Enhancement of textual images classification using their global and local visual content

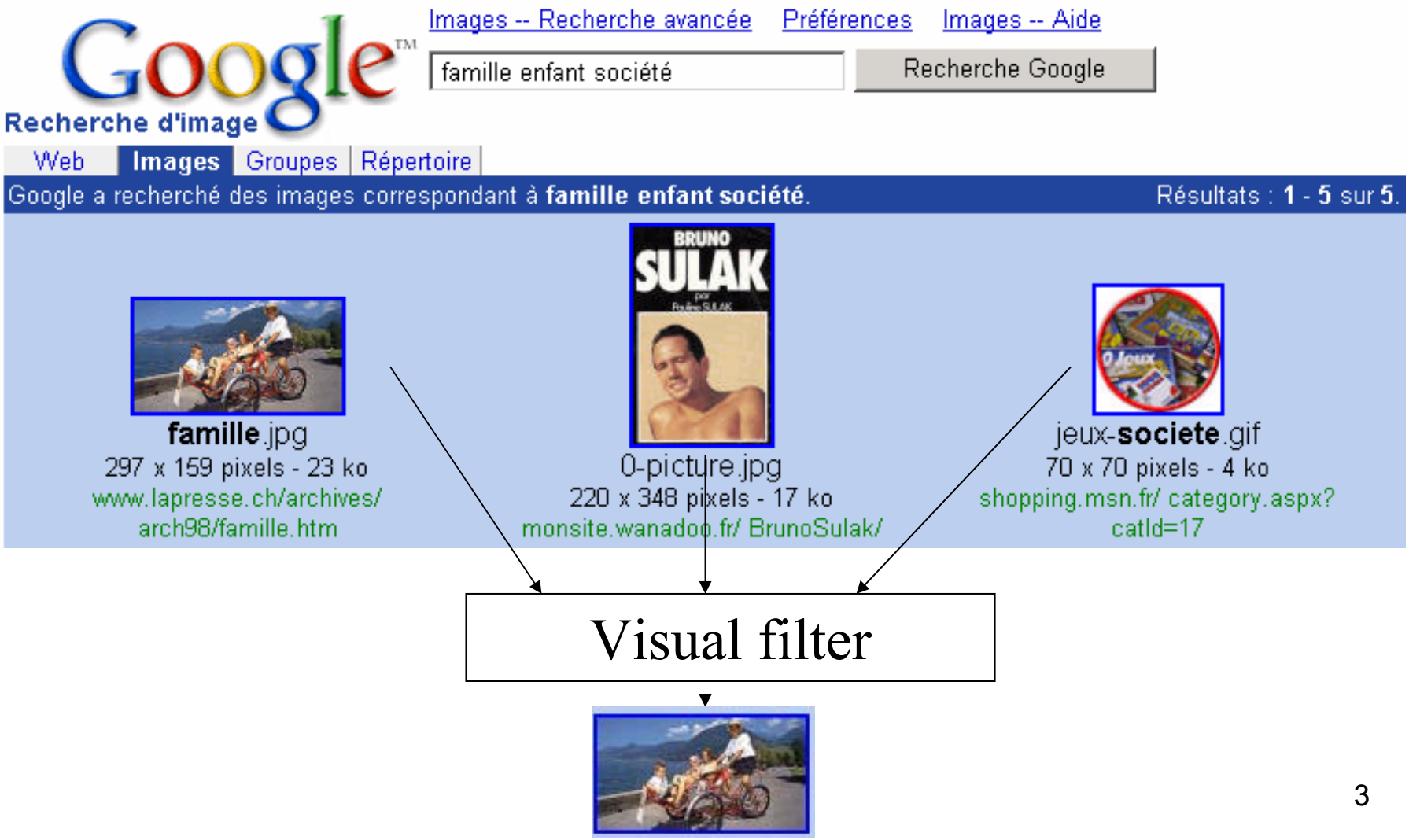
Sabrina Tollari, Hervé Glotin, Jacques Le Maitre
Université de Toulon et du Var
Laboratoire SIS - Équipe Informatique
France

MAWIS 2003

Plan

- Objective
- State of the Art
- Presentation of the corpus
- Presentation ot the system
- 3 experiments
- Conclusion

