



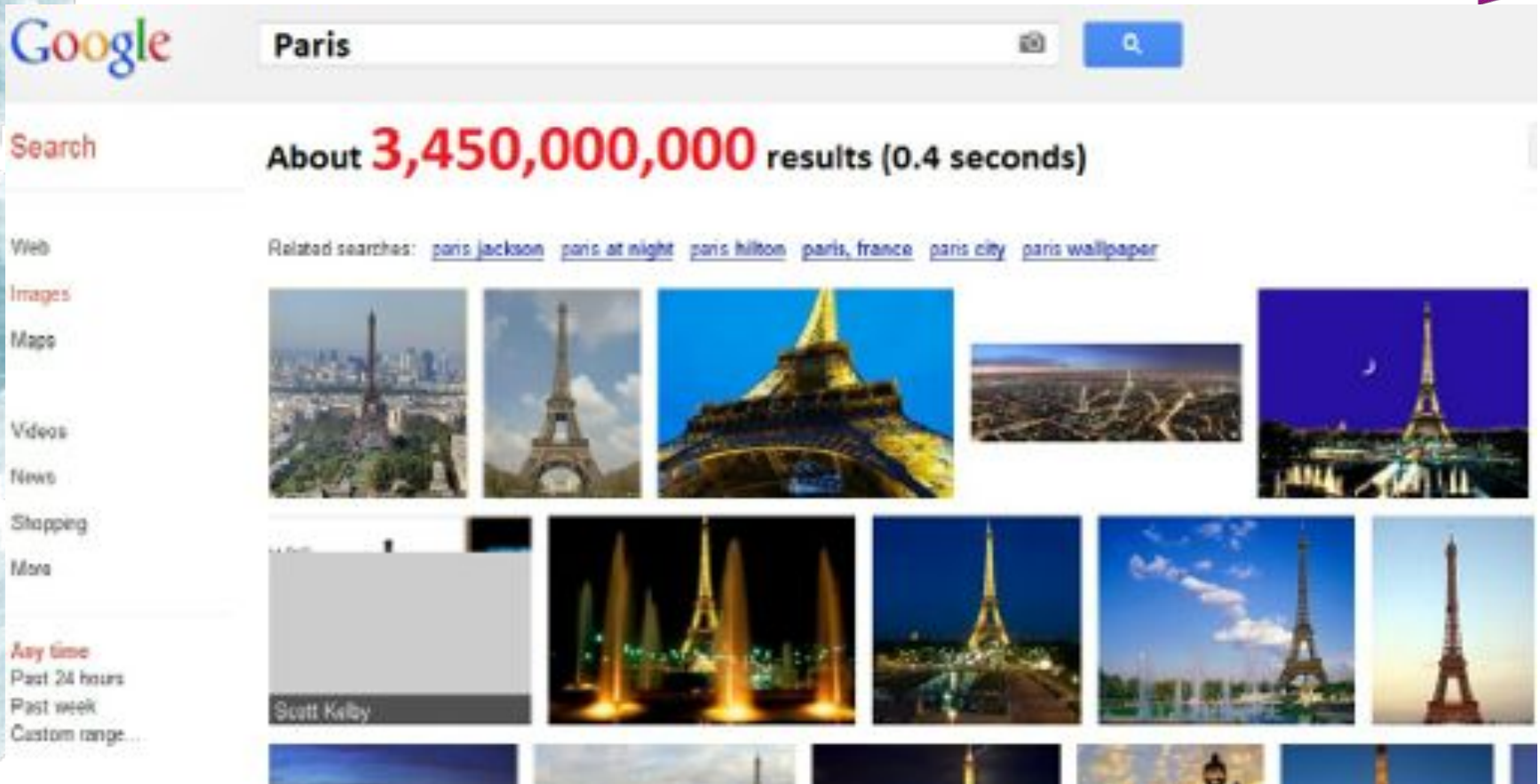
# Finding Groups of Duplicate Images in Very Large Datasets

Winn Voravuthikunchai, Bruno Crémilleux and Frédéric Jurie

22 June 2012



# Motivation: finding duplicates



- Unimaginable number of images on the internet but how high is the redundancy?
- How can we find them?

# Possible solutions?

## ■ Clustering?



- Very expensive
- We don't want to cluster all images but just find duplicates!

## ■ Image search?

Query Image



Results ranked according to similarity between the query image



- Each image has to be taken in turn as a query and the results has to be merged.

