# Multi-modal query expansion for video object instances retrieval


ARTEMIS Department
Institut Mines - Télécom
Télécom SudParis
UMR CNRS 8145 MAP5
9, rue Charles Fourier
91011 Evry Cedex
France

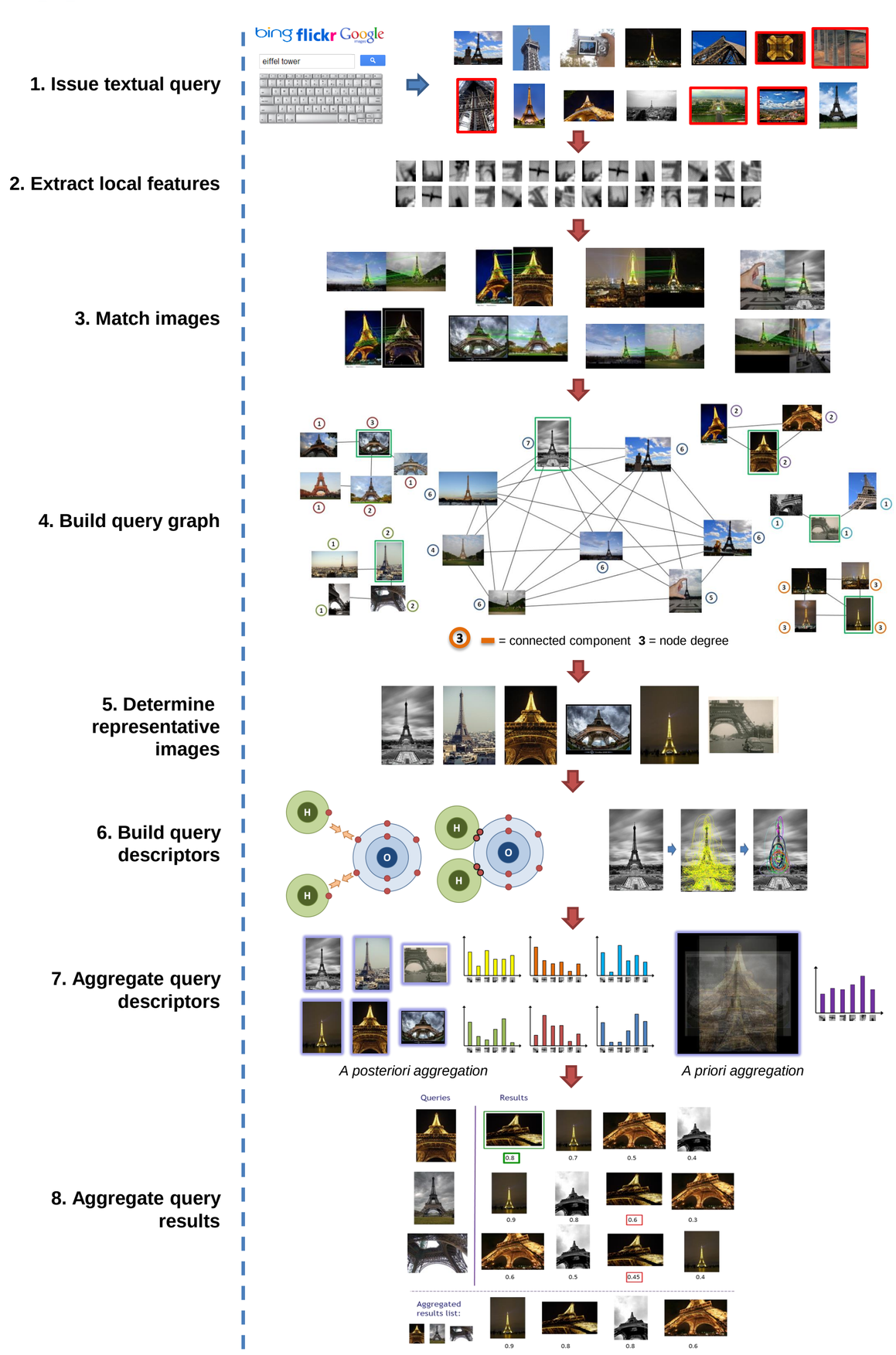
Authors

Andrei BURSUC
Titus ZAHARIA


## Objectives

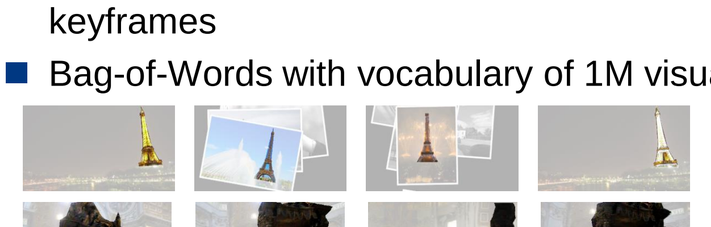
- Retrieve object instances from a large video repository starting from minimum, user-provided textual information
- Leverage on users' affinity for textual queries and crawl images from the Internet
- Remove outliers from retrieved data and identify representative instances for the topic given by the user
- Build visual descriptors from filtered representative instances and use them for querying the video repository

## Approach overview




## Evaluation

- Trecvid 2012 Instance Search Task Flickr dataset
- 74,958 videos mined from Flickr
- 22 query topic s with up to 9 example images with precise object annotation and basic textual description: 102 query images
- Hessian Affine regions + RootSIFT descriptors from 683,433 keyframes
- Bag-of-Words with vocabulary of 1M visual words