Enhancement of image search engine

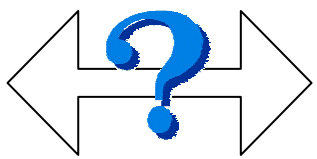


The problem

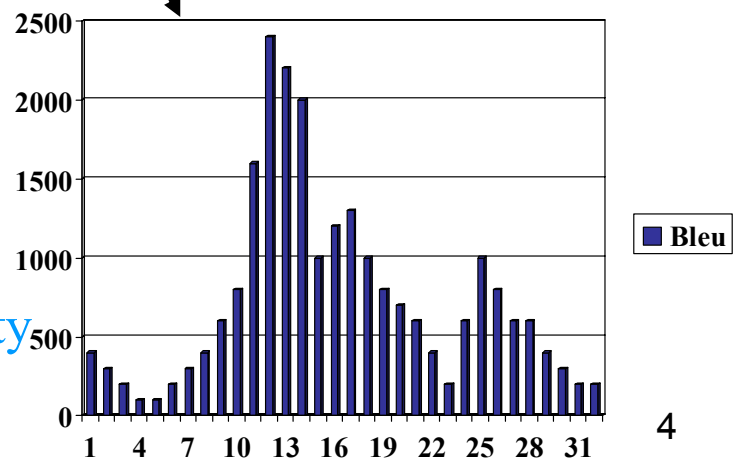
Textual indices

Visual features

Paysage Cameroun
Agriculture



Complementarity



Textual indices
Low or High level visual features

- Textual indices :
  - Manual indexation, Keywords…
  - Auto indexation : Legend, surrounding text…
- Low level Visual features :

  - Color : RGB, HSV, Brightness, Edges
- High level Visual features :

  - Shape, Spectrum analysis (rotation invariance) Fourrier transform, wavelet,texture

- Semantics of textual indice >> semantics of High Level Features

- Low level visual features requieres low computation time and give complementarity information to textual indices

State of the art: text OR visual information ?

Visual information	Visual and/or textual information *
Virage(1996) NeTra(1997) SurfImage(INRIA,1998) IKONA(INRIA, 2001)	Chabot(1995) QBIC(IBM,1995) VisualSeek(1996) MARS(1997) Zhou et Huang (2002)**

* None of them is using text information to enhance visual querying and *vice versa*

** Early fusion merging textual and visual features, or visual / textual user feedback

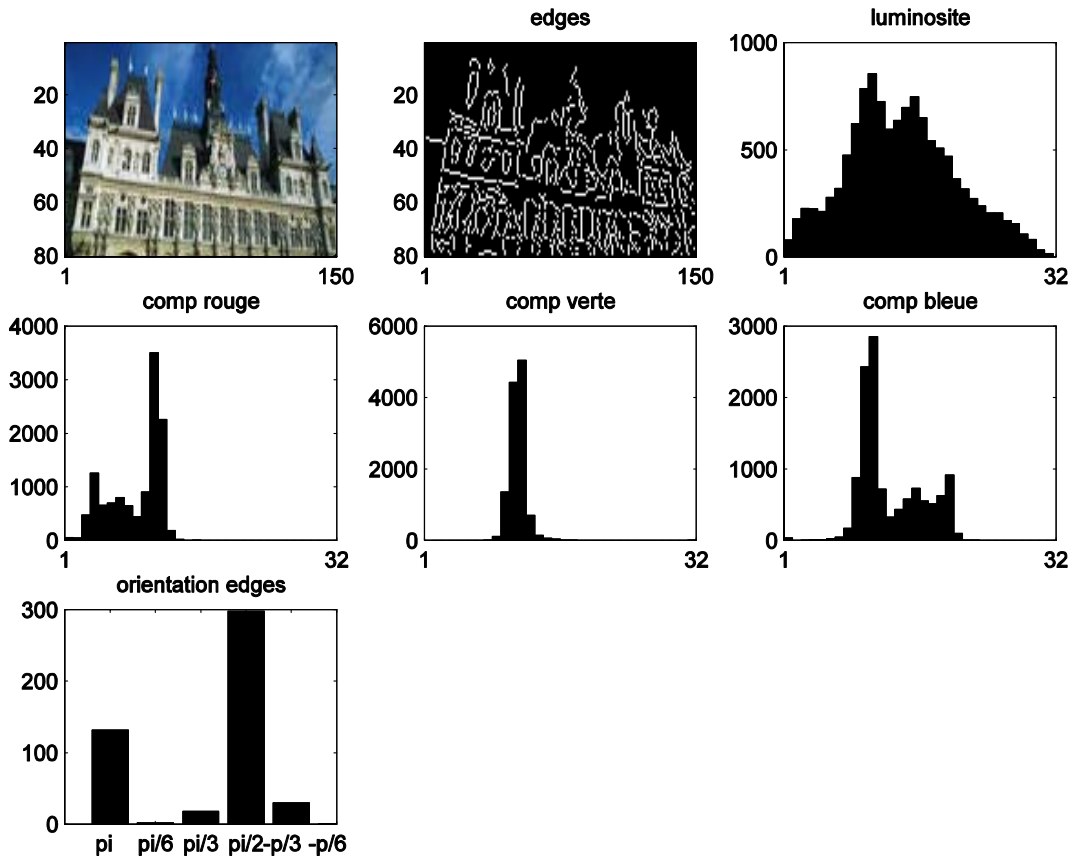
The corpus (1/2)

