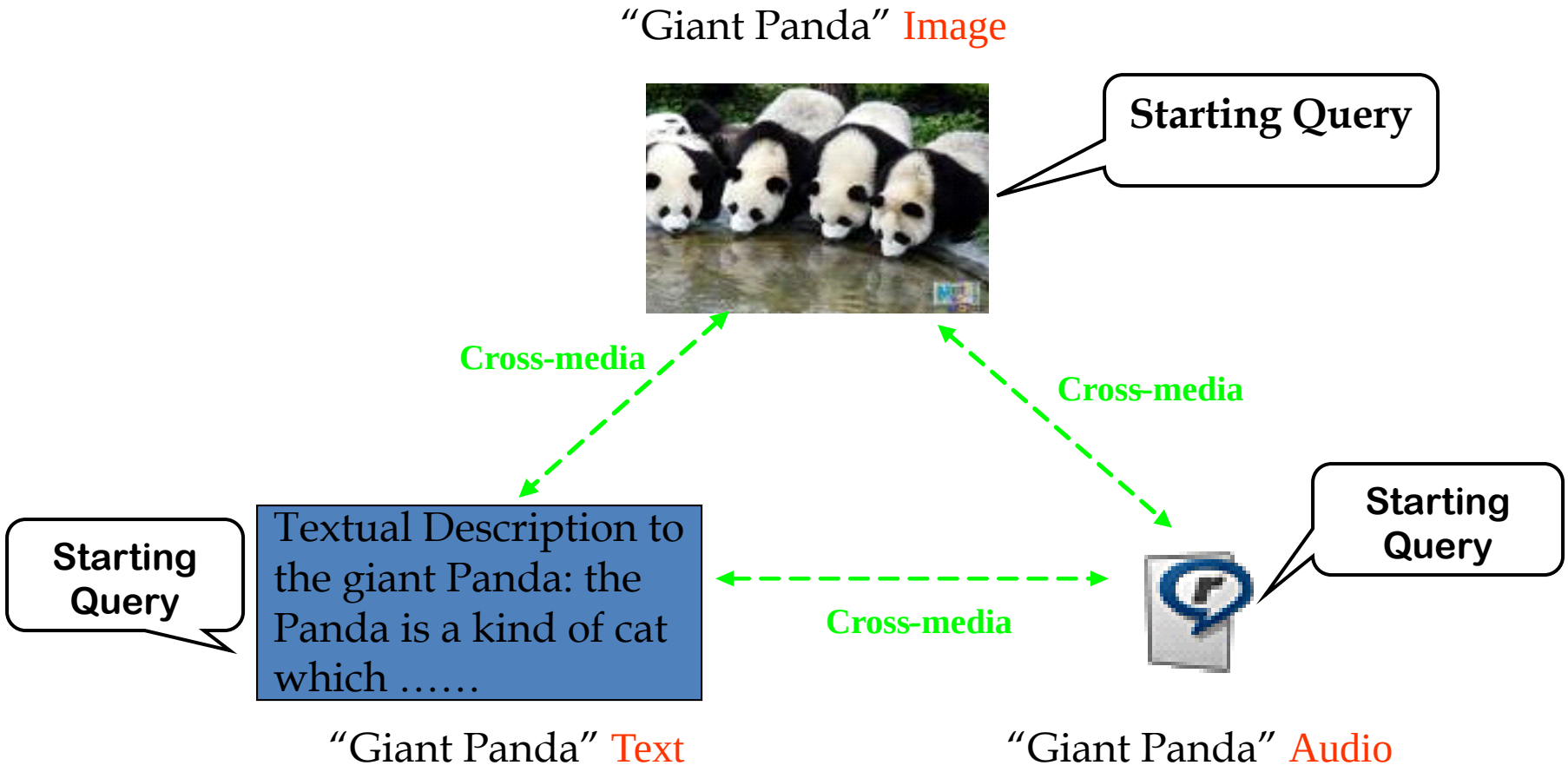
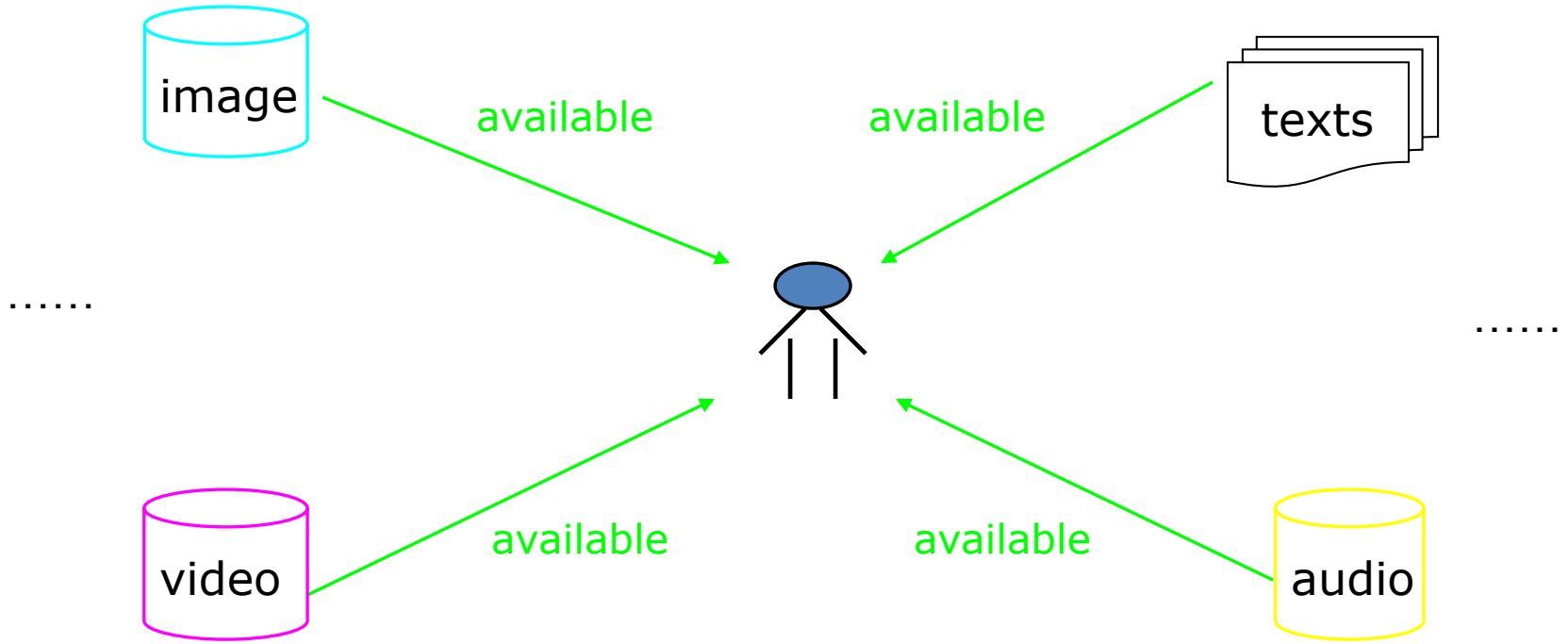
Cross-media Intelligent Searching

Scenario: a simple example of cross-media :



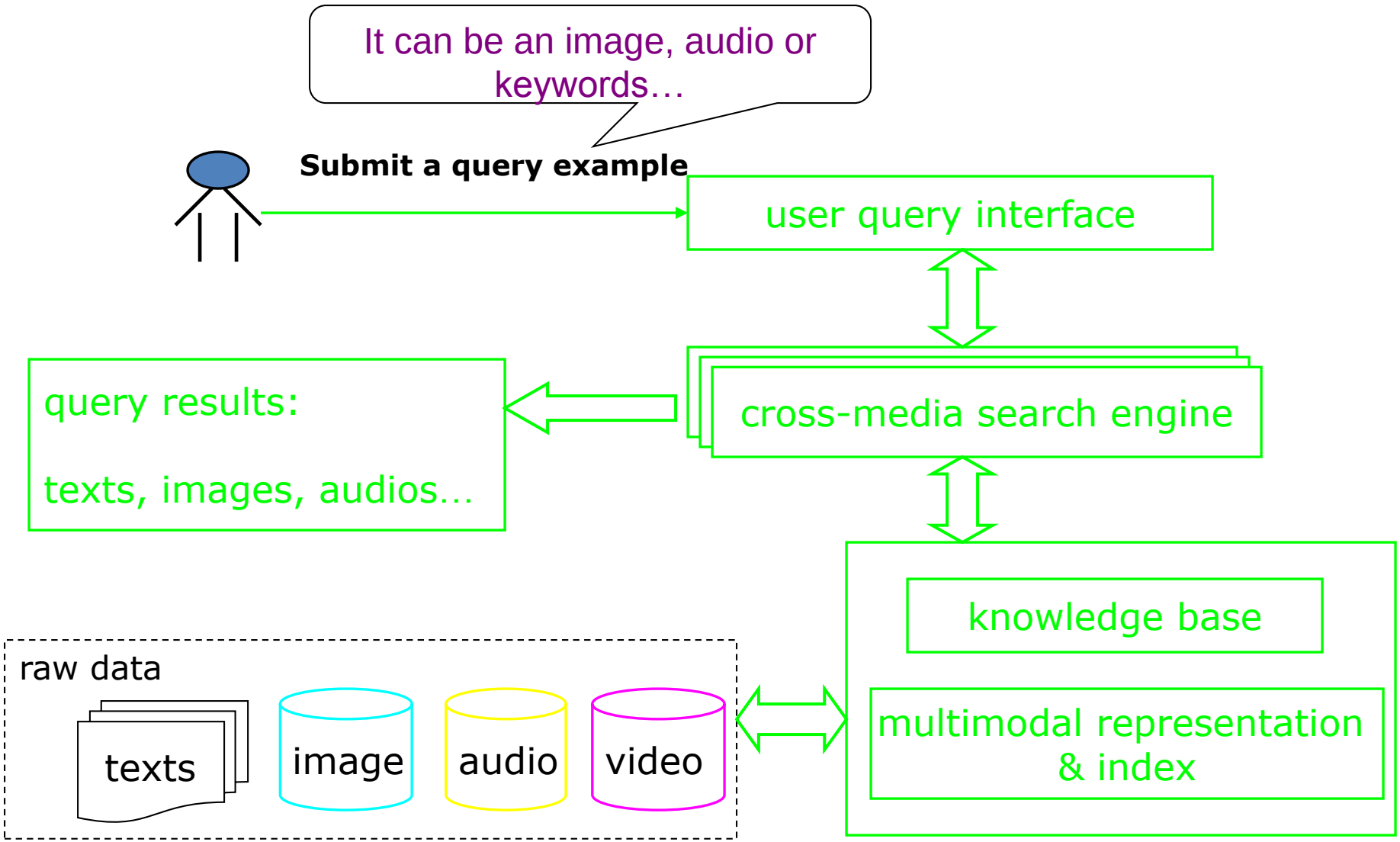
User can start a query from any type of media, and relevant multimedia data would be returned.

Cross-media retrieval is a useful way to access multimodal data:



◆ Cross-media retrieval can be regarded as the simulation of the real world, and it helps us get multimodal data in a more flexible and more informative way!

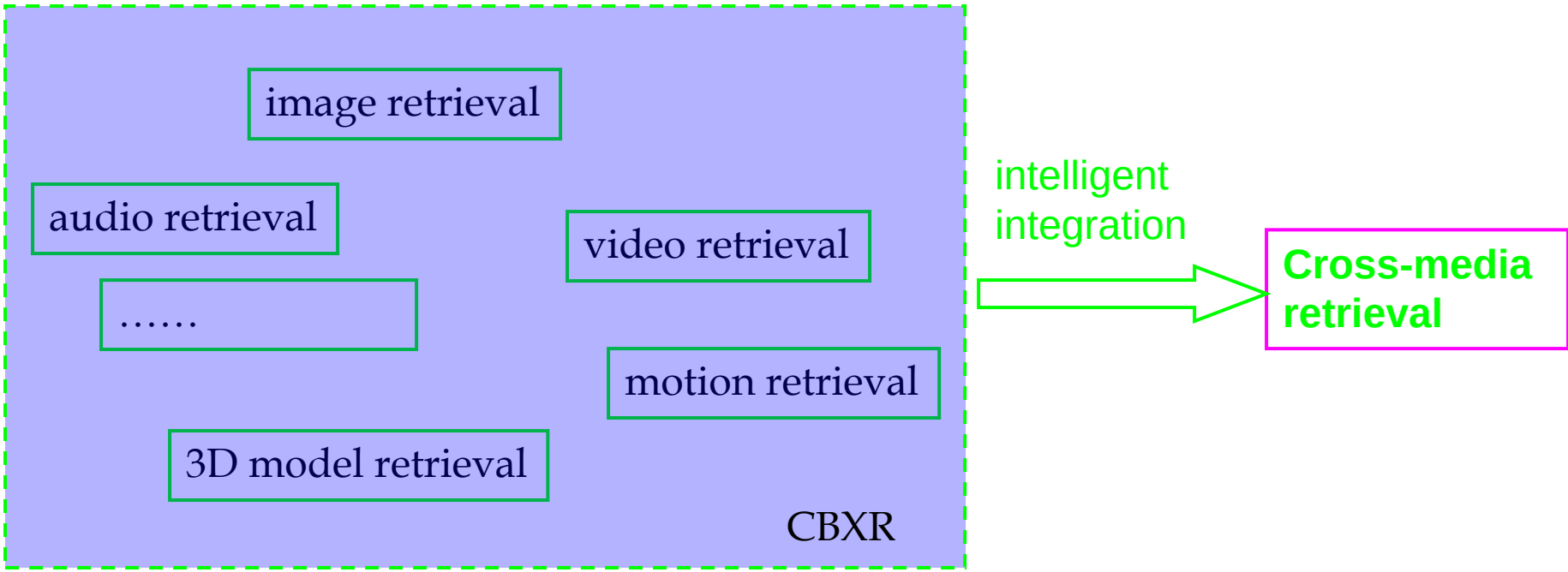
What cross-media retrieval needs to do?



From Multimedia Retrieval to Cross-media Retrieval

1) Image Retrieval: Content-based

- towards cross-media Retrieval
  - **Motivation**



We can provide a more flexible and efficient way to access multimodal data.

