Travel Media Analysis from Multiple Perspectives

Wei-Ta Chu (朱威達)

wtchu@cs.ccu.edu.tw

Assistant Professor

Dept. of Computer Science and Information Engineering

National Chung Cheng University

Talk@Providence University 2009/10/02

Outline

Introduction

- People treasure travel experience, put it into memory, and want to efficiently manage or manipulate it.

	Modalities	**Facets**	**Functions**	**Correlation**	**Access manners**
Rep. selection	Video, photo	**What, where**	Browsing	Single modality	PC, PDA, mobile phone
ROI determination	Video, photo	**What, where**	Browsing	Single modality	PC, PDA, mobile phone
Face clustering	Video, photo	**Who**	Annotation, browsing, retrieval	Single modality	PC, PDA, mobile phone
Video scene detection	Video, photo	**Where**	Annotation, browsing	Multiple modalities	PC

Representative Selection and ROI Determination

Wei-Ta Chu (朱威達)

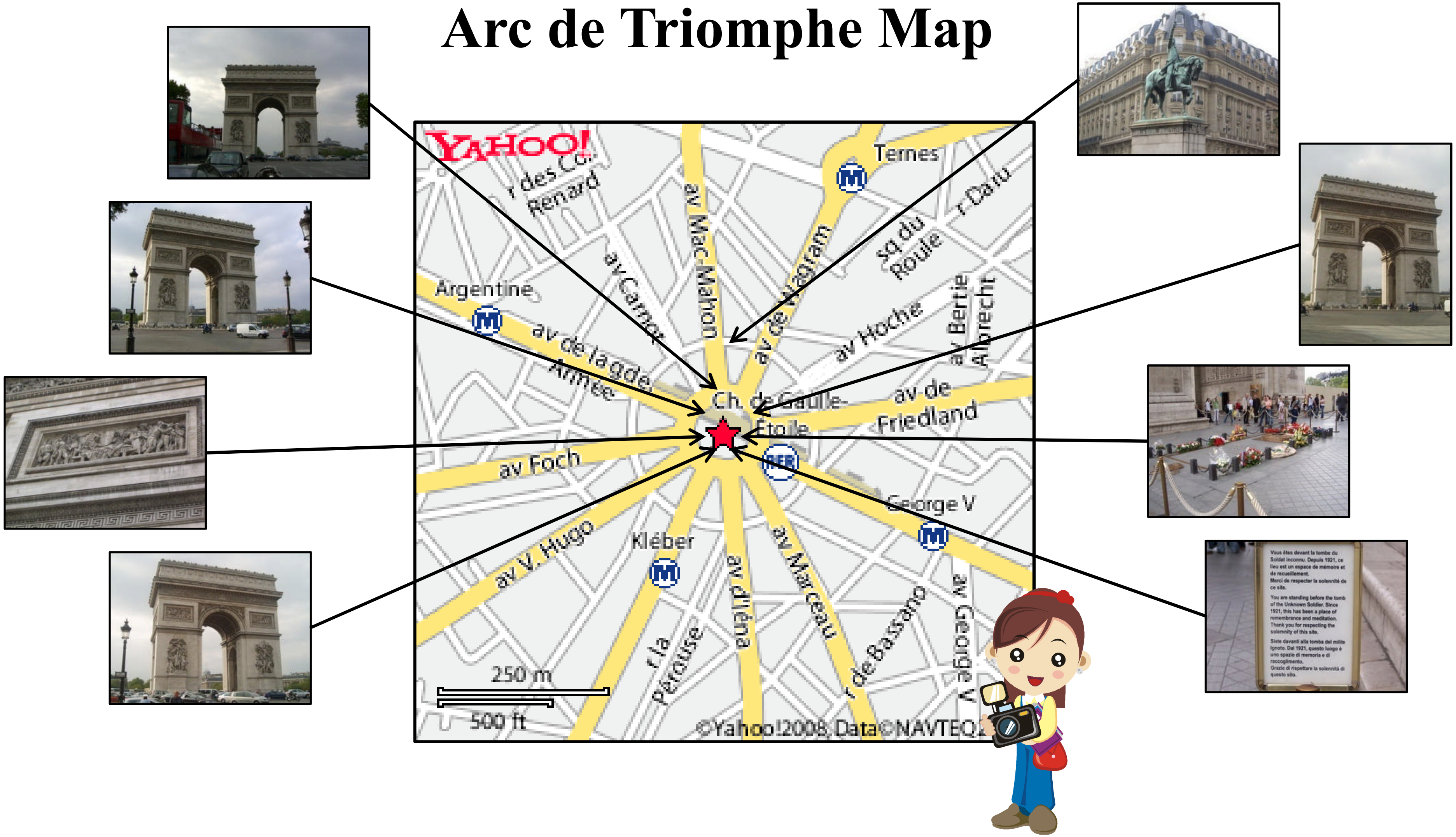
wtchu@cs.ccu.edu.tw

Assistant Professor

Dept. of Computer Science and Information Engineering

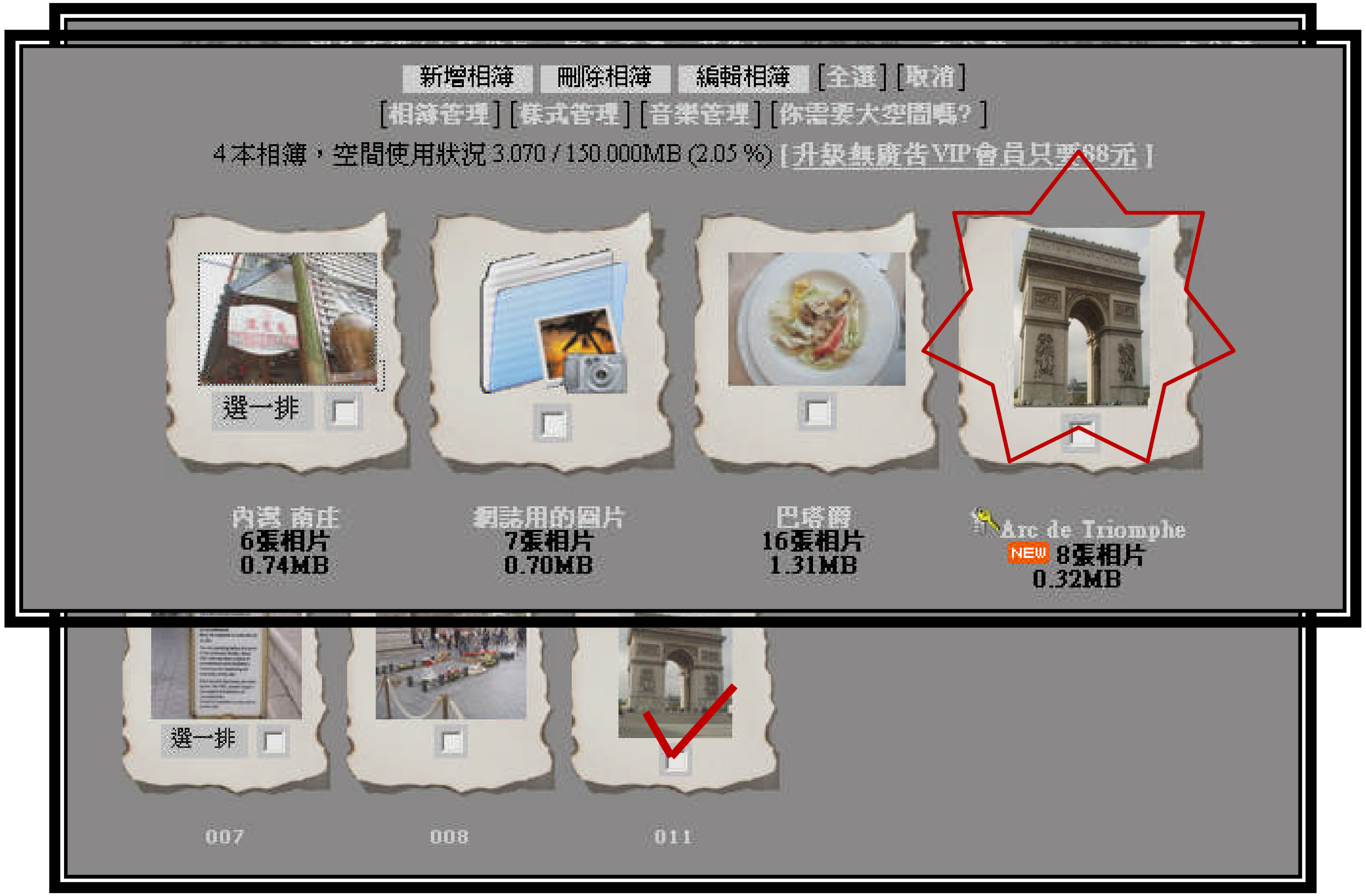
National Chung Cheng University

Motivation


Arc de Triomphe Map
YAHOO!
r des Co.
Renard
Ternes
r Daru
sq du
Roule
av de Wagram
av Carnot
av Mac-Mahon
Argentine
av de la gde.
Armée
av Hoche
av Bertie
Albrecht
Ch. de Gaulle-
Etoile
av de
Friedland
av Foch
RER
George V
av V. Hugo
Kléber
av d'Iéna
av Marceau
r de Bassano
av George V
r la
Pérouse
250 m
500 ft
©Yahoo!2008, Data©NAVTEQ2

Motivation

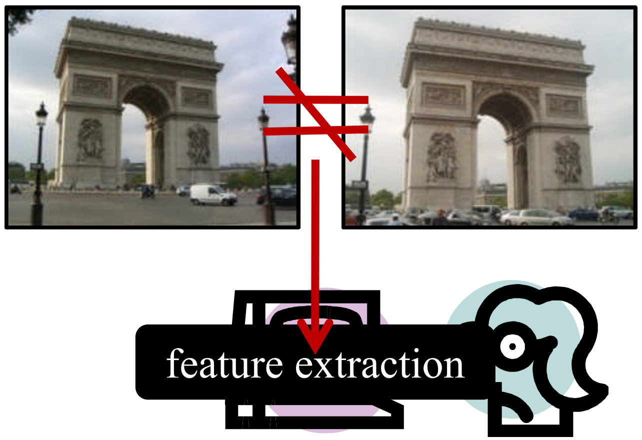
- Wretch album



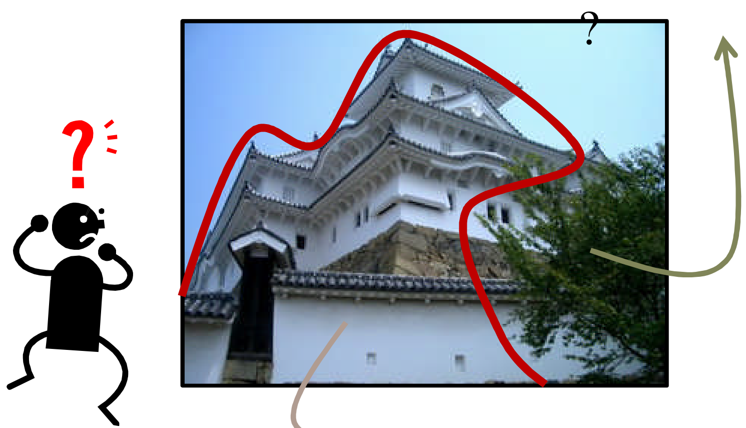
Goals & Challenges

Representative photo

A. Near-duplicate photos



B. Image segmentation



Framework

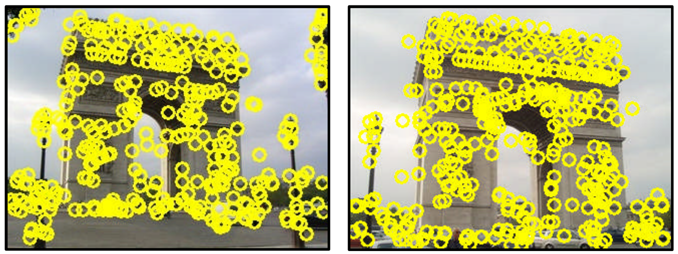
Feature Extraction

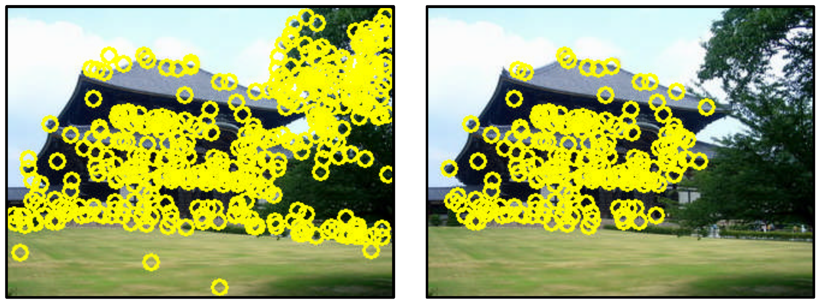
Image Dataset

SIFT Descriptor

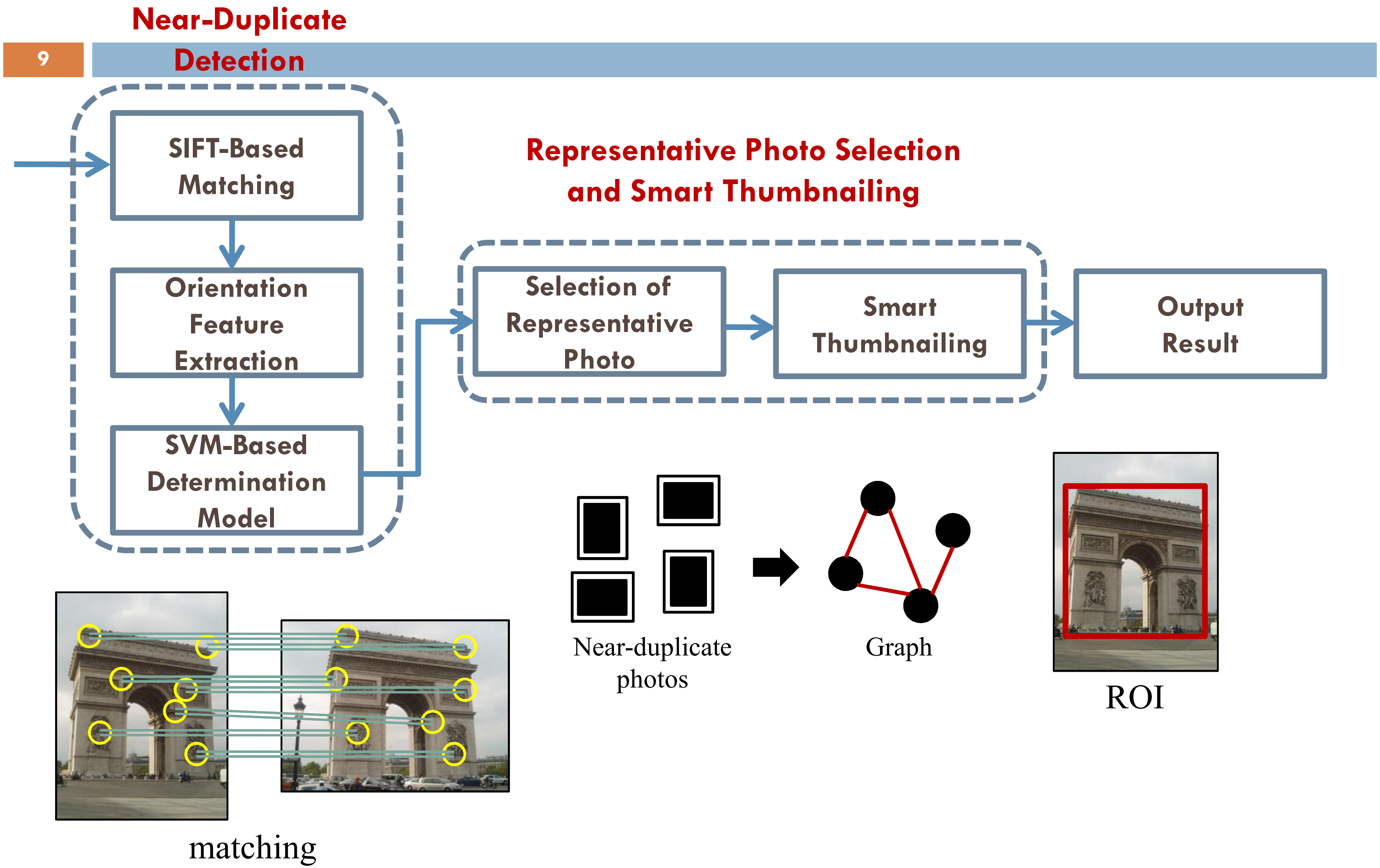
Feature Points Filtering

A. PLSA-Based
B. Point-Based
C. Region-Based





Framework



Near-Duplicate Detection
SIFT-Based Matching
Orientation Feature Extraction
SVM-Based Determination Model
Representative Photo Selection and Smart Thumbnailing
Selection of Representative Photo
Smart Thumbnailing
Output Result
Near-duplicate photos
Graph
ROI
matching

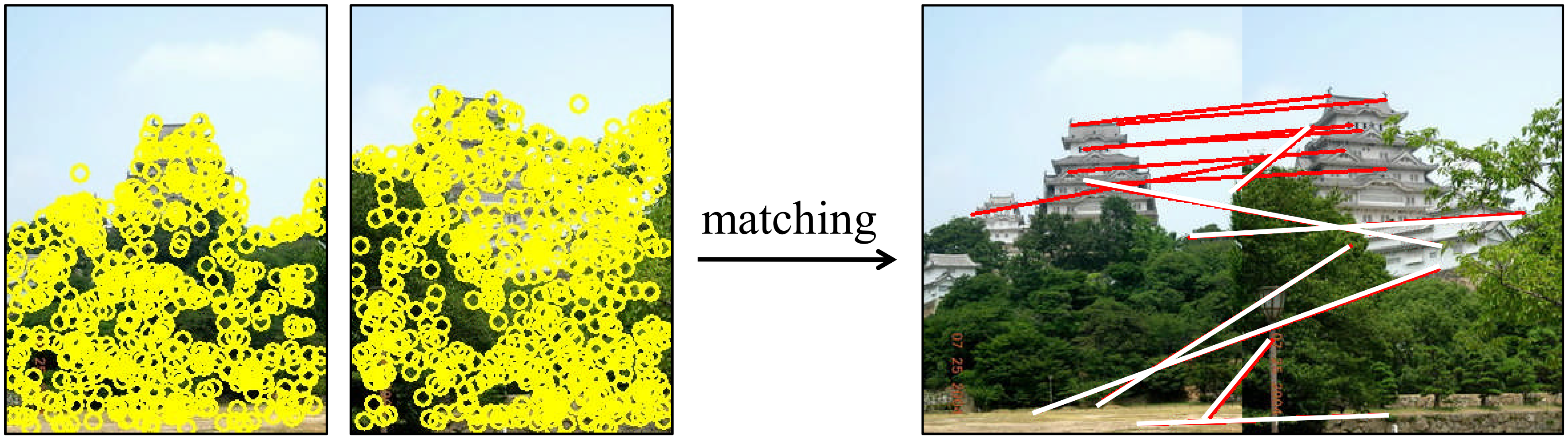
Feature Extraction

- Scale-Invariant Feature Transform (SIFT)
  - The difference of Gaussian (DoG) detector is applied to detect feature points.
  - Each feature point can be described by a 128-dimensional orientation histogram.
- The advantage of SIFT descriptor
  - SIFT feature is invariant to scale and rotation.
  - SIFT descriptor is robust to color change, brightness, and contrast.