- 600 press agency photos
- Textually indexed by a picture researcher with keywords from a hierarchical thesaurus
- Stored as MPEG-7 descriptions in an XML file

```
<?xml version="1.0" encoding="UTF-8"?>
<mpeg7:Mpeg7 xmlns:xsi="http://www.w3.org/1999/XMLSchema-instance"
          xmlns:mpeg7="http://www.mpeg7.org/2001/MPEG-7_Schema">
     <mpeg7:PrivateIdentifier>BAR9501001C-1</mpeg7:PrivateIdentifier>
     <mpeg7:Title>Rhumsiki, Nord Cameroun</mpeg7:Title>
     <mpeg7:KeywordAnnotation>
          <mpeg7:Keyword>Paysage</mpeg7:Keyword>
          <mpeg7:Keyword>Agriculture</mpeg7:Keyword>
     </mpeg7:KeywordAnnotation>
     <mpeg7:MediaUri>BAR9501001C-1.jpg</mpeg7:MediaUri>
</mpeg7:Mpeg7>
```

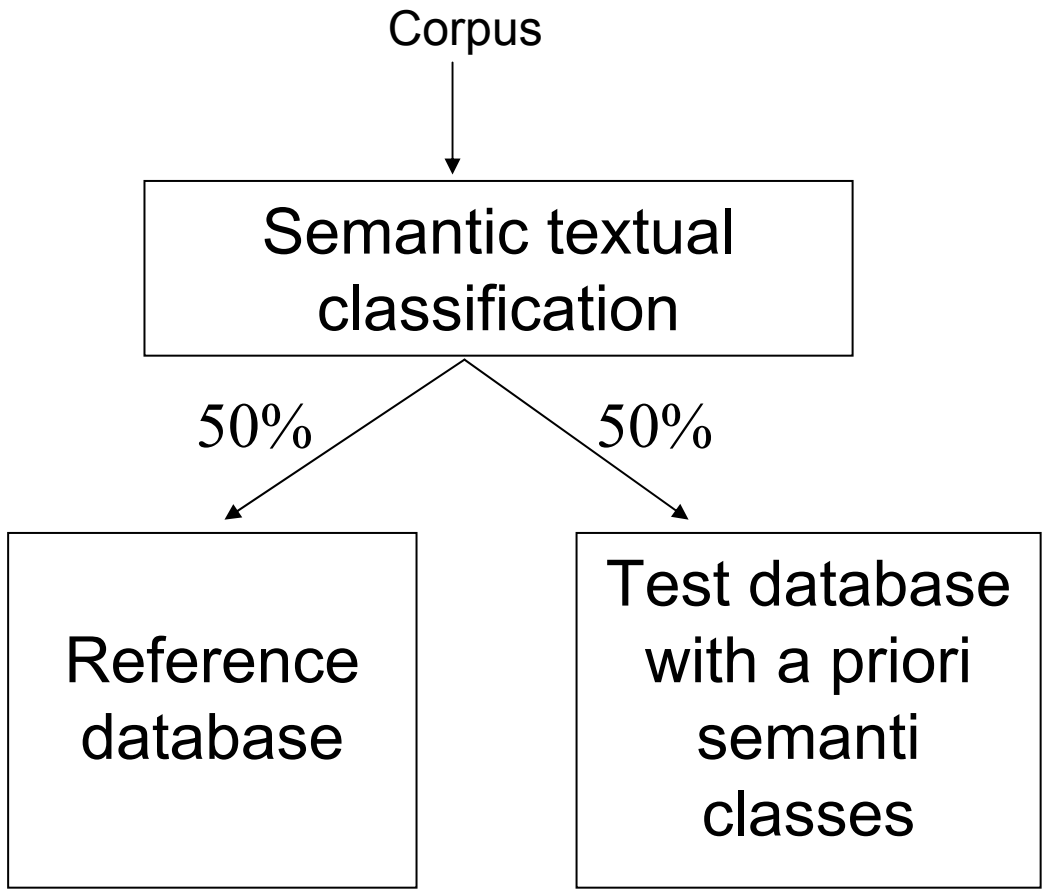
The corpus (2/2)

