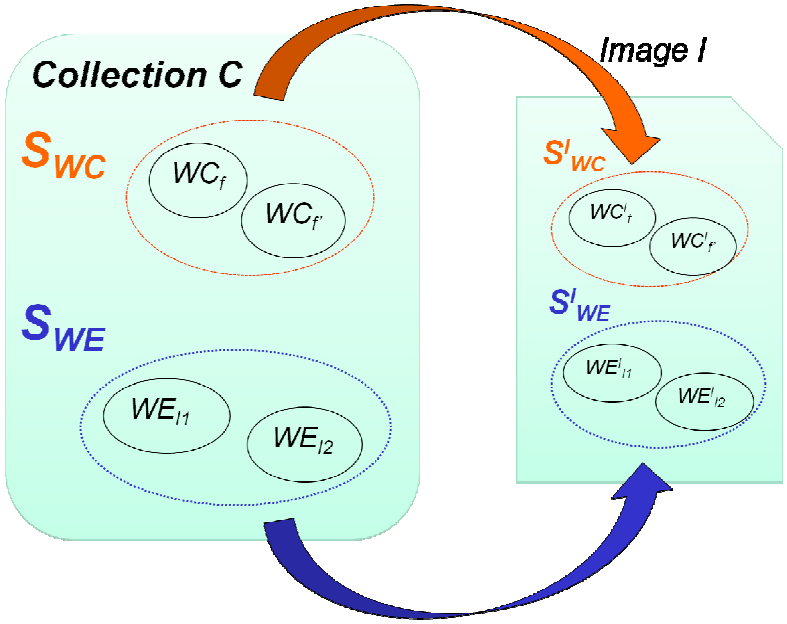
    $S_{WC} = \bigcup WC_f$
  - Set of weighted relation sets

    $S_{WE} = \bigcup WE_l$

- For an image I
  - Set of weighted concept sets

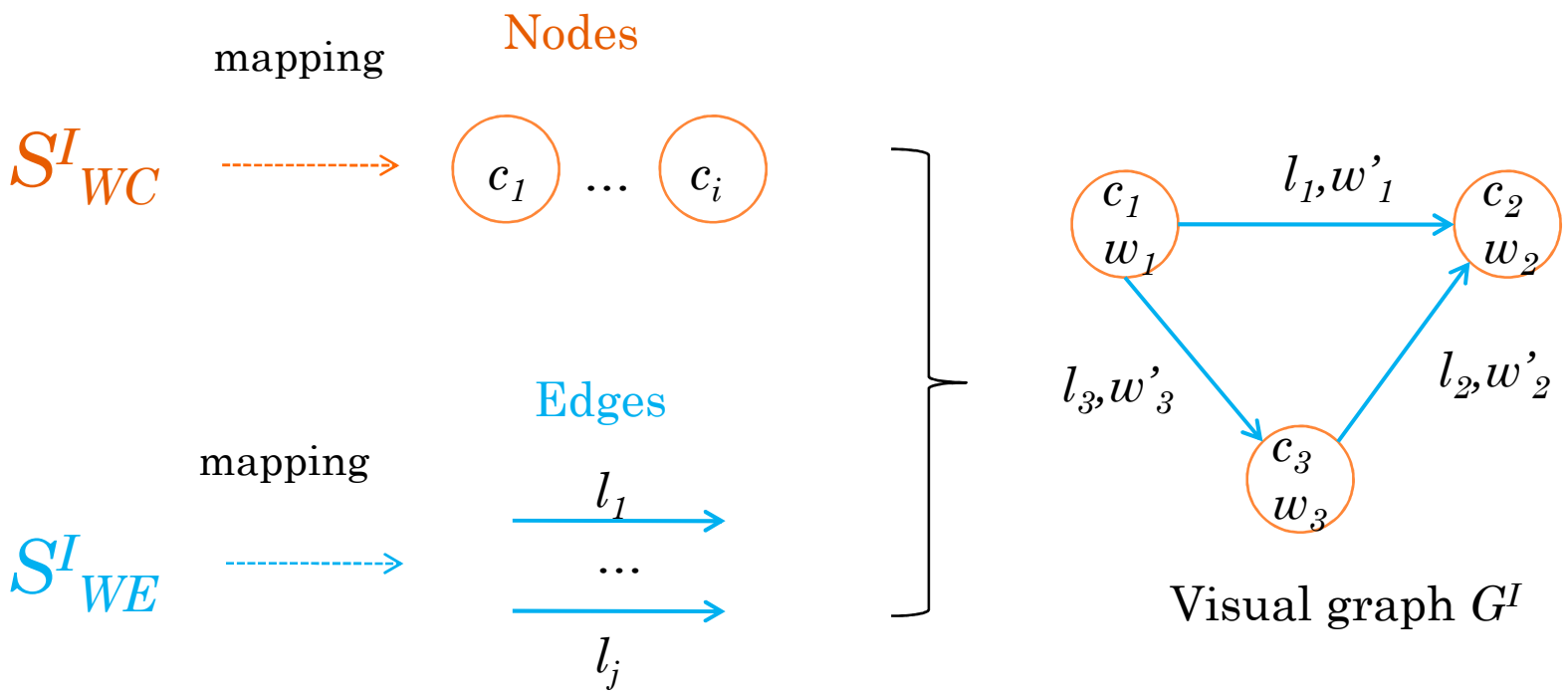
    $S^I_{WC} = \bigcup WC^I_f$
  - Set of weighted relation sets