D. Lowe, "Distinctive Image Feature from Scale-Invariant Keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

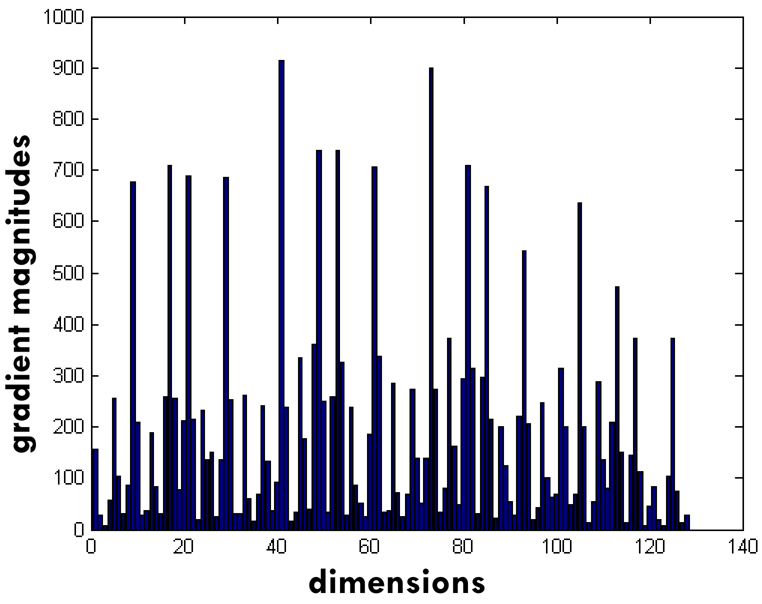
Feature Filtering

- The large number of noisy feature points diminish the performance of near-duplicate detection.
- Three feature filtering methods are proposed, including PLSA-based, point-based, and region-based.

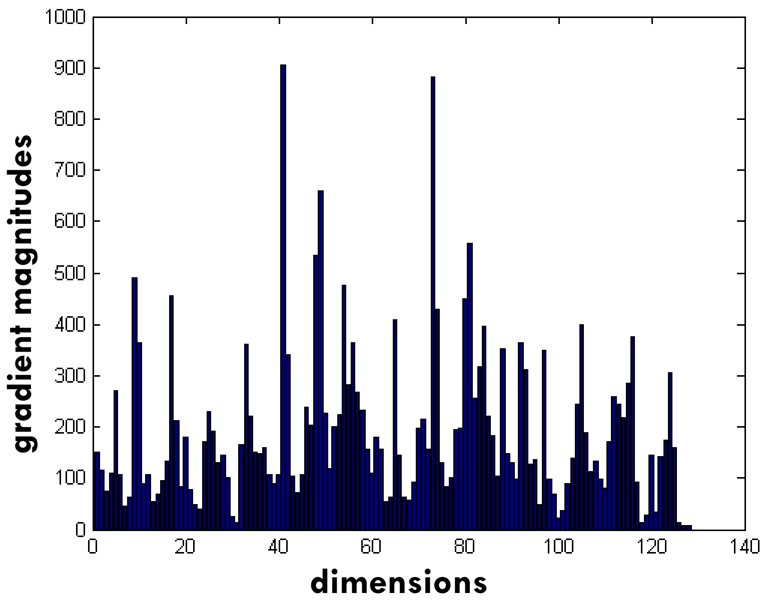


Point-Based Feature Filtering

- Artificial objects often have geometric structure, while natural scenes have relatively random structure.



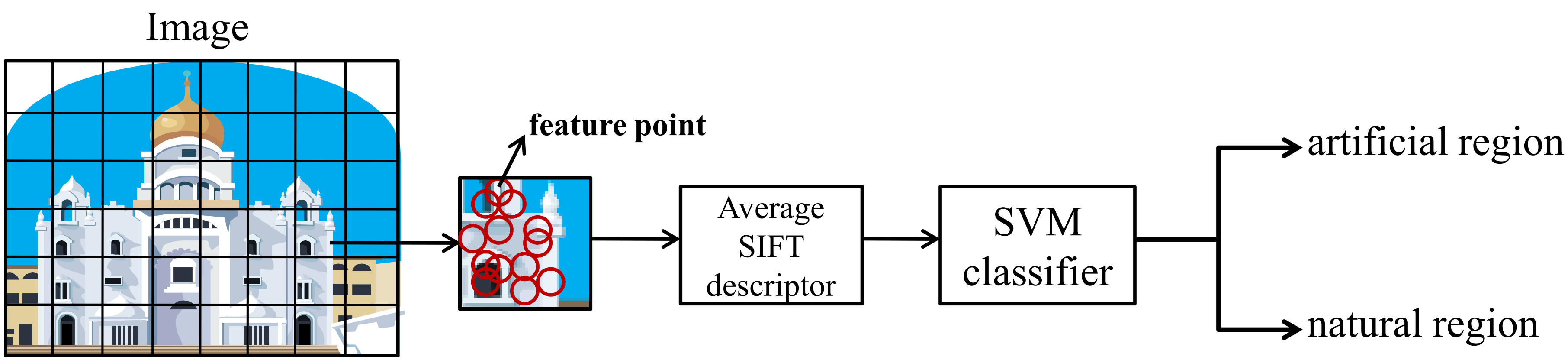
(A) 1000 artificial feature points



(B) 1000 natural feature points

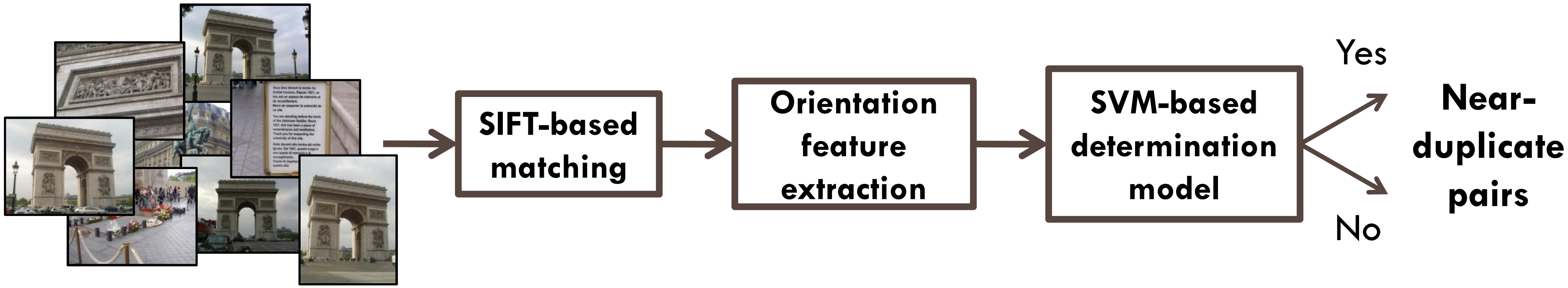
Region-Based Feature Filtering

- Point-based filtering doesn't consider the spatial correlation between feature points in neighborhood.
- We divide each image into several regions, which are represented by the average descriptor of each region.



Near-Duplicate Detection Process

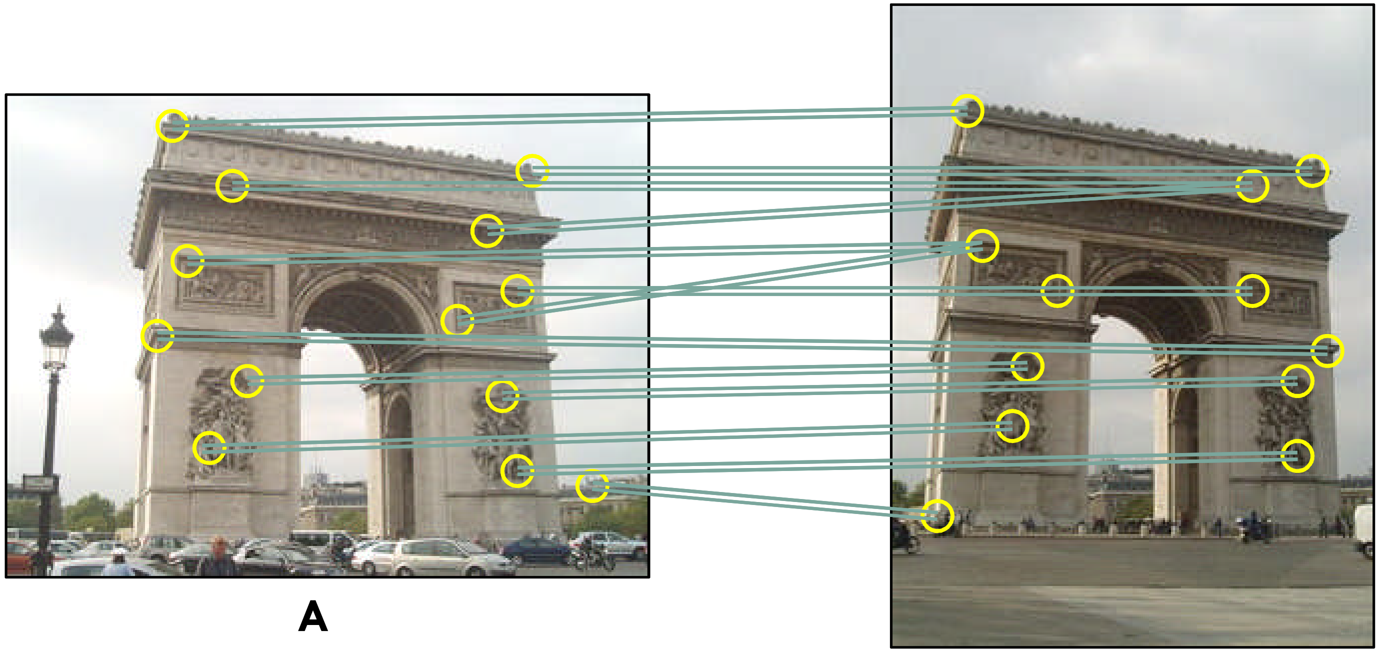
- SIFT-based matching
- Orientation feature extraction
- SVM-based determination model



W.-L. Zhao, C.-W. Ngo, H.-K. Tan, and X. Wu, “Near-Duplicate Keyframe Identification with Interest Point Matching and Pattern Learning,” IEEE Trans. on Multimedia, vol. 9, no. 5, pp. 1037-1048, 2007.

SIFT-Based Matching

- We utilize one-to-one symmetric matching to filter out false matches.



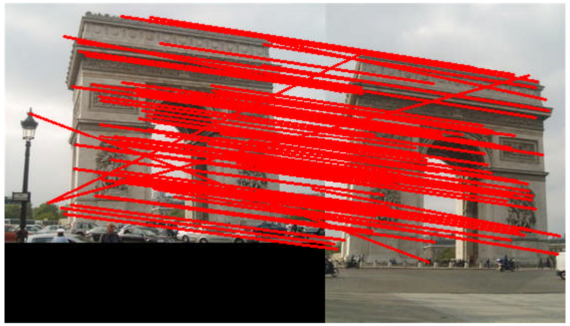
A

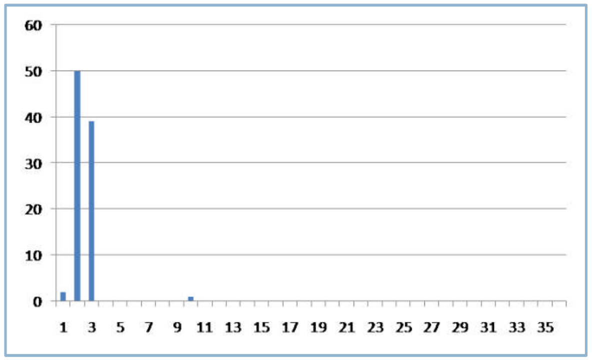
B

One-to-One Symmetric Matching

Orientation Feature Extraction

- We compute angles between matched lines and the horizontal axis, and quantize them into 36 bins, with a step of 5 ° from 0 ° to 180 °.

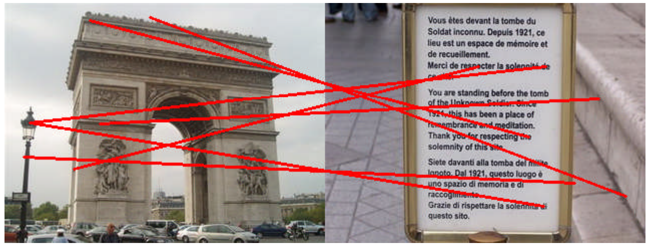


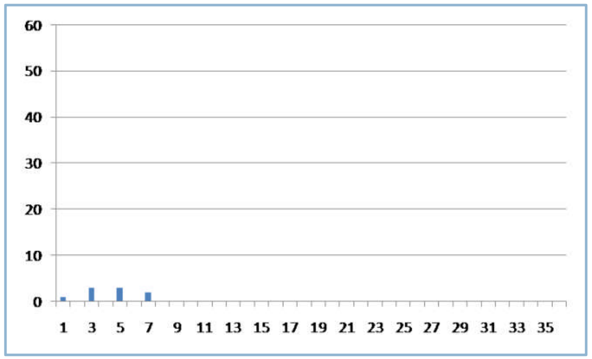


number of matched line

quantized orientation

(A) Near-Duplicate





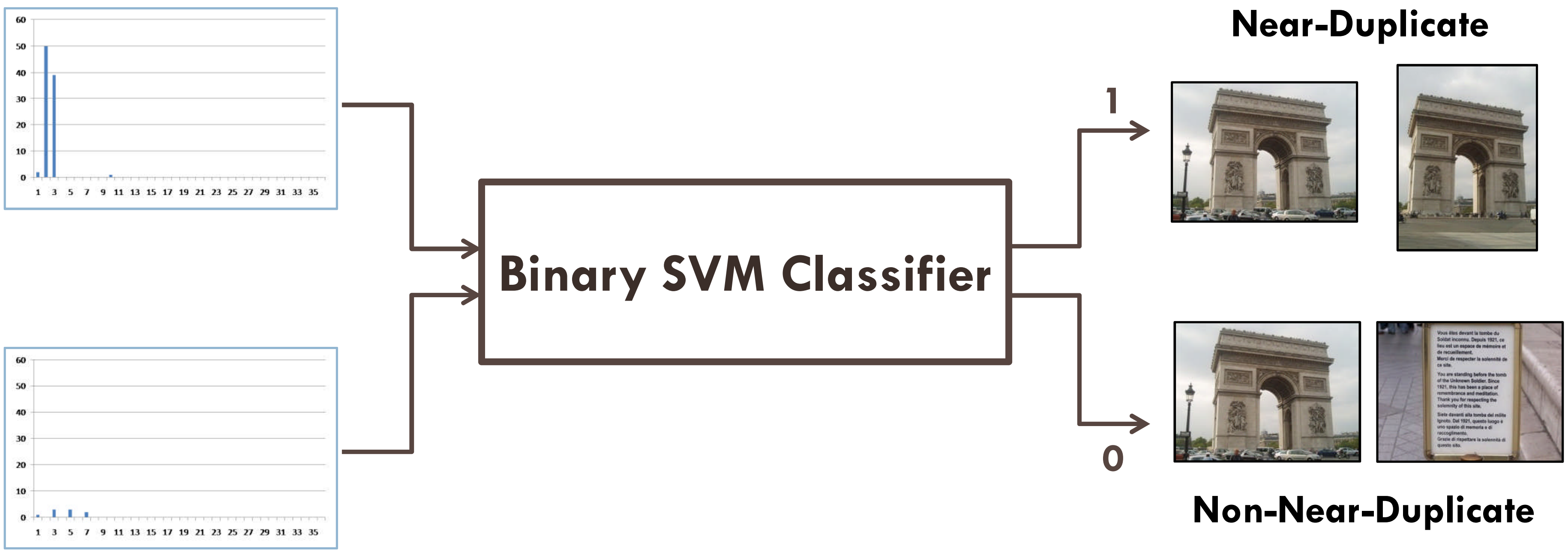
number of matched line

quantized orientation

(B) Non-Near-Duplicate

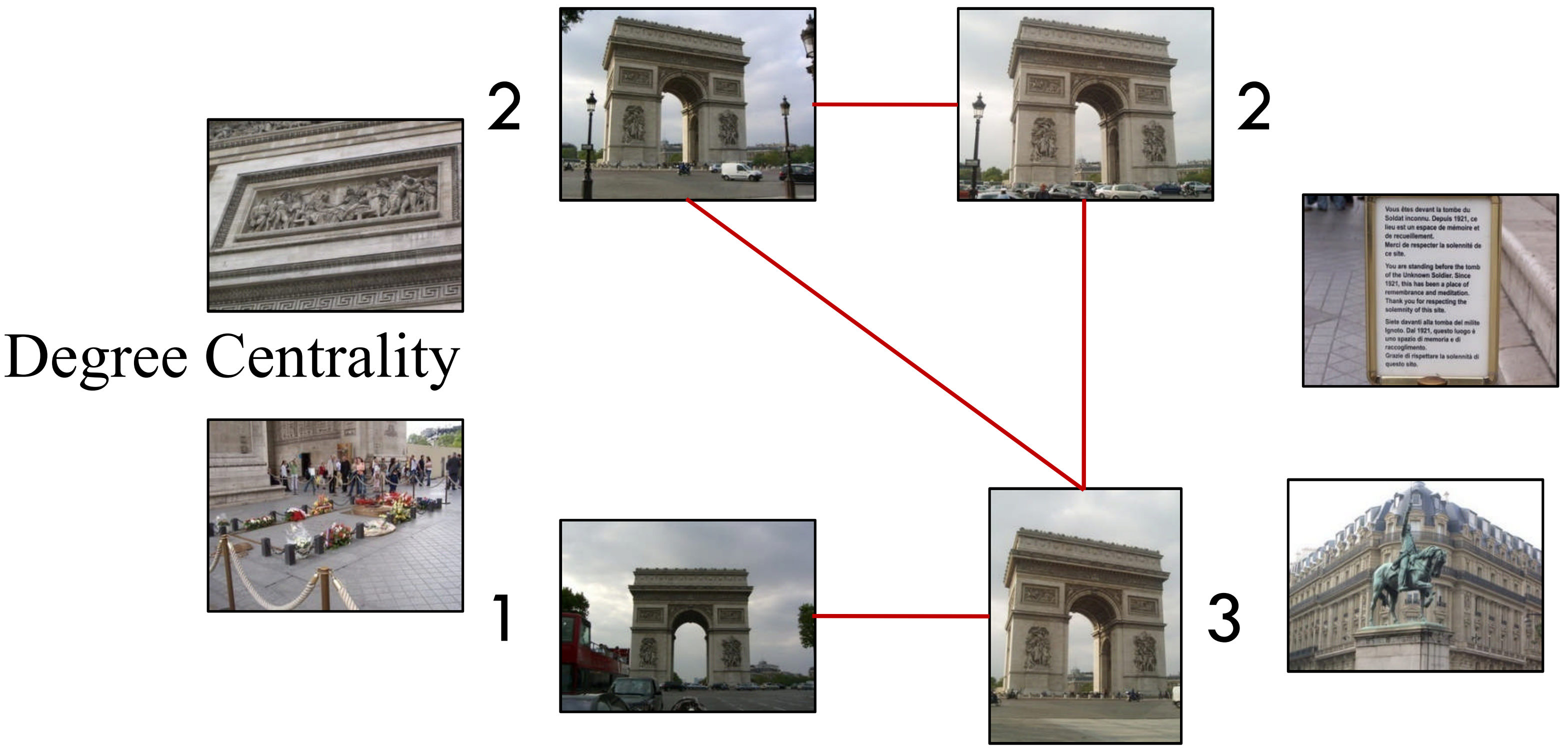
SVM-Based Determination Model

- A binary SVM classifier is used to model the characteristics of orientation histograms.