## Principles:

- Turn images into sets of transactions (i.e. lists of items)
- Mine transactions and discover those that shared by several images

## Key contributions:

- How to represent images with **VERY COMPACT** sets of binary features?
- How to mine the transactions



# Our proposed method

Bag of word extraction

Data mining transaction encoding (TF-IDF normalized & select top-K words)

Mine long patterns with min-support = 2

Transactions containing a long pattern are considered as a group of duplicates.



{3, 4, 5}



{1, 3, 7}



{4, 9, 10}



{1, 2, 10}



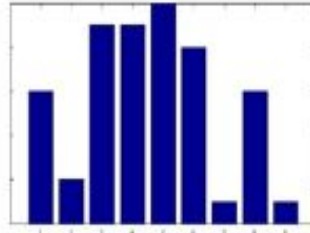
{4, 5, 6}

Closed-Pattern Miner  
Min-Support = 2  
Length Threshold = 2

{4, 5}



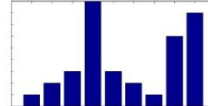
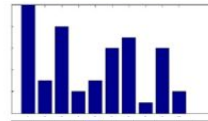
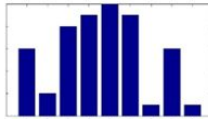
# Transaction encoding



Top-K = 3

Binary feature = 0011100000

- Evaluated in image search scenario using dot product similarity measure compare to baseline Bag of word with Chi-2 distance.

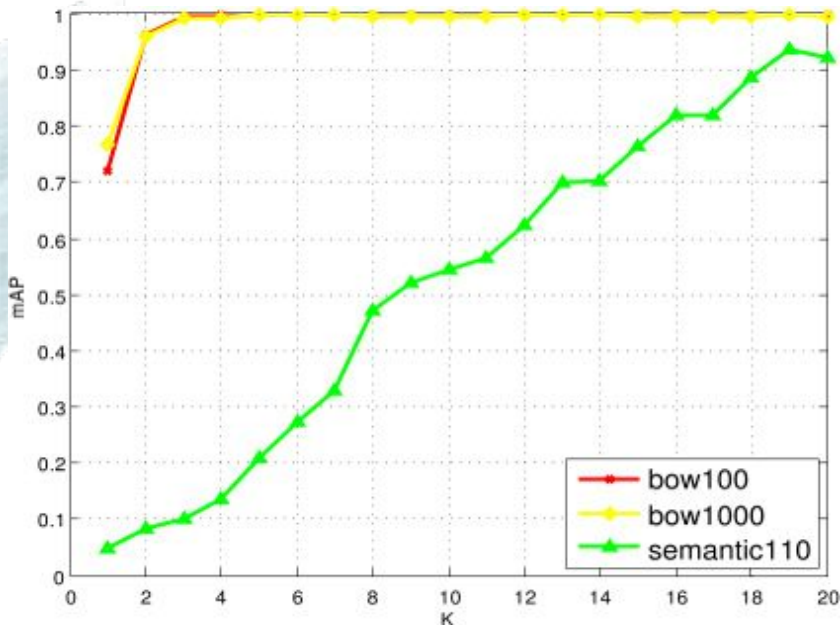


Binary feature	Dot product	Rank
0001110000	2	1
1010001000	1	2
0001000011	0	3
1100000001	0	4

# Transaction encoding

- Use the **copyday** dataset. For evaluating the robustness of image descriptors against artificial image transformations in an image search scenario.
  - 157 original images, 9 Jpeg Attacks, 9 cropping attacks.
  - 1 copy per 1 original image per 1 attack. (1 original image has 18 copies)
  - Retrieve set - The original images + 10,000 noise images.
  - Evaluate by taking each copy image as query, and find it's original from the retrieve set.