    $S^I_{WE} = \bigcup WE^I_l$

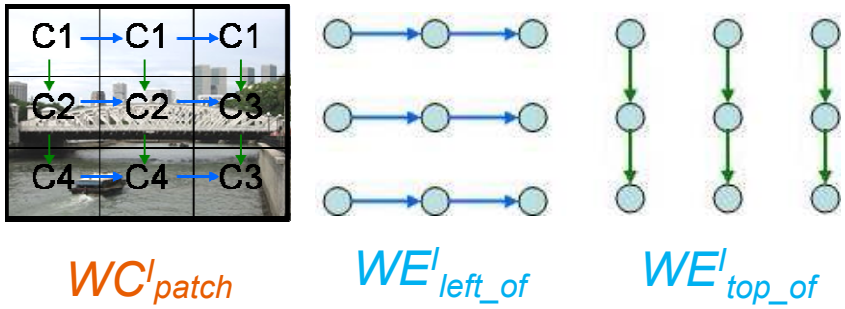


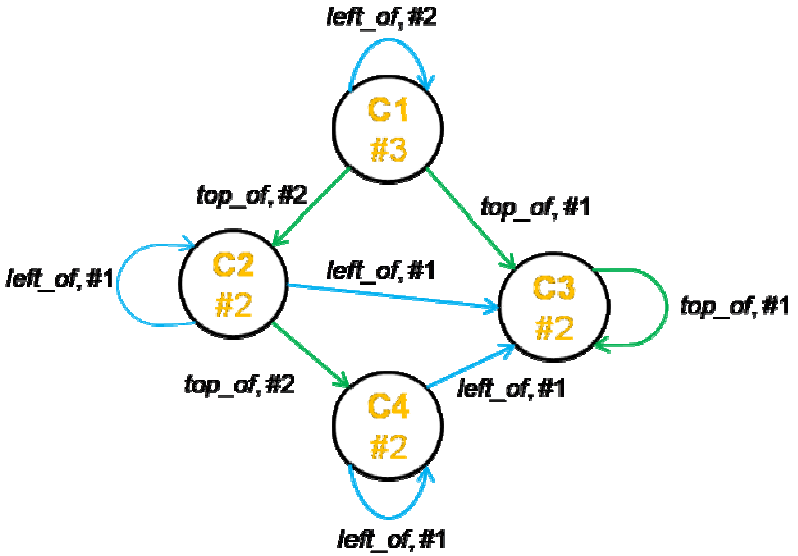
VISUAL GRAPH DEFINITION

- Visual graph for an image I: $G^I = <S^I_{WC}, S^I_{WE}>$



EXAMPLE





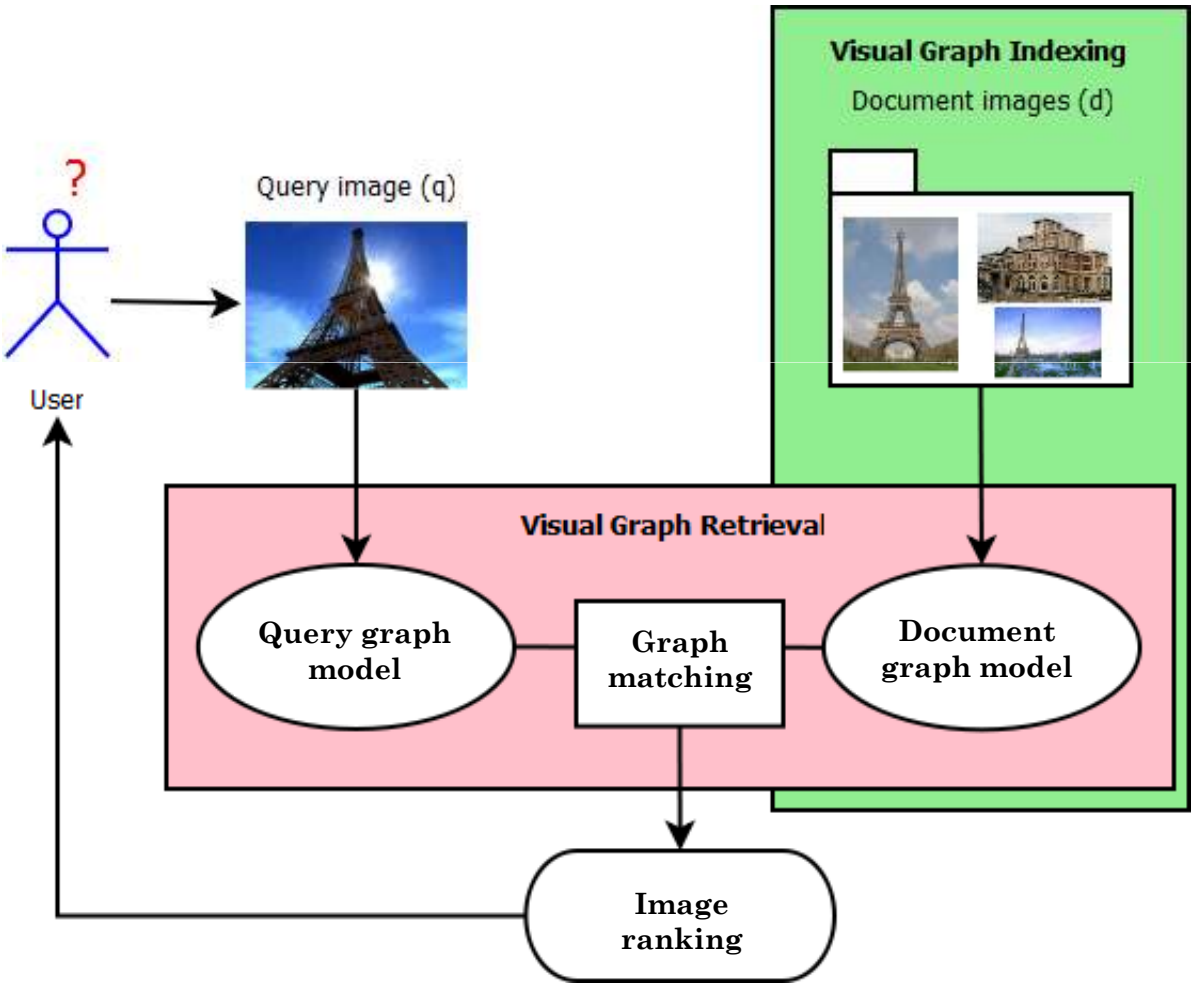
- $S^I_{WC} = \{WC^I_{patch}\}$

  $WC^I_{patch} = \{(c_1,3), (c_2,2), (c_3,2), (c_4,2)\}$

- $S^I_{WE} = \{WE^I_{left\_of}, WE^I_{top\_of}\}$

  $WE^I_{left\_of} = \{(c1,c1,left\_of,2), (c2,c2,left\_of,1),$
  $(c2,c3,left\_of,1), (c4,c3,left\_of,1), (c4,c4,left\_of,1)\}$