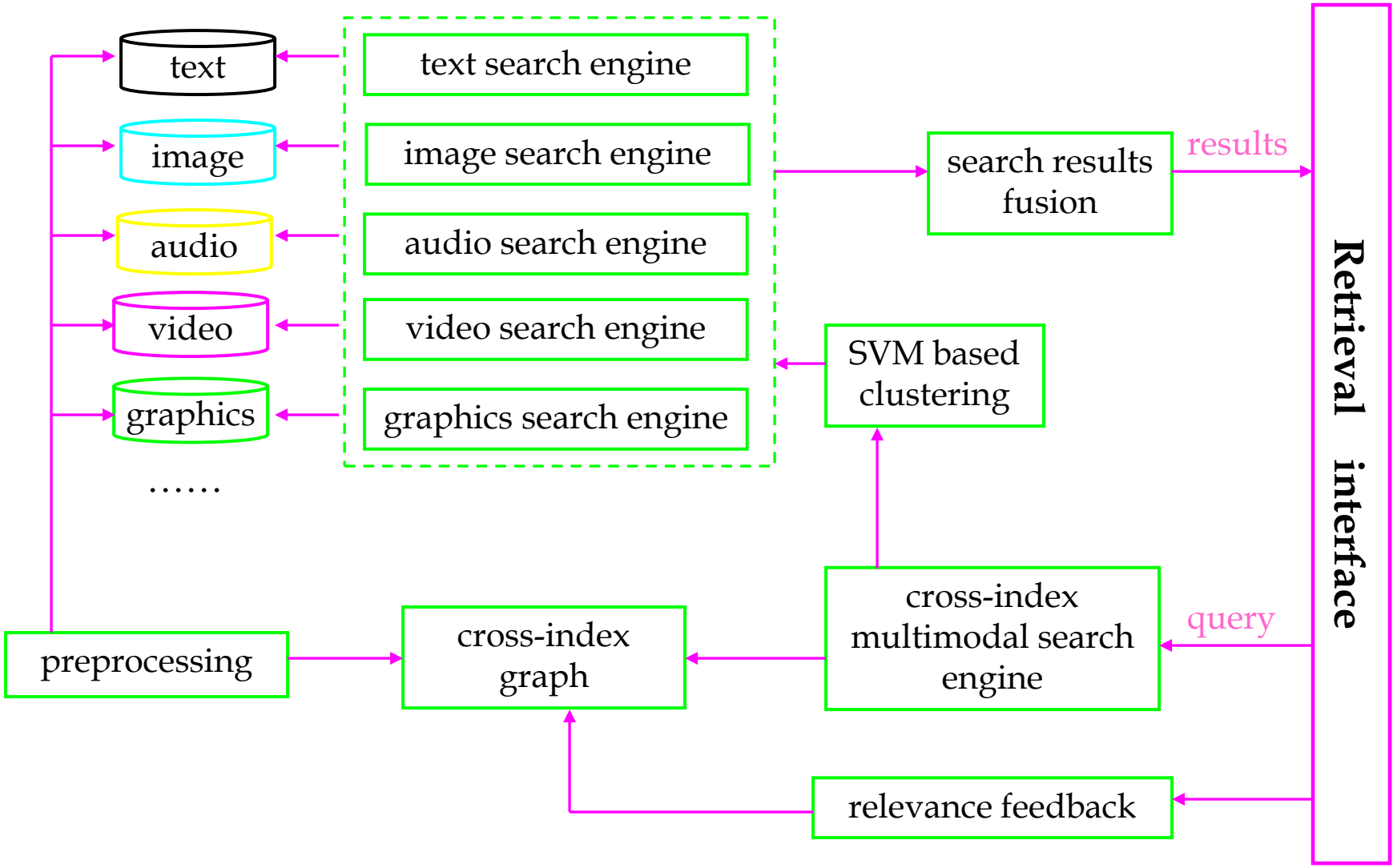
We name it as cross-media retrieval.

- Support multimodal sources
  - smooth integration of multimodal data;
  - query media objects by examples of different modalities;

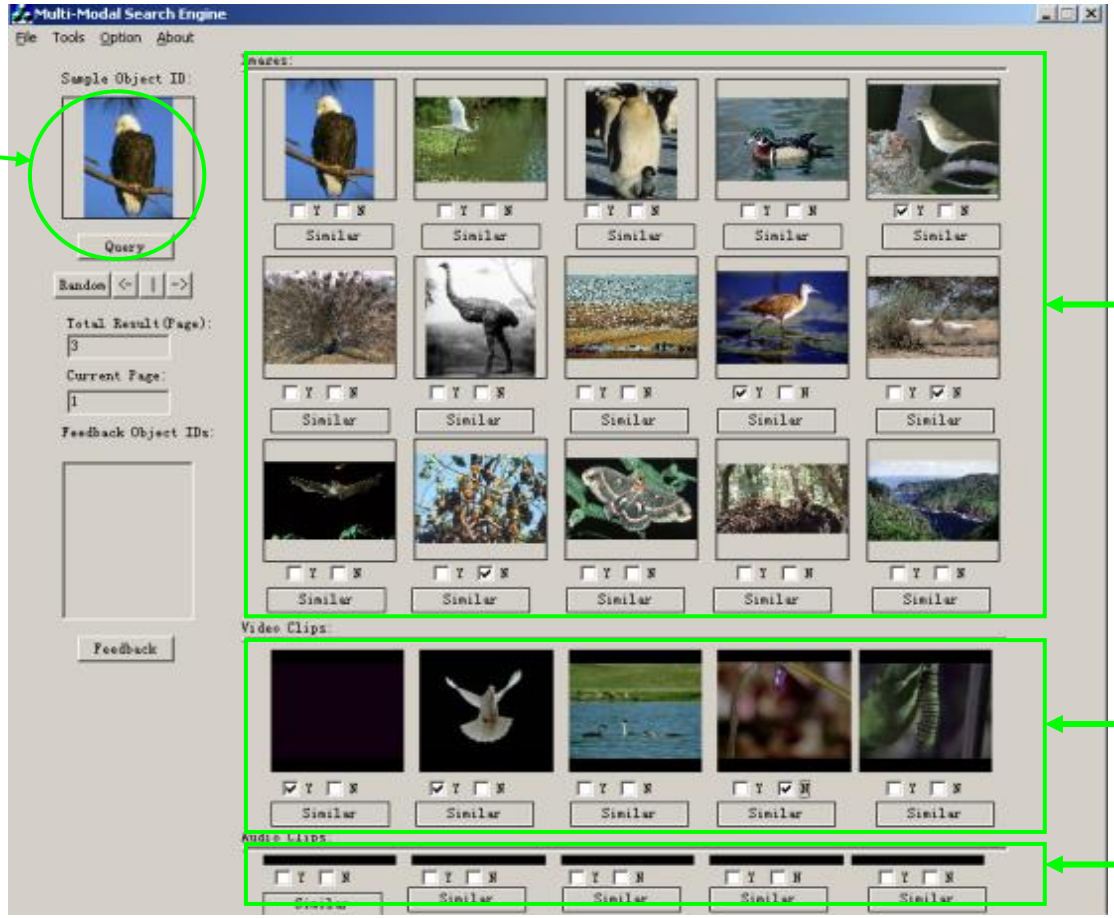
- Challenging issues:
  - texts, images, audios, etc. are represented with different features
  - different features are heterogeneous
  - cross-media similarity can’t be measured by content features
  - there is a semantic gap between low-level features and semantics

- Solution to Cross-media retrieval
  - **build cross-indexing from multimodal data**
  - **organize multimedia document**
  - **explore cross-media correlations**
  - ……

Cross-indexing-based retrieval: General idea



(1) Cross-index retrieval: interface



an image query example

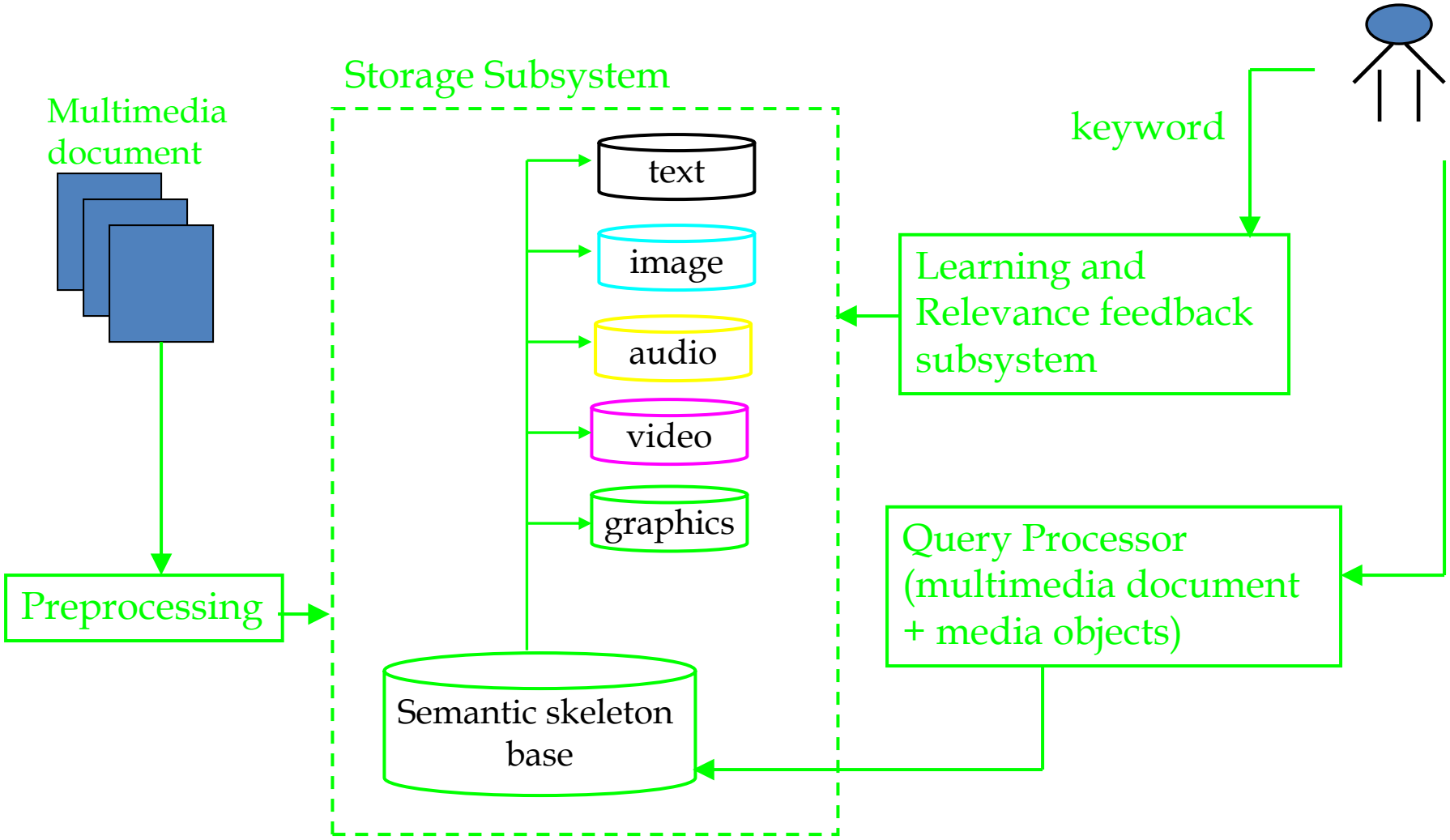
retrieved images

retrieved video

retrieved audio

The system now support images, audios and videos.
Users can submit any of the media objects, and the system returns relevant images, audios and videos.

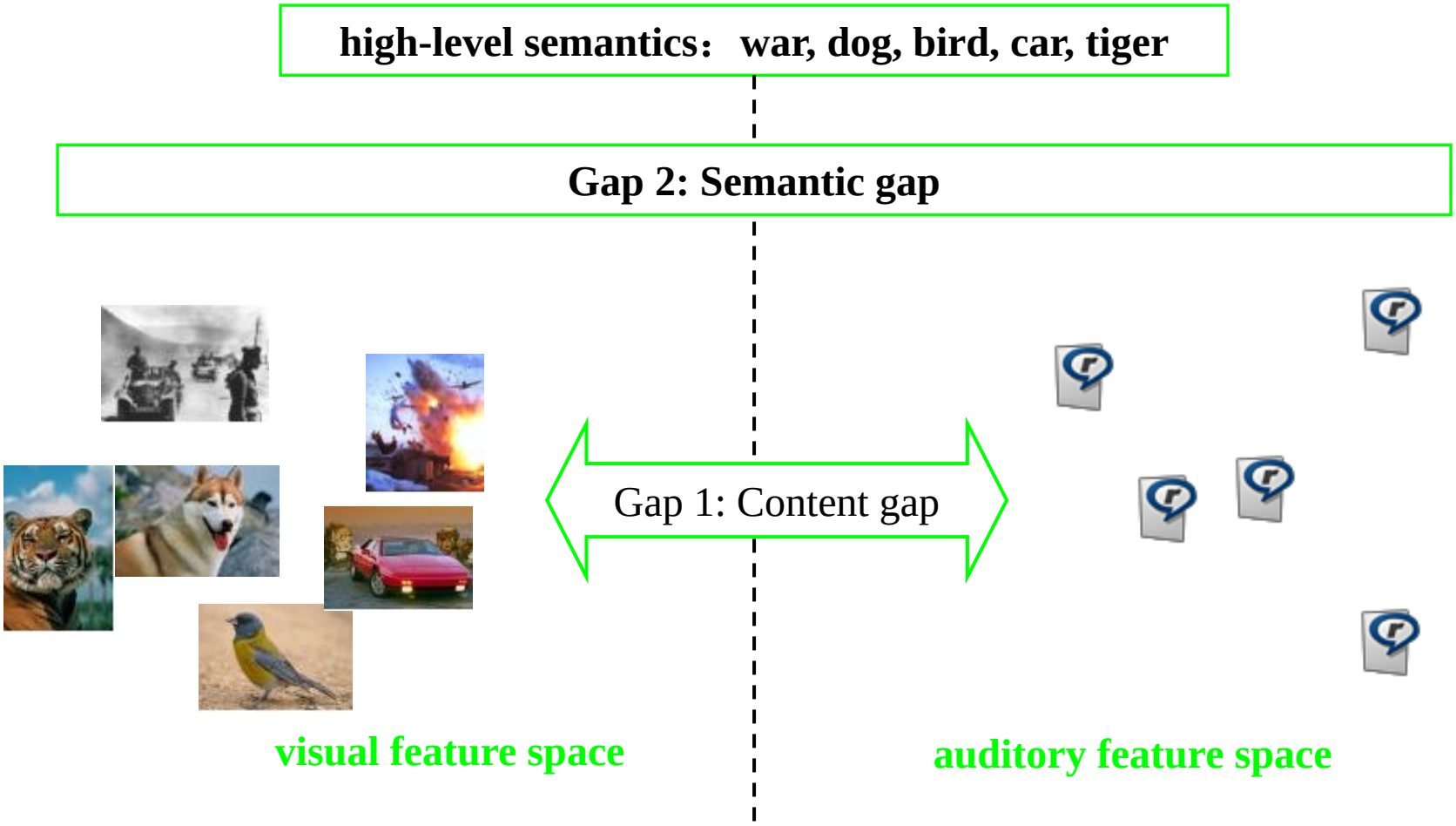
Build multimedia document: framework



Building multimedia document: retrieval interface

- the left figure is the relevant media data retrieved by the query of "water".
- A multimedia document is visualized as its sketch, i.e. text, images and key-frame lists for videos.
- Besides keyword-based search, the user can perform a content-based search with a specific media object as the query example

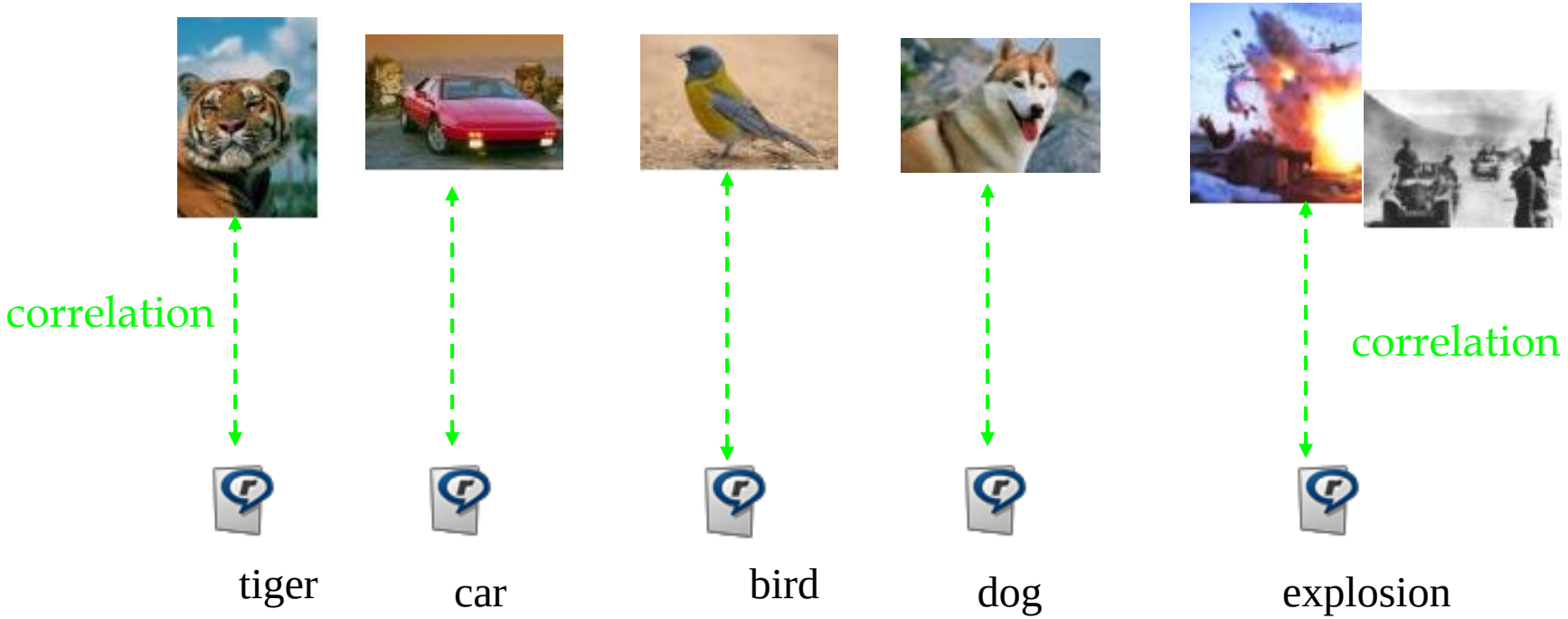
Exploring cross-media correlations: challenges



Challenges: 1. multimodal data reside in heterogeneous feature spaces
2. the semantic gap

Exploring Cross-media Correlations: Solutions

*Image*s and *audio*s represent high-level semantics from different perspectives. If we can find the correlation between different perspectives, we can enable cross-media retrieval with the bridge of correlations.