- Simple automatic low-level visual features (histograms)
  - Red
  - Green
  - Blue
  - Brightness
  - Edge direction



Method

Step A



Step B

Ramdom division

Step C

Evaluation of an automatic classification of the test database using the reference database

Step A : Construction of a textual semantics using an ascendant hierarchical classification

- Lance & Williams, 1967
- Objective : cluster similar images
- Highlight of non trivial semantic classes
- Check the class cardinal

Algorithm

program AHC

input

E: the set of n elements to classify

Dist: the array n*n of distances between elements

output

C: a set of semantic classes

begin

For each element e in E

Add Classe(e) in C

end for

While T do

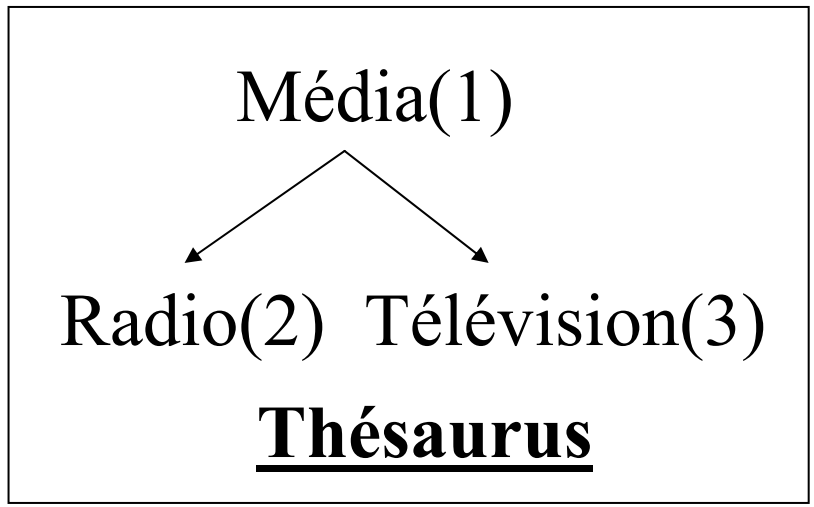
Merge the 2 nearest classes

end while

end.

Vectorial representation

- Salton, 1971



- Ex :

  Let I be the image such that Term(I)={Radio}

  – Vector(I)=(0,1,0)

  – Extended_vector(I)=(1,1,0)

Similarity measure for the AHC

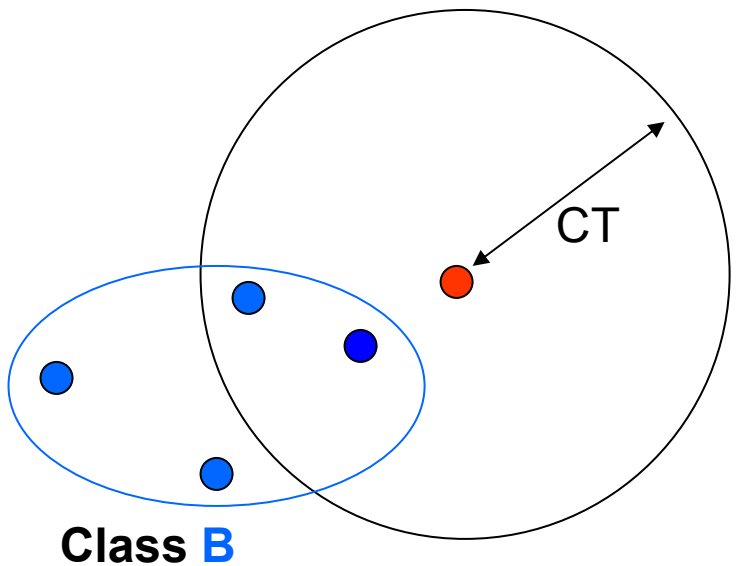
Let x et y the vectors of the images X et Y

$$\text{dist}(X,Y) = 1 - |cos(\vec{x}, \vec{y})|$$

Classical criterion

- **nearest neighbour**
- **farthest neighbour**

New agregation criterion



Step A : construction of the semantic classes

- 24 classes
  - Each contains 8 to 98 images

3 most frequent keywords of some classes :

