

# Web Image Retrieval on ImagEVAL: Evidences on visualness and textualness dependency in fusion model

Sabrina TOLLARI and Hervé GLOTIN  
UMR CNRS 6168 LSIS  
Université du Sud Toulon-Var  
tollari@univ-tln.fr, glotin@univ-tln.fr

# Overview

- ImagEVAL Task2 description
- Visual and textual features extraction
- Image retrieval by combining textual and visual information
- Results
  - ImagEVAL Task2 campaign results
  - Global results
  - Textualness and visualness fusion behavior
- Conclusion

# ImagEVAL Task2 description

- The corpus data: 700 URLs
  - 700 web pages
  - about 10 000 web images
- 25 queries. Each query is composed of:
  - a list of keywords
  - and between 1 and 9 query images
- Aim: find relevant web images for 25 queries
  - Between 10 and 100 relevant images by query
- For the official runs:
  - relevant images are unknown
  - 300 images are answered by query
  - MAP scores are calculated by TREC eval tools

# Query examples

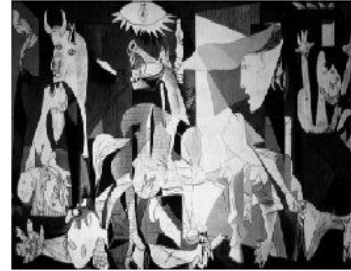
« Poplar tree »

+



« Picasso Guernica »

+



# ImagEVAL Task2 query information

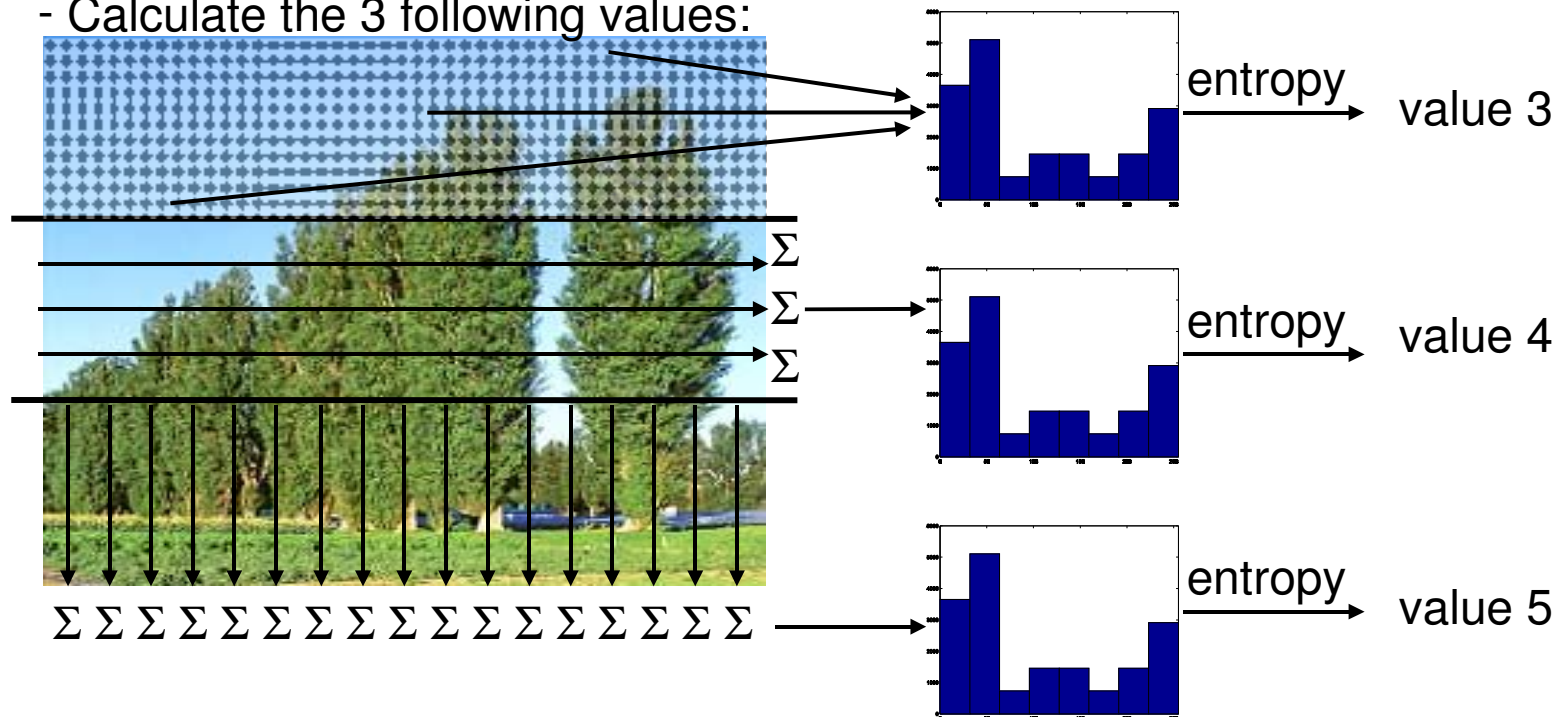
	Query	Number of query images	Number of relevant images
1	bee	7	39
2	avocado	7	39
3	tennis ball	4	20
4	lemon	6	94
5	Ladybird	6	19
6	Etiopian flag	1	13
7	European flag	1	31
8	Picasso Guernica	3	19
9	Joconde	2	14
10	Lava flow	7	66
11	Delacroix Liberty	3	11
12	Great Wall China	6	88
13	Perce Rock	7	33