*Eiffel Tower*

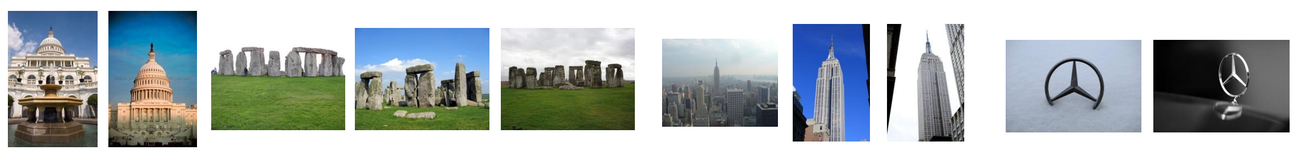
*Baldachin in Saint Peter's Basilica*

## Results

- Retrieval performance (mean Average Precision)
  - **Centered representative query**: Consists of the union of all points from the respresentative image that have been matched/shared with at least one neighboring image from the query graph.
  - **Distributed representative query:** Consists of the union of all points from every neighboring image that have been matched with points from the representative image

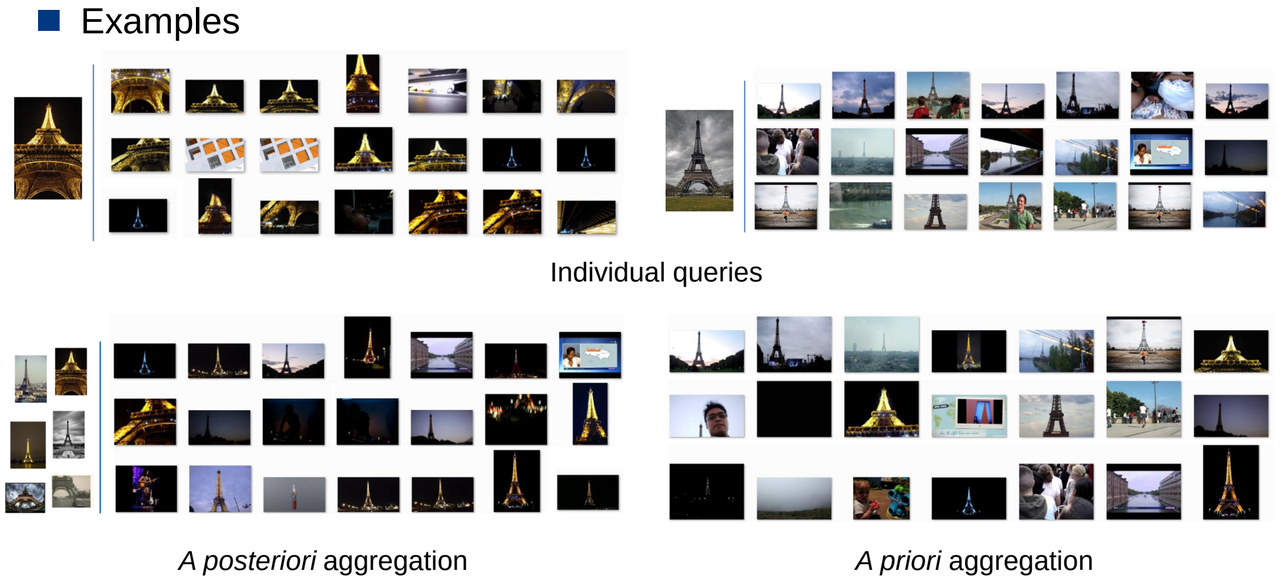
| Expansion method | Number of mined images | Aggregation strategy | mean Average Precision |
|---|---|---|---|
| Centered representative query | 25 | *A posteriori* | 0.0455 |
| | | *A priori* | 0.0476 |
| | 50 | *A posteriori* | 0.0585 |
| | | *A priori* | 0.0583 |
| | 100 | *A posteriori* | 0.0689 |
| | | *A priori* | 0.0688 |
| Distributed representative query | 25 | *A posteriori* | 0.0540 |
| | | *A priori* | 0.0558 |
| | 50 | *A posteriori* | 0.0756 |
| | | *A priori* | 0.0787 |
| | 100 | *A posteriori* | 0.0871 |
| | | *A priori* | **0.0967** |
| TRECVID 2012 Median mean Average Precision | | | 0.0795 |
| Baseline Bag-of-Words | | | 0.095 |

- Representative images



U.S. Capitol exterior — Stonehenge — Empire State Building — Mercedes star

- Examples



Individual queries

*A posteriori* aggregation — *A priori* aggregation

## Conclusion and perspectives

- Novel multi-modal query definition and expansion method: text ⇨ image ⇨ video
- Good object retrieval performance even when using only textual data
- Distributed query descriptors with *a priori* aggregation provide better results while reducing the number of query operations
- Extend method for multiple Internet sources
- Use an ad-hoc SVM classifier on representative images
- Integrate other image metadata vor validating positive instances (geotags, image popularity, uploader reputation)


May 2013 Machine Vision Applications Conference



Contact {Andrei.Bursuc, Titus.Zaharia}@telecom-sudparis.eu Site web http://artemis.telecom-sudparis.eu