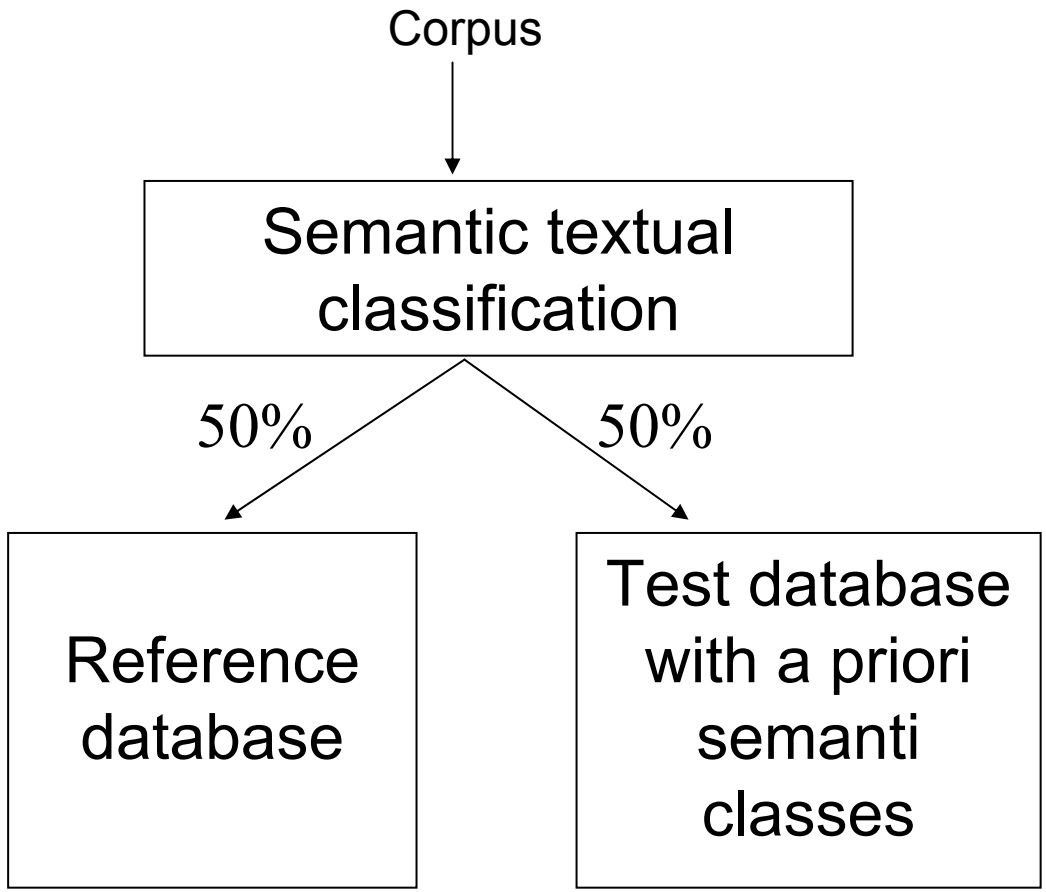
Class	Frequency 1	Frequency 2	Frequency 3
1	Femme	Ouvriers	Industrie
2	Cameroun	Agriculture	Paysage
3	Constructeurs	Transport	Automobile
4	Contemporaine	Portrait	Rhône
5	Société	Famille	Enfant

Method

Step A

Step B

Ramdom division



Step C: Evaluation of an automatic classification of the test database using the reference database

Image of the test database (original class **Co**)

C_1

Paysage, agriculture, Cameroun

C_2

Femme, Ouvrier, Industrie

Reference database

Evaluated class **Ce**

If **Co**≠**Ce** then classification error

Step C: 3 different classifications

1. Text Only Classification
2. Visual Only Classification
3. Text and Visual Classification

Kullback-Leibler distance (1951)