Exploring cross-media correlations: mathematical realization

Basic idea:

Input: image feature matrix: Audio feature matrix:

Canonical correlation analysis

X and Y are of different dimension !

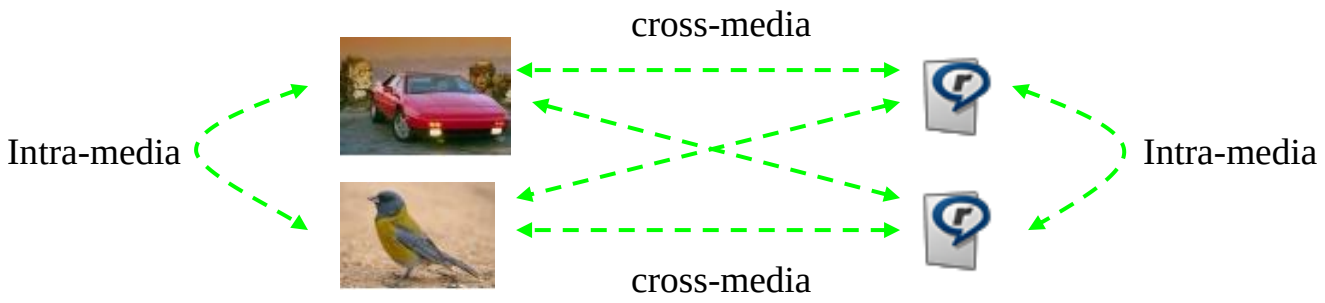
Output:

At the same time, the correlation between X and Y maximally coincides with the correlation between X' and Y'

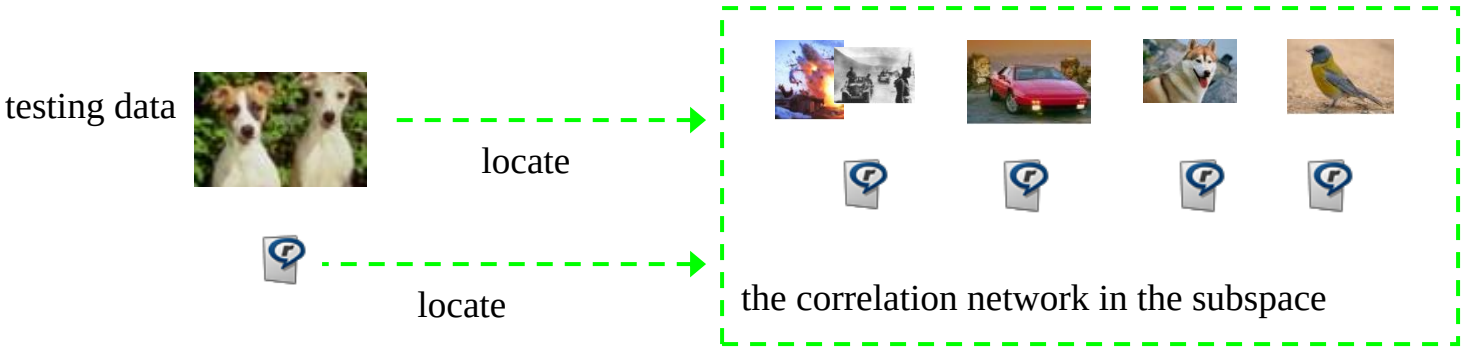
X and Y are of the same dimension !

Exploring cross-media correlations: subsequent challenges

1. how to measure both intra- and inter-media correlations ?



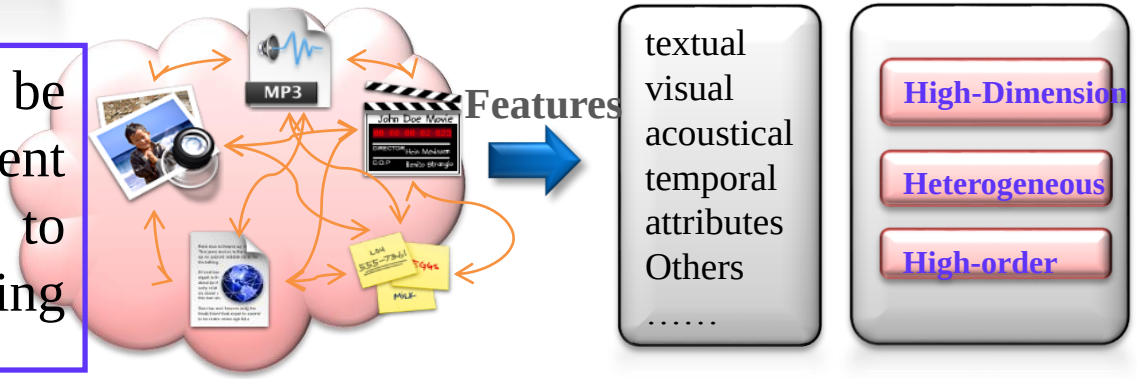
2. how to introduce new media objects into the system?



Three Properties of *Cross-media*

Cross-Modality

Many kinds of features can be obtained and they have different intrinsic discriminative power to characterize the corresponding semantic.

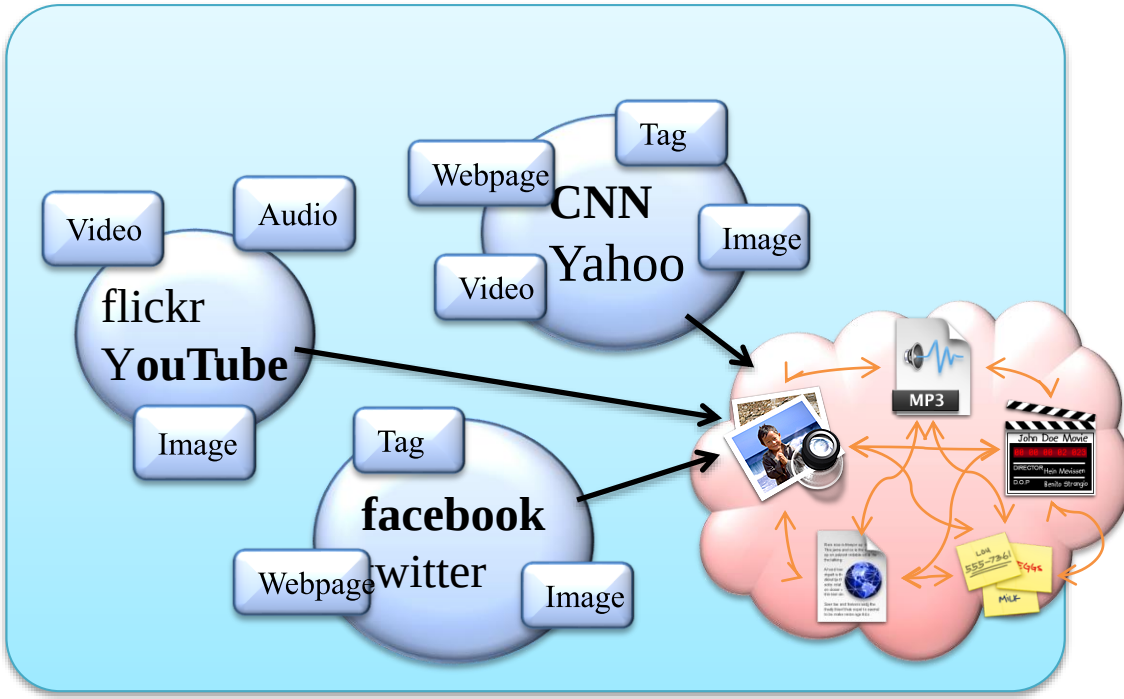


- **Issues:**
  - Feature fusion; Heterogeneous feature selection; Cross-modal metric learning
  - …

Three Properties of *Cross-media*

Cross-domain/ Cross-collections

The data about a same topic/event may be obtained from multiple sources.



- **Issues:**
  - Near-duplicated detection; Cross-domain learning; Transfer Learning
  - …

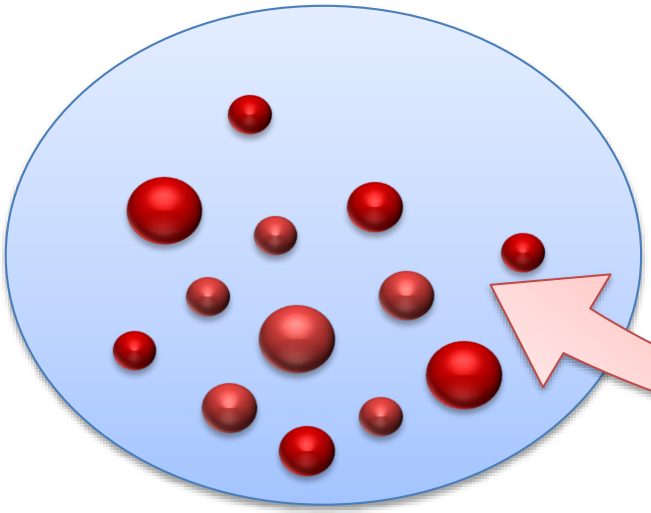
Motivation and Background:

Three Properties of *Cross-media*

Cross Space

From Cyberspace to Reality

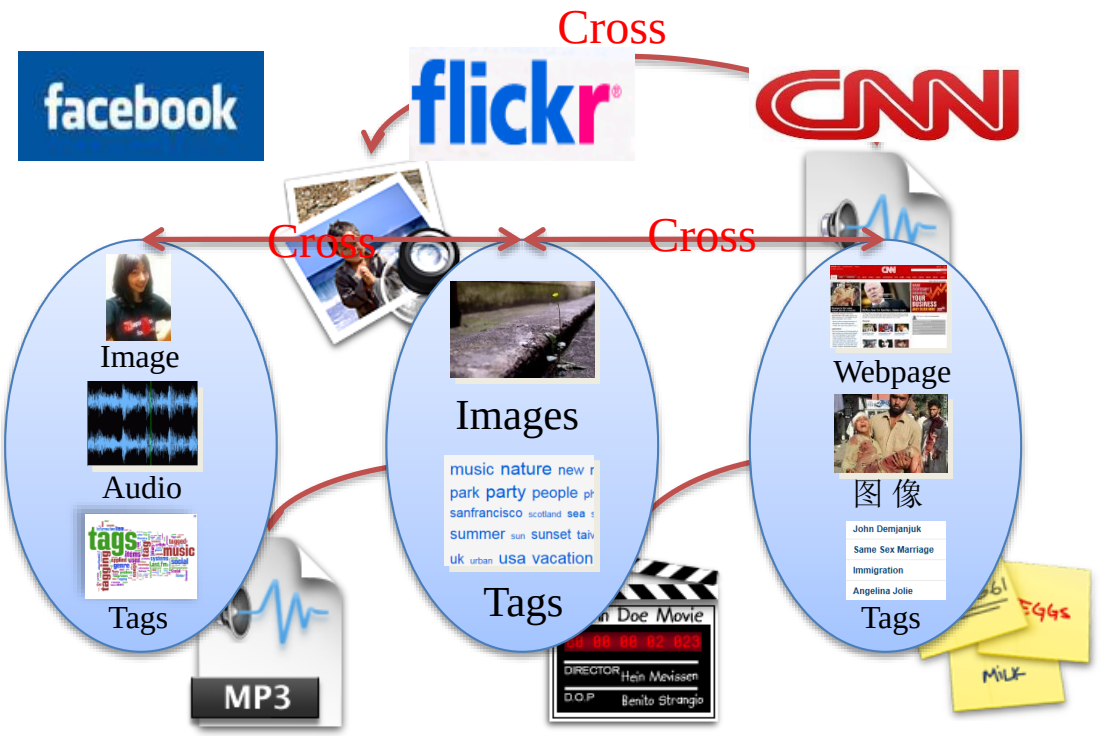
The virtual world (cyberspace) and the real-world (reality) complement each other, such as *Google Flutrends*



Cyberspace Complement **Reality**

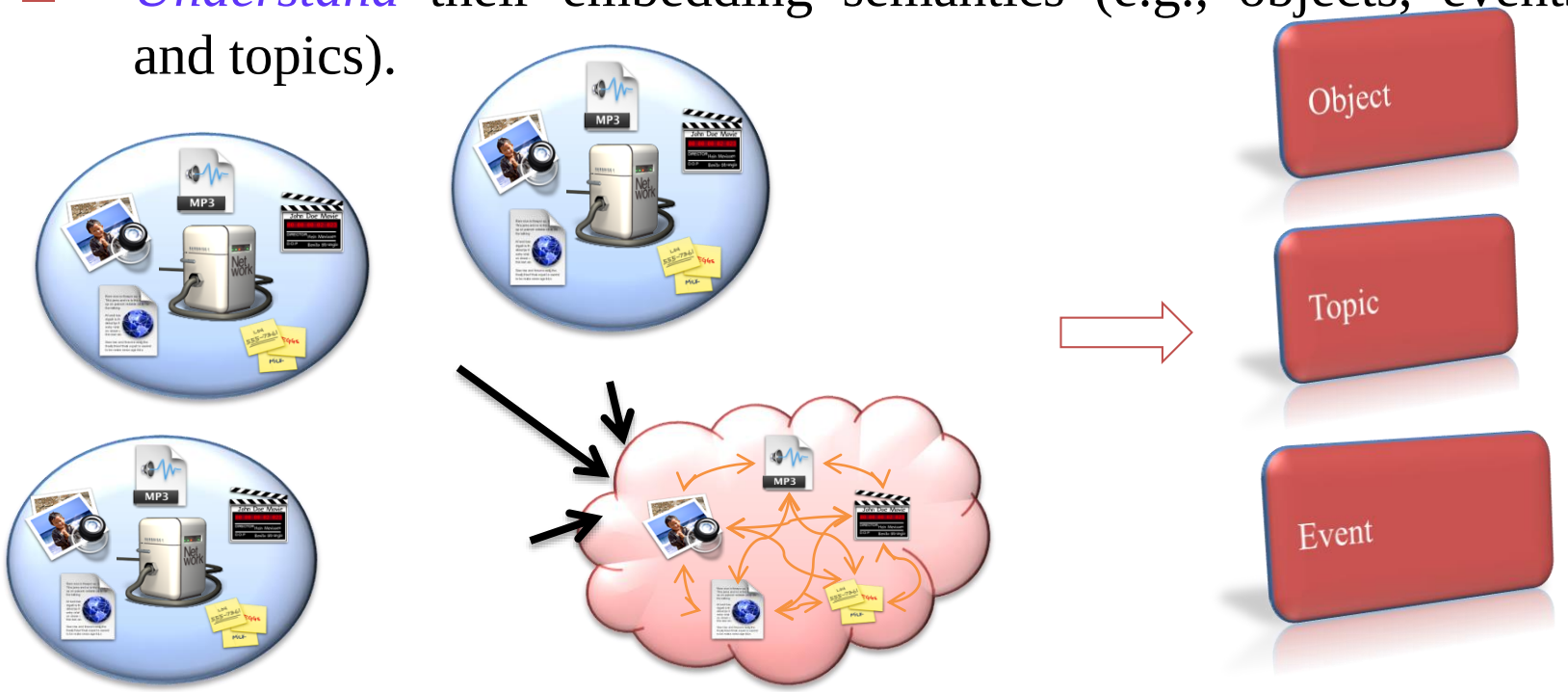
The illustration of the concept of *cross*

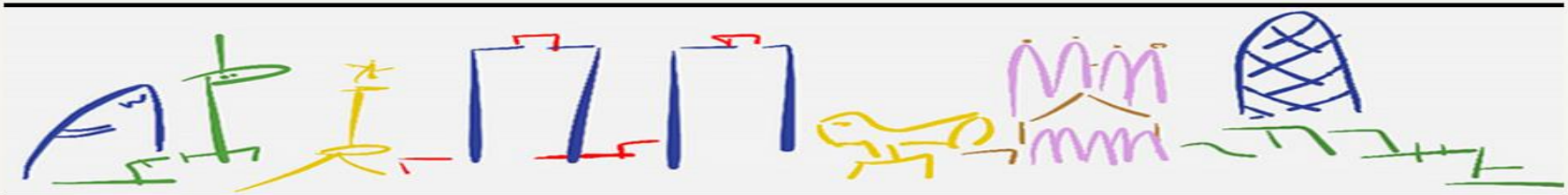
- Multi-modal data is ***connected*** due to their correlations.
- The multiple source data is ***connected*** due to their correlations.