	Query	Number of query images	Number of relevant images
14	clown fish	7	51
15	Siamese cat	6	33
16	tennis playground	9	40
17	Ayers Rock	6	41
18	zebra	6	30
19	Eiffel Tower	5	53
20	Statue Liberty	4	18
21	Niagara Falls	6	51
22	teddy bear	6	9
23	screwdriver	5	20
24	poplar tree	5	19
25	map Norway	6	8

# Visual feature extraction

- Each image is split into 3 horizontal bands
- For each band  $b$ 
  - For each color  $L=R+G+B$ ,  $r=R/L$ ,  $g=G/L$  do
    - Calculate the mean (value 1) and the std (value 2) of pixel values

- Calculate the 3 following values:



Finally, for each band and each feature there are 5 values.  
The total number of visual features by image is  $3 \times 3 \times 5 = 45$ .

# Image retrieval by visual information only

- For each query  $Q$ 
  - For each data image  $I$ 
    - For each query image  $q_j$  of  $Q$  do
      - $d_j = \text{L2norm}(q_j, I)$
    - $D_V(Q, I) = \mathbf{average\_operator}(d_1, d_2, \dots, d_j, \dots)$ 
      - where the `average_operator` is:
        - the arithmetic mean (mean),
        - the geometric mean (gm),
        - the harmonic mean (hm)
        - or the minimum (min).
- The first 300 images which have the smallest  $D_V$  are returned.

# Image retrieval by textual features only

- First, the following classical treatment is applied to web pages and query texts:
  - All HTML tags (such as <H1>, <table>, <IMG>... ) are deleted
  - Text are normalized (uppercase are replaced by lowercase, ...)
  - Classical stop words are removed
- Second, we count each word occurrence in web pages
- Third, for each query:
  - We suppose that each query word has the same importance
  - We seek for text that include at least one word of the query
  - We calculate standard tfidf values
  - We add tfidf values to images of each web page
- Finally, we sort images by decreasing order of their tfidf values

# Visuo-textual fusion

We merge visual and textual information using a weighted average of the visual and textual distance:

$$D(Q,I,t) = t \times D_T(Q,I) + (1 - t) \times D_V(Q,I)$$

where  $D_T$  is a normalised distance based on the tfidf values and  $D_V$  is the distance between a query  $Q$  and a data image.

The weight  $t$  is the textual fusion rate.