Let x and y be two probability distributions

Kullback-Leibler divergence:

$$KL(x,y) = \sum_{j=1}^{n} x_j \log \frac{x_j}{y_j}$$

Kullback-Leibler distance:

$$DKL(x,y) = KL(x,y) + KL(y,x)$$

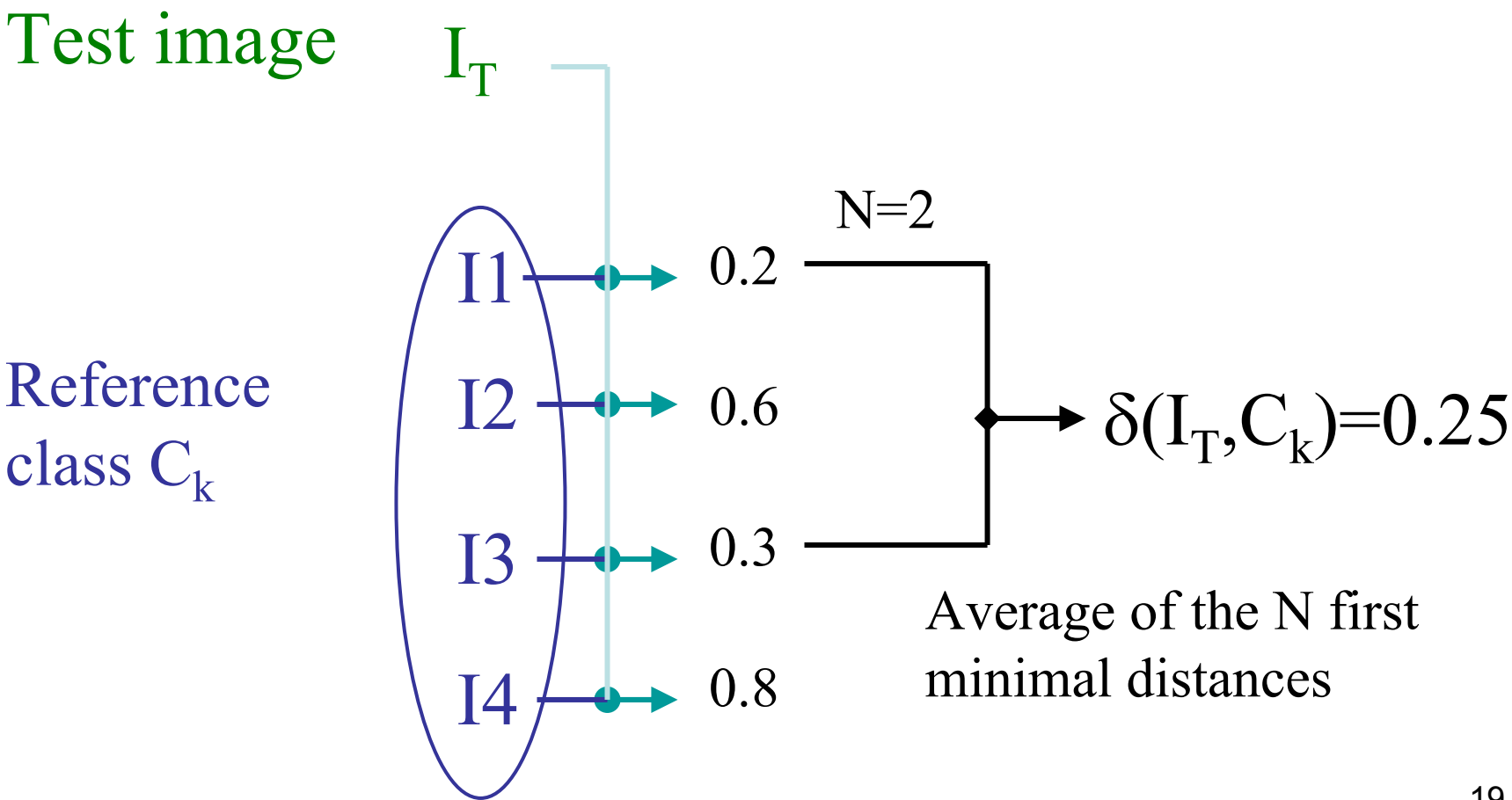
1. Results of Text Only Classification

- Average vector for each class $\vec{C}_k^t$
- Textual class of the image I_T:

$$C^t(I_T) = \text{argmin}_{k \in \{1,2,\ldots,c\}} DKL(\vec{I}_T^t, \vec{C}_k^t)$$

Results	Textual vector extended with thesaurus	Textual vector without extension
Error rate	1.17 %	13.72 %