Selection of Representative Photo

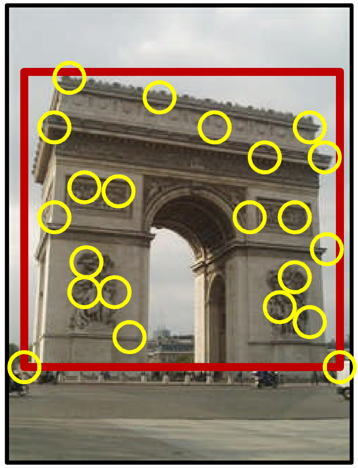
- The relationships between near-duplicate photos are represented as a graph.



Smart Thumbnailing

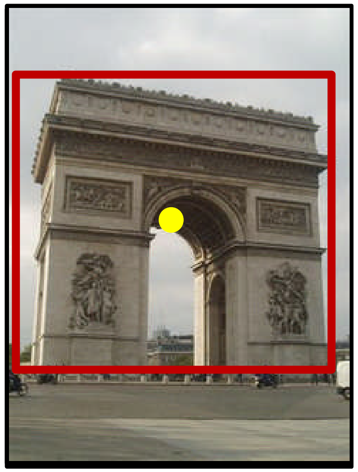
Our approach

- We exploit spatial distribution of matched feature points in the representative photo to find the most prominent region.



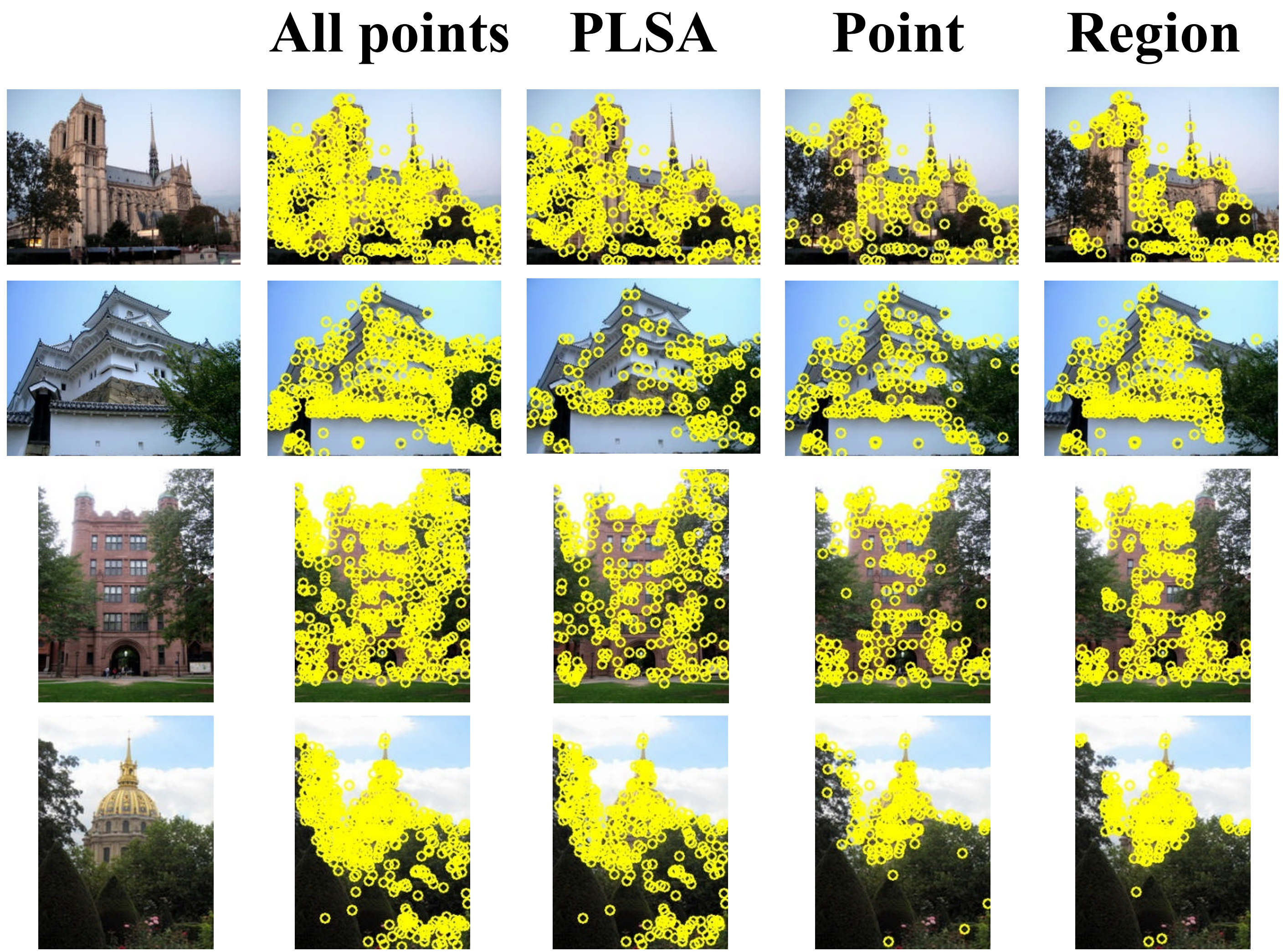
Saliency-based approach

- Saliency values reflect the visual stimuli to human vision system, such as color contrast, intensity contrast, and orientation contrast.

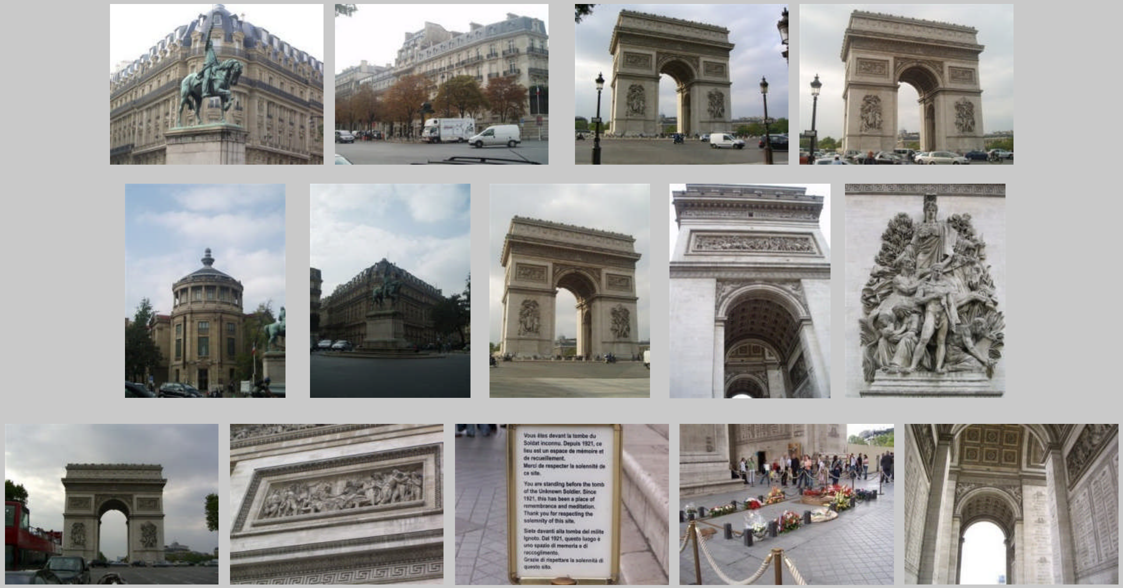


D. Walther and C. Koch, "Modeling attention to salient proto-objects," Neural Networks, pp. 1395-1407, 2006.

Performance of Feature Classification



Performance of Representative Selection

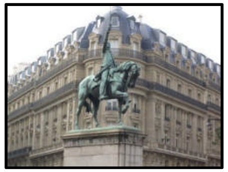
Scenic spot	Representative photo	Original photo set
Arc de Triomphe		

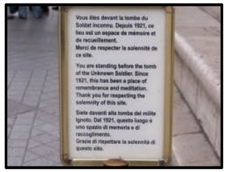
Performance of Representative Selection

Scenic spot	Representative photo	Original photo set
Statue of Liberty		

Performance of Representative Selection

- Guidelines of giving scores to each photo.





Score	Description
5	The image shows the most representative object you know for this scenic spot.
4	Although the most representative object shows on the image, it’s not good in shooting angles or in lighting conditions.
3	Although the image doesn’t show the most representative object, some other buildings or specific objects are shown.
2	There are specific objects without specific topic in this image.
1	I totally don’t know the purpose of this image.

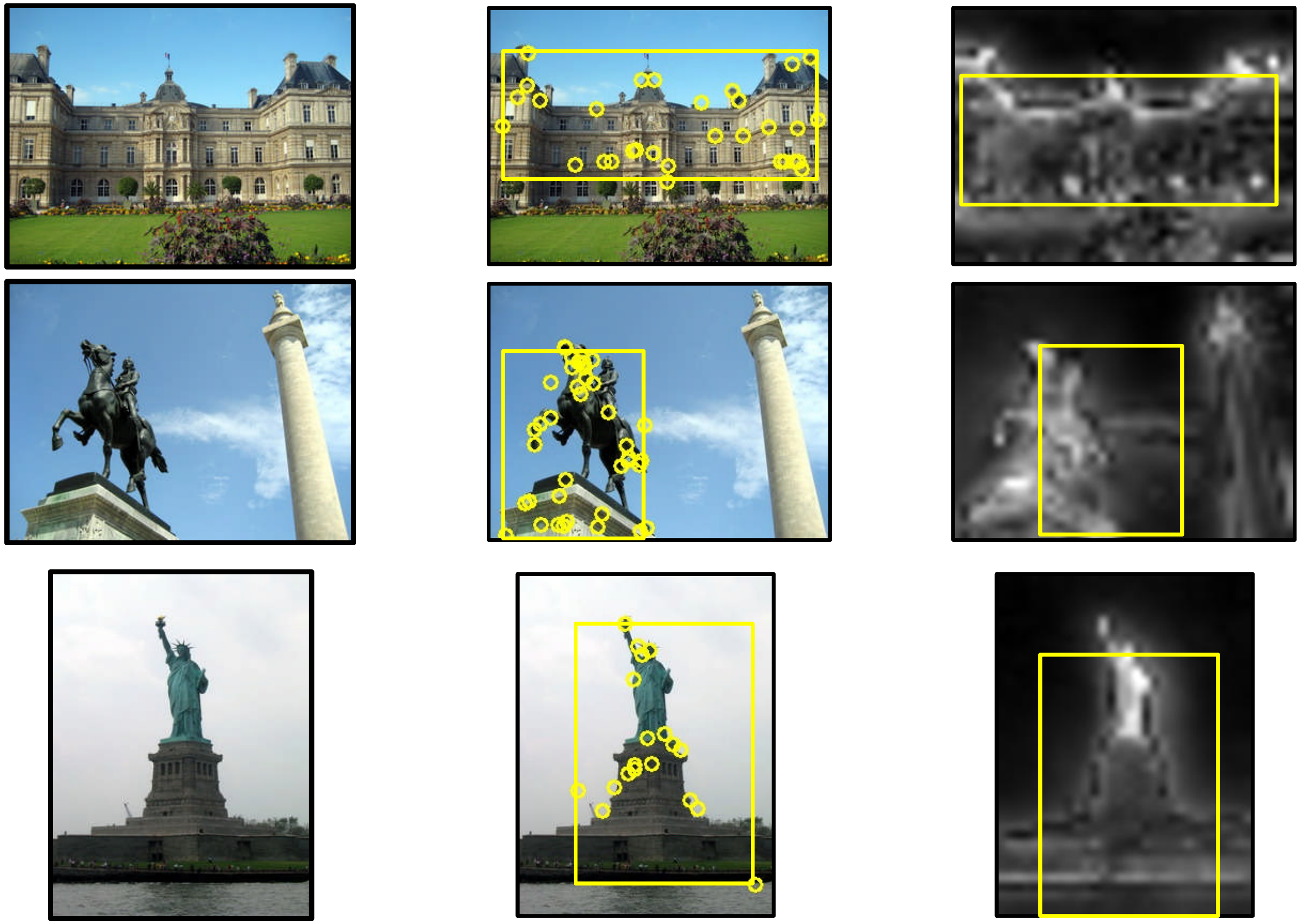
Performance of Representative Selection

- Fifty-two scenic spot photo sets

Scenic	Without Filtering	PLSA –based Filtering	Point-based Filtering	Region-based Filtering
Arc de Triomphe	4.89	5	5	4.78
State of Liberty	4.33	2.22	4.67	4.56
Luxembourg	4.22	4.44	3.22	4
Time Square	4	3.78	4.11	4
. . .	. . .	. . .	. . .	. . .
Rokuon-ji	4.33	1.67	3	4.11
Westminster	3.78	2.67	3.89	4.22
Overall	**3.58**	**3.32**	**3.61**	**3.63**