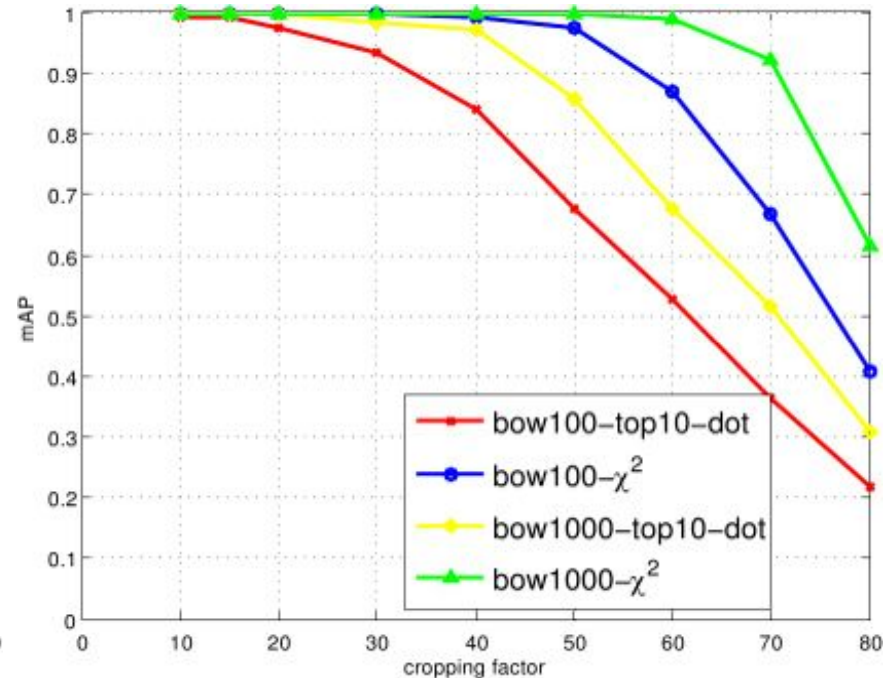
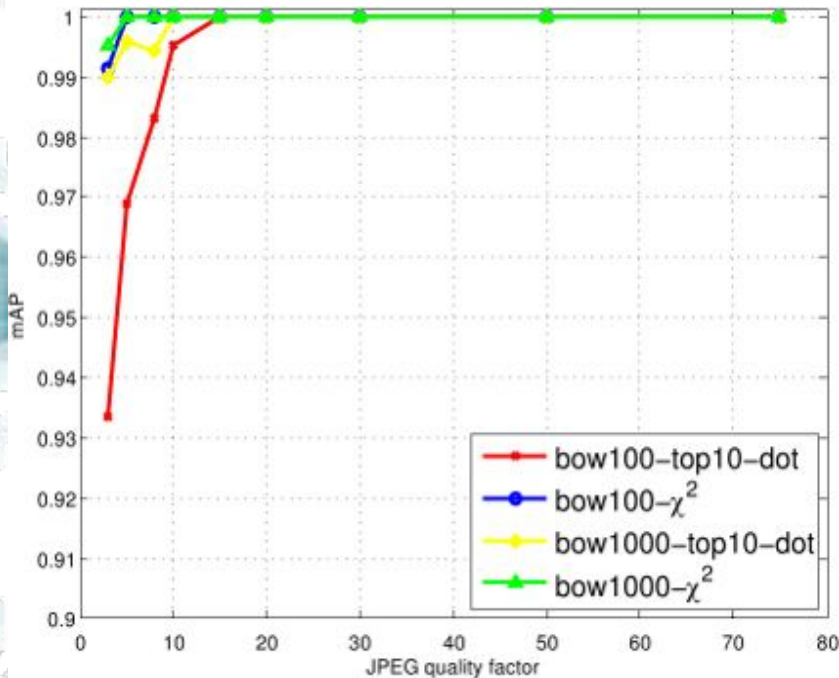
## ■ Results tune Top-K (JPEG75)



- Top-K visual words performs better than top-K semantic concepts.
- Optimum results when  $K \geq 6$  for bag of word (100 and 1,000 dimension)
- Select  $K=10$ , for other experiments in order to be robust against more difficult attacks.

# Transaction encoding

## Results Jpeg and crop attacks retrieval



- Larger vocab performs better.
- The top-K binary feature performs as good as the baseline for Jpeg attacks and crop attacks below 30% cropping factor.