Early fusion of visual features



Results of Visual Classification

$$C^v(I_T) = \text{argmin}_{k \in \{1,2,\ldots,c\}} \delta(I_T, C_k)$$

N	1	2	3	4
Rouge*	75.68	74.50	**71.76**	**71.76**
Vert*	79.60	78.03	76.86	**76.07**
Bleu*	78.03	77.64	78.03	**77.25**
Luminance*	79.21	78.03	**76.07**	77.64
Direction*	84.70	78.03	**76.86**	**76.86**

* Error rate in %

Theoritical error rate: 91.6%

The late visuo-textual fusion

- Evaluation of the probability that image I_T belongs to class C_k by late visuo-textual fusion

$$P_{I_T}^{v \vee t}(C_k) = \sum_{j=1}^{5} P_{I_T}^{v}(C_k|A_j) \times \omega'(A_j) + P_{I_T}^{t}(C_k) \times \omega'(A_6)$$

$$C^{v \vee t}(I_T) = \underline{\text{argmax}}_{k \in \{1,2,...,c\}} P_{I_T}^{v \vee t}(C_k)$$

Class Probability definitions

$$P^{t}_{I_T}(C_k) = 1 - \frac{DKL(\vec{I_T}, \vec{C_k})}{\sum_k DKL(\vec{I_T}, \vec{C_k})}$$

$$P^{v}_{I_T}(C_k|A) = 1 - \frac{\delta_A(I_T, C_k)}{\sum_k \delta_A(I_T, C_k)}$$

$A \in$ {Red, Green, Blue, Brightness, Edge direction}

Weighting definition

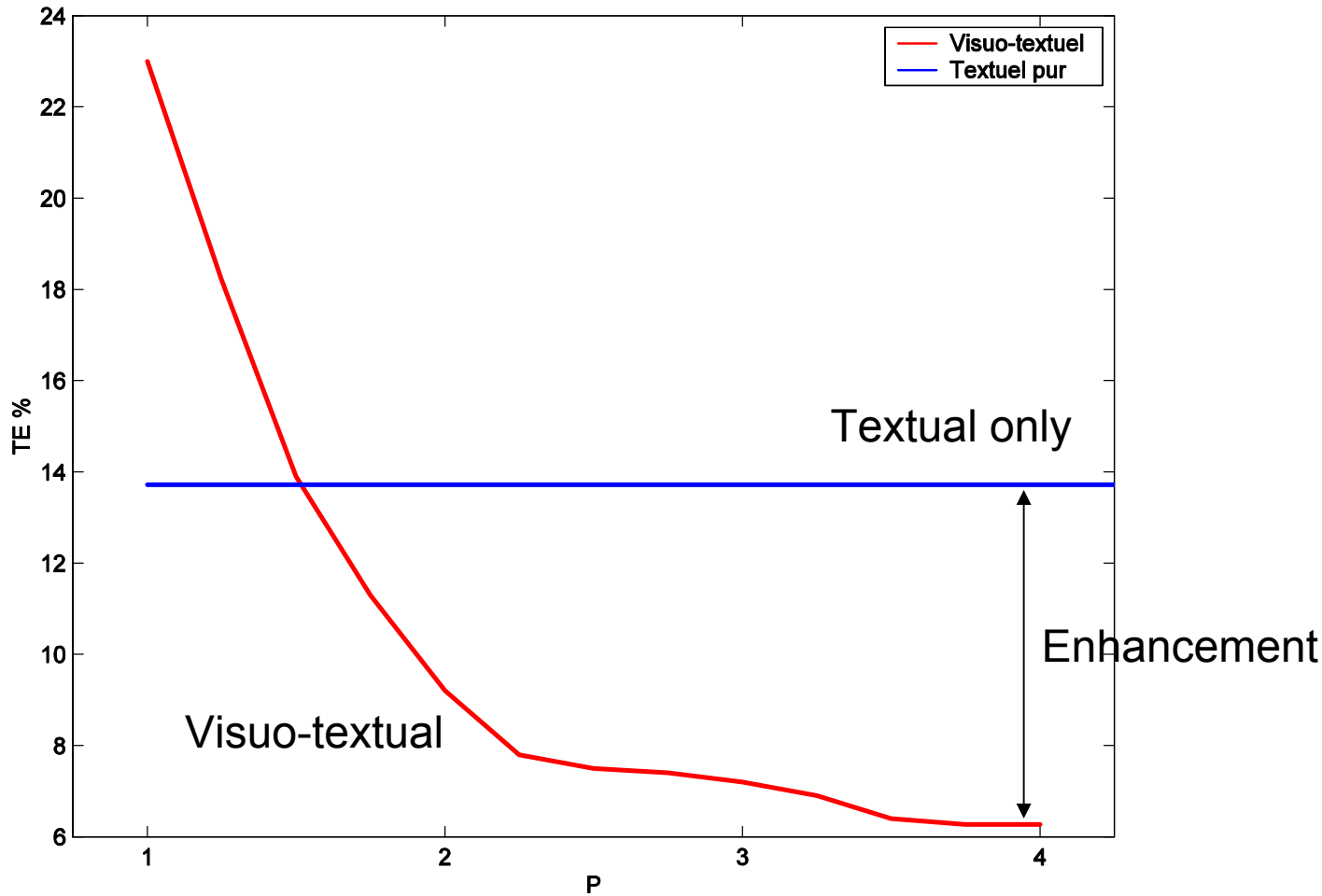
- Let TE(j) be the error rate using visual features A_j