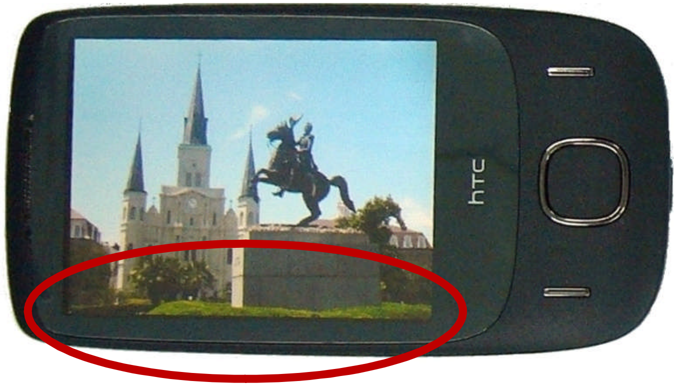
Performance of ROI Determination

Original **Region-Based** **Saliency Map**

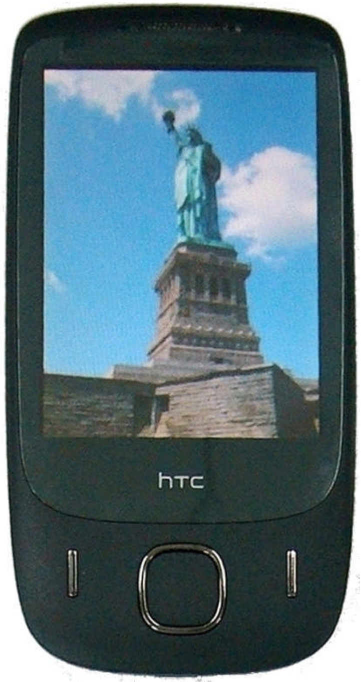


Displaying ROIs on Mobile Devices

Original



ROI

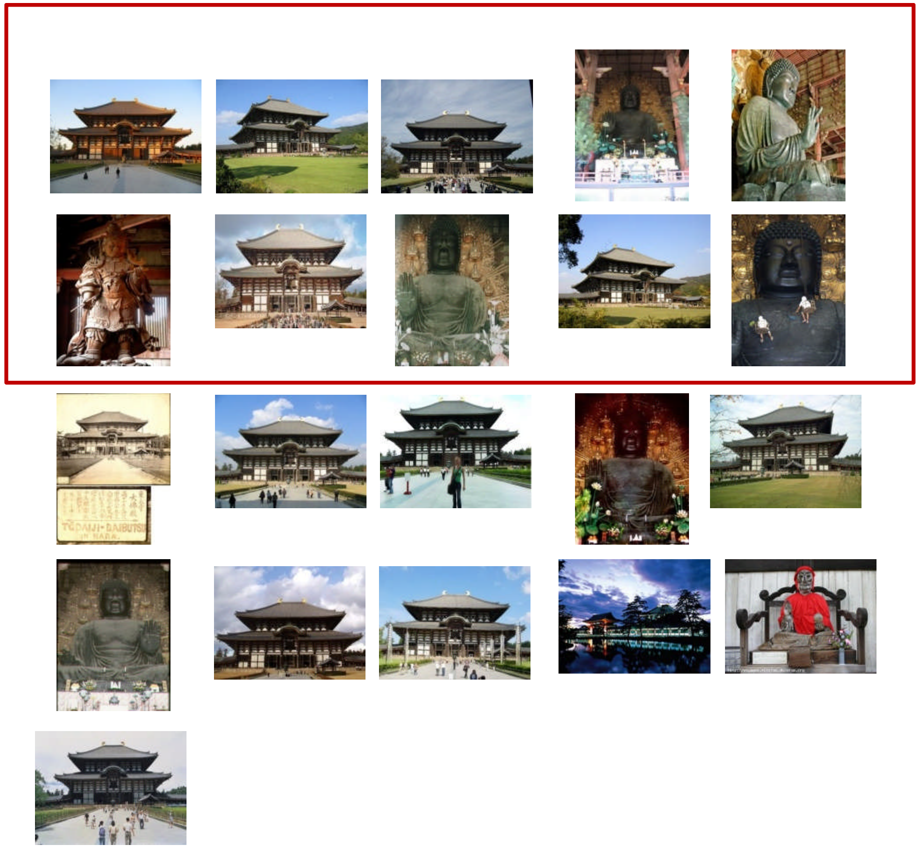


Original

ROI

Image Re-ranking

Google image search results of Tadaji



Re-ranking results

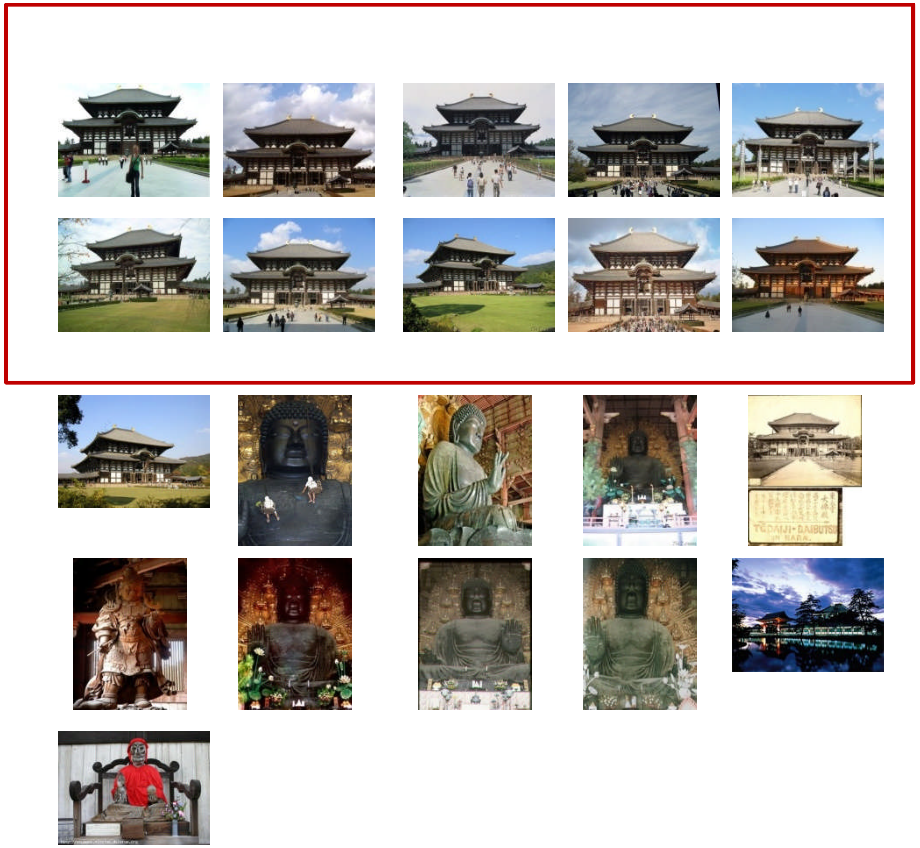


Photo Summarization

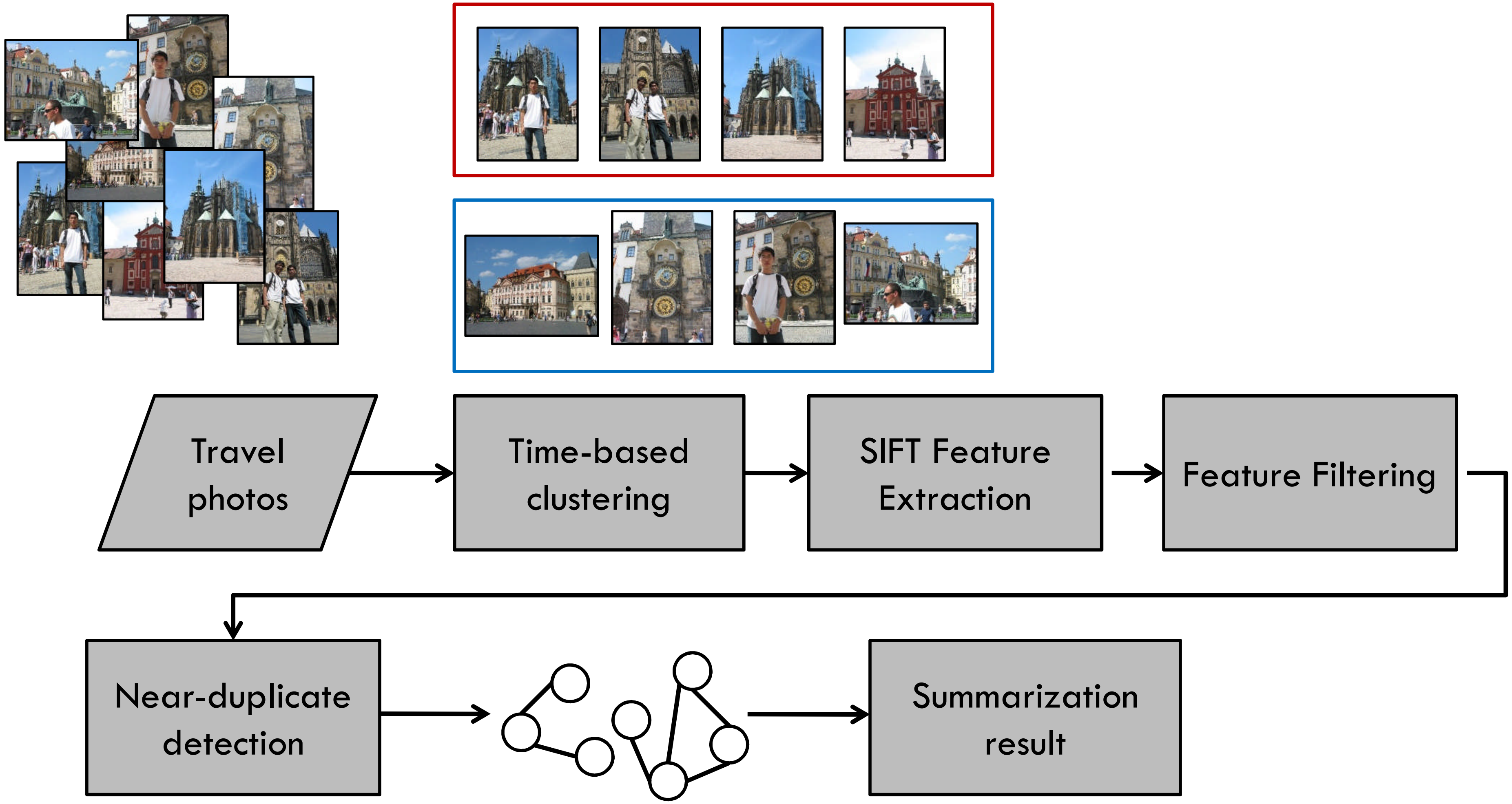
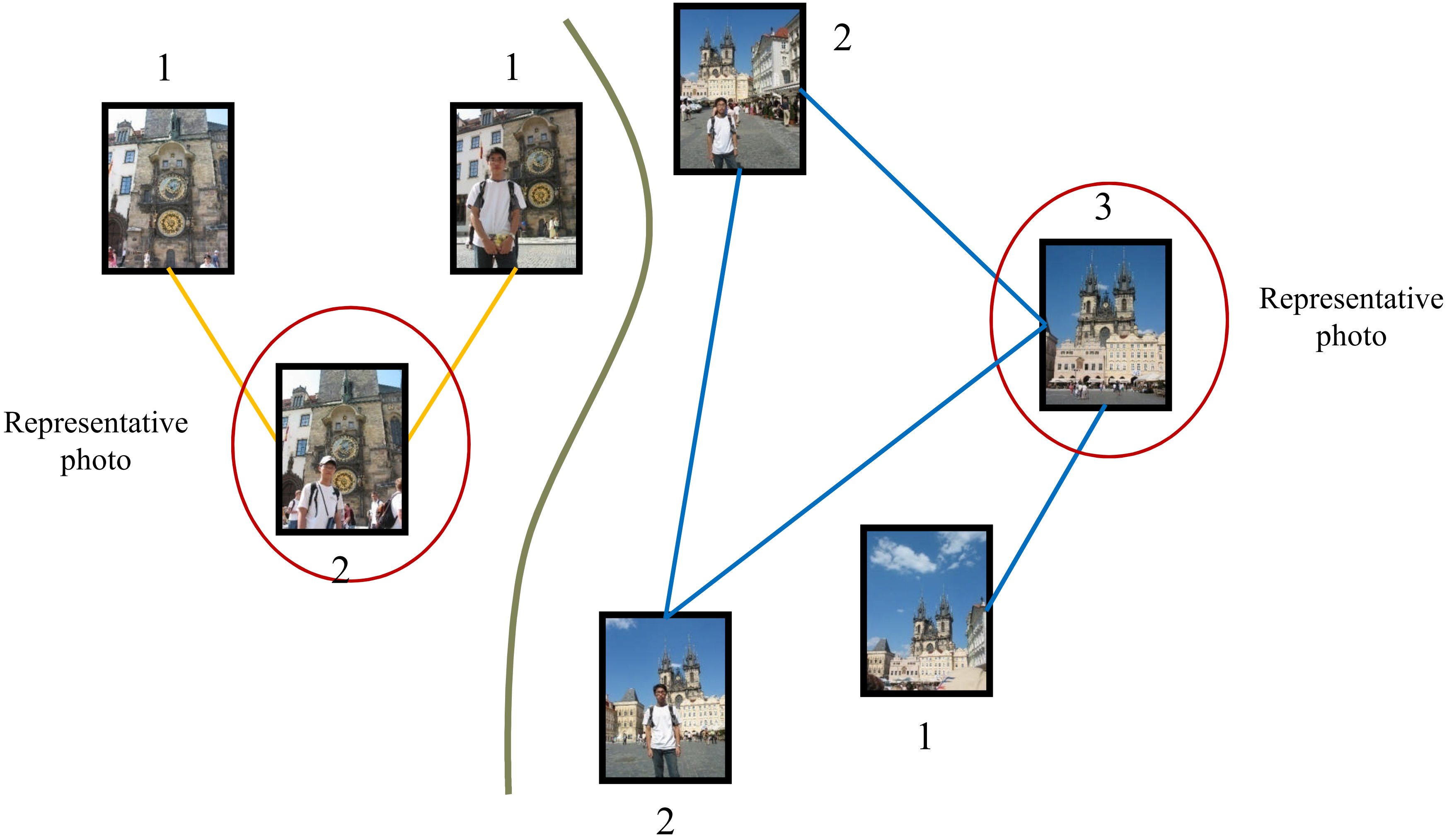


Photo Summarization



Summary

- We propose an approach to facilitate users to efficiently manage and browse photos.
- Three feature filtering methods are developed, and the region-based one is found to be the most effective.
- We utilize the spatial distribution of matched feature points to determinate ROI, and demonstrate that it's better than conventional saliency-based approaches.

Face Clustering in Consumer Photos

Wei-Ta Chu (朱威達)

wtchu@cs.ccu.edu.tw

Assistant Professor

Dept. of Computer Science and Information Engineering

National Chung Cheng University

Motivation

- With the thriving market of digital cameras, people could record their daily life with ease.
- How do people manage their digital photographs?

"To be more useful in this domain,
Content-Based Image Retrieval
would need to give more meaningful results,
for example by providing face recognition."

~ K. Rodden, K. R. Wood.
How do people manage their digital photographs?. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 409-416, 2003.

Challenges

- Large variations in lighting

Challenges

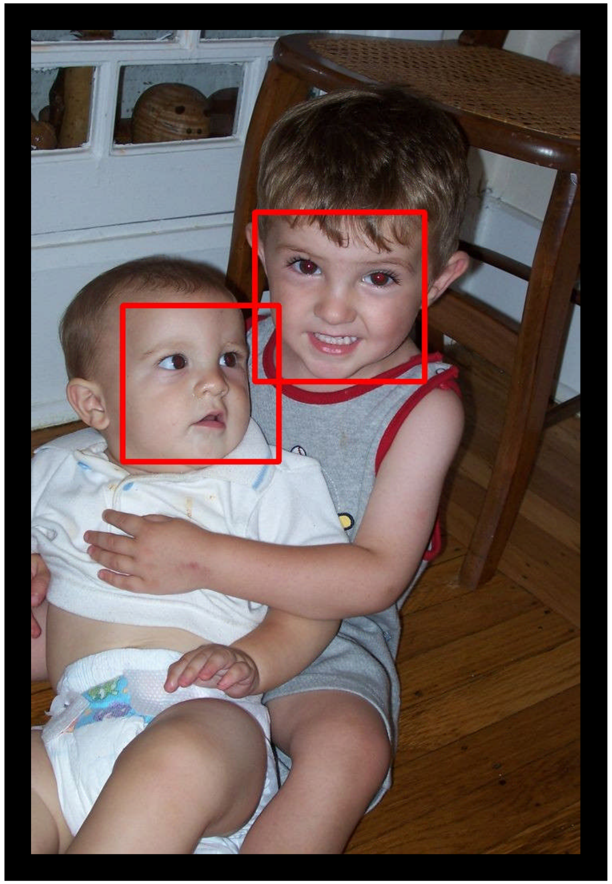
- Large variations in scale, pose and expression

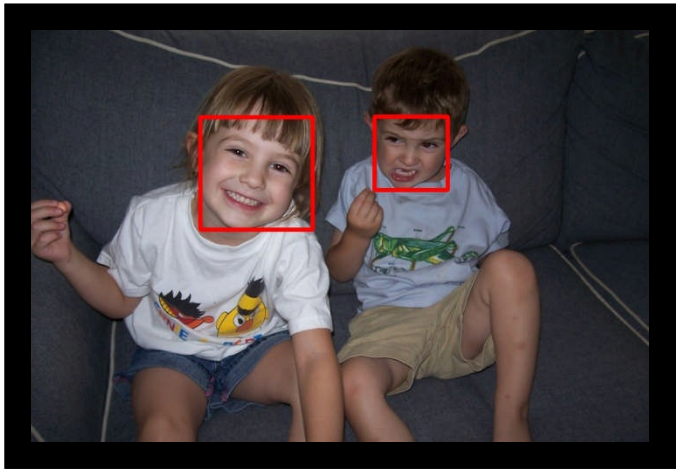
Challenges

- Occlusions

Face Detection

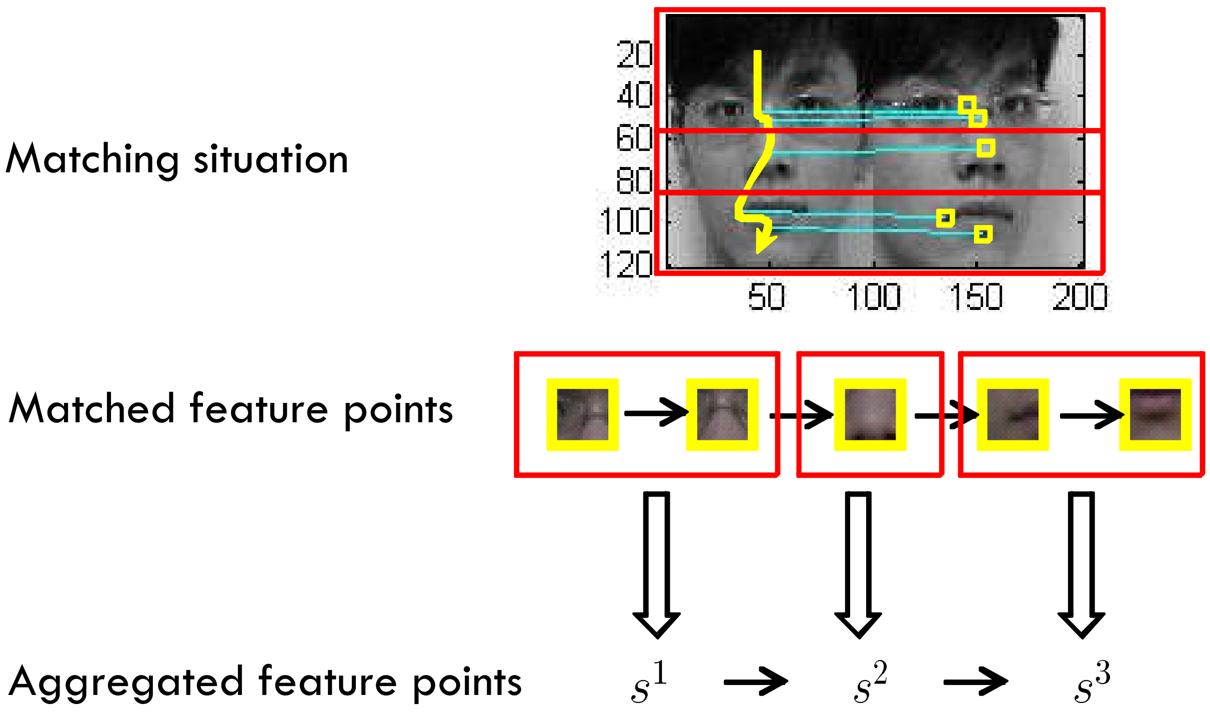
- We use AdaBoost-Based method to detect faces





Face Matching based on Local Feature Points

- We exploit SIFT (Scale-Invariant Feature Transform) features to match face images.
- *We leverage visual language models to describe matching situations between two face images rather than a face itself.*



Representation of Face Matching Situation

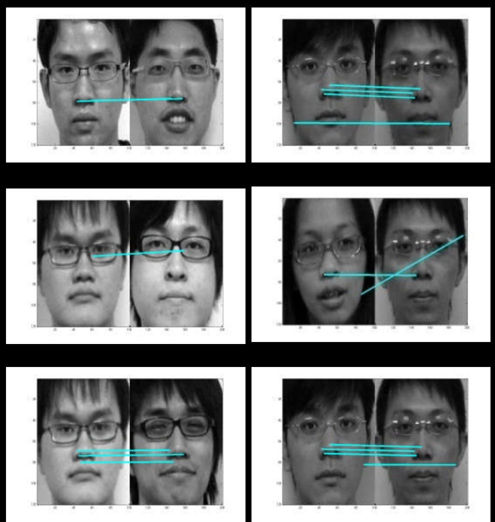
- Visual Vocabulary Construction
  - Based on the AT&T face database, we collect aggregated feature points from 40 different people with distinct face features.
  - 1800 pairs of faces are conducted. For a pair of faces, we finally obtain three aggregated feature points.
  - The k-means algorithm is applied to cluster similar features into groups, where each group represents a visual word.

Sivic, J. and Zisserman, A. 2003. Video Google: A text retrieval approach to object matching in videos. In Proc. of ICCV, 2, pp. 1470-1477.

Representation of Face Matching Situation

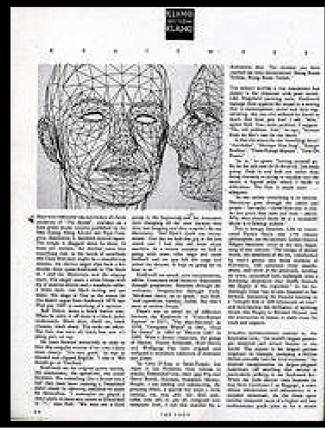
- Visual sentence

The matching situations between faces of two distinct individuals



Visual Vocabulary

A set of visual sentence of matching situations between faces of different individuals



Training of Visual Language Model

- An introduction to Language Model
  - The visual word proximity of a matching situation is measured by the following probability form.

$$p(v_k|v_1v_2\ldots v_N) = p(v_k|v_1v_2\ldots v_{k-1})$$

  - Example:

p(中正大學) = p(中) p(正 | 中) p(大 | 中正) p(學 | 中正大)

p(國立中正大學資訊工程學系) = ?

Training of Visual Language Model

- To further simplify this conditional probability, techniques of conventional language model suggest that each word only depends on its immediate neighbors, called n-gram.

- Unigram

$$P(中正大學) = P(中)\,P(正)\,P(大)\,P(學)$$

- Bigram

$$P(中正大學) = P(中)\,P(正\,|\,中)\,P(大\,|\,正)\,P(學\,|\,大)$$

- Trigram

$$P(中正大學) = P(中)\,P(正\,|\,中)\,P(大\,|\,中正)\,P(學\,|\,正大)$$

Training of Visual Language Model

- We make the following assumptions in model construction.
  - Each visual word in the same visual sentence is correlated.
  - The dependency between visual words is generated from top to bottom.
- The first visual language model (VLM) describes the matching situations between faces of the same individuals, while the second visual language model describes the matching situation between two distinct individuals.

Wu, L., Li, M., Li, Z., Ma, W.-Y., and Yu, N. 2007. Visual language modeling for image classification. In Proc. of MIR, pp. 115-124.