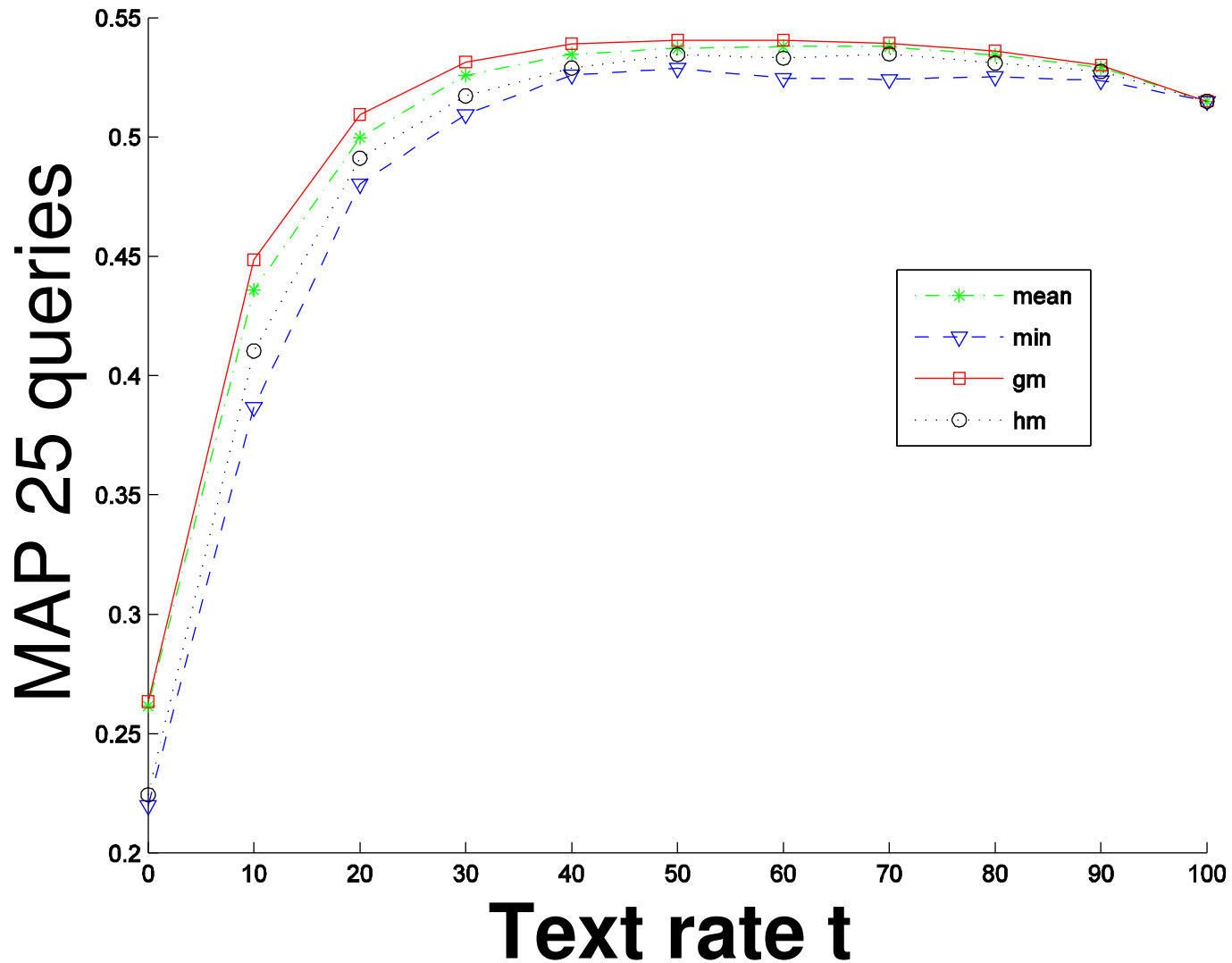
We keep the first 300 images that have the smallest  $D$  values.