APPROACH OVERVIEW

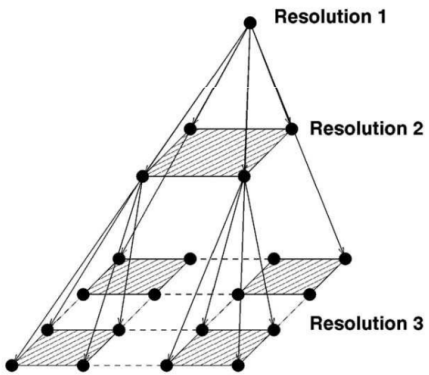


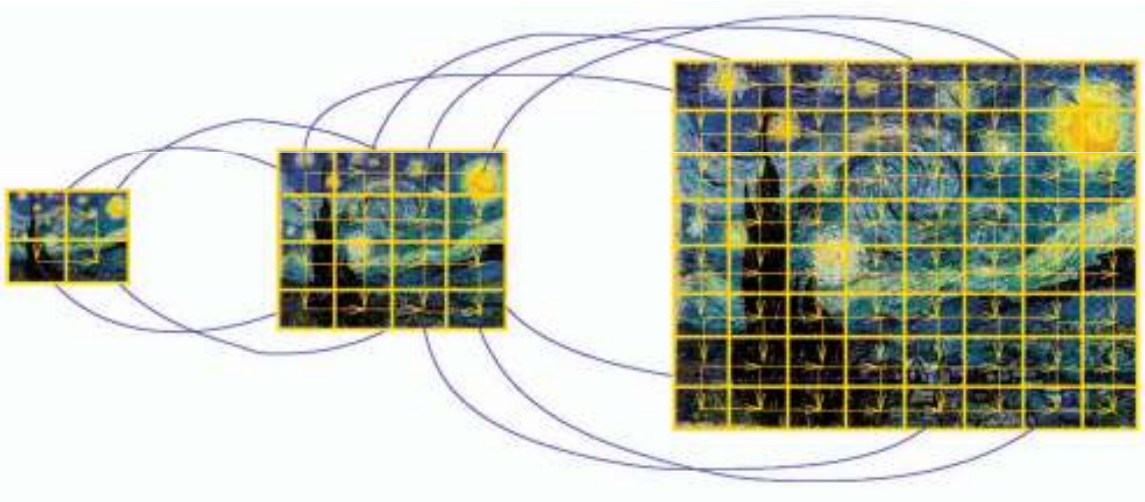
PART II: VISUAL GRAPH RETRIEVAL

1. State of the art
2. Visual graph retrieval

CURRENT MATCHING METHODS

- Inexact graph matching using **2D Multi-resolution Hidden Markov Model**



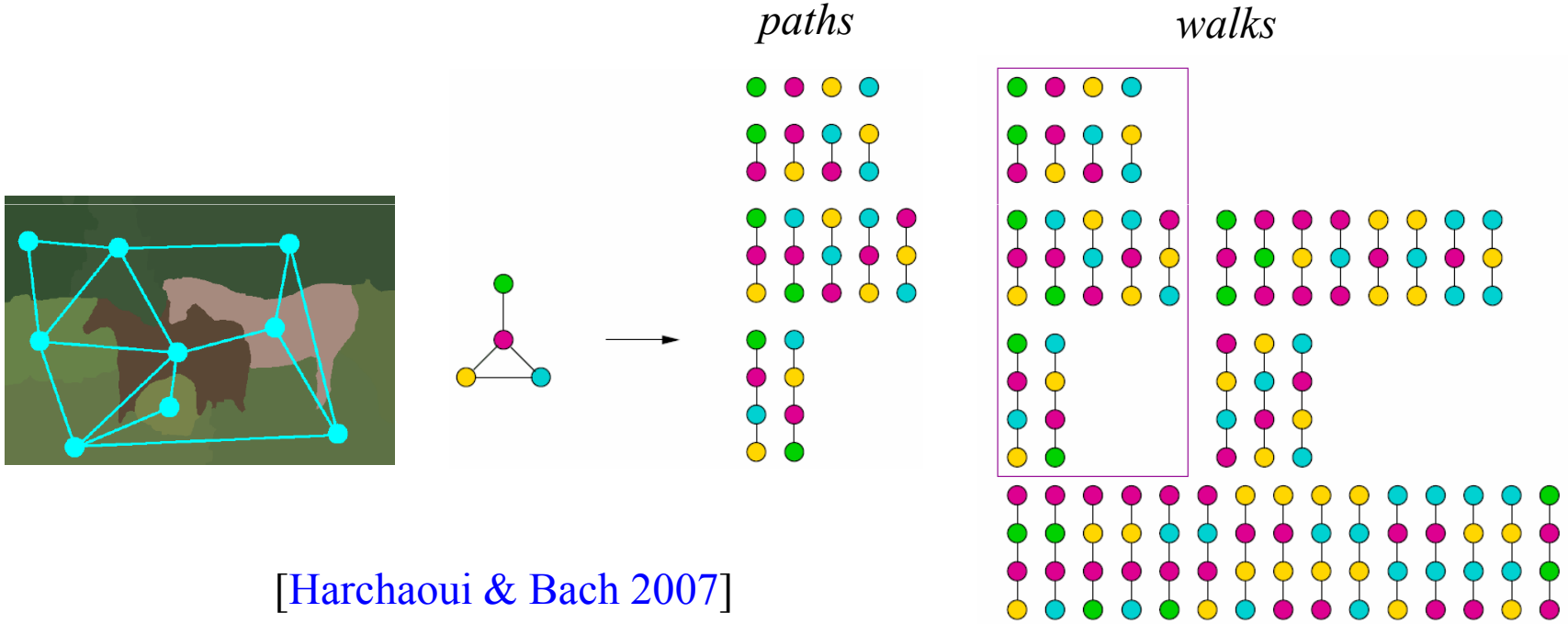


[Li & Wang 2003]

→ Estimation of Hidden Markov Models are time consuming

CURRENT MATCHING METHODS (CONT.)

- **Kernel-based graph clustering** based on *paths* and *walks*

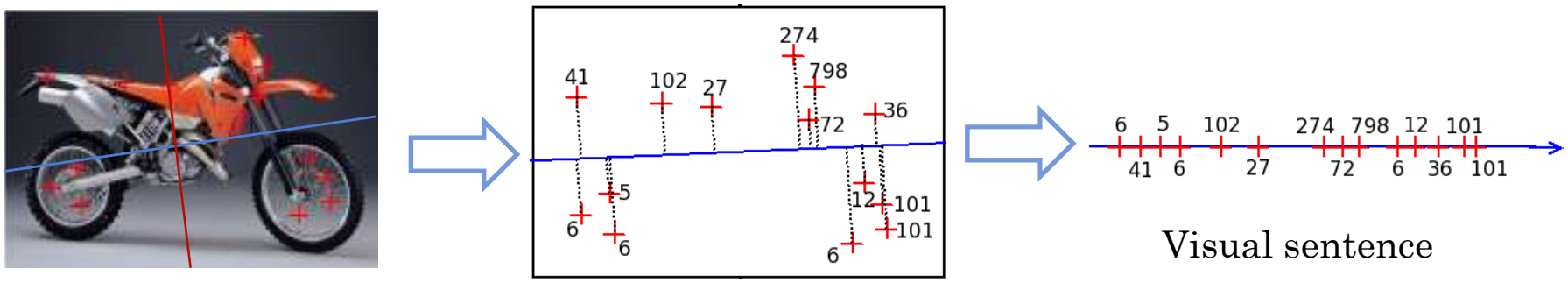


[Harchaoui & Bach 2007]

→ Applicable only for planar graph

CURRENT MATCHING METHODS (CONT.)