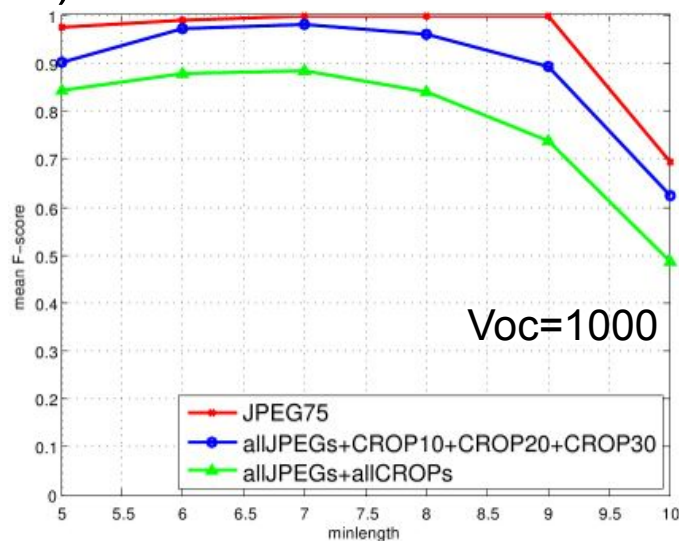
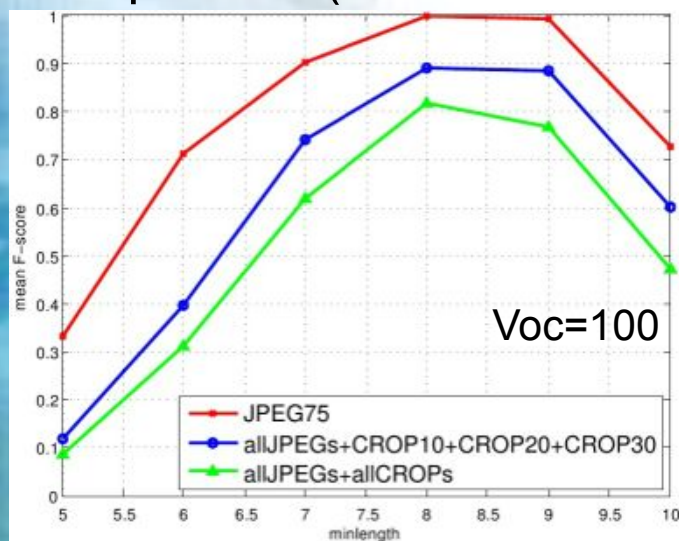
■ **Conclusion – the new representation is sufficient for detecting duplicates having a very compact size of only ~13 bytes.**



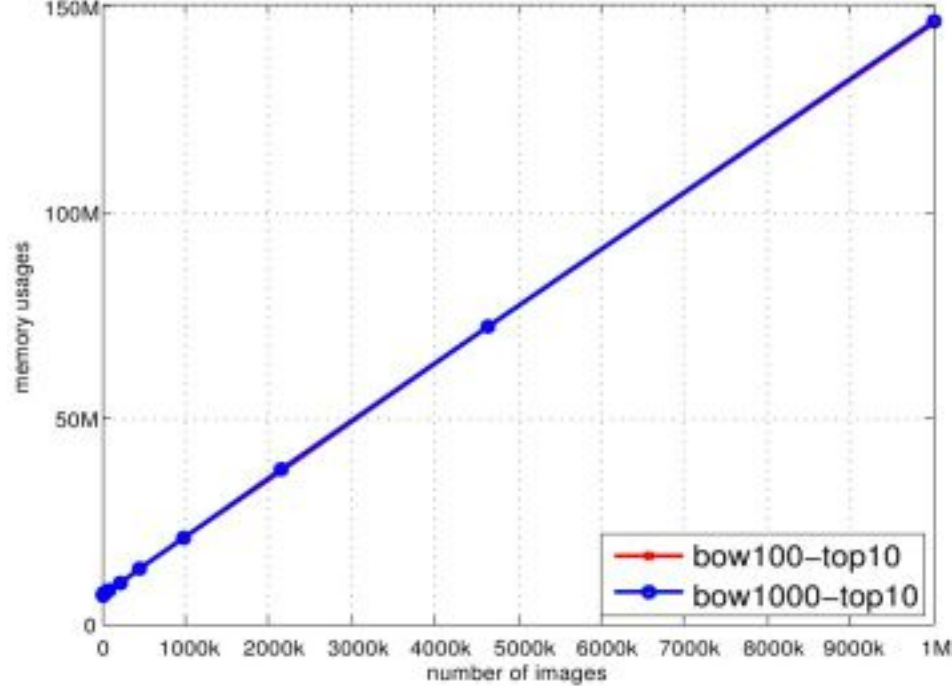
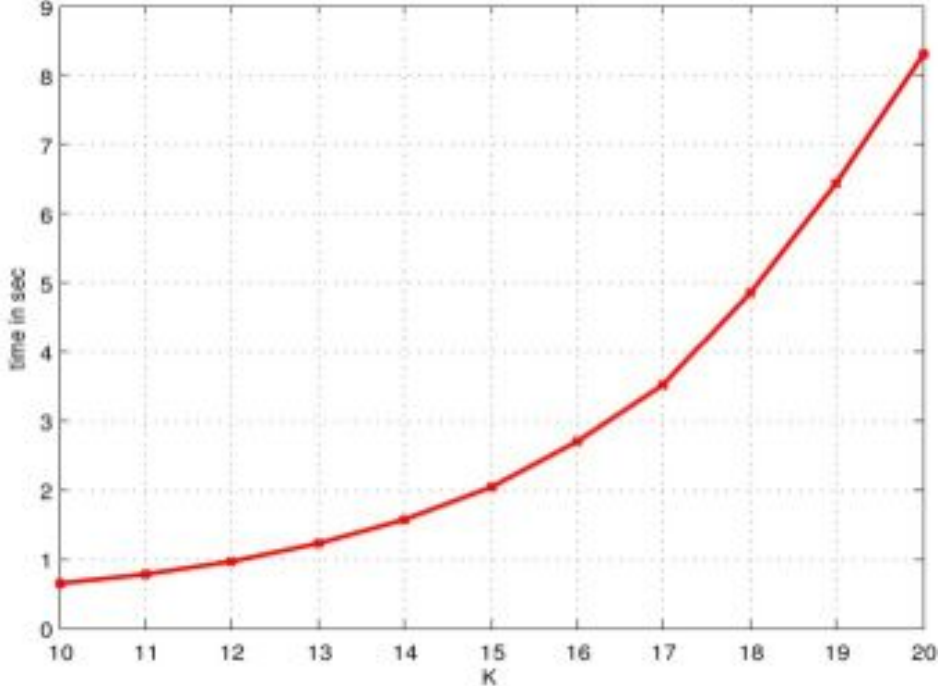
# Mining duplicate groups