# ImagEVAL Official MAP results

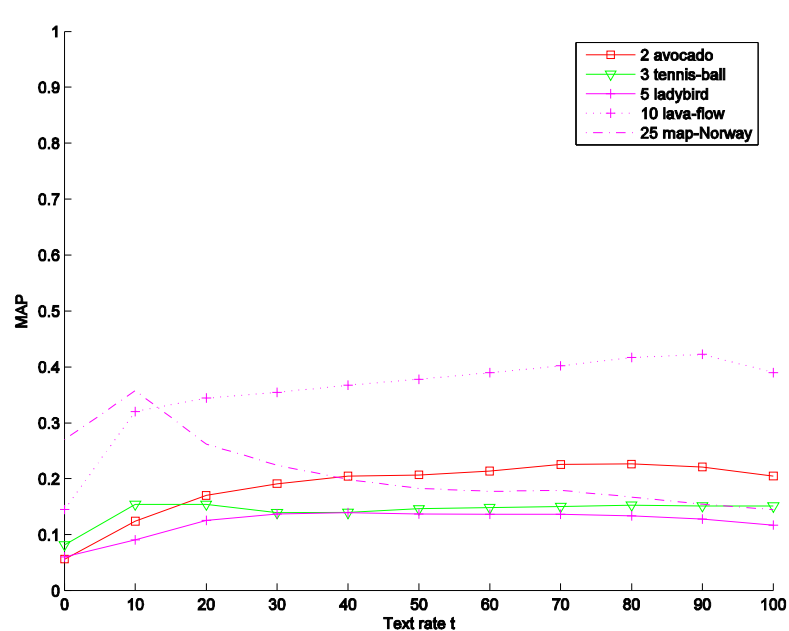
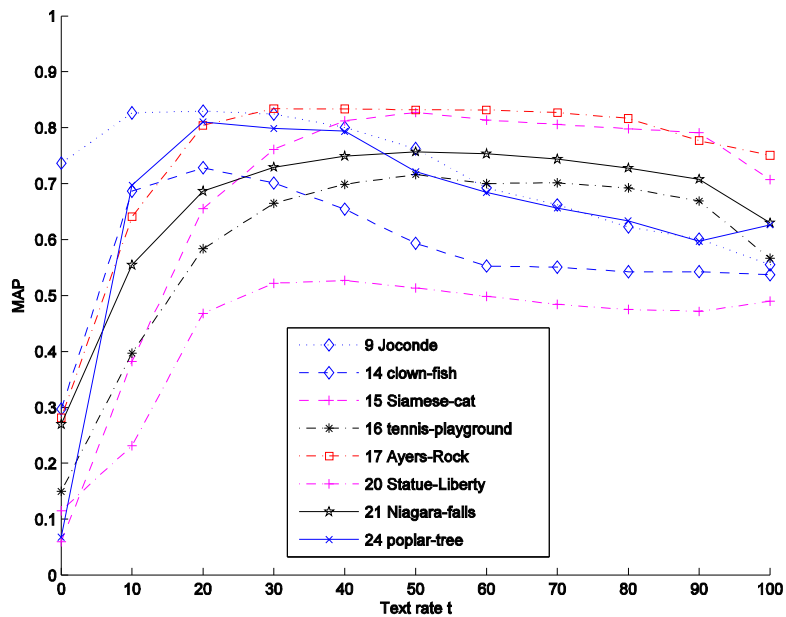
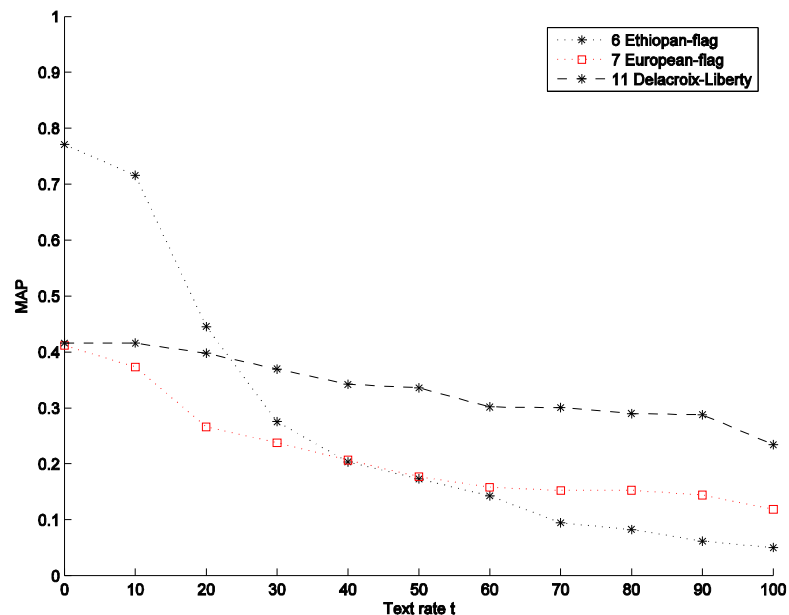
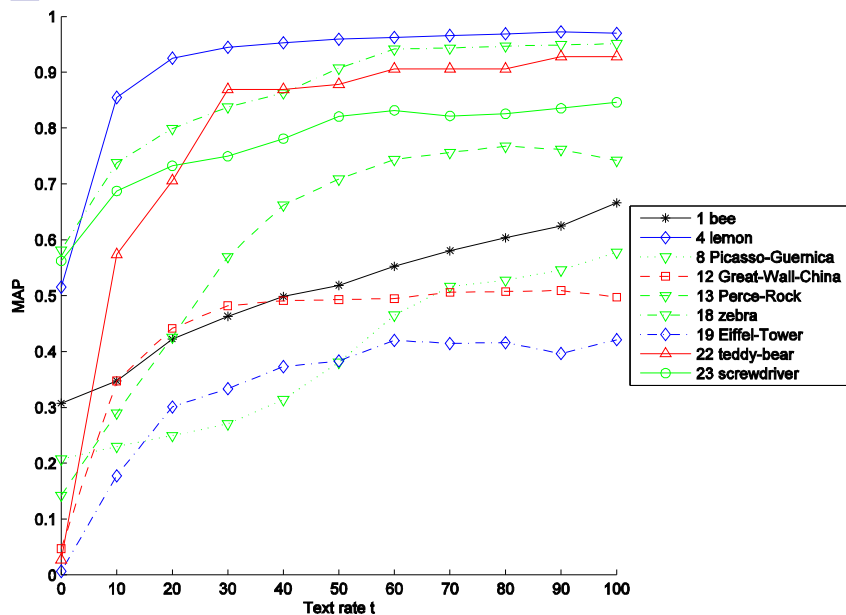
Rank	Text only	Visual only	Fusion
1	0.559 (Run 5)	0.271 (Run 16)	<b>0.613</b> (Run 1)
2	0.513* (Run 9)	0.261* (Run 17)	0.536* (Run 7)
3	0.455 (Run 12)	0.181 (Run 20)	0.517 (Run 8)