- Matching with **language modeling**
  - Unigram model (*bag of visual words*)
  - *n*-grams models (*n* = 2, 3, 4)

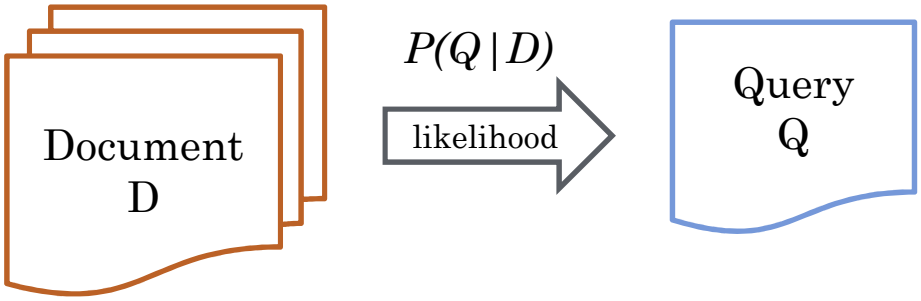


[Tirilly *et al.* 2008]

→ Spatial relationships are defined **implicitly** by *n*-grams sequence

LANGUAGE MODEL IN INFORMATION RETRIEVAL

- Query likelihood probability



[Ponte & Croft 1998]

- Unigram model & multinomial distribution

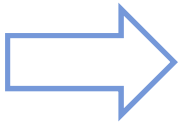
$$P(Q|D) = \prod_{q_i \in Q} P(q_i|D) = \frac{\#(q_i|D)}{\#(*|D)}$$

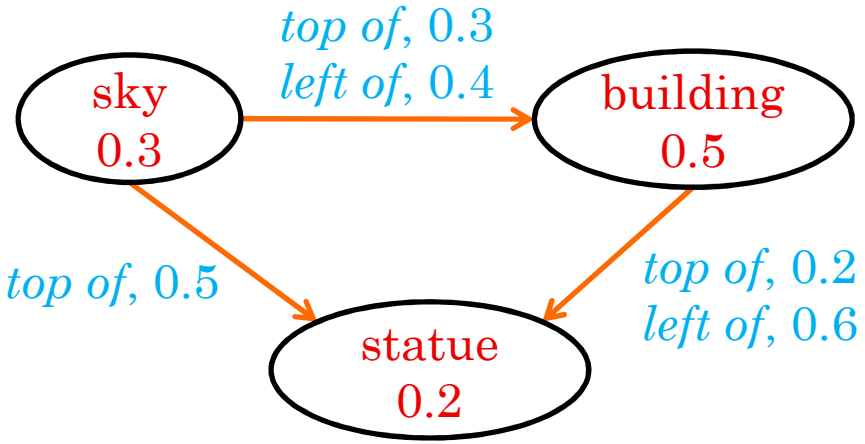
Smoothing techniques

→ Efficient method for text retrieval in IR

OUR PROPOSAL: VISUAL GRAPH MATCHING

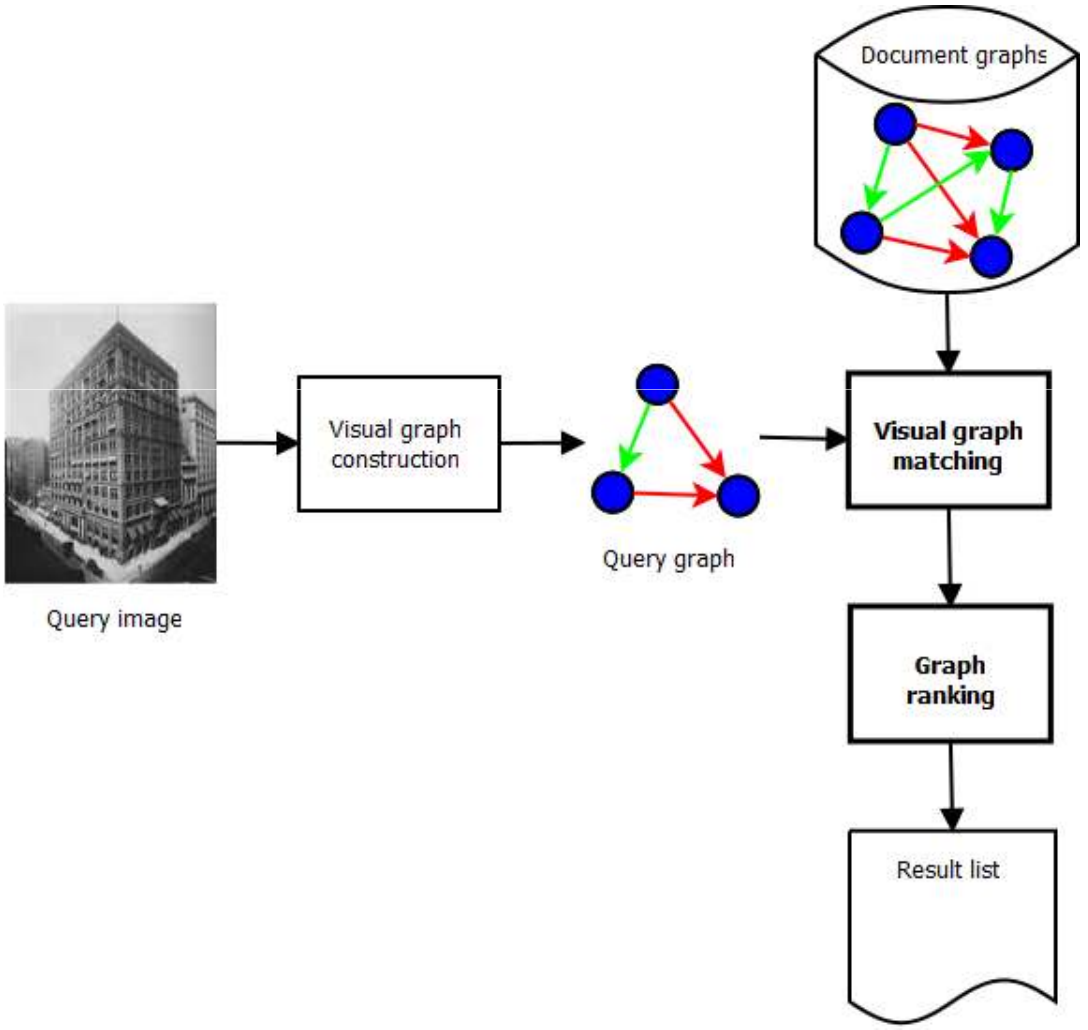
- Graph matching algorithm based on **language modeling** that takes into account:
  - Multiple type of *visual concepts* (nodes)
  - Multiple type of *relations* (edges)
  - *Weight/probability* of concept and relation