The utilization of cross-media

- Three steps:
  - *Discover* the correlations between data objects
  - *Collect* all of correlated data together
  - *Understand* their embedding semantics (e.g., objects, events and topics).





6. Multimedia Analysis Description

Advances in multimedia analysis have helped enable us to capture, create, and consume multimedia information with unprecedented ease and frequency. In turn, the size of personal and shared multimedia collections and the availability of associated rich contextual and usage information are both growing quickly. Multimedia analysis must evolve to support interaction with substantial personal and shared multimedia collections, often across mobile and desktop environments. The increasingly multi-modal nature of multimedia data collections affords new opportunities for multimedia and cross-media analysis to progress and address the changing demands of multimedia consumers.

This track seeks submissions that contribute to continued progress in information extraction and processing from multimedia data. We actively encourage submissions that incorporate new modalities, sensors, and information sources into traditional multimedia analysis problems. Topics of interest include but are not restricted to:

- Multimedia feature extraction
- Semantic concept detection
- Cross-media analysis
- Multi-modal information processing and fusion
- Temporal or structural analysis of multimedia data
- Machine learning for multimedia analysis
- Scalable processing and scalability issues in multimedia content analysis
- Advanced descriptors and similarity metrics for multimedia data
- Object recognition/detection/segmentation
- 3D content analysis
- Cross-camera content analysis

Cross-media analysis is taken as a main track of
ACM Multimedia 2013

The recent research about cross-media learning

- Cross-media Retrieval
- Cross-media Ranking
- Cross-media Hashing
- Cross-collection Topic Modeling

The recent research about cross-media learning

Cross-media Retrieval

- ***Mission***: support similarity search for multi-modal data, e.g., the retrieval of images in response to a query textual document or vice versa.

Fanno Creek passes through or near 14 parks in several jurisdictions. The Portland Parks and Recreation Department manages three: Hillsdale Park, with picnic tables and a dog park near the headwaters; Albert Kelly Park, with unpaved paths, picnic tables, play areas, and Wi-Fi north of the creek about from the mouth, and the Fanno Creek Natural Area, north of the creek about from the mouth.

Query Textual Document

Retrieved Images

The recent research about cross-media learning

Cross-media Ranking

- ***Mission***: learn one appropriate metric for ranking multi-modal data to preserve the orders of relevance. For example, The retrieved images are ranked in term of their relevance to the query textual document in a listwise manner.

"A Hansom cab is a kind of horse-drawn carriage first designed and patented in 1834 by Joseph Hansom, an architect from Leicestershire, England. Its purpose was to combine speed with safety, with a low center of gravity that was essential for safe cornering. The Hansom Cab was introduced to the United States during the late 19th century, and was most commonly used there in New York City."

The night skyline of Frankfurt, showing the Commerzbank Tower (centre) and the Maintower (right of centre). Frankfurt is the fifth-largest city in Germany, and the surrounding Frankfurt Rhein-Main Region is Germany's second-largest metropolitan area.

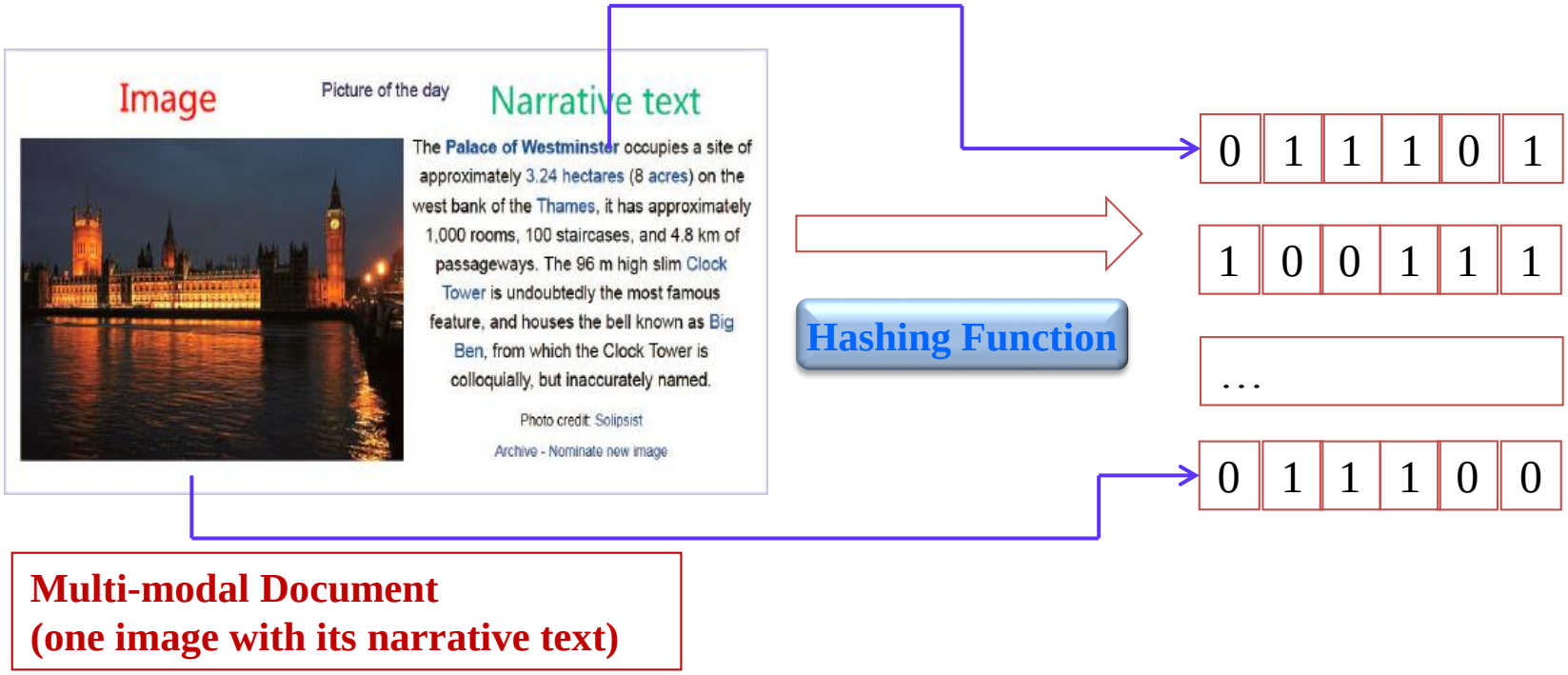
Query Textual Document

Ranked Listwise Image Results

The recent research about cross-media learning

Cross-media Hashing

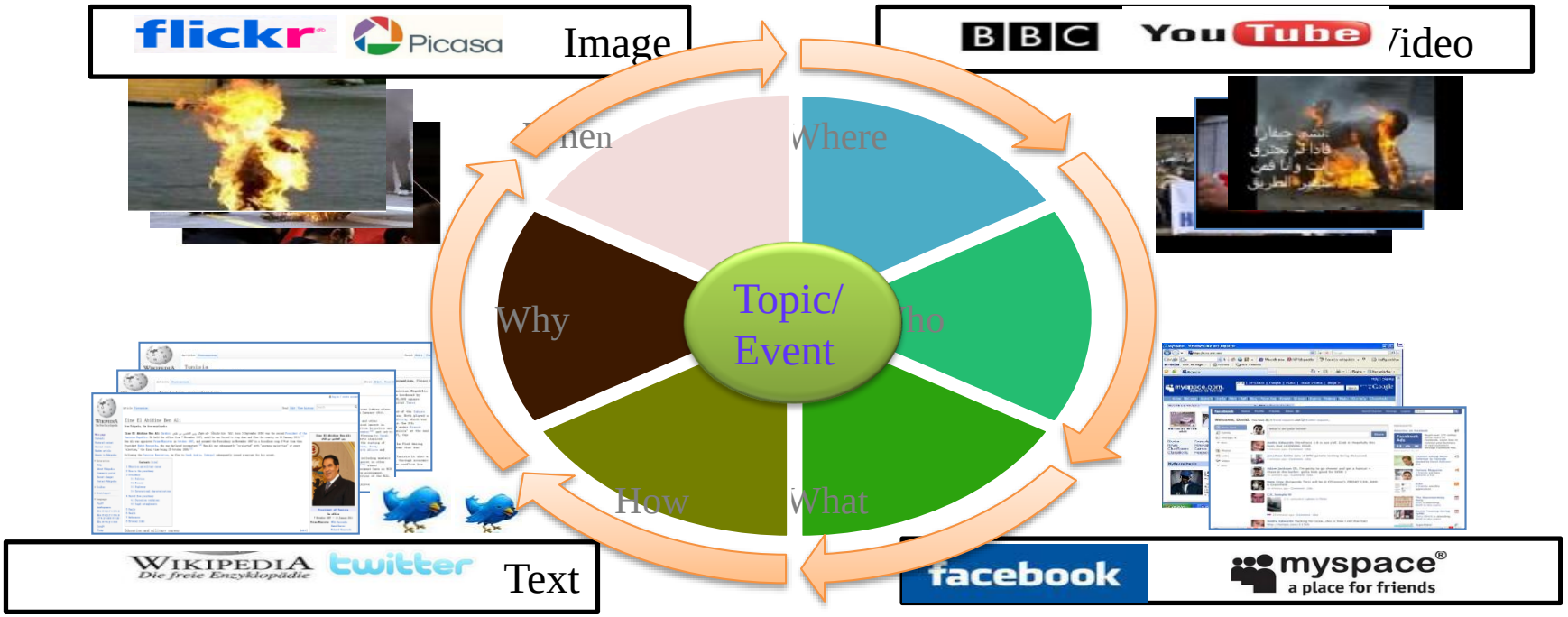
- ***Mission***: attempt to learn hashing function(s) to faithfully preserve the intra-modality and inter-modality similarities and map the high-dimensional multi-modal data to compact binary codes.



The recent research about cross-media learning

Cross-collection topic modeling

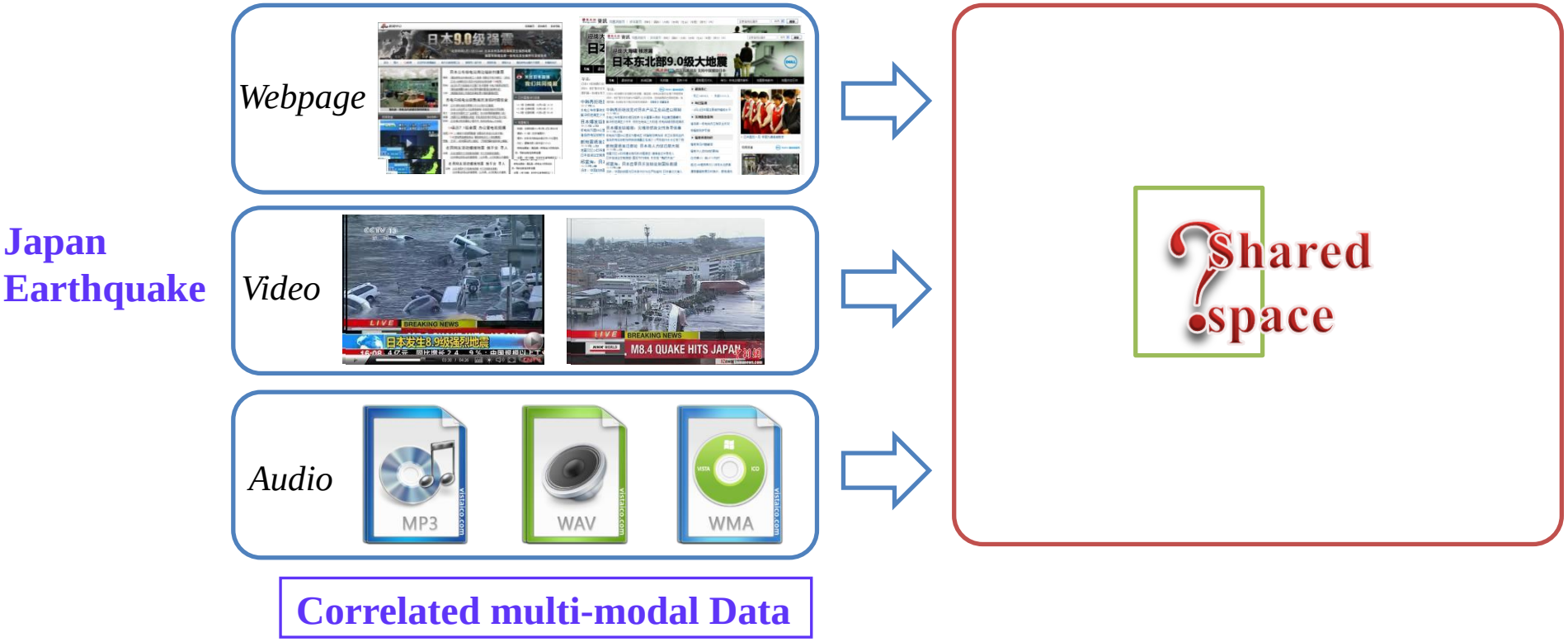
- ***Mission***: describe one topic/event with aspect-oriented (e.g., who-what-how) multi-modal data (e.g., representative images or topical words).



The recent research about cross-media learning

The Challenge

- How to bridge both semantic-gap and heterogeneity gap?



Cross-media retrieval :

Supervised coupled dictionary learning with group structures

- Finding relevant textual documents that best match a given image;
- Or finding a set of images that visually best illustrate a given text description.