\* LSIS MAP scores. For our fusion results, t=50% and average\_operator=mean.

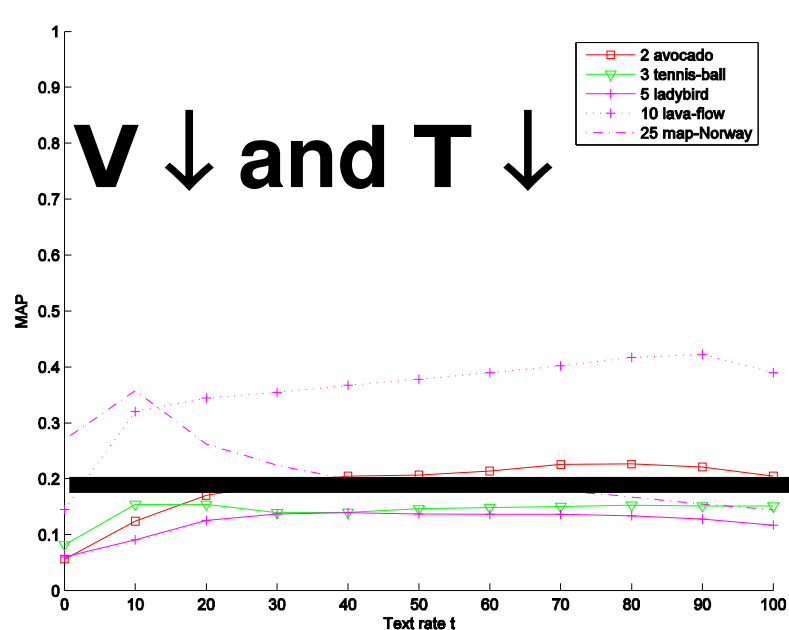
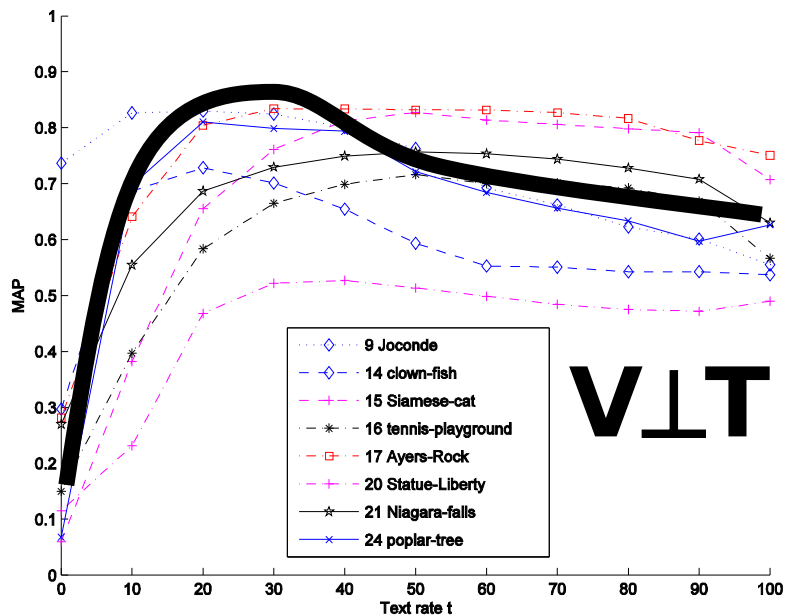
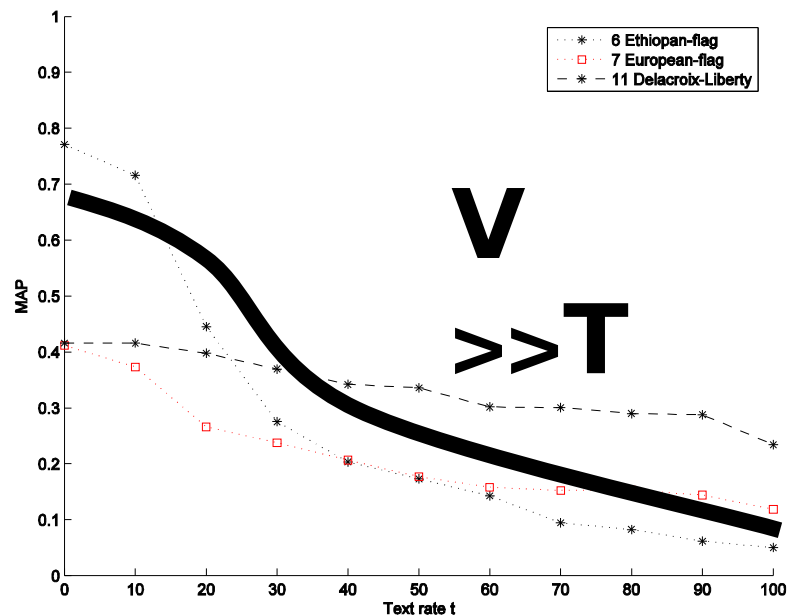
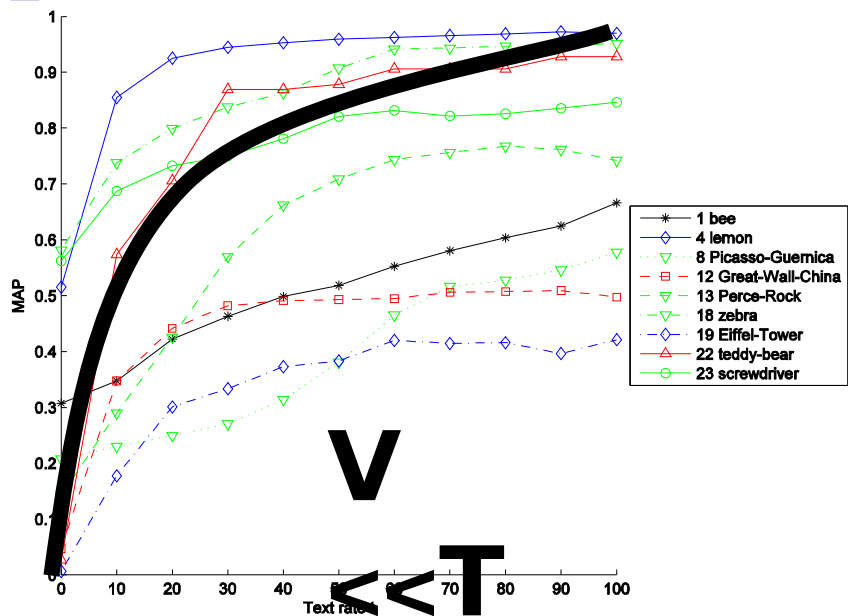
# Impact of the average operator