Training of Visual Language Model

- Unigram construction

$$p(v_k|C_i) = \frac{\mathrm{Count}(v_k|C_i)}{\sum_{v \in V} \mathrm{Count}(v|C_i)}, \quad i = 1, 2.$$

- Bigram construction

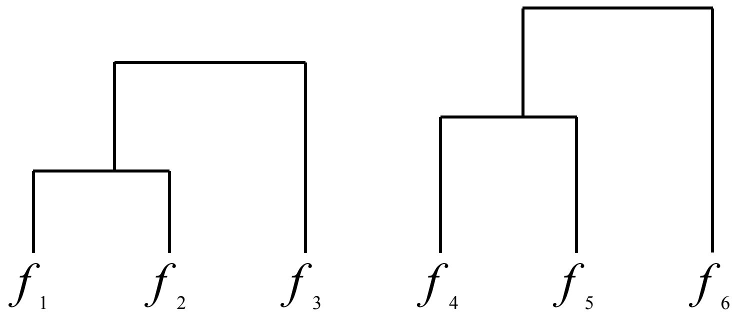
$$p(v_k|v_{k-1}, C_i) = \frac{\mathrm{Count}(v_{k-1}v_k|C_i)}{\mathrm{Count}(v_{k-1}|C_i)}, \quad i = 1, 2.$$

- Trigram construction

$$p(v_k|v_{k-2}v_{k-1}, C_i) = \frac{\mathrm{Count}(v_{k-2}v_{k-1}v_k|C_i)}{\mathrm{Count}(v_{k-2}v_{k-1}|C_i)}, \quad i = 1, 2.$$

Face Clustering Using VLM

- An example of bottom-up clustering algorithm to cluster faces



- Face likelihood ratio for a pair of faces
- Modified Hausdorff distance for two face clusters

Face Clustering Using VLM

- Face likelihood ratio for a pair of faces

$$r_{i,j} = \frac{p(S_{i,j}|M_1)}{p(S_{i,j}|M_2)}$$

Where
- $S_{i,j}$ is the visual sentence representing the matching situation between f_i and f_j
- M_1 is the visual language describing matching situations between the same individual's faces
- M_2 describes matching situations between different individuals' faces

- Modified Hausdorff distance for two face clusters

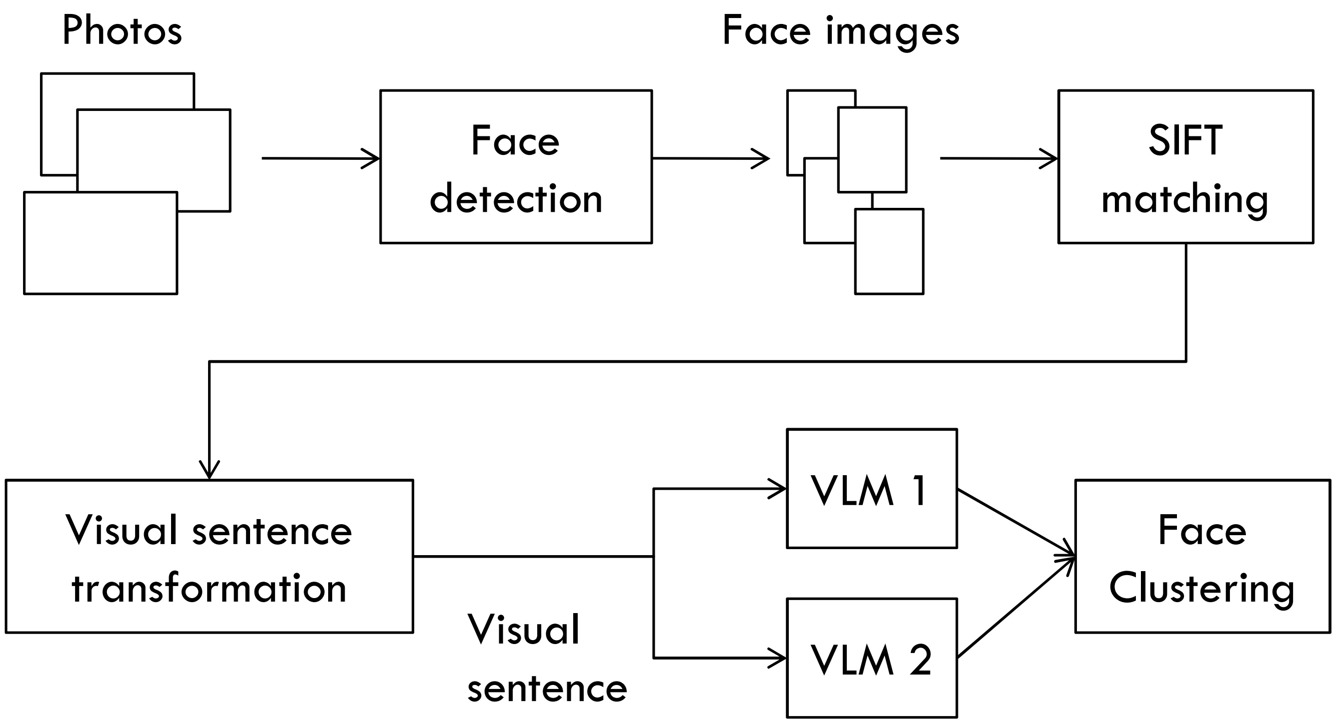
$$\mathcal{H}(F_i, F_j) = \max(h(F_i, F_j), h(F_j, F_i))$$

$$h(F_i, F_j) = \frac{1}{|F_i|} \sum_{f_p^{(i)}} \min_{f_q^{(j)}} \boxed{(1 - p(S_{p,q}|M_1))}$$

distance

Summary of VLMs for Face Clustering

- Face clustering based on Visual Language Models



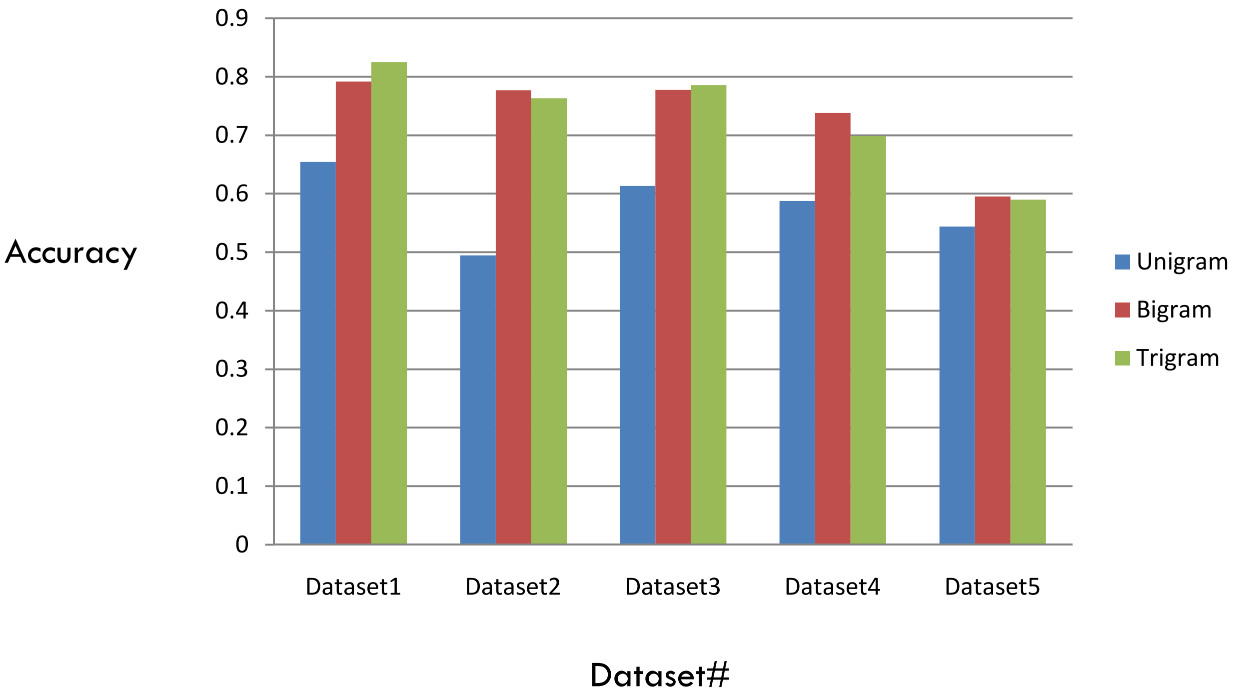
Experiments

- Choices of VLM
  - For testing, we evaluate Unigram, Bigram and Trigram by clustering five sets of face images.

Test dataset		# face images	# clusters	Description
1	AT&T	400	40	sLV, sEV, sPV
2	Lab faces	368	10	sLV, sEV, sPV
3	Lab daily	89	7	sLV, lEV, lPV
4	A family	42	5	lLV, sEV, lPV
5	B family	56	4	lLV, lEV, lPV

LV: lighting variation; EV: expression variation; PV: pose variation; s: slight, l: large.

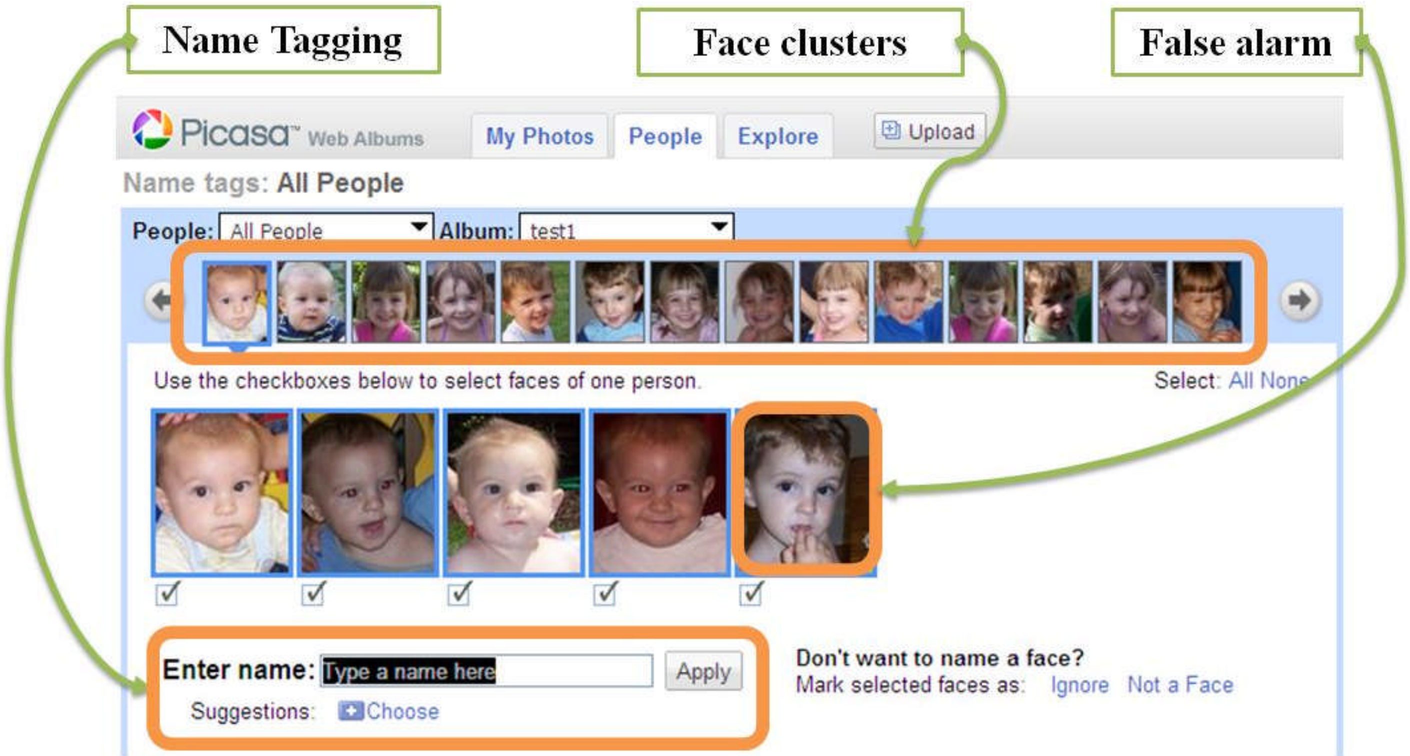
Experiments



The average accuracy values over these five datasets for unigram, bigram, and trigram models are 0.58, 0.74, and 0.73

Experiments

- Observation from Google Picasa

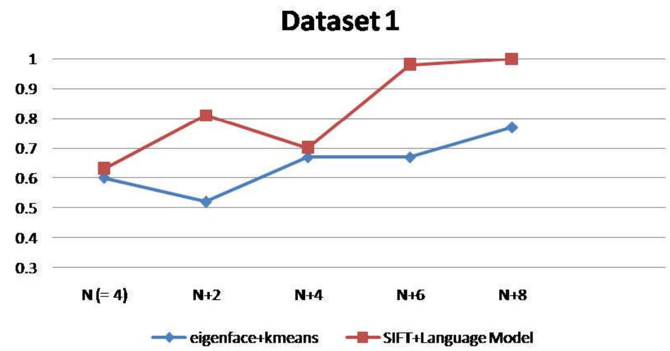


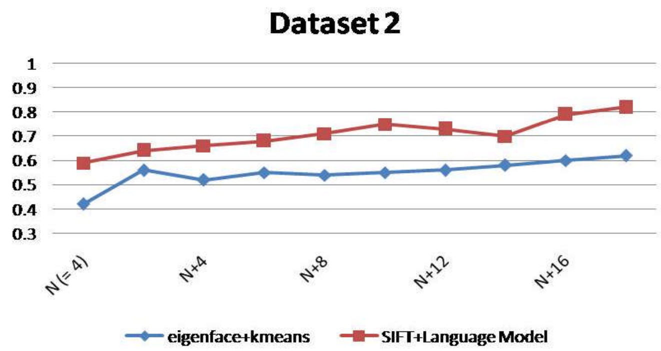
Goal: **Lower clustering numbers**

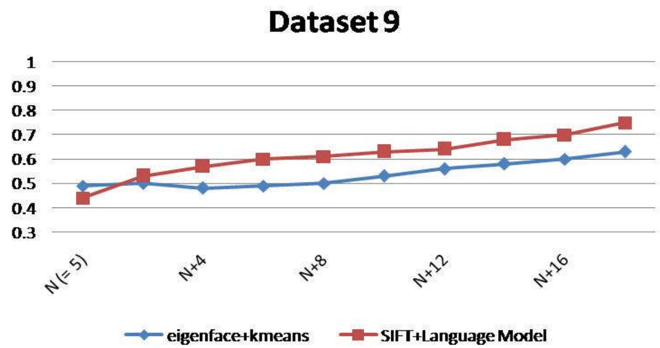
Higher clustering accuracy

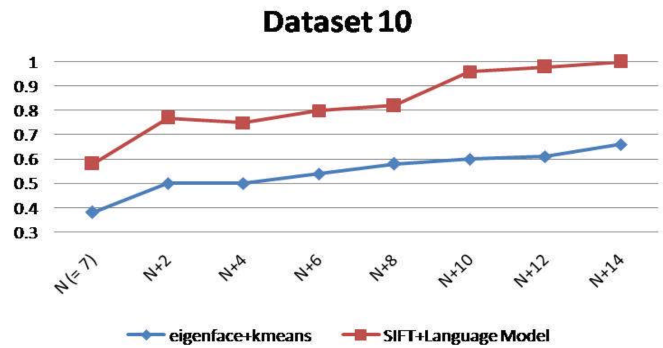
Experiments

- Face clustering performance evaluation
  - There are 16 datasets containing totally 1199 images.
  - The number of persons in a dataset range from two to seven.









Experiments

- To quantitatively measure the clustering performance, we calculate a ratio by considering the number of face clusters when a specific clustering accuracy is achieved:

$$R = |F_{eig}|/|F_{VLM}|$$

  - where $|F_{eig}|$ and $|F_{VLM}|$ are the numbers of face clusters obtained by the eigenface approach and our method that first time achieve at least 80% face clustering accuracy.

- After evaluating the 16 datasets, we finally get the average ratio $\bar{R} = 1.58$

Summary

- A new viewpoint is proposed to effectively address face clustering for consumer photos.
- We elaborately transform matching situations between faces into visual sentence representation, and construct visual language models to describe the dependency of different parts of faces.
- Based on the probabilistic framework, an agglomerative clustering approach is used to group the same individual's faces into the same cluster.
- The experimental results demonstrate superior performance.

Scene Detection in Travel Videos

Wei-Ta Chu (朱威達)

wtchu@cs.ccu.edu.tw

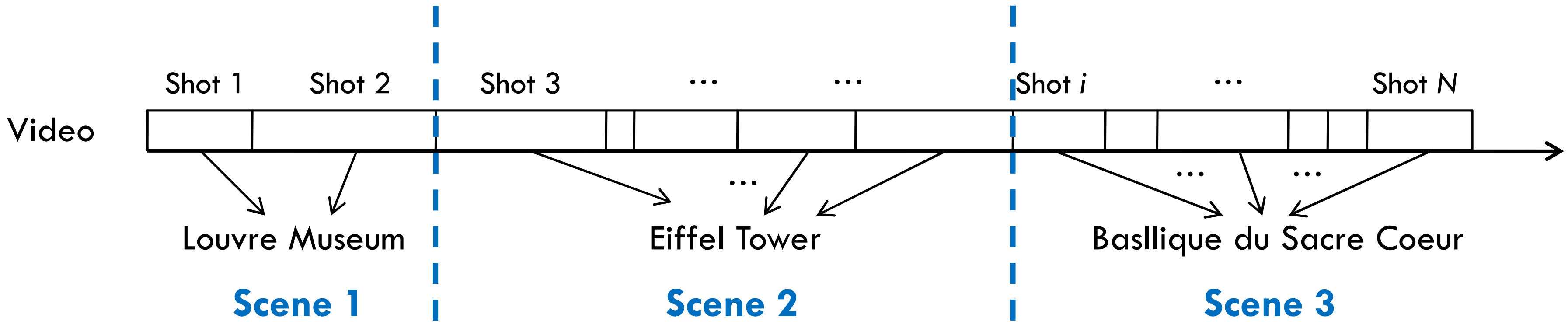
Assistant Professor

Dept. of Computer Science and Information Engineering

National Chung Cheng University

Introduction

- People get used to record daily life and travel experience by digital cameras and camcorders.
- We focus on videos captured in journeys, and address the problem of scene detection.
- A scene in travel videos means *a cluster of video shots that correspond to a scenic spot*.



Challenges

- Travel videos captured in uncontrolled environments are often suffered from annoying effects.

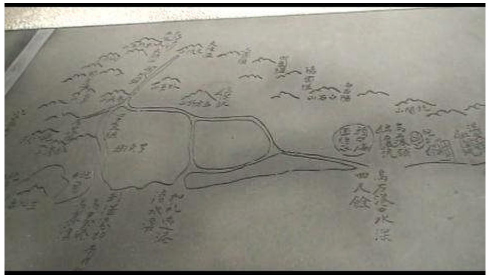
Overexposure

Underexposure

Hand shaking

- There is no clear structure in travel videos.