- Our Approach:
  - Supervised Coupled Dictionary Learning with Group Structures for Multi-modal Retrieval (**SliM2**)

  - Yueting Zhuang, Yanfei Wang, Fei Wu, Yin Zhang, Weiming Lu, *Supervised Coupled Dictionary Learning with Group Structures for Multi-modal Retrieval*, Proceeding of the Twenty-Seventh Conference on Artificial Intelligence (AAAI), 1070-1076,2013,2013 (**Oral Paper)**

Cross-media retrieval :

Supervised coupled dictionary learning with group structures

Cross-modal metric learning methods can be mainly classified into two kinds of approaches

CCA and its extensions ⇒ Seek subspaces to maximize the correlations between two sets of multidimensional variables.
CCA (*Hotelling 1936*) kernel CCA (Akaho 2006) , sparse CCA (*Hardoon 2011*), structured sparse CCA (*Chen 2012*) and GMA(*Sharma 2012*)

The extensions of LDA ⇒ Model the similarity across different modality through topic proportion which include
Corr-LDA (*Blei 2003*) , tr-mmLDA (Putthividhy 2010), mf-CTM (Salomatin, 2009), MDRF (Jia 2011) etc.

Cross-media retrieval :

Supervised coupled dictionary learning with group structures

- Too restricted to uncontrolled multi-modal data

CCA and its extensions

- Assume the different modality data have a common or a shared subspace

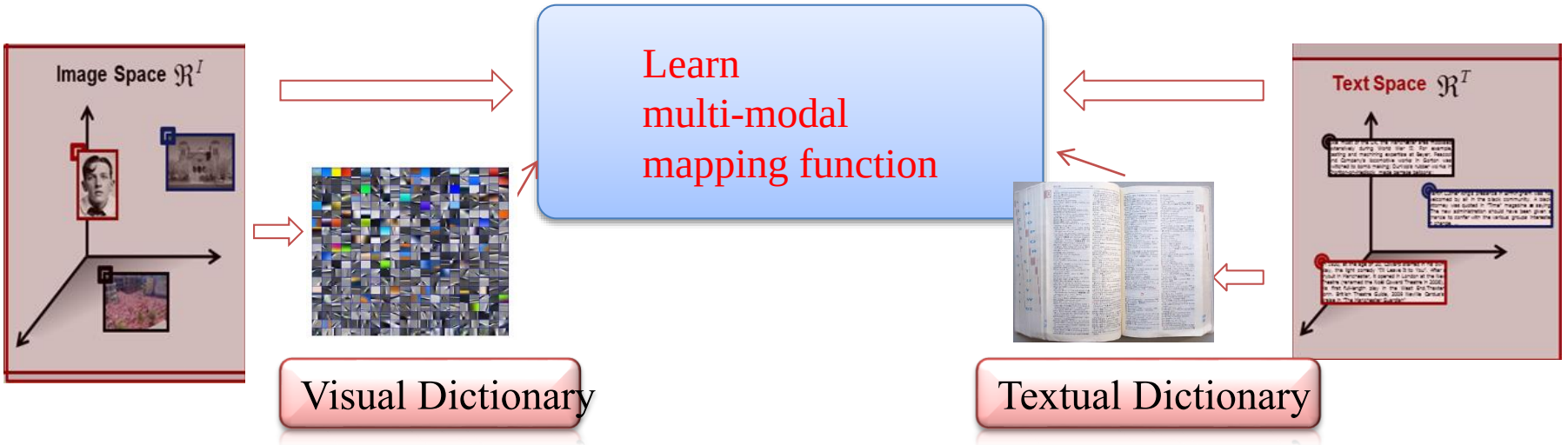
The extensions of LDA

- Assume the different modality data have same topic proportions or Same topic numbers or have one-to-one topic correspondences.

Cross-media retrieval :

Supervised coupled dictionary learning with group structures

- **The proposed Slim2**:
  - Motivated by the fact that *dictionary learning* (DL) methods have the intrinsic power of dealing with the heterogeneous features by generating different dictionaries for multi-modal data.



Cross-media retrieval :

Supervised coupled dictionary learning with group structures

The optimization of Slim2 :

$$\min \sum_{m=1}^{M} \|\mathbf{X}^{(m)} - \mathbf{D}^{(m)}\mathbf{A}^{(m)}\|_F^2 + \sum_{m=1}^{M}\sum_{l=1}^{J} \lambda_m \|\mathbf{A}^{(m)}_{:,\Omega_l}\|_{1,2} + \beta \sum_{m=1}^{M}\sum_{n\neq m} \|\mathbf{A}^{(n)} - \mathbf{W}^{(m)}\mathbf{A}^{(m)}\|_F^2$$

Sparse Reconstruction

Group-structure Preserving with L2-1 norm

Multi-modal Correlation-preserving Mapping

Algorithm 1 The optimization of SliM²

Input The labeled training set of N pairs data with M modalities from J classes $\{(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(M)}, l_i)\} \in \{(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(M)}, \mathbf{L})\}$.

1: Initialize $D = \{\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \cdots, \mathbf{D}^{(M)}\}$ and $W = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \cdots, \mathbf{W}^{(M)}\}$,

2: Optimize $A = \{\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(M)}\}$ by Eq.(6),

3: Update $D = \{\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \cdots, \mathbf{D}^{(M)}\}$ with other variables fixed using Eq.(7),

4: Update $W = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \cdots, \mathbf{W}^{(M)}\}$ with other variables fixed using Eq.(9),

5: Repeat 2-4 till convergence.

Output multi-modal dictionaries D and a set of mapping functions W

Algorithm 2 The multi-modal retrieval by SliM²

Input The leaned multi-modal dictionaries $D = \{\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \cdots, \mathbf{D}^{(M)}\}$ and a set of mapping functions $W = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \cdots, \mathbf{W}^{(M)}\}$ from training data and query data $\mathbf{x}_q^{(m)} \in R^{P_m}$ in the m-th modality

1: Initialize $\alpha_q^{(m)}$,$\alpha_r^{(n)}$ and corresponding retrieval $\mathbf{x}_r^{(n)}$ using equation(10),

2: Optimize $\hat{\alpha}_q^{(m)}$,$\hat{\alpha}_r^{(n)}$ with other variables fixed using equation (11),

3: Update $\hat{\mathbf{x}}_r^{(n)}$ using equation (12),

4: Repeat 2-3 till convergence.

5: the ranked neighbors of $\hat{\mathbf{x}}_r^{(n)}$.

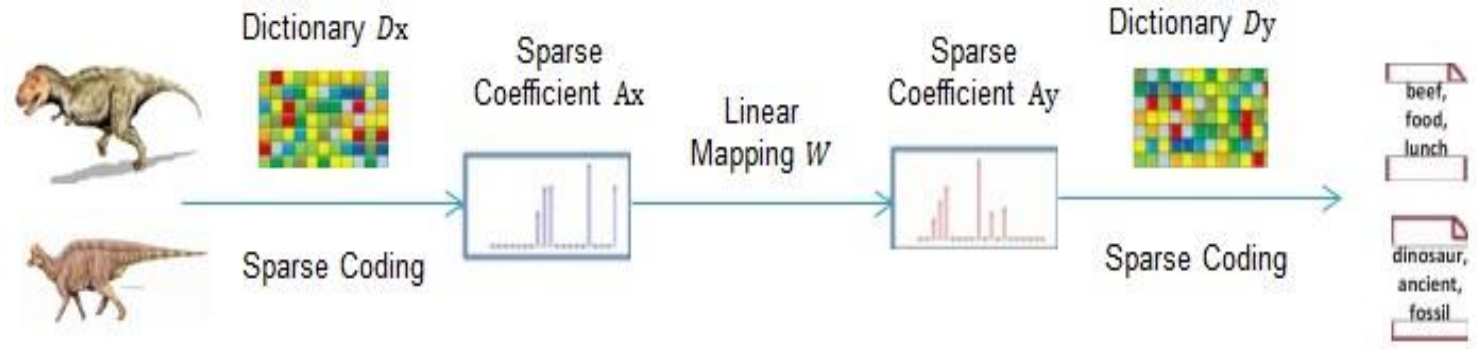
Output The retrieved similar data in the n-th modality

Cross-media retrieval :

Supervised coupled dictionary learning with group structures

- **The advantages of the proposed Slim2 :**
- ***Group-structures preserving***: *encourage* the reconstruction of data from the same group(e.g., class) by the same dictionary elements.

- ***Multi-modal correlation mapping***: *learn* a relatively simple mapping function across modalities.



Cross-media retrieval :

Supervised coupled dictionary learning with group structures

Wiki	BoVW(500D),BoW(1000D)		BoVW(1000D),BoW(5000D)	
	Image Query Text	Text query Image	Image Query Text	Text query Image
CCA	0.1767	0.1809	0.1994	0.1859
GMA	0.2245	**0.2148**	0.2093	**0.2267**
SCDL	0.2341	0.1988	0.2527	0.1981
SliM^2	**0.2399**	0.2025	**0.2548**	0.2021

The performance comparison in terms of MAP scores on Wiki data set

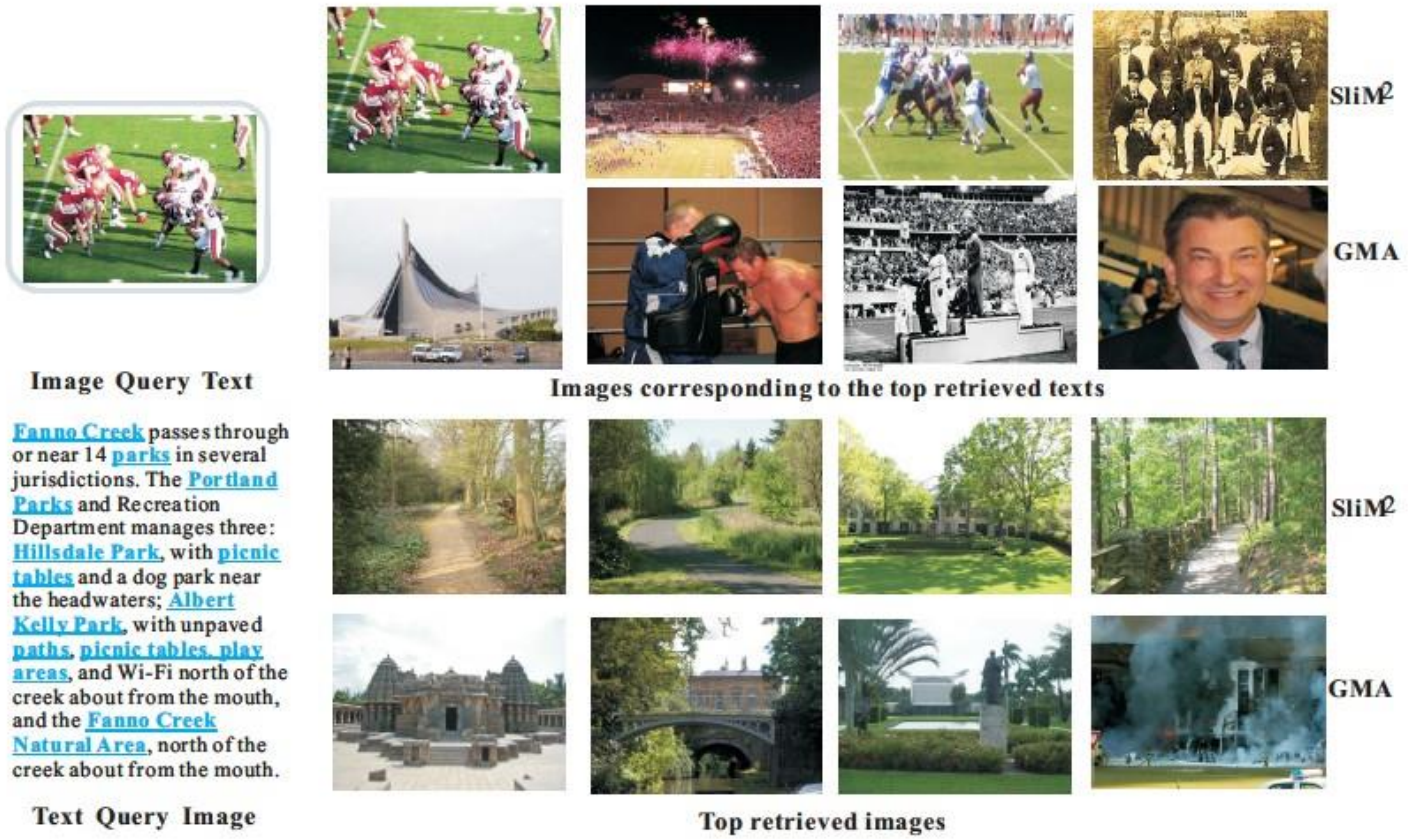
Wiki	BoVW(500D),BoW(1000D)		BoVW(1000D),BoW(5000D)	
	Image Query Text	Text Query Image	Image Query Text	Text Query Image
CCA	0.2236	0.2340	0.3054	0.2845
GMA	0.2877	0.2548	0.3002	0.2496
SCDL	0.3709	0.2790	0.3857	0.3037
SliM^2	**0.3899**	**0.2842**	**0.4084**	**0.3106**

The performance comparison in terms of Percentage scores on Wiki data set

CCA(Hotelling 1936) GMA(Sharma et al. 2012) SCDL(Wang et al. 2012)

Cross-media retrieval :

Supervised coupled dictionary learning with group structures



Examples of image query text and text query image over Wiki data set by SliM2 (top row) and GMA (bottom row)

Cross-media Ranking :

Bi-directional Structural Learning to Rank

- Learn a multi-modal ranking function to preserve the orders of relevance of multi-modal data.

- Our Approach:
  - Bi-directional Structural Learning to Rank

    - Xinyan Lu, Fei Wu, Siliang Tang,Zhongfei Zhang, Xiaofei He, Yueting Zhuang, A low rank structural large-margin method for cross-modal ranking, *SIGIR 2013* (**Full Paper**)
    - Fei Wu, Xinyan Lu, Yin Zhang, Zhongfei Zhang, Shuicheng Yan, Yueting Zhuang,Cross-Media Semantic Representation via Bi-directional Learning to Rank, *Proceedings of the 2013 ACM International Conference on Multimedia (ACM Multimedia*, **Full Paper**),2013

Cross-media Ranking :

Bi-directional Structural Learning to Rank

- Bi-directional structural learning to rank means that both text-query-image and image-query-text ranking examples are utilized in the training period.

- This is a general cross-media ranking algorithm to optimize the bi-directional listwise ranking loss with a latent space embedding.

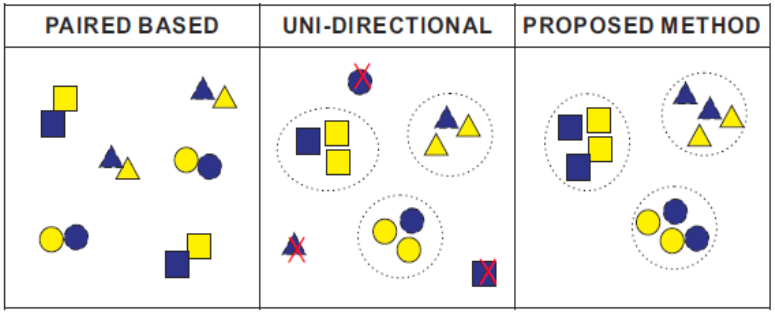


Figure 1: A simple demonstration of the latent spaces learned by different approaches. The same shape indicates relevant semantics. Colors represent modalities (i.e., text and imagery). The paired-based methods like CCA try to unite paired samples only. The uni-directional-ranking-based methods like PAMIR and SSI only capture the relationship between two modalities from one direction of retrieval but their generalization performances are limited since they do not capture the latent structure of the query modality, which is represented as blue queries with red cross in the figure. The proposed method Bi-CMSRM is trained with bi-directional training examples by which it can be applied to both directions of retrieval and the generalization performance is improved.