[Maisonnasse *et al.* 2009]
[Pham *et al.* 2010]

VISUAL GRAPH RETRIEVAL SCHEME



VISUAL GRAPH MATCHING

- Inspired by LM, **probability likelihood $P(G^q \mid G^d)$** of generating query graph G^q from document graph G^d

$$P(G^q \mid G^d) = P(S^q_{WC} \mid G^d) \times P(S^q_{WE} \mid S^q_{WC}, G^d)$$

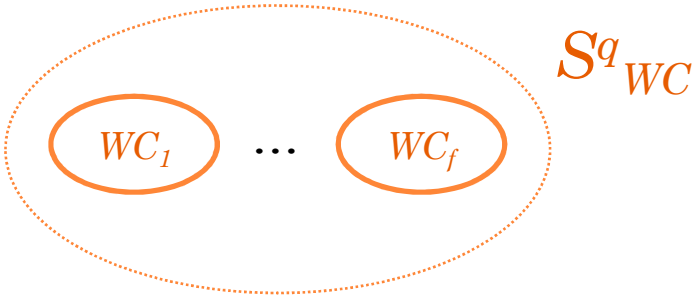
Concept sets — Relation sets

$G^q = \langle S^q_{WC}, S^q_{WE} \rangle$ query graph
$G^d = \langle S^d_{WC}, S^d_{WE} \rangle$ document graph

[Pham *et al.* 2010]

PROBABILITY OF CONCEPT SETS

- Concept set *independent hypothesis* : $\cap_f WC^q = \emptyset$



$$P(S^q_{WC} \mid G^d) = \prod_{WC^q \in S^q_{WC}} P(WC^q \mid G^d)$$

- *Multinomial distribution* model for WC^q

$$P(WC^q \mid G^d) \propto \prod_{c \in C} P(c \mid G^d)^{\#(c,q)}$$

SMOOTHING TECHNIQUES

- Problem: "missing concept" from the documents

  $P(c \mid G^d) = 0 \rightarrow P(G^q \mid G^d) = 0$

- Solution: give a small probability from the collection C for that "missing concept"

- Our proposal: Jelinek-Mercer smoothing in IR

$$P(c \mid G^d) = (1-\lambda_C)\frac{\#(c,d)}{\#(*,d)} + \lambda_C \frac{\#(c,C)}{\#(*,C)} \text{ with } \lambda_C \in [0,1]$$

PROBABILITY OF RELATION SETS

- Relation set *independent hypothesis* : $\cap_l WE^q = \emptyset$

$$P(S_{WE}^q \mid S_{WC}^q, G^d) = \prod_{WE^q \in S_{WE}^q} P(WE^q \mid S_{WC}^q, G^d)$$

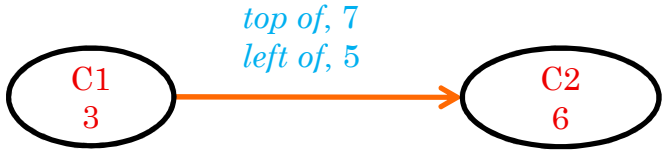
- Multinomial distribution model

$$P(WE^q \mid S_{WC}^q, G^d) \propto \prod_{(c,c',l) \in C \times C' \times L} P(L(c,c') = l \mid WC^q, WC'^q, G^d)^{\#(c,c',l,q)}$$

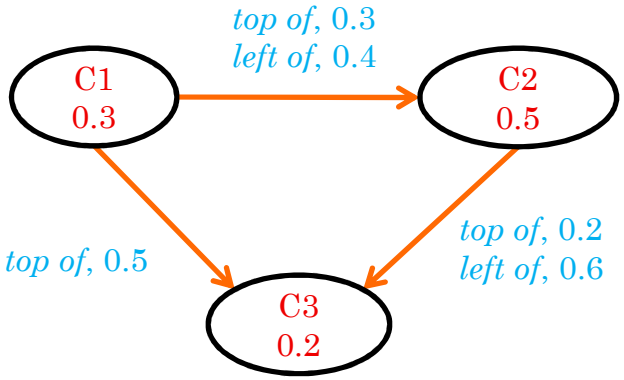
- Jelinek-Mercer smoothing

$$P(L(c,c') = l \mid WC^q, WC'^q, G^d) = (1-\lambda_L)\frac{\#(c,c',l,d)}{\#(c,c',*,d)} + \lambda_L \frac{\#(c,c',l,C)}{\#(c,c',*,C)}$$