Hand Cover



Unknown

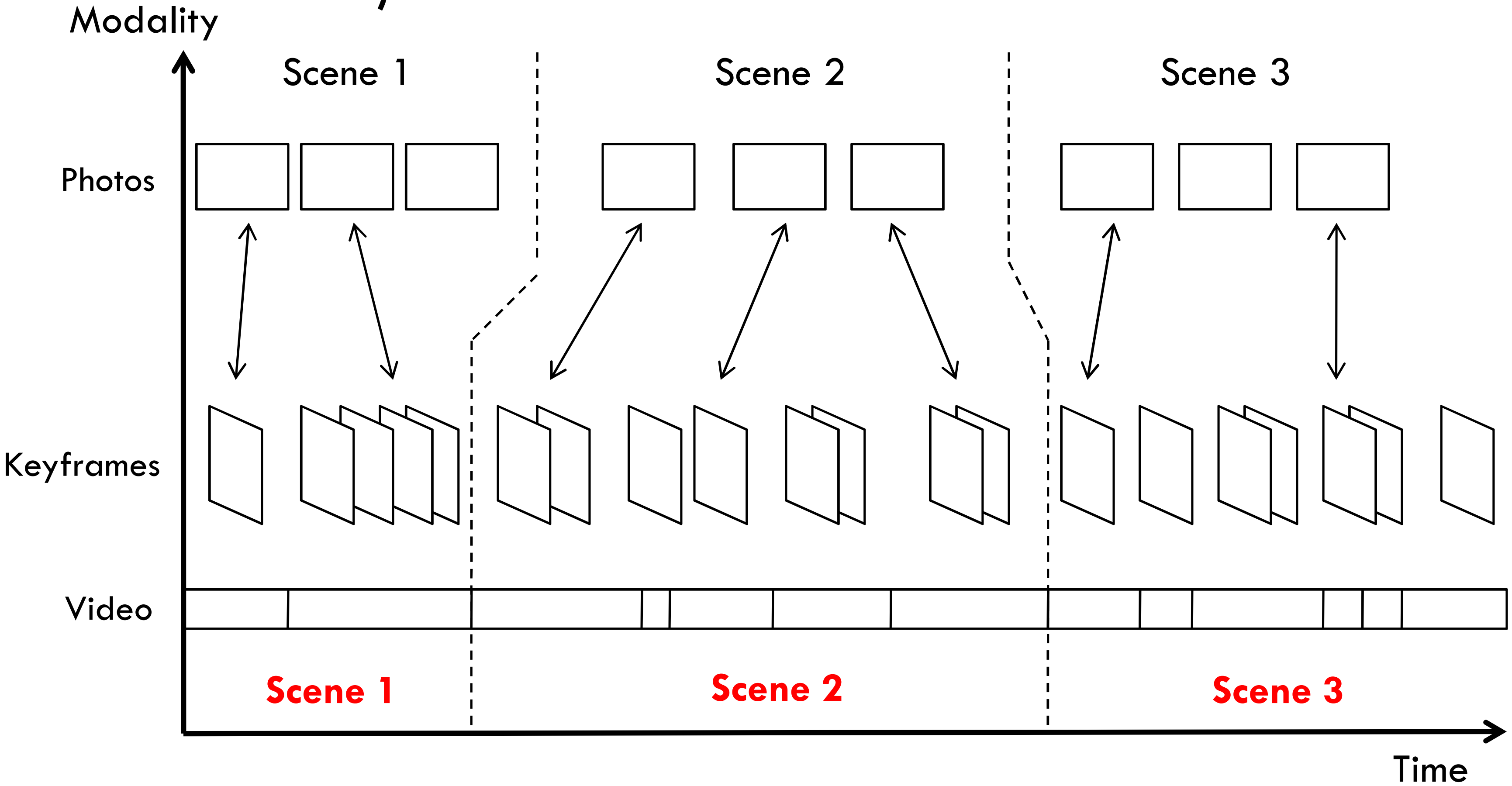
Unknown

Related Works

- Structuring home videos [Gatica-Perez'03][Pan'04]
- Automatic home video editing [Hua'04][Lee'05][Peng'08][Shipman'08]
- User intent modeling [Achanta'06][Mei'05]
- Scene detection
  - Cluster video segments based on similarity of visual features. [Yeung'98][Rasheed'03]
  - Problems described above harm conventional approaches because videos shots at the same scene may have significantly different appearance.

Essence of The Idea

- Cross-Media Correlation (Photo-Keyframe Correlation)



Overview of Framework

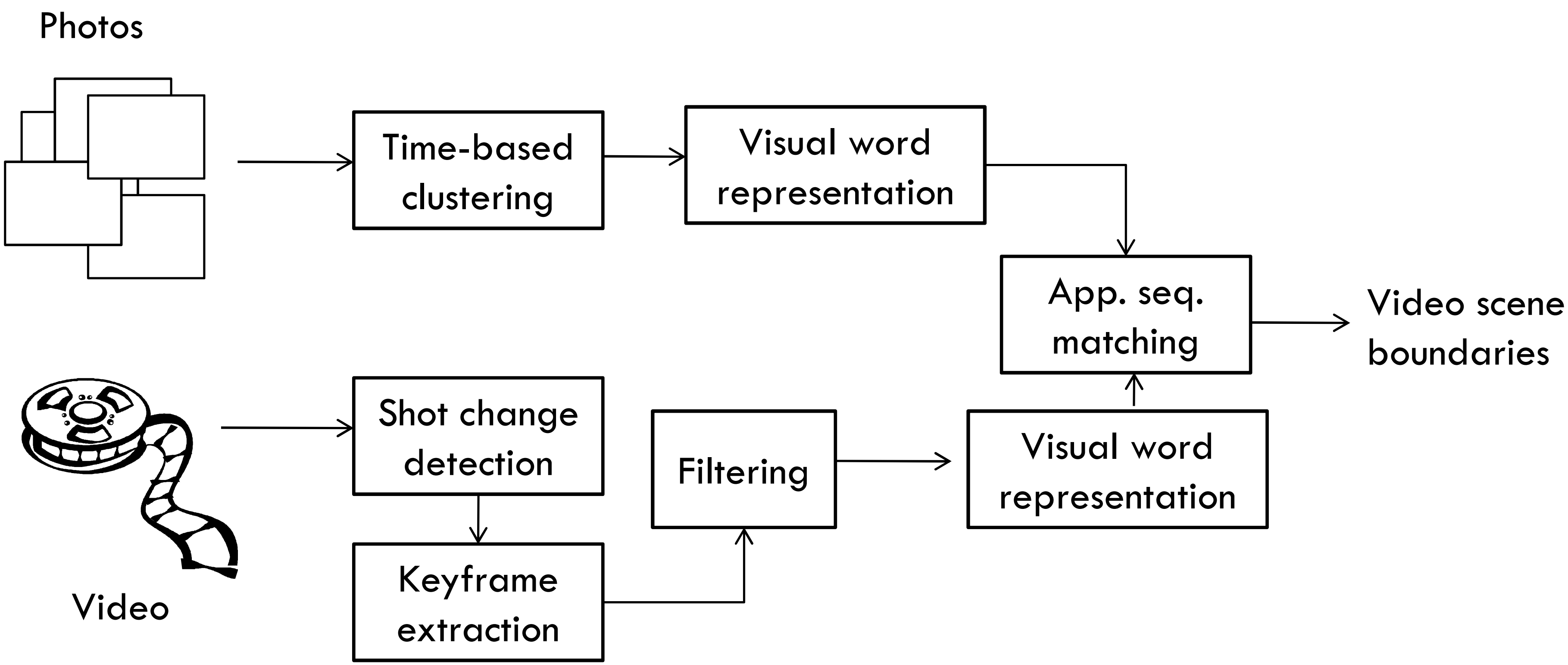
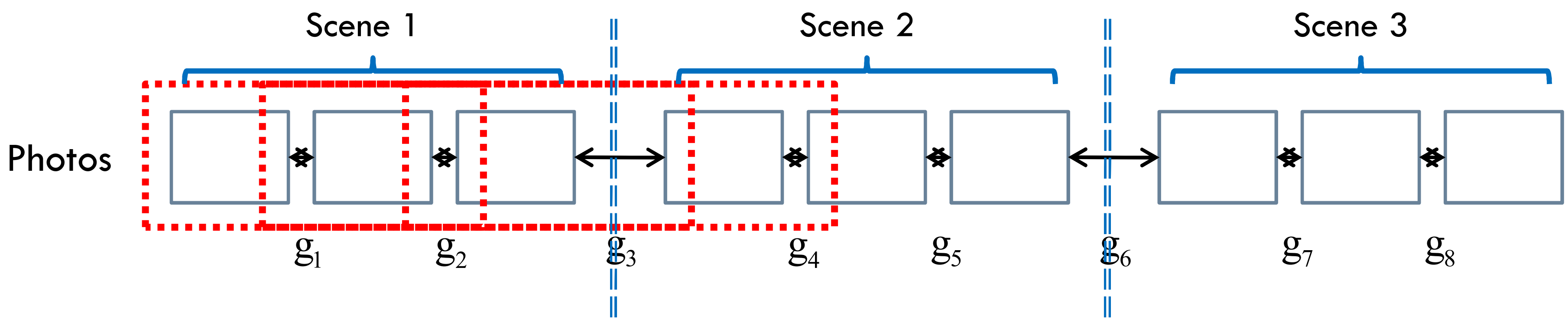


Photo Scene Detection

- There are large time gaps between photos in different scenic spots because of transportation.

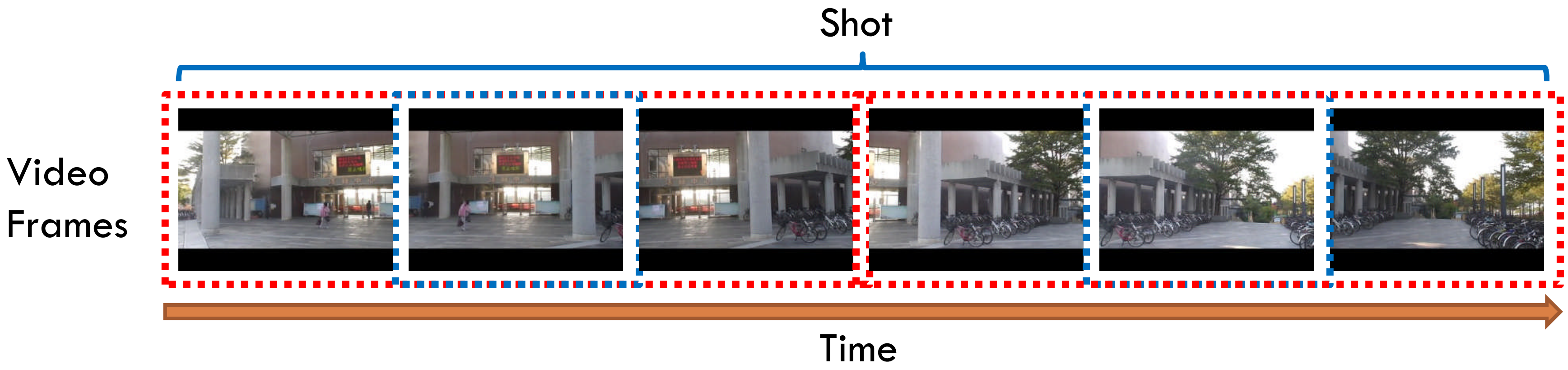


$g_i = t_{i+1} - t_i$ Scene boundary 1 Scene boundary 2

$\log(g_N) \geq K + \frac{1}{2d+1} \sum_{i=-d}^{d} \log(g_{N+i})$ ***A scene boundary exists !***

Platt, J.C., Czerwinski, M., and Field, B.A. 2003. PhotoTOC: automating clustering for browsing personal photographs. In Proc. of IEEE Pacific Rim Conference on Multimedia, 6-10.

Keyframe Extraction



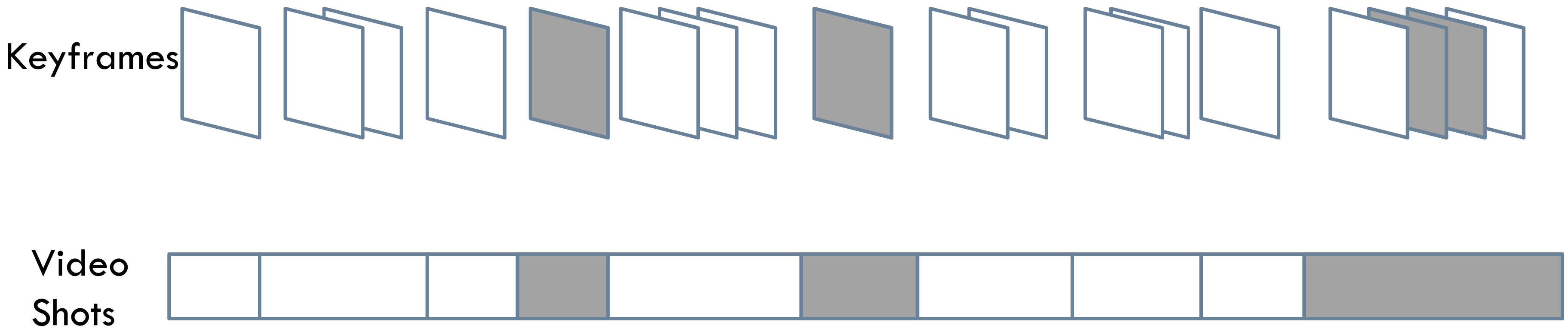
1. We use the global k-means algorithm [Likas'03] to determine how many groups should be in this shot.

2. Then, we choose the centroid of each group as the keyframe.

Keyframes:

Likas, A., Vlassis, N., and Verbeek, J.J. 2003. The global k-means clustering algorithm. Pattern Recognition, vol. 36, 451-461.

Filtering



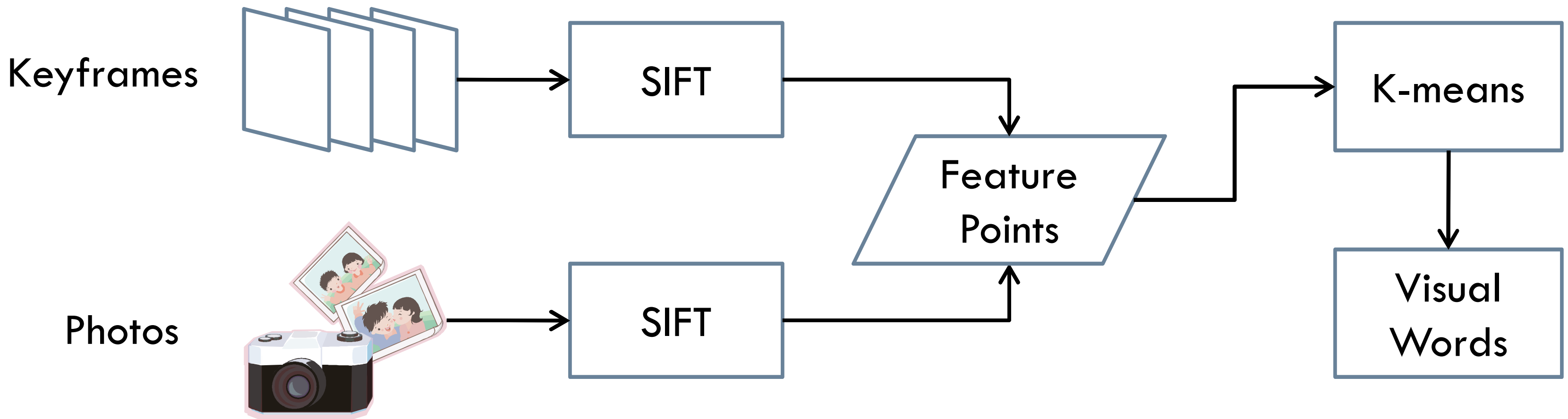
- Using edge information in different resolutions to detect blurred keyframes [Tong'04].
- Filter out video shots with blurred keyframes.

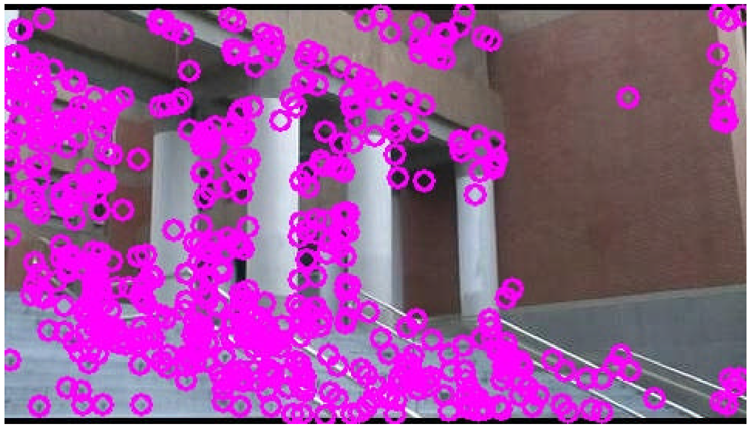
Advantages:

1. Reduces time complexity of cross-media matching
2. Eliminates the influence of bad-quality shots.

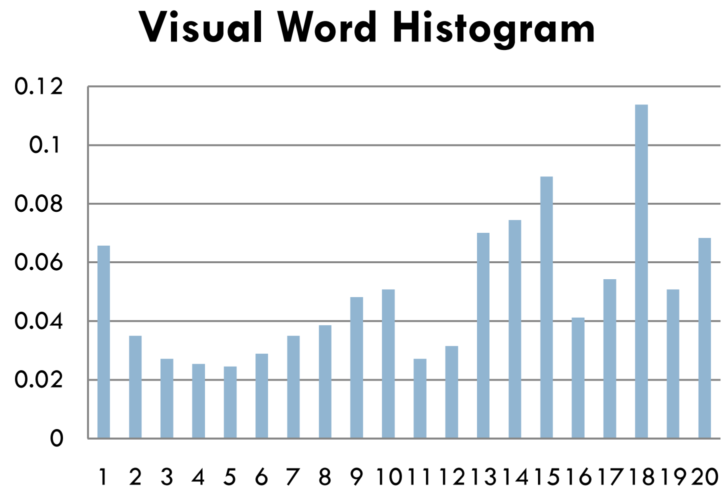
Tong, H., Li, M., Zhang, H.-J., and Zhang, C. 2004.
Blur detection for digital images using wavelet transform.
In Proc. of IEEE International Conference on Multimedia & Expo, 17-20.

Visual Word Representation





K = 20



A sequence of photos (keyframes) has been transformed to a sequence of visual word histograms.