Cross-media Ranking :

Bi-directional Structural Learning to Rank

Image-query-text direction

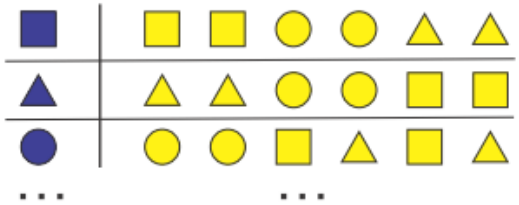
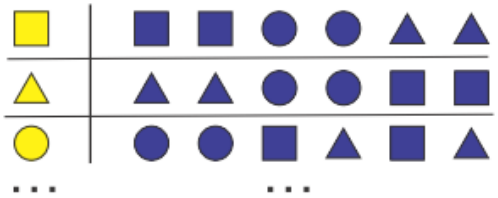


Image queries Ranked text docum

Text-query-image direction



Text queries Ranked images

Text-to-image correlation

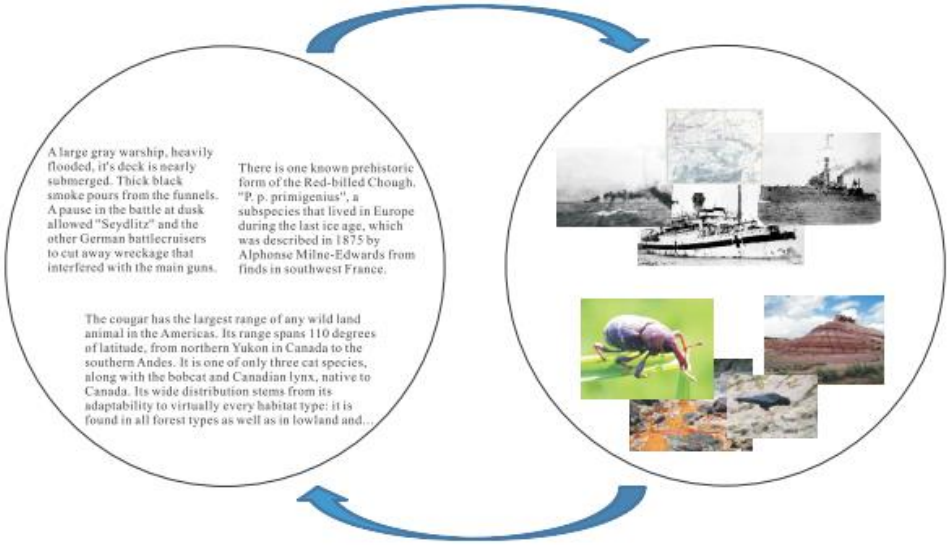
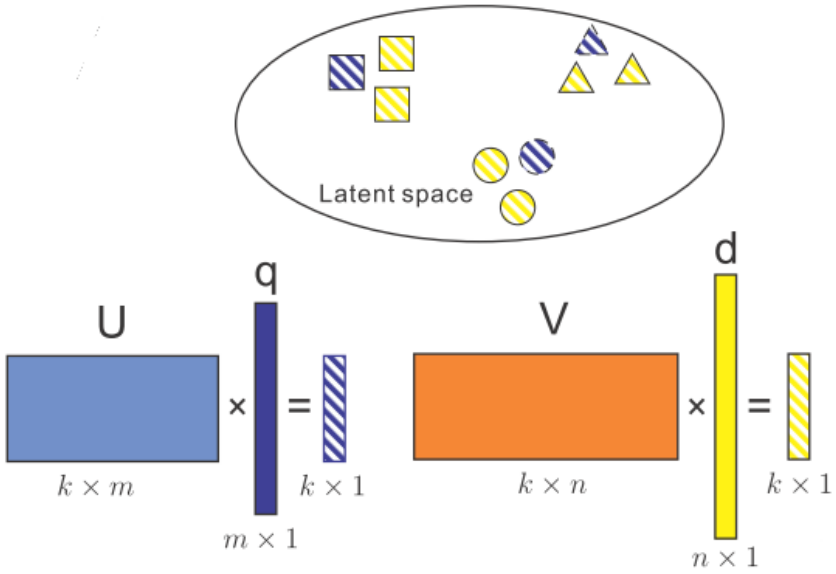


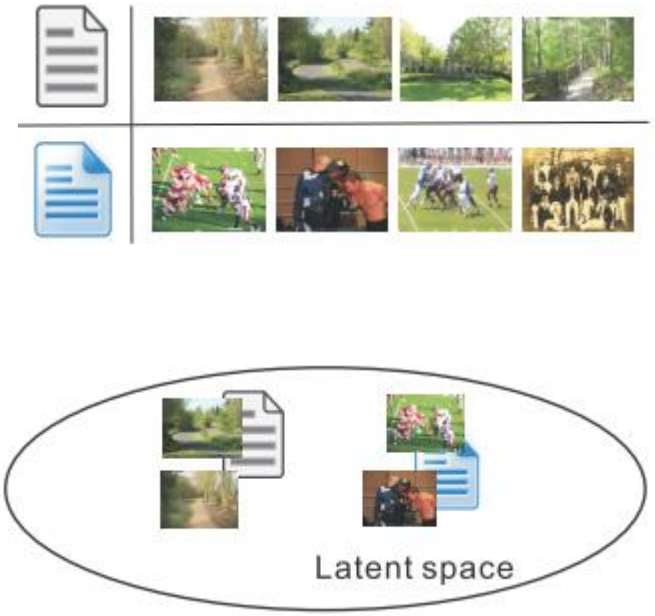
Image-to-text correlation

Modeling of Multi-modal Correlation

Cross-media Ranking :

Bi-directional Structural Learning to Rank





All the ***m***-dimensional queries **q** and ***n***-dimensional target documents **d** are mapped to a ***k***-dimensional latent space by **U** and **V** respectively, in which those data objects with the same semantics are grouped to minimize certain listwise ranking loss (e.g., MAP) directly

Cross-media Ranking :

Bi-directional Structural Learning to Rank

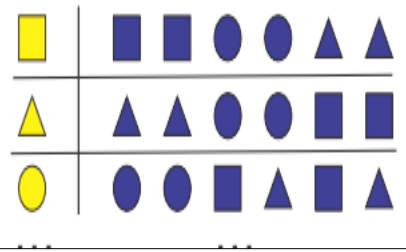
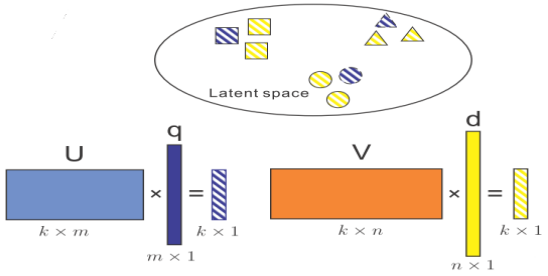
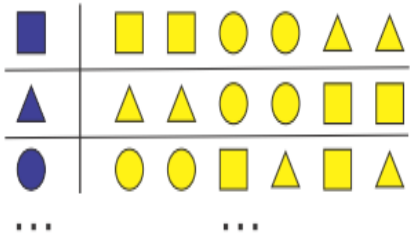


Image-query-text direction



Latent semantic embedding



Text-query-image direction

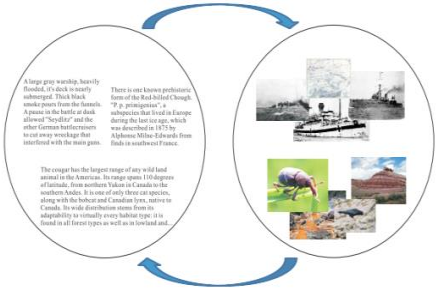
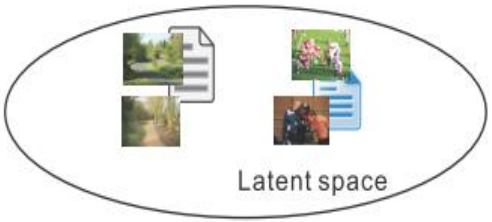


Image-to-text correlation
and Text-image correlation



bi-directional ranking examples

bi-directional structural Learning with low-rank penalty

Ranking function

Cross-media Ranking :

Bi-directional Structural Learning to Rank

- The objective of *Bi-directional structural learning to rank* tends to maximize the margins between the true ranking and all the other possible rankings of the target documents for each query in the other modality

Structural risk

Bi-directional empirical risk

$$\min_{U,V,\xi_1,\xi_2} \frac{\lambda}{2}\|U\|_F^2 + \frac{\lambda}{2}\|V\|_F^2 + \frac{1}{N}\sum_{i=1}^{N}\xi_{1,i} + \frac{1}{M}\sum_{j=N+1}^{N+M}\xi_{2,j}$$

$$s.t. \quad \forall i \in \{1,\ldots,N\}, \forall \mathbf{y} \in \mathcal{Y}:$$

$$\delta F(t_i, \mathbf{p}_i, \mathbf{y}) \geq \Delta(\mathbf{y}_i^*, \mathbf{y}) - \xi_{1,i}$$

The penalty for text-query-image direction

$$\forall j \in \{N+1,\ldots,N+M\}, \forall \mathbf{y} \in \mathcal{Y}$$

$$\delta F(p_j, \mathbf{t}_j, \mathbf{y}) \geq \Delta(\mathbf{y}_j^*, \mathbf{y}) - \xi_{2,j}\,.$$

The penalty for image-query-text direction

Cross-media Ranking :

Bi-directional Structural Learning to Rank

Algorithm 1 Bi-directional Cross-Media Semantic Representation Model (Bi-CMSRM).

Input: text-query samples $(t_i, \mathbf{p}_i, \mathbf{y}_i^*), i = 1, \ldots, N$, image-query samples $(p_j, \mathbf{t}_j, \mathbf{y}_j^*), j = N+1, \ldots, N+M$, trade-off control parameter $\lambda > 0$, accuracy tolerance threshold $\epsilon > 0$

Output: mapping parameters U and V, slack variables $\xi_1 \geq 0$ and $\xi_2 \geq 0$

1: $\mathcal{W}_1 \leftarrow \emptyset$, $\mathcal{W}_2 \leftarrow \emptyset$

2: **repeat**

3: Solve for the optimal U, V and slack ξ_1, ξ_2:

$$\min_{U,V,\xi_1,\xi_2} \quad \frac{\lambda}{2}\|U\|_F^2 + \frac{\lambda}{2}\|V\|_F^2 + \xi_1 + \xi_2$$

$$\text{s.t.} \quad \forall (\mathbf{y}_1, \ldots, \mathbf{y}_N) \in \mathcal{W}_1 :$$

$$\frac{1}{N}\sum_{i=1}^{N} \delta F(t_i, \mathbf{p}_i, \mathbf{y}_i) \geq \frac{1}{N}\sum_{i=1}^{N} \Delta(\mathbf{y}_i^*, \mathbf{y}_i) - \xi_1$$

$$\forall (\mathbf{y}_{N+1}, \ldots, \mathbf{y}_{N+M}) \in \mathcal{W}_2 :$$

$$\frac{1}{M}\sum_{j=N+1}^{N+M} \delta F(p_j, \mathbf{t}_j, \mathbf{y}_j) \geq \frac{1}{M}\sum_{j=N+1}^{N+M} \Delta(\mathbf{y}_j^*, \mathbf{y}_j) - \xi_2$$

4: **for** $i = 1$ **to** N **do**

5: $\hat{\mathbf{y}}_i \leftarrow \underset{\mathbf{y} \in \mathcal{Y}}{\text{argmax}} \Delta(\mathbf{y}_i^*, \mathbf{y}) + F(t_i, \mathbf{p}_i, \mathbf{y}_i)$

6: **end for**

7: $\mathcal{W}_1 \leftarrow \mathcal{W}_1 \cup (\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_N)$

8: **for** $j = N+1$ **to** $N+M$ **do**

9: $\hat{\mathbf{y}}_j \leftarrow \underset{\mathbf{y} \in \mathcal{Y}}{\text{argmax}} \Delta(\mathbf{y}_j^*, \mathbf{y}) + F(p_j, \mathbf{t}_j, \mathbf{y}_j)$

10: **end for**

11: $\mathcal{W}_2 \leftarrow \mathcal{W}_2 \cup (\hat{\mathbf{y}}_{N+1}, \ldots, \hat{\mathbf{y}}_{N+M})$

12: **until**

$$\frac{1}{N}\sum_{i=1}^{N} \Delta(\hat{\mathbf{y}}_i^*, \hat{\mathbf{y}}_i) - \frac{1}{N}\sum_{i=1}^{N} \delta F(t_i, \mathbf{p}_i, \hat{\mathbf{y}}_i) \leq \xi_1 + \epsilon$$

and

$$\frac{1}{M}\sum_{j=N+1}^{N+M} \Delta(\hat{\mathbf{y}}_j^*, \hat{\mathbf{y}}_j) - \frac{1}{M}\sum_{j=N+1}^{N+M} \delta F(p_j, \mathbf{t}_i, \hat{\mathbf{y}}_i) \leq \xi_2 + \epsilon$$

13: **return** U, V, ξ_1, ξ_2;

Cross-media Ranking :

Bi-directional Structural Learning to Rank

	Text Query (R=50)	Text Query (R=all)	Image Query (R=50)	Image Query (R=all)
CCA	0.2343	0.1433	0.2208	0.1451
PAMIR	0.3093	0.1734	0.1797	0.1779
SSI	0.2821	0.1664	0.2344	0.1759
Uni-CMSRM	0.3663	0.2021	0.2570	0.2229
Bi-CMSRM	**0.3981**	**0.2123**	**0.2599**	**0.2528**

Wikipedia dataset in terms of MAP@R

	Text Query (R=50)	Text Query (R=all)	Image Query (R=50)	Image Query (R=all)
CCA	0.1497	0.0851	0.1523	0.0883
PAMIR	0.2046	0.1184	**0.5003**	0.2410
SSI	0.2156	0.1140	0.4101	0.1992
Uni-CMSRM	0.2781	0.1424	0.4997	**0.2491**
Bi-CMSRM	**0.3224**	**0.1453**	0.4950	0.2380

NUS-WIDE dataset in terms of MAP@R

Cross-media Ranking :

Bi-directional Structural Learning to Rank

Figure 3: Exemplar retrieval comparison between the proposed Bi-CMSRM and Uni-CMSRM on the Wiki dataset. For text-query-image direction, the query text is shown with its corresponding image and selected words. For the image-query-text direction, the retrieved documents are shown with their corresponding images.

Cross-media Hashing

Hashing is promising way to speed up the ANN (approximate nearest neighbor) similarity search, which makes a tradeoff between accuracy and efficiency.