VISUAL GRAPH MATCHING EXAMPLE



Query G^q

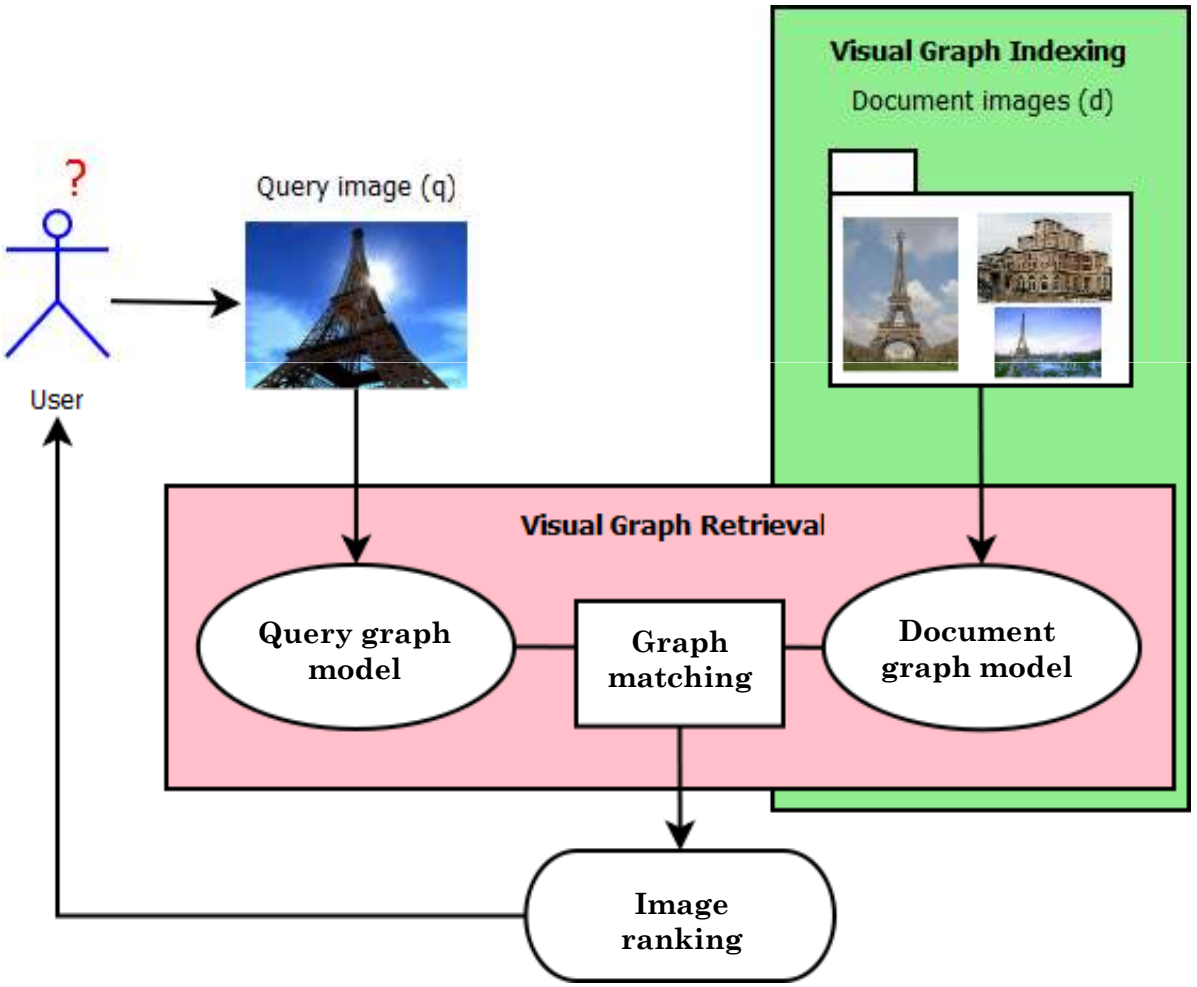


Document G^d

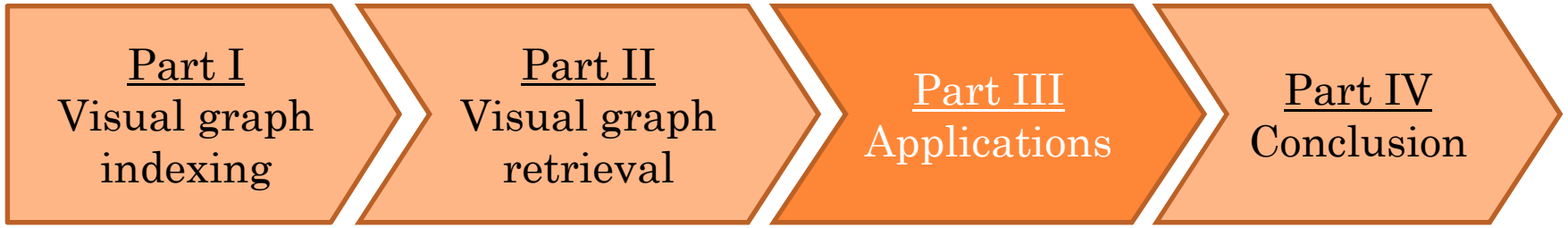
$$P(G^q \mid G^d) = P(C1 \mid G^d)^3 . P(C2 \mid G^d)^6 \times P(L(C1,C2)=top\_of \mid WC^q, G^d)^7 . P(L(C1,C2)=left\_of \mid WC^q, G^d)^5$$
$$= (0.3)^3 . (0.5)^6 . (0.3)^7 . (0.4)^5$$
$$= 0.59049.10^{-8}$$