Approximate Sequence Matching

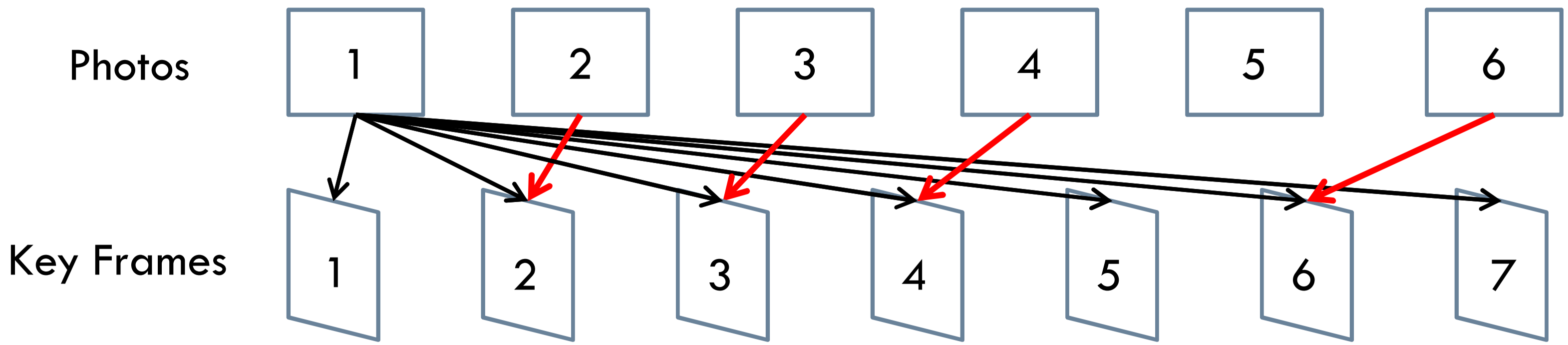
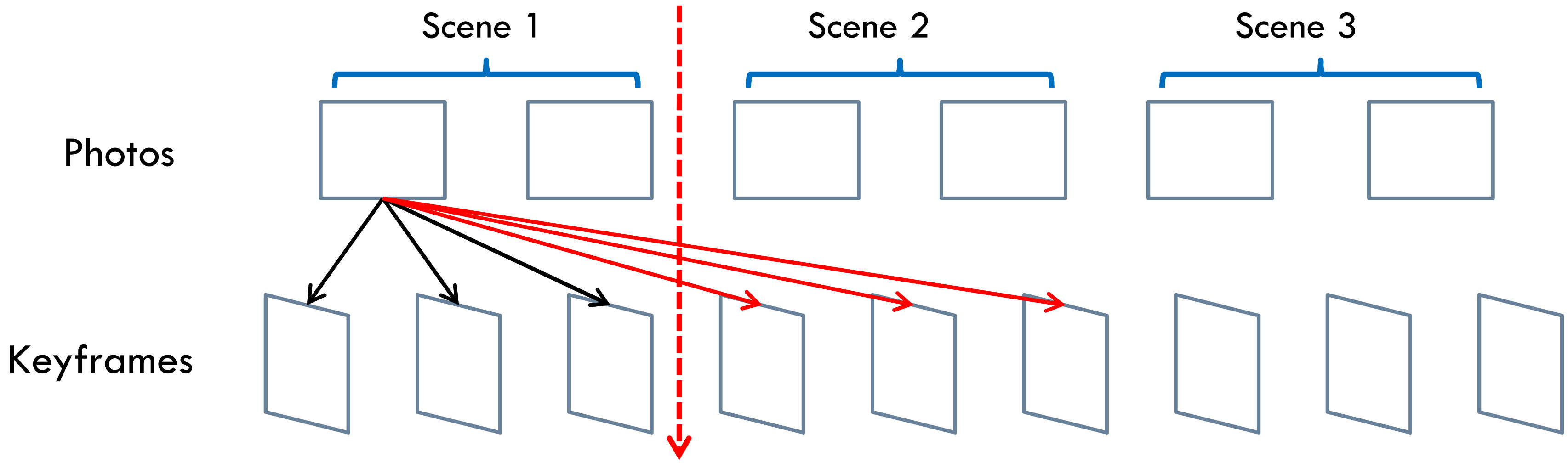


		Photo 1	Photo 2	Photo 3	Photo 4	Photo 5	Photo 6
	0	0	0	0	0	0	0
Frame 1	0	0 ↑	1 ↖	1 ←	1 ←	1 ←	1 ←
Frame 2	0	1 ↖	1 ↖	1 ↑	1 ↑	1 ↑	1 ↑
Frame 3	0	1 ↑	1 ↑	2 ↖	2 ←	2 ←	2 ←
Frame 4	0	1 ↑	1 ↑	2 ↑	3 ↖	3 ←	3 ↖
Frame 5	0	1 ↑	1 ↑	2 ↑	3 ↑	3 ↑	3 ↑
Frame 6	0	1 ↑	1 ↑	2 ↑	3 ↑	4 ↖	4 ↖
Frame 7	0	1 ↑	1 ↑	2 ↑	3 ↑	4 ↖	4 ↑

$$\sum_{k=0}^{N-1} |(h_i(k) - h_j(k))| < \delta$$

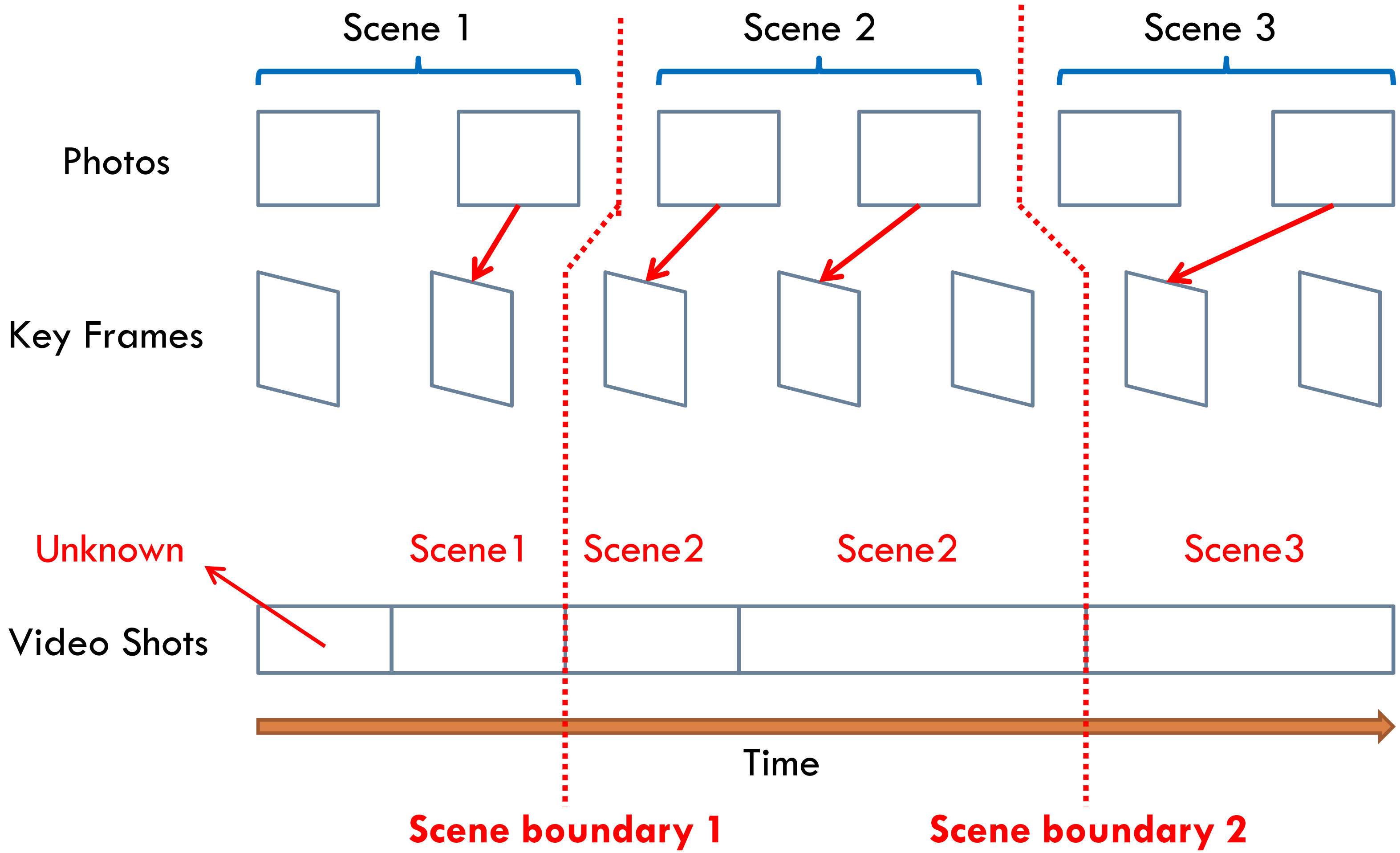
Time Constraint

- In a journey, we sequentially visit scenic spots and take photos and videos in the same time order.



Extra range: The Number of Key Frames/The Number of Photo Scenes

Video Scene Detection



Postprocessing

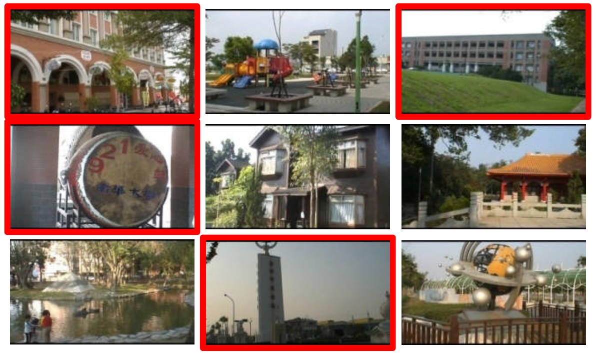
- Assigned by the labels the closest matched shots or interpolation

	$shot_{j-i}$	…	$shot_{j-1}$	$shot_j$	$shot_{j+1}$	…	$shot_{j+k}$
Case 1	Scene a	…	Scene a	Scene a	Scene a	…	Scene a
Case 2	Scene a	…	Scene a	Scene a	Scene b	…	Scene b
Case 3				Scene a	Scene a	…	Scene a
Case 4	Scene a	…	Scene a	Scene a			

Evaluation Data

	# scenes	Length	# of KFs	# photos
Video1	6	12:57	176	101
Video2	4	10:20	113	20
Video3	3	15:07	73	41
Video4	5	8:29	74	46
Video5	5	11:03	127	126

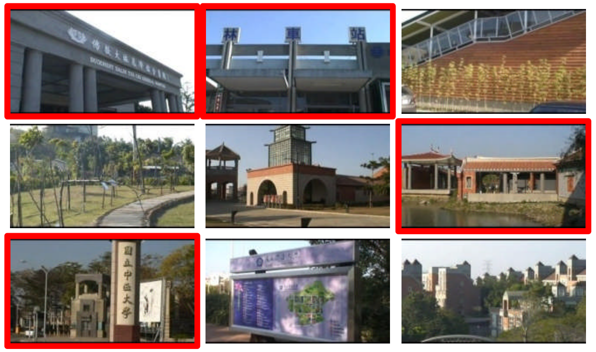
Sample video keyframes



Sample photos

Data set 1

Data set 5



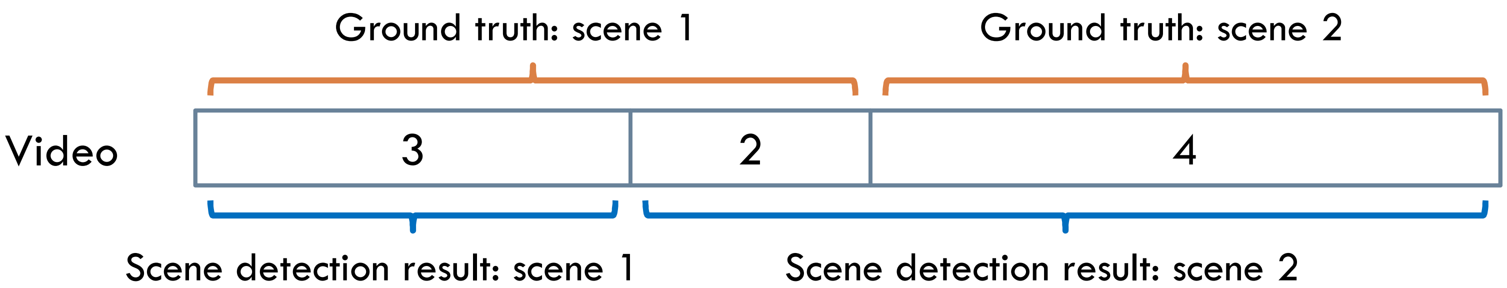
Evaluation Metric – Purity Value

$$\rho = \left(\sum_{i=1}^{Ng} \frac{\tau(s_i)}{T} \sum_{j=1}^{Nv} \frac{\tau^2(s_i, s_j^*)}{\tau^2(s_i)} \right) \cdot \left(\sum_{j=1}^{Nv} \frac{\tau(s_j^*)}{T} \sum_{i=1}^{Ng} \frac{\tau^2(s_i, s_j^*)}{\tau^2(s_j^*)} \right)$$

$\tau(s_i, s_j^*)$ is the length of overlap between the scene s_i and s_j^*

$\tau(s_i)$ is the length of the scene

T is the total length of all scenes.

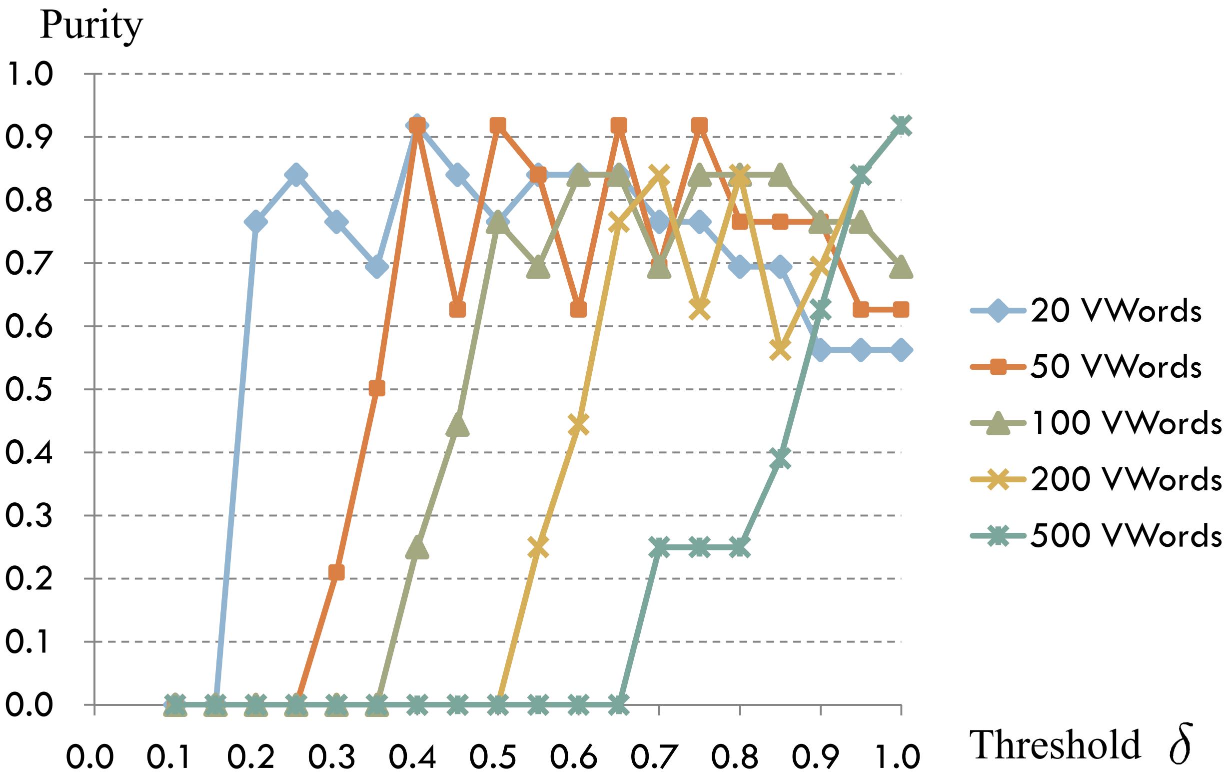


Purity=(5/9*(9/25+4/25)+4/9*16/16)*(3/9*9/9+6/9*(4/36+16/36))
=0.5160493827160494

Vinciarelli, A. and Favre, S. 2007.
Broadcast news story segmentation using social network analysis and hidden Markov models.
In Proc. of ACM Multimedia, 261-264.

Performance of Scene Detection

- The best performance occurs in different settings for different visual words.
- 20 visual words are used to present photos and keyframes in the following experiments.

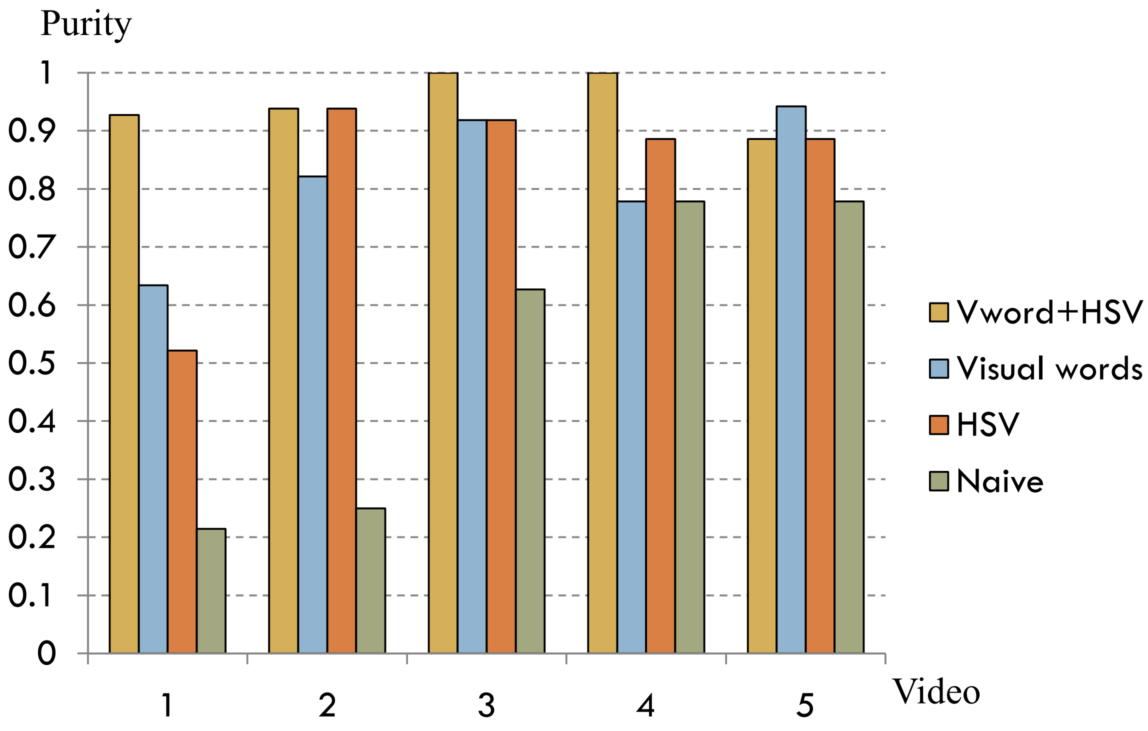


$$\sum_{k=0}^{N-1} |(h_i(k) - h_j(k))| < \delta$$

Performance of Scene Detection

- Visual word histograms work better in Videos 1 and 5 -- describe *what* are in an image.
- HSV histograms work better in Videos 2 and 4 -- describe color information.
- The best performance is obtained by combining them.