## Quantitative results

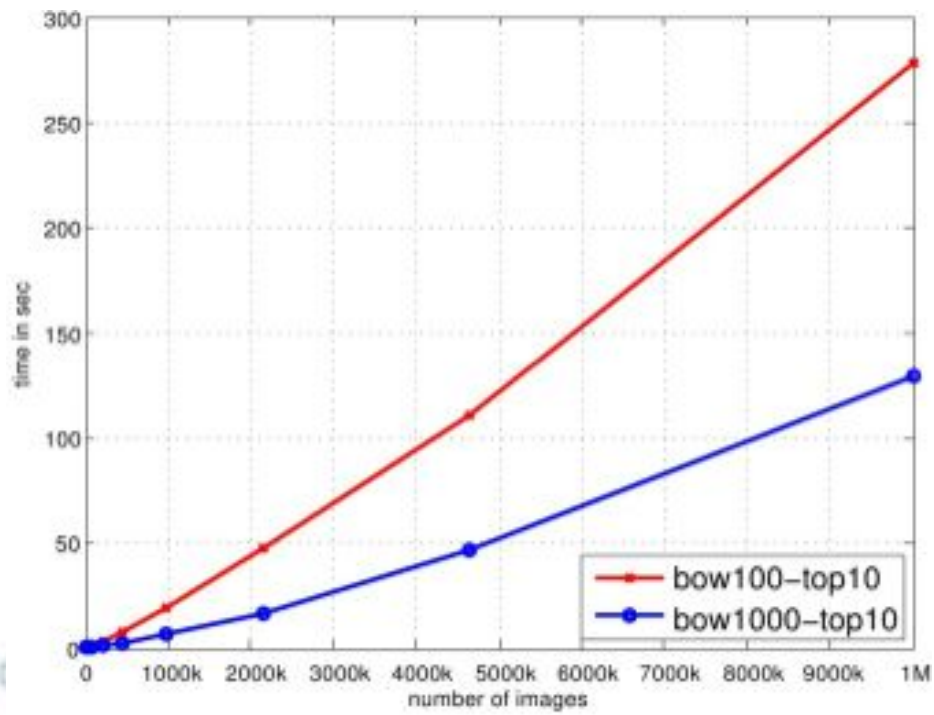
- Use copyday dataset mix 157 original images with sets of attacks images + 1,000,000 artificial noise images.
- Optimum results is to correctly detect 157 groups of duplicates (mean F-score = 1)



- Minlength = 7 and 1,000 visual words dictionary gives optimal results.
- For the light attacks, the groups of images are perfectly detected. Even for the strongest attacks the results are still very good



- Time complexity grows exponentially according to top-K
- Time complexity grows linearly according to number of images
- Memory usage grows linearly according to number of images



# Mining duplicate groups

## ■ Qualitative results

- Use the “One million random web images database”.
- ~80,000 groups of duplicate images found in less than 3 minutes.





Thank you!