→ Images are ranked based on their **probability likelihoods**

APPROACH OVERVIEW



PART III: APPLICATIONS



1. Scene recognition
2. Robot localization

SUMMARY

1. Outdoor scene recognition

2. Indoor robot localization

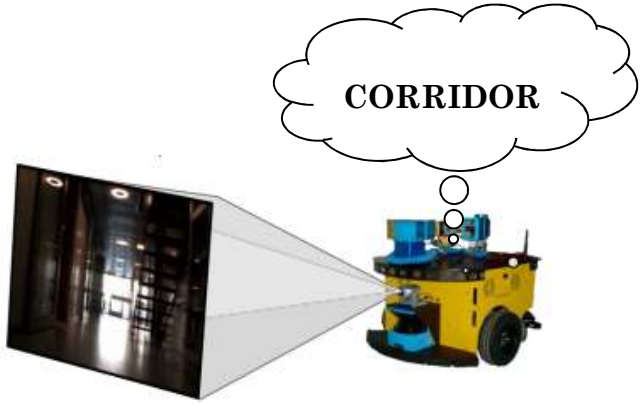
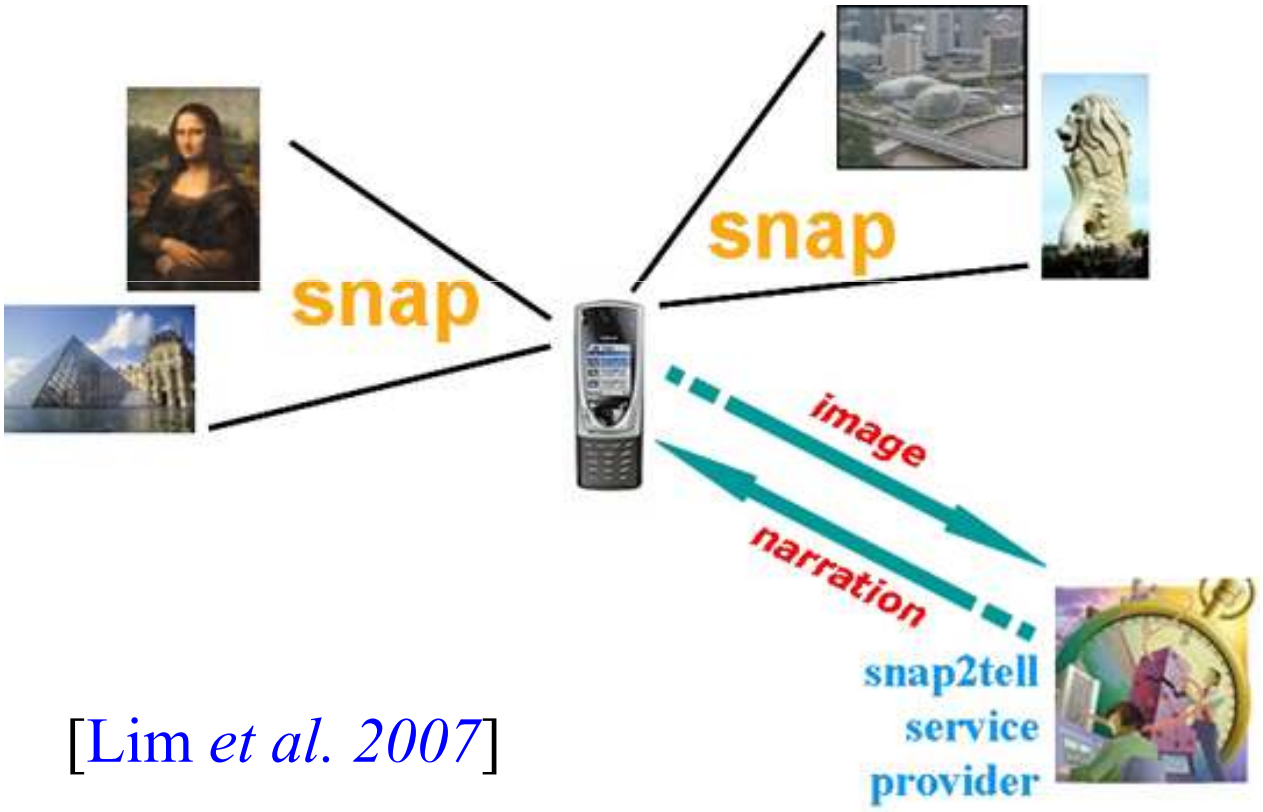


IMAGE-BASED MOBILE TOUR GUIDE



[Lim *et al. 2007*]

STOIC-101 COLLECTION

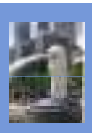
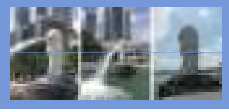
	Training	Test	Overall
Image	3189	660	**3849**
Scene	101	101	**101**

- Difficulties
  - Occlusion and moving objects
  - Variation of viewpoints, scales
  - Variation of lighting conditions

The Singapore Tourist Object Identification Collection

EVALUATION METHODS

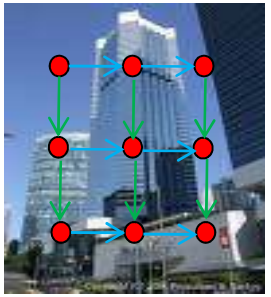
- Several scenarios for training and querying

	Trained by I 	Trained by S 
Query by I 	✓	✓
Query by S 	✓	✓

VISUAL GRAPH MODELS

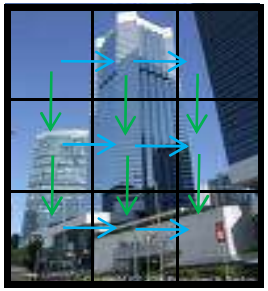
- **Summary**
  - 500 visual concepts
  - 1 concept set WC_{mg} or WC_{gg}
  - 2 intra-relation sets WE_{left_of} , WE_{top_of}

mg concepts



- **Implemented models**
  1. ***mg-LM*** $= < \{WC_{mg}\}, \emptyset >$
  2. ***mg-VGM*** $= < \{WC_{mg}\}, \{WE_{left\_of}, WE_{top\_of}\} >$
  3. ***gg-LM*** $= < \{WC_{gg}\}, \emptyset >$
  4. ***gg-VGM*** $= < \{WC_{gg}\}, \{WE_{left\_of}, WE_{top\_of}\} >$

gg concepts



EXPERIMENTAL RESULTS

Classification accuracy:

Image accuracy = TP_i / N_i (N_i = 660)
Scene accuracy = TP_s / N_s (N_s = 101)

Train	Query	*mg-LM*	***mg-VGM***	*gg-LM*	***gg-VGM***
I	I	0.789	**0.794** (+0.6%)	0.484	0.551 (+13.8%)
I	S	0.822	**1.00** (+21.6%)	0.465	0.762 (+63.8%)
S	I	0.529	0594 (+12.3%)	0.478	**0.603** (+26.1%)
S	S	**1.00**	**1.00**	0.891	0.920 (+3.2%)

→ **VGMs** (with spatial relations) outperform **LMs**

→ Significant impact of **multiple querying images (S)**

COMPARISON WITH THE STATE-OF-THE-ART

- SVM* method: RBF kernel with cross validation

	#class	**SVM**	**LM**	**VGM**
mg-concepts	101	0.744	0.789 (+ 6.0%)	**0.794** (+ 6.3%)

VGM outperforms both **LM** and **SVM** methods

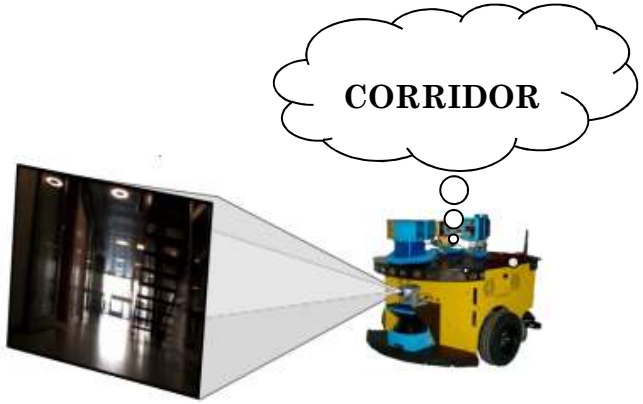
- Implementation
  - **C/C++** with the **LTI-Lib** on Linux platform
  - **3.0 GHz quad-core** CPU and **8.0 Gb** of memory
  - Execution time: **0.22 seconds** per image