Average purity value = 0.95



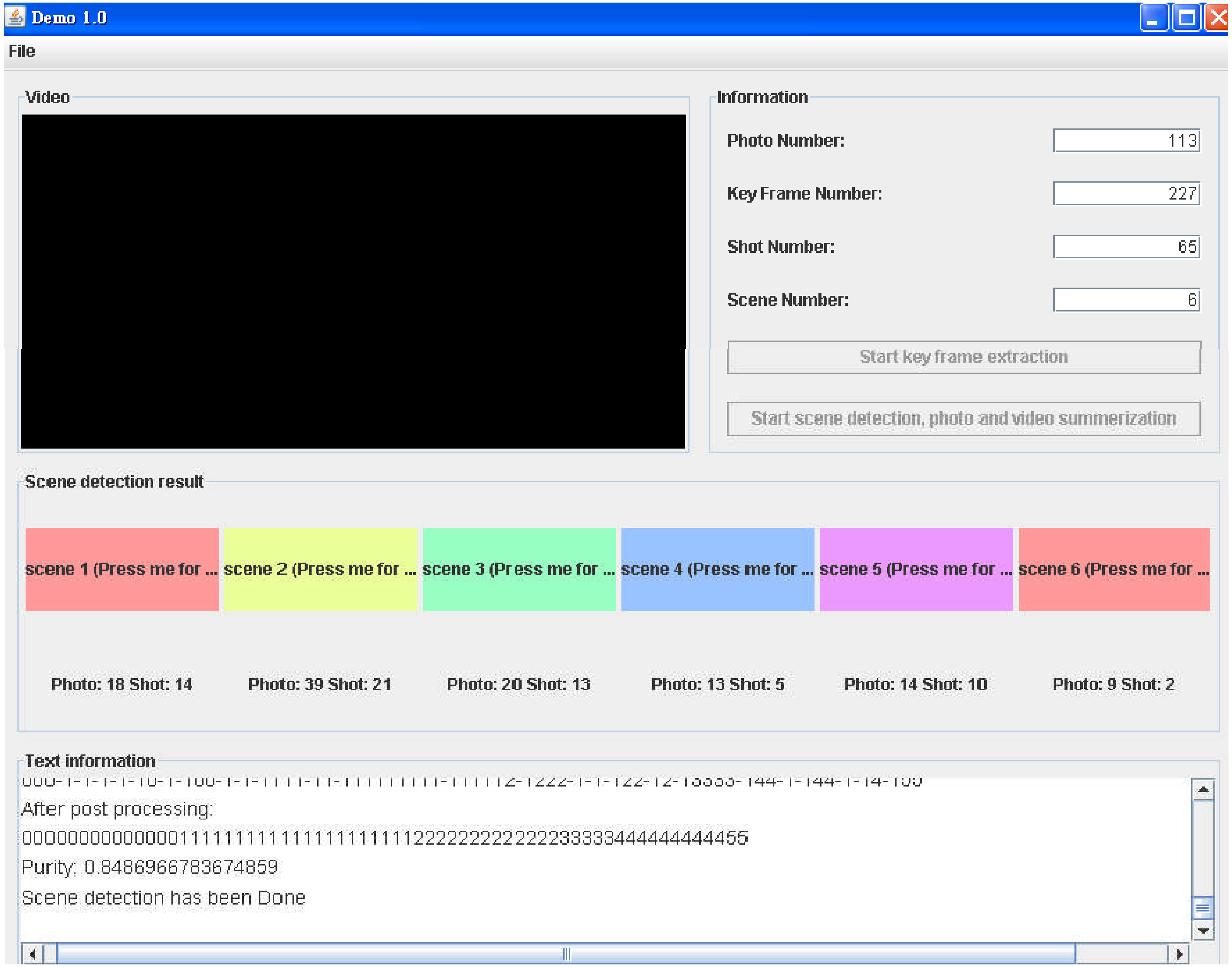
Performance Comparison

- Measure over-segmentation situation
- (*m*,*n*) denotes a scene is segmented into *m* and *n* scenes, by the method in [Chasanis'07] and ours.
- The method in [Chasanis'07] doesn't take advantage of cross-media correlation.

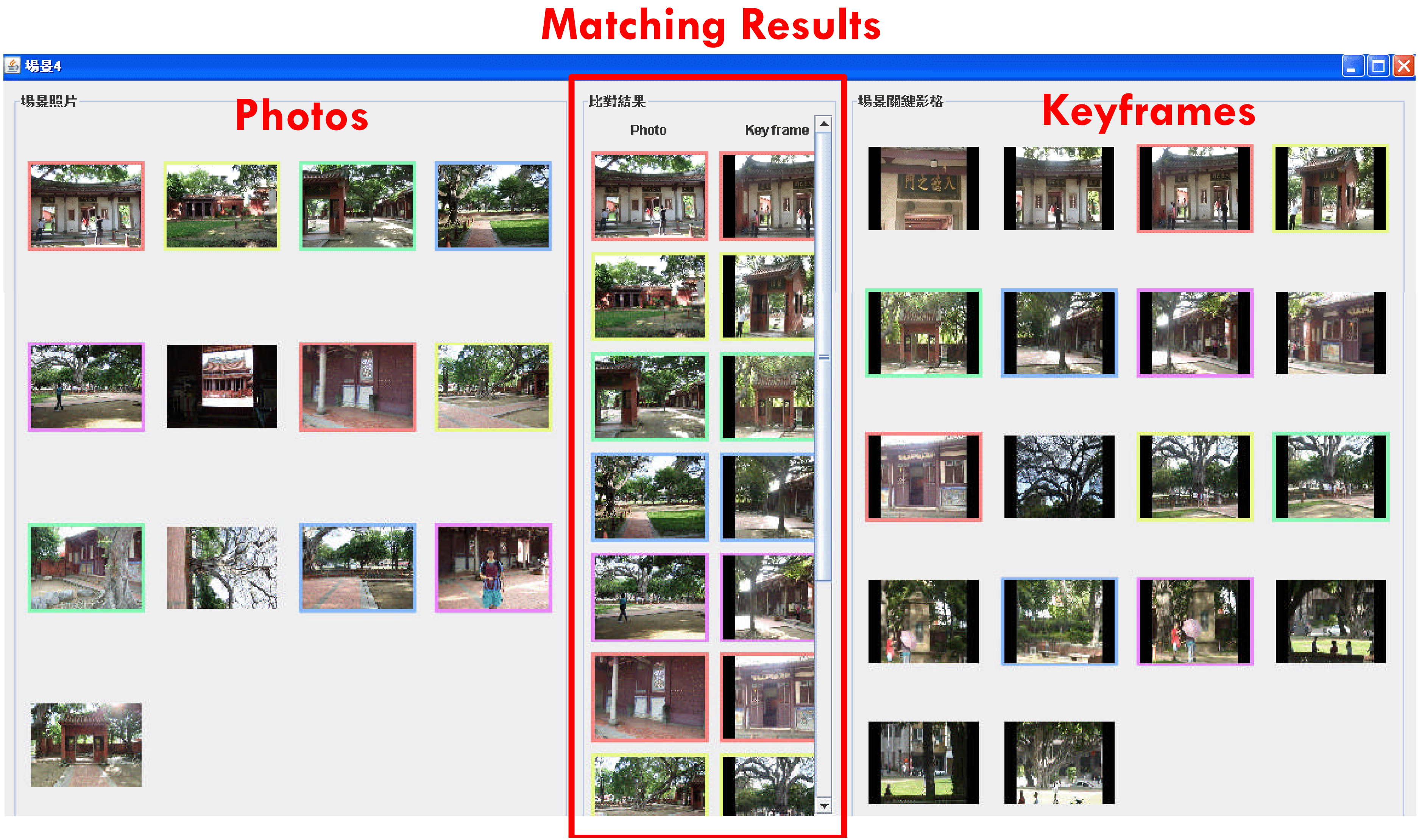
	S1	S2	S3	S4	S5	S6	Overall
Video1	(1,1)	(4,1)	(7,2)	(3,1)	(9,2)	(3,1)	(27,8)
Video2	(2,2)	(8,1)	(1,1)	(1,1)			(12,5)
Video3	(6,1)	(3,1)	(1,1)				(10,3)
Video4	(1,1)	(1,1)	(1,1)	(3,1)	(2,1)		(8,5)
Video5	(1,1)	(2,2)	(1,1)	(5,2)	(1,1)		(10,7)

Chasanis, V., Likas, A., and Galatsanos, N. 2007. Scene detection in videos using shot clustering and symbolic sequence segmentation, in Proc. of MMSP, 187-190.

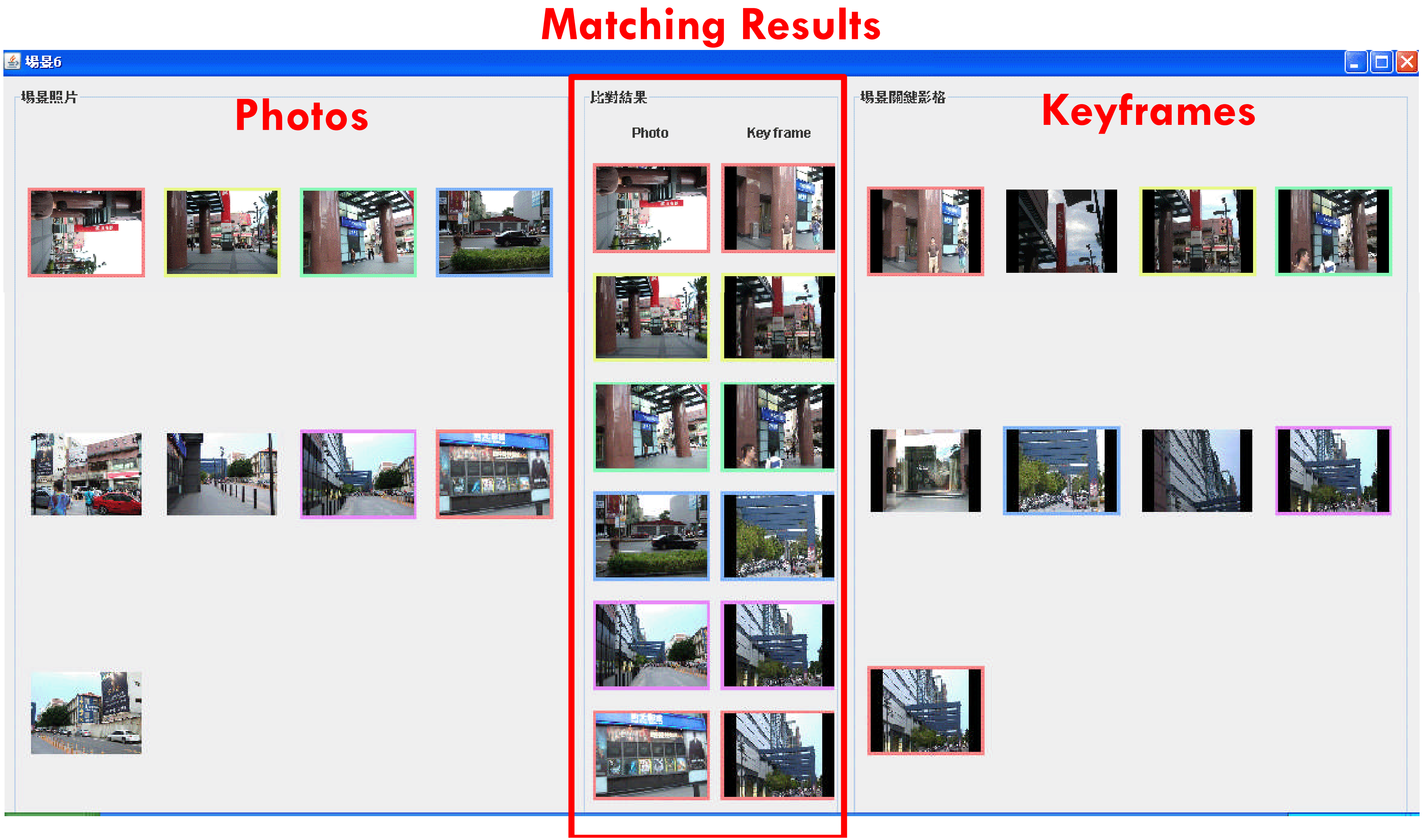
System Interface



System Interface



System Interface



Summary

- Contributions:
  - Using cross-media correlation to facilitate scene detection for travel videos.
  - Study of performance achieved by the proposed method and conventional approaches.
- Using cross-media correlation is an interesting and effective approach in analyzing travel videos. It may be extended to other domains, such as correspondence news videos and print media.

76

Conclusion

Wei-Ta Chu (朱威達)

wtchu@cs.ccu.edu.tw

Assistant Professor

Dept. of Computer Science and Information Engineering

National Chung Cheng University

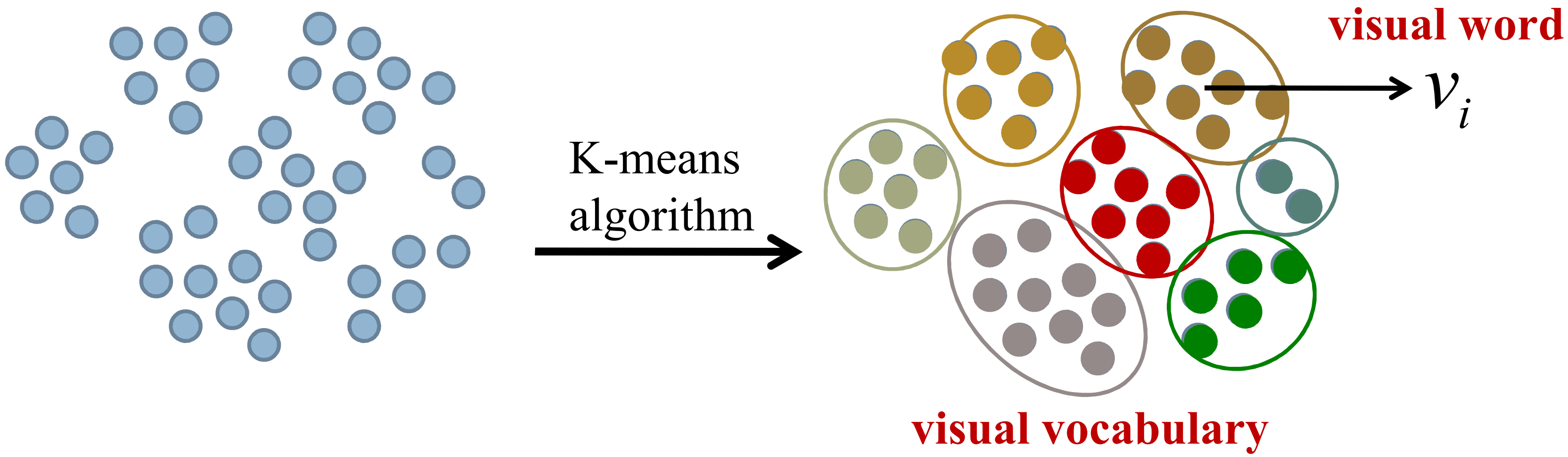
Conclusion

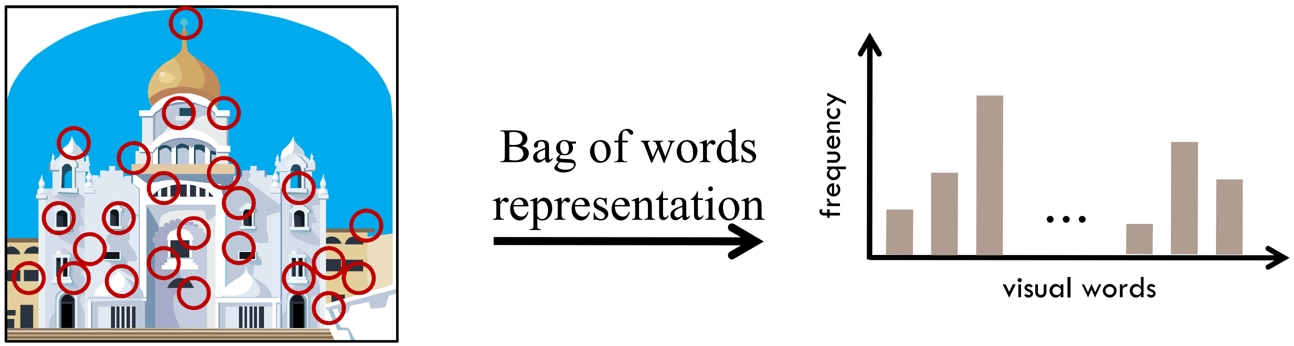
- Perspectives of travel media analysis:
  - processing modalities, access facets, active functions, correlation between different modalities, and access manners.
- Representative selection and ROI determination
- Face clustering in consumer photos
- Scene detection in travel videos
- Exploiting cross-media correlation and more elaborately utilizing characteristics of travel media would be an emerging research topic.

Backup Slides

PLSA-Based Feature Filtering

- Bag of words representation

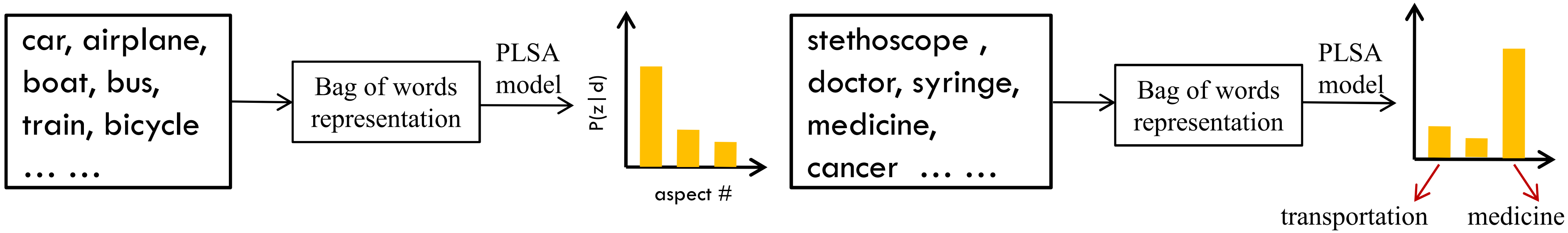




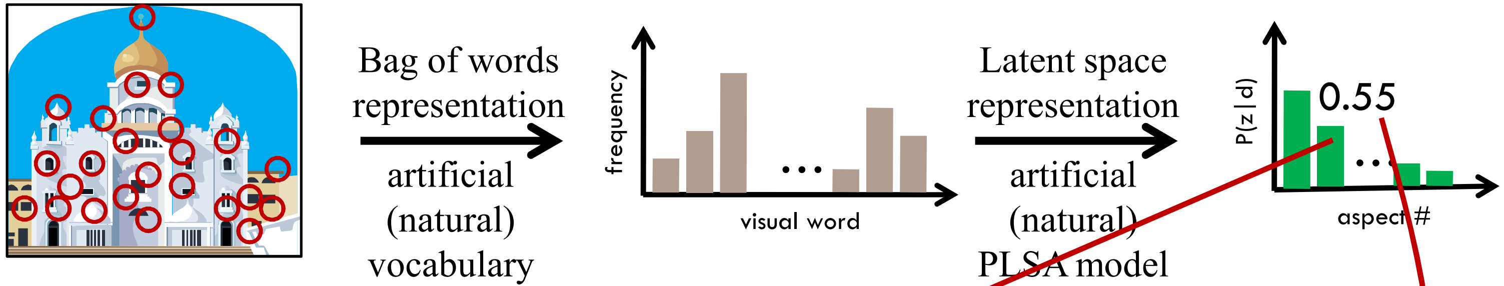
PLSA-Based Feature Filtering

PLSA model

- A document(image) d_j is modeled as a mixture of latent aspects z_k .



- Artificial and natural PLSA model



- A feature point $S(v_i)$ is claimed to be an artificial feature point if

$$\frac{P(z_{v_i^a} \mid v_i^a, d_j)}{P(z_{v_i^n} \mid v_i^n, d_j)} > \alpha$$

T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," Machine Learning, pp. 177-196, 2001.