Hashing Function

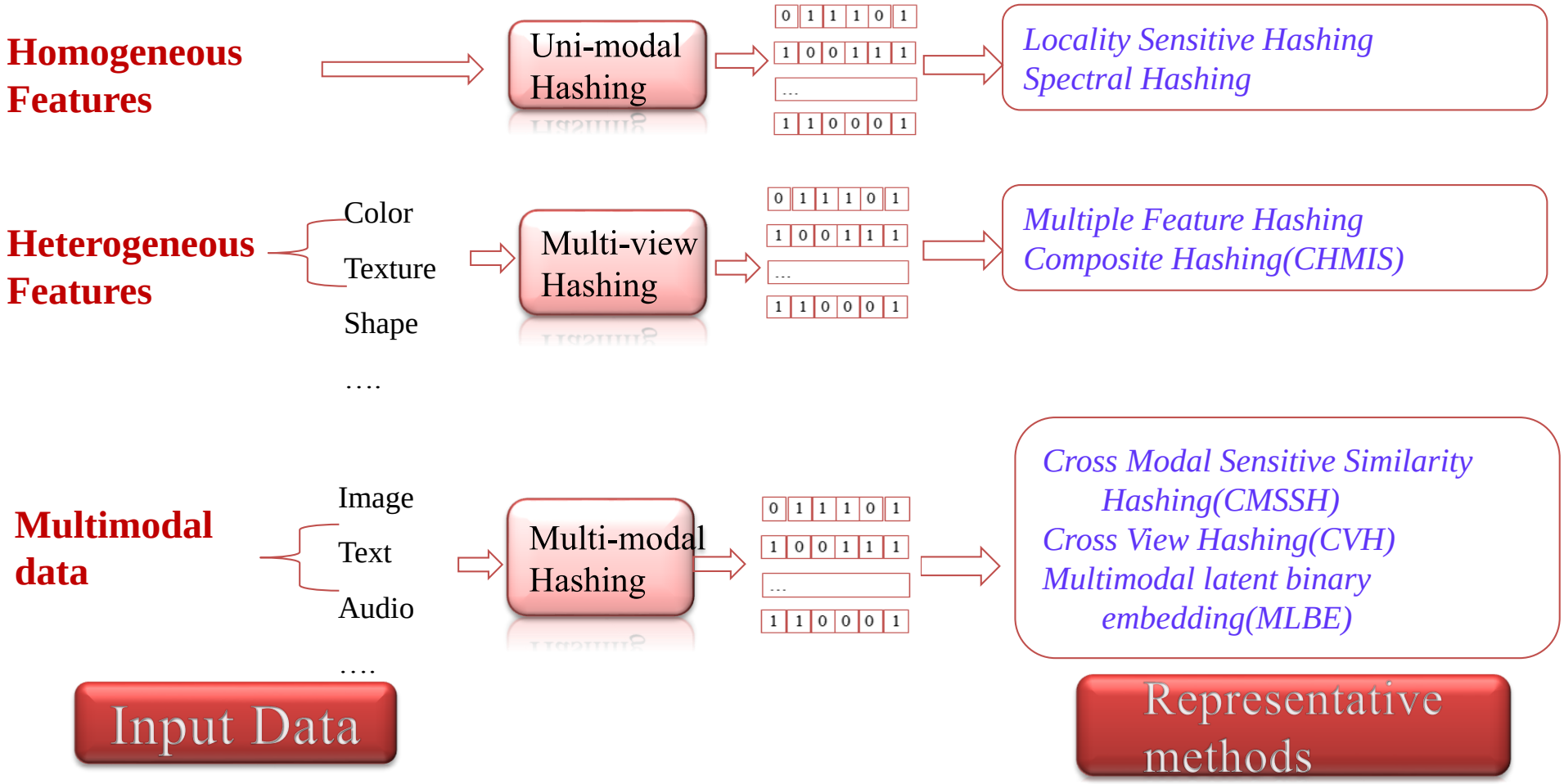
0	1	1	1	0	1
1	0	0	1	1	1
…					
1	1	0	0	0	1

High-dimensional features

Compact binary codes

Cross-media Hashing

- Three kinds of hashing approaches



Cross-media Hashing:

Sparse Multi-modal Hashing

- Multi-modal hashing tends to utilize the intrinsic intra-modality and inter-modality similarity to learn the appropriate relationships of the data objects and provide efficient search across different modalities

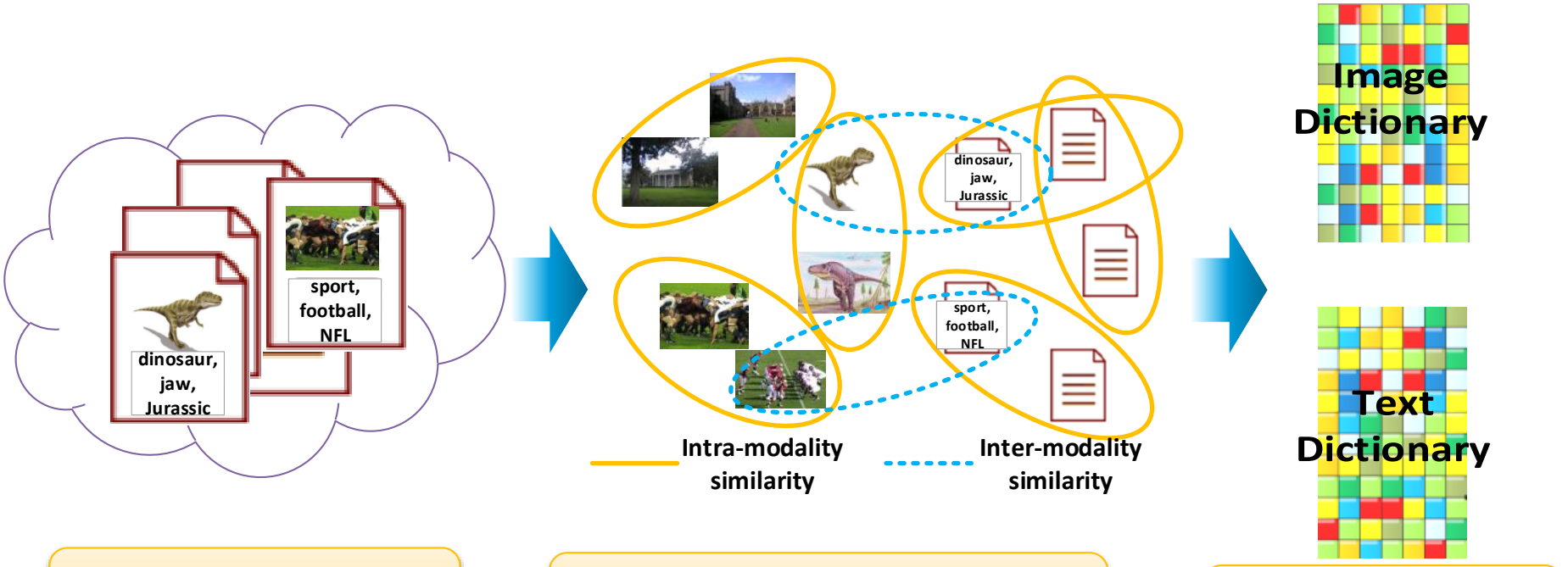
- Approach: Sparse Multi-modal Hashing

- Fei Wu, Zhou Yu, Yi Yang, Siliang Tang, Yueting Zhuang, Sparse multi-modal hashing, IEEE Transactions Multimedia

Cross-media Hashing:

Sparse Multi-modal Hashing

- Step 1: The Joint Learning of Multi-modal Dictionaries



Cross-media Hashing:

Sparse Multi-modal Hashing

- Our approach is formulated by coupling the multi-modal dictionary learning (in terms of approximate reconstruction of each data object with a weighted linear combination of a small number of "basis vectors") and a regularized hypergraph penalty (in terms of the modeling of multi-modal correlation).

Sparse Reconstruction

Hypergraph Laplacian Penalty

$$\min_{\mathbf{A},\mathbf{D}^x,\mathbf{D}^y} \quad \|\mathbf{X} - \mathbf{D}^x\mathbf{A}^x\|_F^2 + \|\mathbf{Y} - \mathbf{D}^y\mathbf{A}^y\|_F^2 + \Omega(\mathbf{A})$$

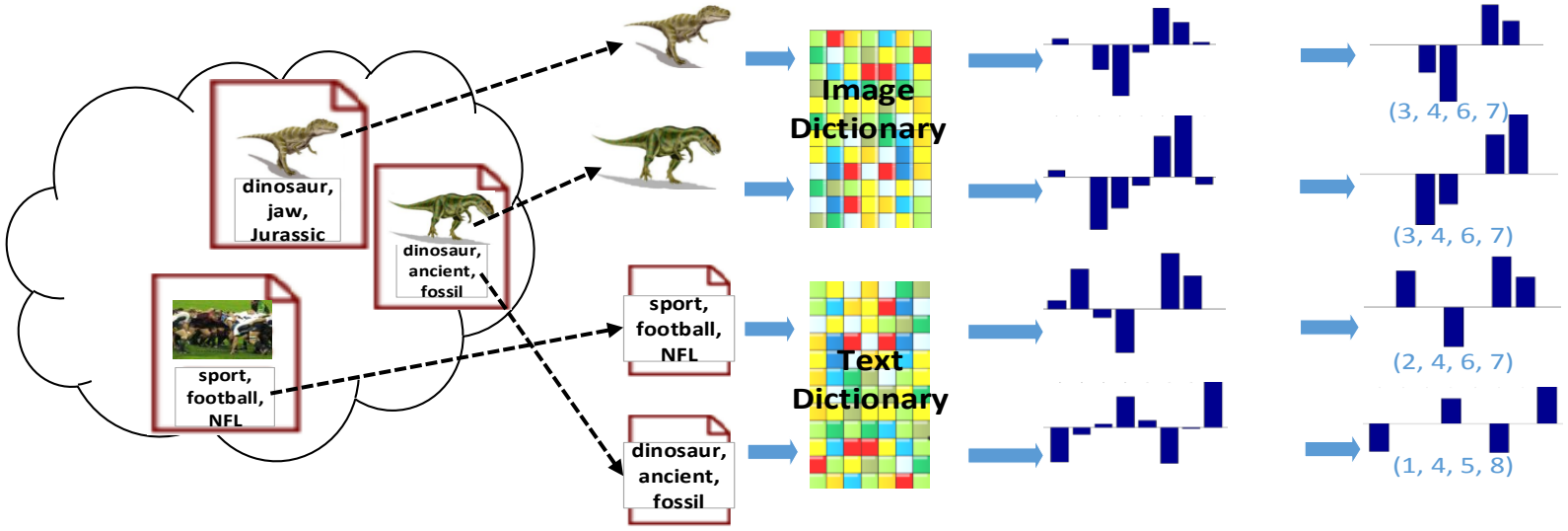
$$s.t. \quad \|\mathbf{d}_k^x\|_F^2 \le 1, \ \|\mathbf{d}_k^y\|_F^2 \le 1 \quad \forall k = 1, 2, \ldots K$$

$$\begin{aligned}\Omega(\mathbf{A}) &= \lambda\|\mathbf{A}\|_1 + \frac{\alpha}{2}\sum_{e\in E}\sum_{\{i,j\}\subseteq e}\frac{w(e)}{\delta(e)}\|a_i - a_j\|^2 \\ &= \lambda\|\mathbf{A}\|_1 + \alpha\mathrm{Tr}(\mathbf{A}\mathbf{L}_h\mathbf{A}^T)\end{aligned}$$

Cross-media Hashing:

Sparse Multi-modal Hashing

- Step 2: The Generation of Sparse Codesets



Multi-modal objects | Sparse Reconstruction | Sparse codesets

Both intra-modality and inter-modality similarities are preserved. For examples, two "dinosaur" images have the same sparse codeset, and two "dinosaur" images have similar sparse codesets with their relevant text (dinosaur, ancient and fossil, *etc*). On the contrary, two "dinosaur" images have apparently different sparse codesets with their irrelevant text(sport, football, *etc*).

Cross-media Hashing:

Sparse Multi-modal Hashing

- activate the most relevant component and induce a compact codeset for each data from its corresponding sparse coefficients

$$\min_{a_q} \ \|x_q - \mathbf{D}^x a_q\|_F^2 + \lambda_x \|a_q\|_1$$

$$\begin{aligned} \mathbf{SC}_+(a^x) &= \{\, i \mid \forall i \in 1,2,...,K, \ \text{if } a_i^x > \sigma \} \\ \mathbf{SC}_-(a^x) &= \{\, i \mid \forall i \in 1,2,...,K, \ \text{if } a_i^x < -\sigma \} \end{aligned}$$

$$Sensitive - Sim(a^x, a^y) = \frac{1}{2}\left(\frac{|\mathbf{SC}_+(a^x) \bigcap \mathbf{SC}_+(a^y)|}{|\mathbf{SC}_+(a^x) \bigcup \mathbf{SC}_+(a^y)|} + \frac{|\mathbf{SC}_-(a^x) \bigcap \mathbf{SC}_-(a^y)|}{|\mathbf{SC}_-(a^x) \bigcup \mathbf{SC}_-(a^y)|} \right)$$

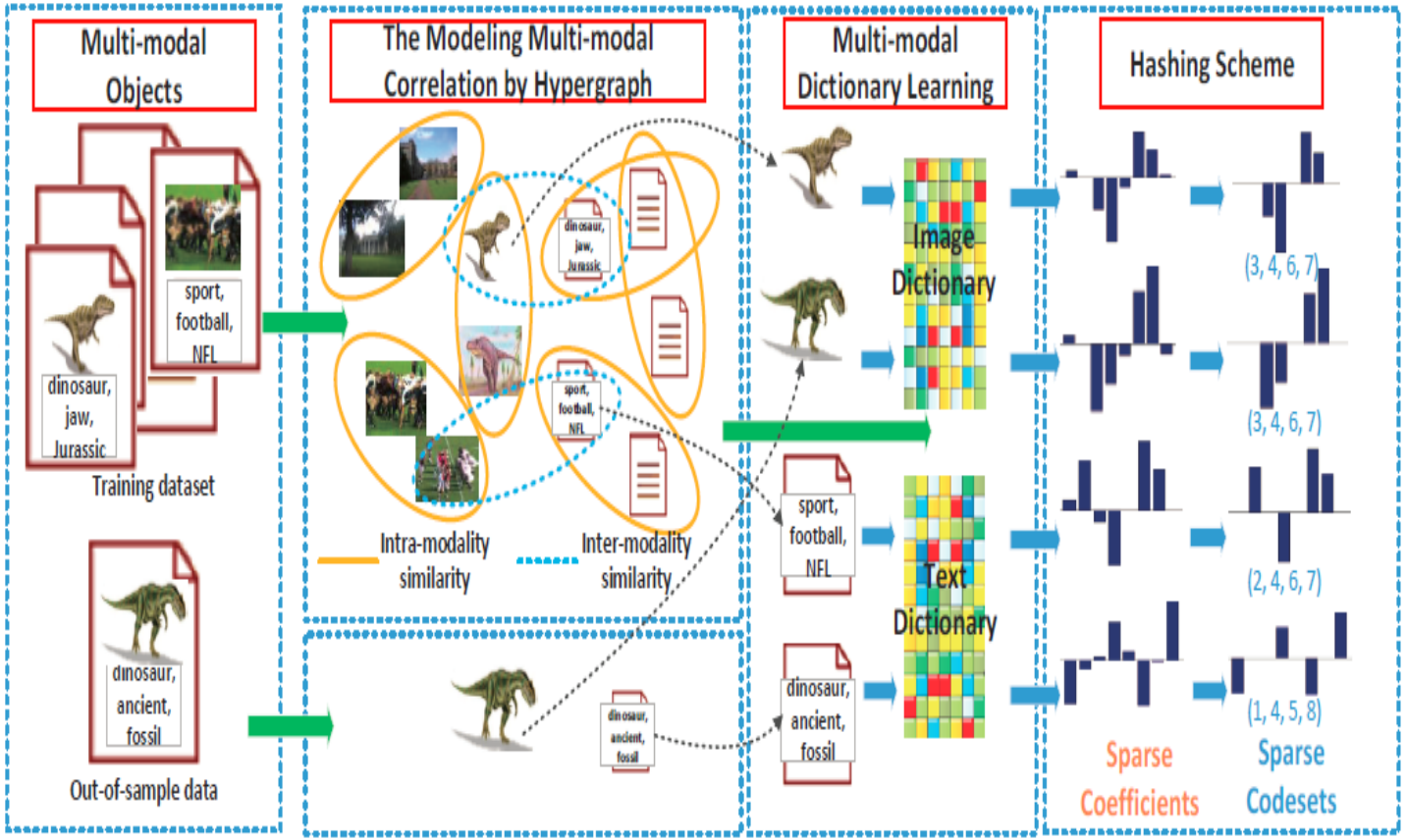


Fig. 1. The algorithmic flowchart of our proposed SM^{2}H. For the sake of illustrative simplicity, we assume only two kinds of data objects (i.e., images and texts) here. A hypergraph is first constructed to model the correlations between multi-modal data objects, then the multi-modal dictionaries are jointly learned to obtain one image dictionary and one text dictionary respectively. Each data object can be succinctly represented using a limited corresponding dictionary atoms and the corresponding sparse coefficients. Finally, the hashing scheme is conducted to identify those significantly informative component(i.e., the sparse codes with large coefficients). The selected component indices are used to construct a sparse codeset for each data object. We can observe the sparse codesets well preserve both intra-modality similarity and the inter-modality similarity. For examples, two "dinosaur" images have the same sparse codeset, and two "dinosaur" images have similar sparse codesets with their relevant text (dinosaur, ancient and fossil, *etc*). On the contrary, two "dinosaur" images have apparently different sparse codesets with their irrelevant text(sport, football, *etc*).