# Fusion behaviors



# Fusion behaviors

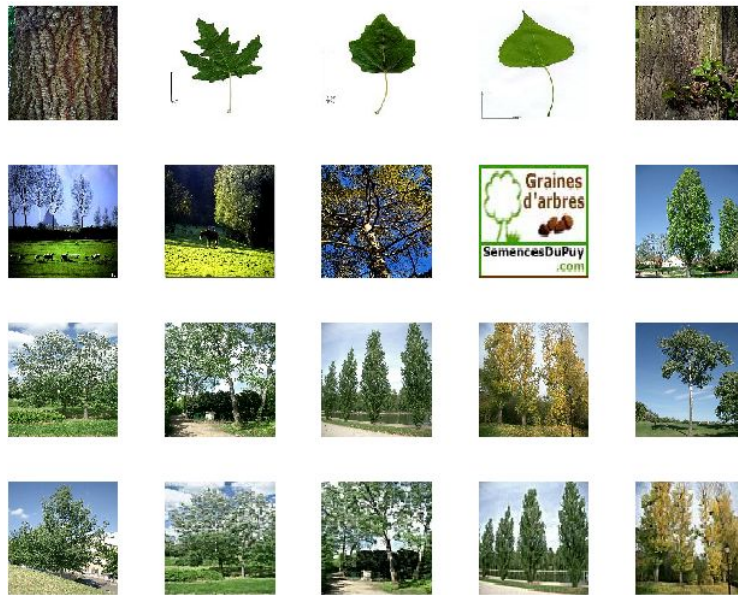


**V : visualness      T : textualness**

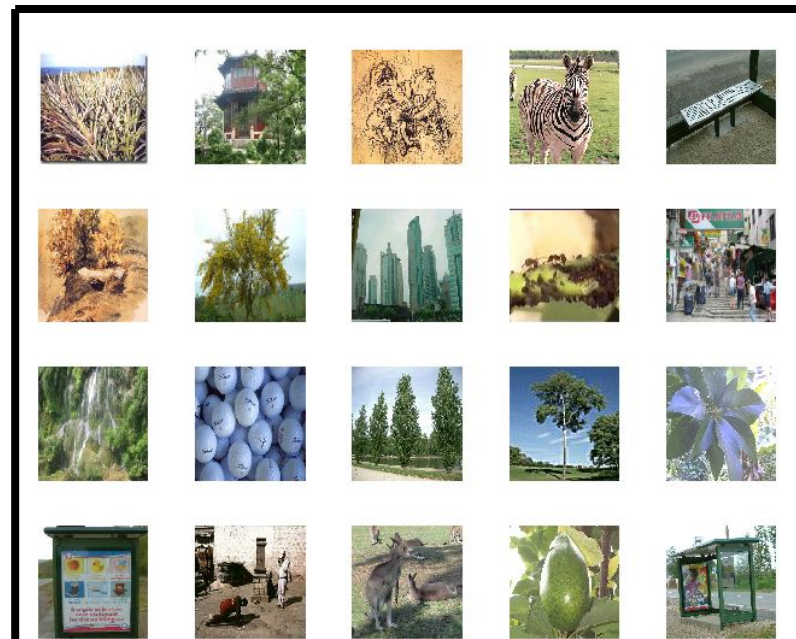
# Example of V<sub>L</sub>T query

Query

poplar  
tree +

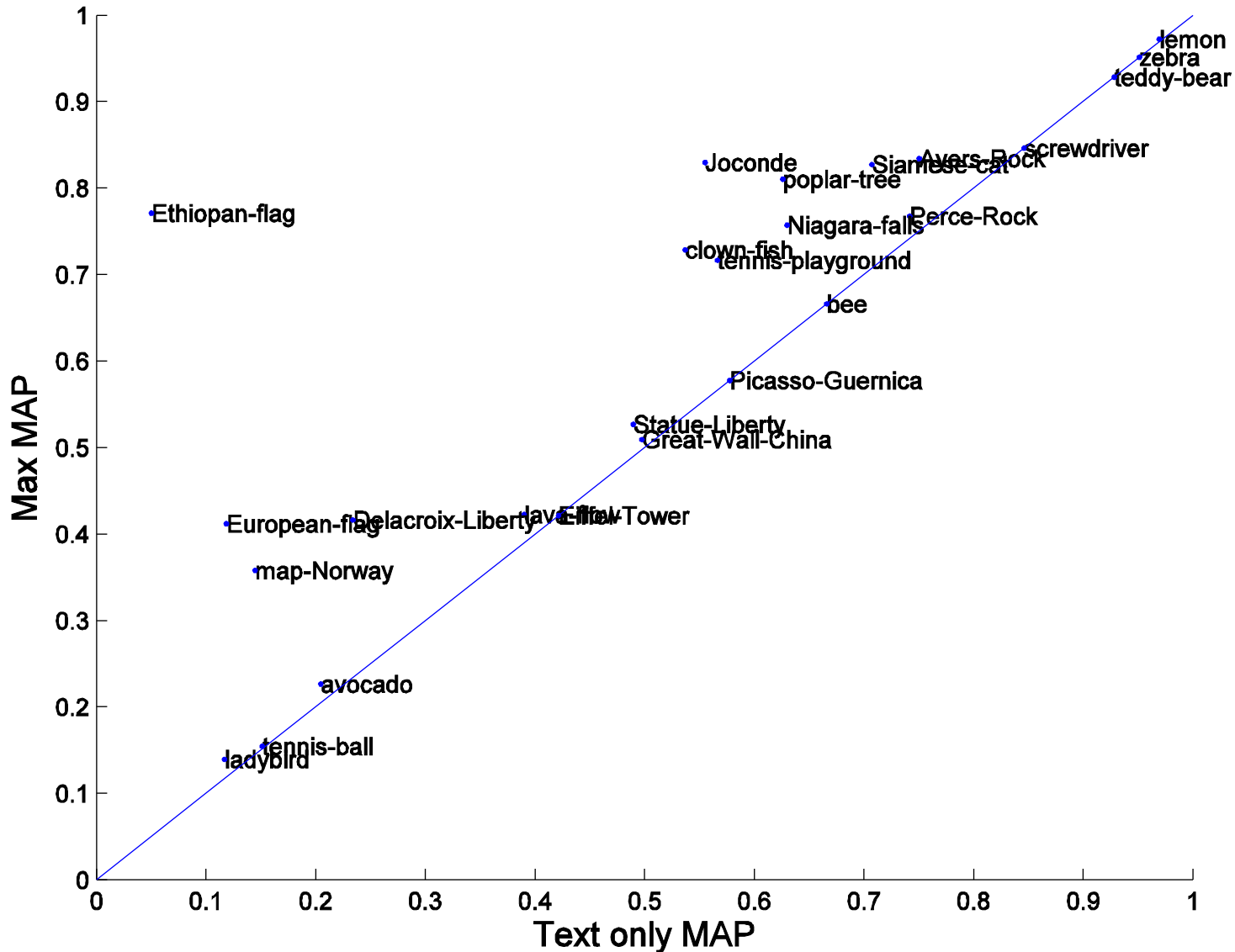


Text only



Visual only

# Max MAP on all fusion versus text only MAP



# Conclusion

- Using simple:

- Visual and textual feature extraction
- Text and visual fusion model

we can improve image retrieval

- We are giving evidences on visualness and textualness concept dependency
- Futur work: we want to improve our results by a concept dependant feature selection



Thank you for your attention





# Impact of the text rate $t$ on recall/precision