*SVM: Support Vector Machine

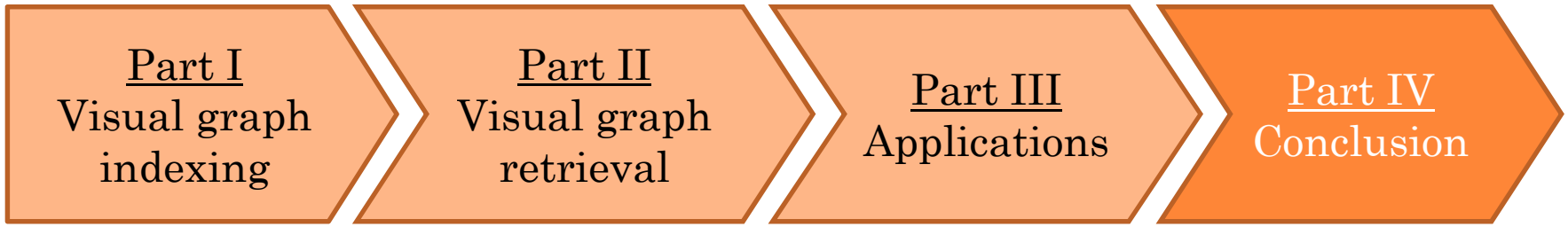
SUMMARY

1. Outdoor scene recognition

2. Indoor robot localization

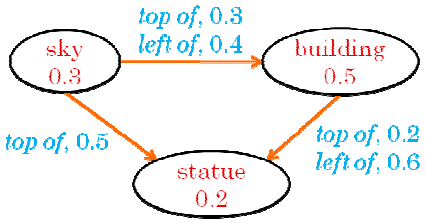


Part IV: Conclusion



1. Contributions
2. Perspectives

CONTRIBUTIONS



- A **graph-based image representation** for image indexing and retrieval
  - multiple *visual concept sets*
  - multiple *relation sets*
  - *weight/probability* of visual concept and relation

- A **simple** and **effective** graph matching process
  - based on the *language modeling* in IR
  - *multinomial distribution* and *independent hypotheses*
  - *generality* and *expendability* in different contexts

Language Modeling

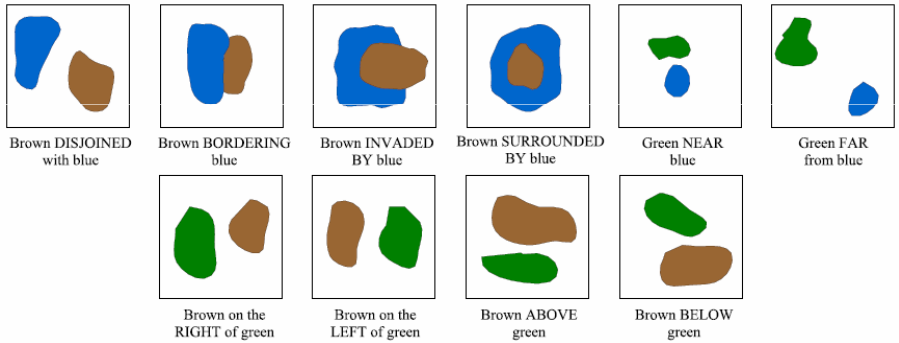
CONTRIBUTIONS (CONT.)

- Application to the problem of real-life **scene recognition**
  - visual graph models adapt to *mobile device*
  - improved the accuracies with the *spatial relations*

- Experiment on the **robot localization** using only visual information
  - visual graph models are robust with the *illumination and environment changes*
  - outperformed the performance of the *state-of-the-art SVM method*

SHORT-TERM PERSPECTIVES

- Combination of textual graph model and visual graph model in a common framework
- Further study on visual concepts and spatial relations



[Aksoy 2006]

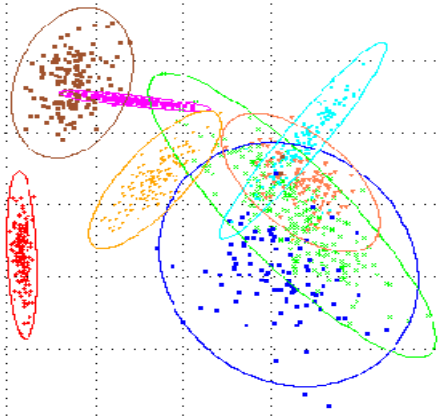
- Evaluation of the proposed approach on large image collections
  - Object classification, VOC
  - Video retrieval, TRECVID

LONG-TERM PERSPECTIVES

- **Relevance modeling** using information divergence
  - *Kullback-Leibler (KL)* divergence model measures the divergence between query models and document models

- Extension of the current **probabilistic framework**
  - Definition of *"soft"* visual concept based on *fuzzy c-means* or *Expectation-Maximization clustering*

EM clustering



PUBLICATIONS

- **Journal Peer-reviewed Articles**
  1. **Trong-Ton Pham**, Philippe Mulhem, Loic Maisonnasse, Eric Gaussier, Joo- Hwee Lim. *Visual Graph Modeling for Scene Recognition and Robot Localization*. Journal on Multimedia Tools and Applications, 20 pages, Springer, January 2011.
  2. **Trong-Ton Pham**, Loic Maisonnasse, Philippe Mulhem, Eric Gaussier. *Modèle de graphe et modèle de langue pour la reconnaissance de scènes visuelles*. Numéro spécial du revu Document Numérique, Vol 13 (211-228), Lavoisier, Juin 2010.

- **International Peer-reviewed Conference Articles**
  1. **Trong-Ton Pham**, Philippe Mulhem, Loic Maisonnasse. *Spatial Relationships in Visual Graph Modeling for Image Categorization*. Proceedings of the 33rd ACM SIGIR'10, pages 729-730, Geneva, Switzerland, 2010.
  2. **Trong-Ton Pham**, Philippe Mulhem, Loic Maisonnasse, Eric Gaussier. *Integration of Spatial Relationship in Visual Language Model for Scene Retrieval*. IEEE 8th International Workshop on Content-Based Multimedia Indexing (CBMI), 6 pages, Grenoble, France, 2010.
  3. **Trong-Ton Pham**, Loic Maisonnasse, Philippe Mulhem, Eric Gaussier. *Visual Language Model for Scene Recognition*. Singaporean-French IPAL Symposium (SinFra'09), 8 pages, Singapore, 2009.
  4. **Trong-Ton Pham**, Nicolas Maillot, Joo-Hwee Lim, Jean-Pierre Chevallet. *Latent Semantic Fusion Model for Image Retrieval and Annotation*. ACM 16th Conference on Information and Knowledge Management (CIKM), pages 439-444, Lisboa, Portugal, 2007.

THANK YOU

- Questions or comments ?

©Pixar