Cross-media Hashing:

Sparse Multi-modal Hashing

TABLE II. THE PERFORMANCE COMPARISON IN TERMS OF MAP SCORES ON NUS-WIDE DATA SET WITH CODE LENGTH m EQUALS TO 16, 32, 48 AND 64, AND THE EQUIVALENT SIZE OF SPARSE CODESETS IN SM^{2}H. THE ITEMS SHOWN IN BOLD ARE THE TWO BEST RESULTS, THE RESULTS WITH ASTERISK ARE THE BEST

Task	Methods	code lengths or equivalent sizes of sparse codesets			
		$m = 16$ or C(25,5)	$m = 32$ or C(50,10)	$m = 48$ or C(100,10)	$m = 64$ or C(150,15)
image-query-texts	CMSSH	**0.4410**	**0.4364**	0.3878	0.3865
	CVH	0.3673	0.3726	0.3652	0.3573
	MLBE	0.3989	0.3737	0.3546	0.3485
	SM^{2}H$_1$	0.4279	0.3855	**0.4506**	**0.4460**
	SM^{2}H$_2$	**0.4496***	**0.4529***	**0.4801***	**0.4520***
text-query-images	CMSSH	0.4013	0.389	0.3699	0.3622
	CVH	0.3771	0.3629	0.3427	0.3400
	MLBE	0.3989	0.3678	0.3451	0.3453
	SM^{2}H$_1$	**0.4689***	**0.4660***	**0.4879***	**0.4457**
	SM^{2}H$_2$	**0.4577**	**0.4465**	**0.4487**	**0.4685***

TABLE IV. THE PERFORMANCE COMPARISON IN TERMS OF MAP SCORES ON WIKI DATA SET WITH CODE LENGTH m EQUALS TO 16, 32, 48 AND 64, AND THE EQUIVALENT SIZE OF SPARSE CODESETS IN SM^{2}H. 500-D BAG-OF-VISUAL-WORDS (BOVW) AND 1,000-D BAG-OF-TEXTUAL-WORDS (BOW), AS WELL AS 1,000-D BAG-OF-VISUAL-WORDS (BOVW) AND 5,000-D BAG-OF-TEXTUAL-WORDS (BOW), ARE USED TO REPRESENT EACH IMAGE AND TEXT RESPECTIVELY. THE ITEMS SHOWN IN BOLD ARE THE TWO BEST RESULTS, THE RESULTS WITH ASTERISK ARE THE BEST

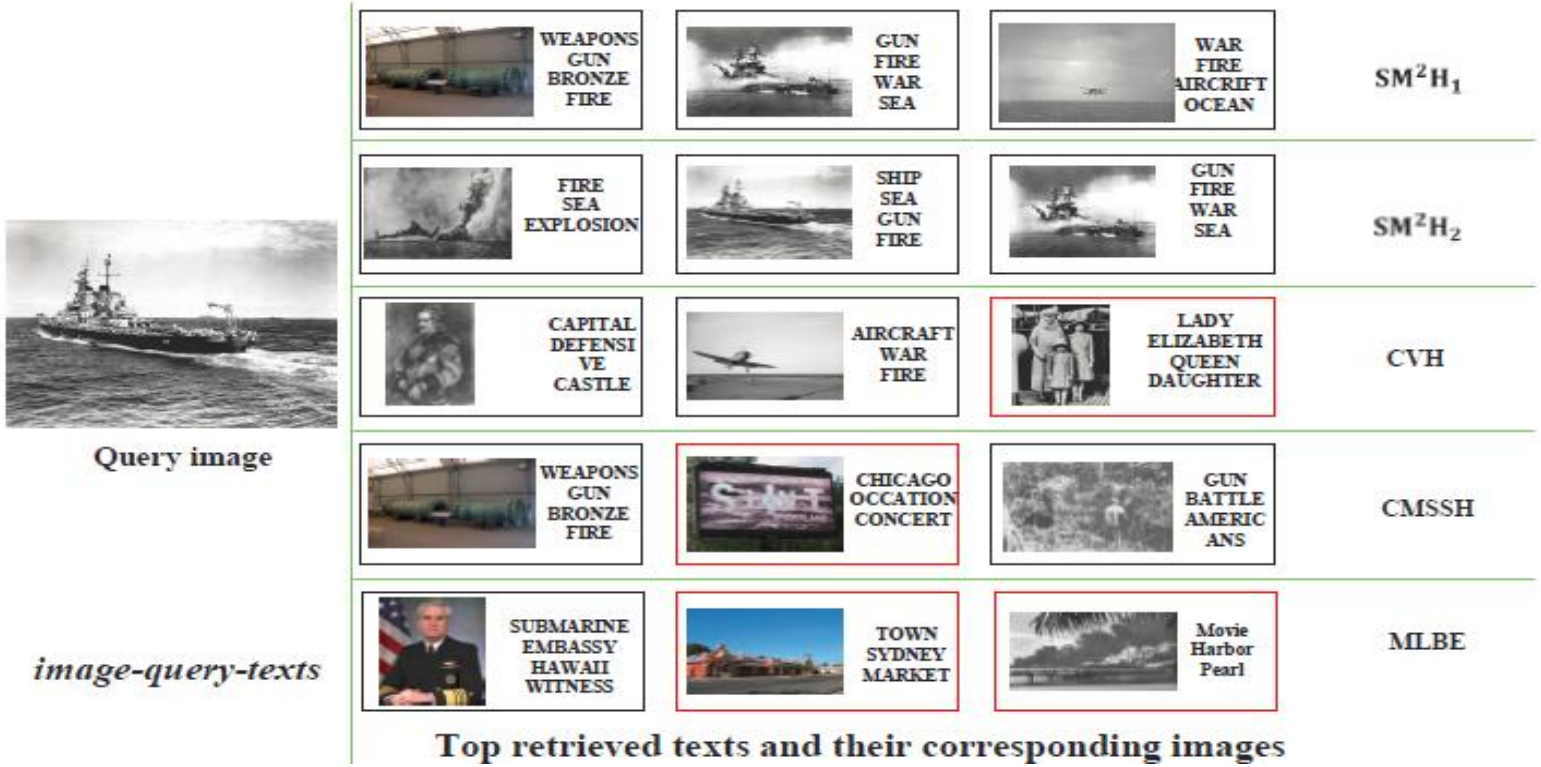
Task	Methods	BoVW(500D),BoW(1000D) code lengths or equivalent sizes of sparse codeset				BoVW(1000D),BoW(5000D) code lengths or equivalent sizes of sparse codesets			
		$m = 16$ or C(25,5)	$m = 32$ or C(50,10)	$m = 48$ or C(100,10)	$m = 64$ or C(150,15)	$m = 16$ or C(25,5)	$m = 32$ or C(50,10)	$m = 48$ or C(100,10)	$m = 64$ or C(150,15)
image-query-texts	CMSSH	0.1697	**0.1895***	0.1806	0.1701	0.1914	0.1890	0.1921	0.1905
	CVH	0.1897	0.1828	0.1850	0.1858	**0.2064**	0.2040	0.2017	0.1962
	MLBE	0.1737	0.1759	0.1709	0.1751	0.2030	0.2027	0.1889	0.1804
	SM^{2}H$_1$	**0.1912**	0.1703	**0.2029**	**0.2254**	0.2036	**0.2050**	**0.2120***	**0.1964**
	SM^{2}H$_2$	**0.1937***	**0.1847**	**0.2038***	**0.2273***	**0.2113***	**0.2110***	**0.2065**	**0.1980***
text-query-images	CMSSH	0.1985	0.1992	0.2003	0.2030	0.2050	0.1998	0.1947	0.2003
	CVH	**0.2021**	0.1768	0.1705	0.1804	0.2206	0.2070	0.2085	0.2076
	MLBE	0.1819	0.1712	0.1766	0.1698	0.1970	0.1876	0.1816	0.1737
	SM^{2}H$_1$	0.2102	**0.2082**	**0.2233***	**0.2200**	**0.2257**	**0.2139**	**0.2287**	**0.2260**
	SM^{2}H$_2$	**0.2139***	**0.2156***	0.2205	**0.2239***	**0.2311***	**0.2164***	**0.2370***	**0.2336***

mAP score on NUS-WIDE

mAP score on WIKI

Cross-media Hashing:

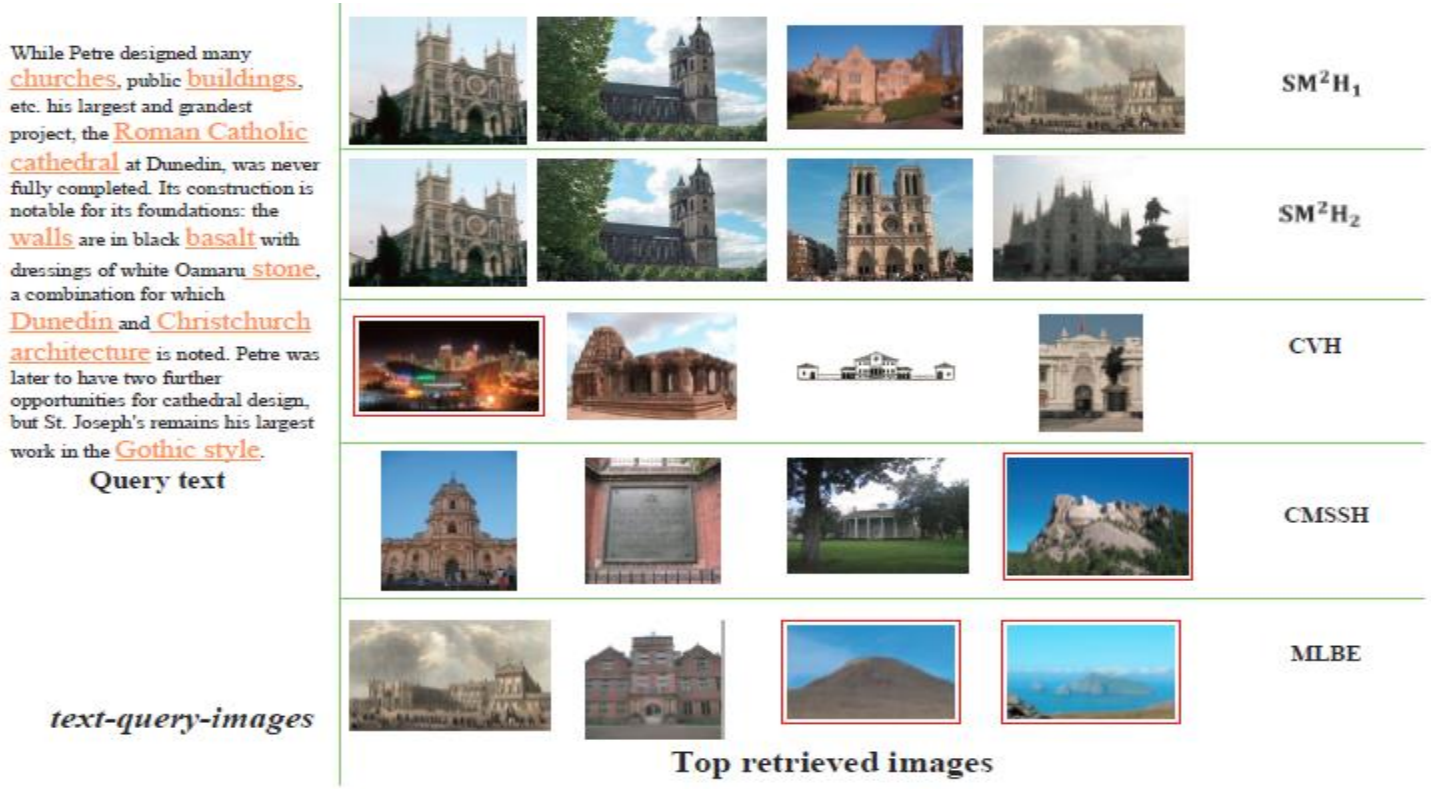
Sparse Multi-modal Hashing



For the image-query-texts direction, we can find that given a ‘battleship’ image belongs to the “warfare” category, the top retrieved texts of our SM2H are clearly relevant while the three counterparts all produce some irrelevant results.

Cross-media Hashing:

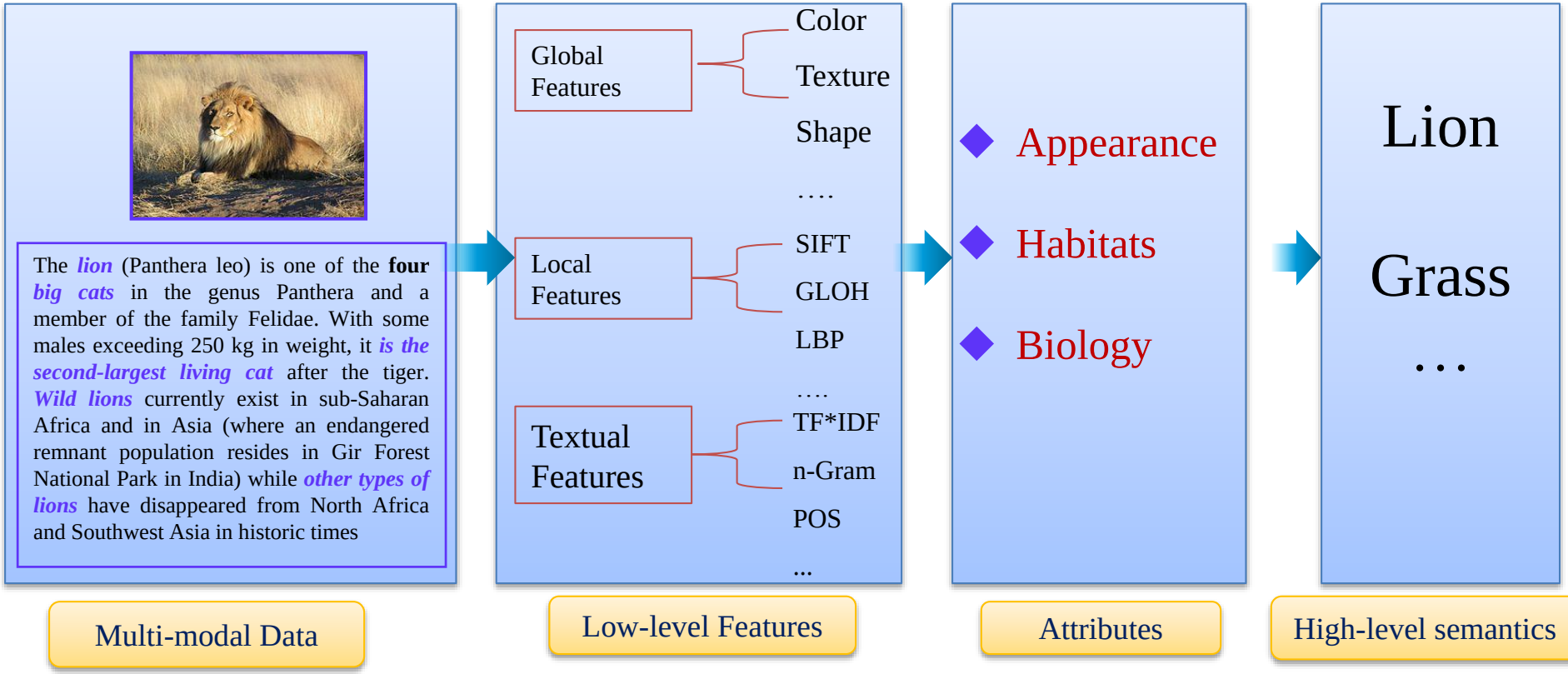
Sparse Multi-modal Hashing



For the text-query-images direction, given a text about “church”,“building”,etc, the top retrieved images of SM2H are the most relevant compared with the counterparts. Moreover, the retrieved result of SM2H2 is more accurate than SM2H1 in human understanding.

Conclusion

- The appropriate utilization of contextual information is a key for cross-media understanding!



Thanks !