$$\omega(A_j) = \frac{1 - TE(j)}{\sum_{i=1}^{6} 1 - TE(i)}$$

- Weighting distorsion using power p

$$\omega'(A_j) = \frac{\omega(A_j)^p}{\sum_{i=1}^{6} \omega(A_i)^p}$$

Result : Enhancement of textual classification increases with p



Visual Only Error Rate : 71 %

Summary of the visuo-textual enhancement using global content

Results	Textual without thesaurus	Visuo-textual fusion	Gain
Error rate	13.72%	6.27%	+54.3%

Low-level visual features improve textual classification.

Conclusion

- We presented a simple system for unifying textual and visual informations.
- We showed that visual information reduces the errors of the textual information without thesaurus of about 50%
- Our corpus being only of 600 images, our method must be tested on a database of more significant data.

Discussion

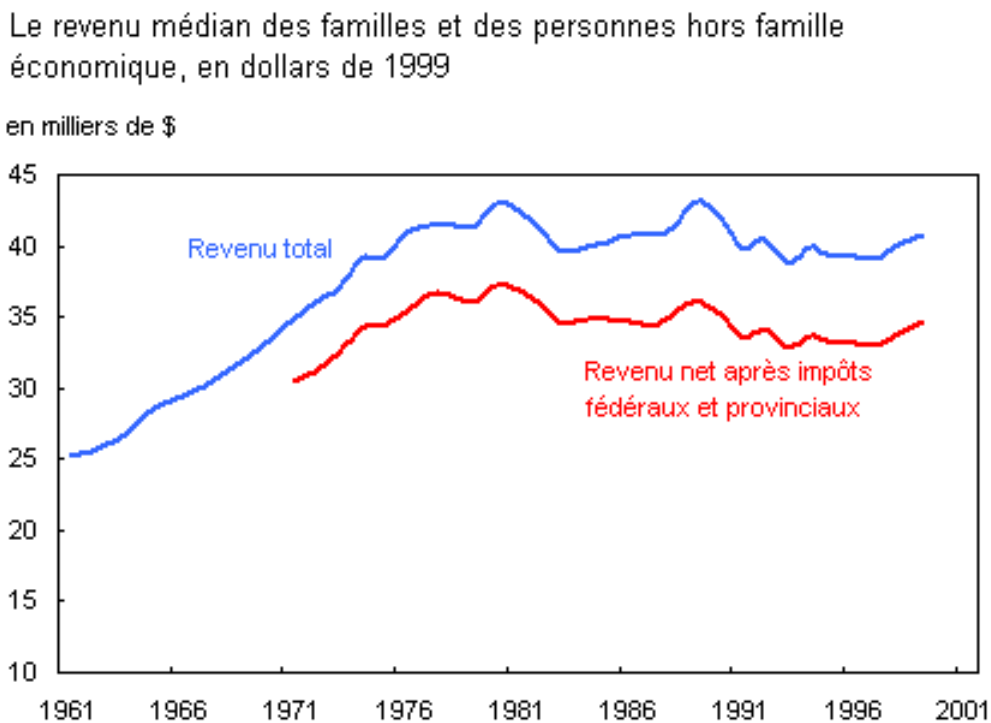
- Other visual attributes as texture or form could be used.

- Many criteria and parameters remain to be studied to improve visual description, as the influence of the size of the visual content .

- Our system can be added as fast visual filter on the result of a request of images on a search engine (such as *Google*).

Perspectives

- One can reverse the experiment to correct a bad textual indexing using the visual content.



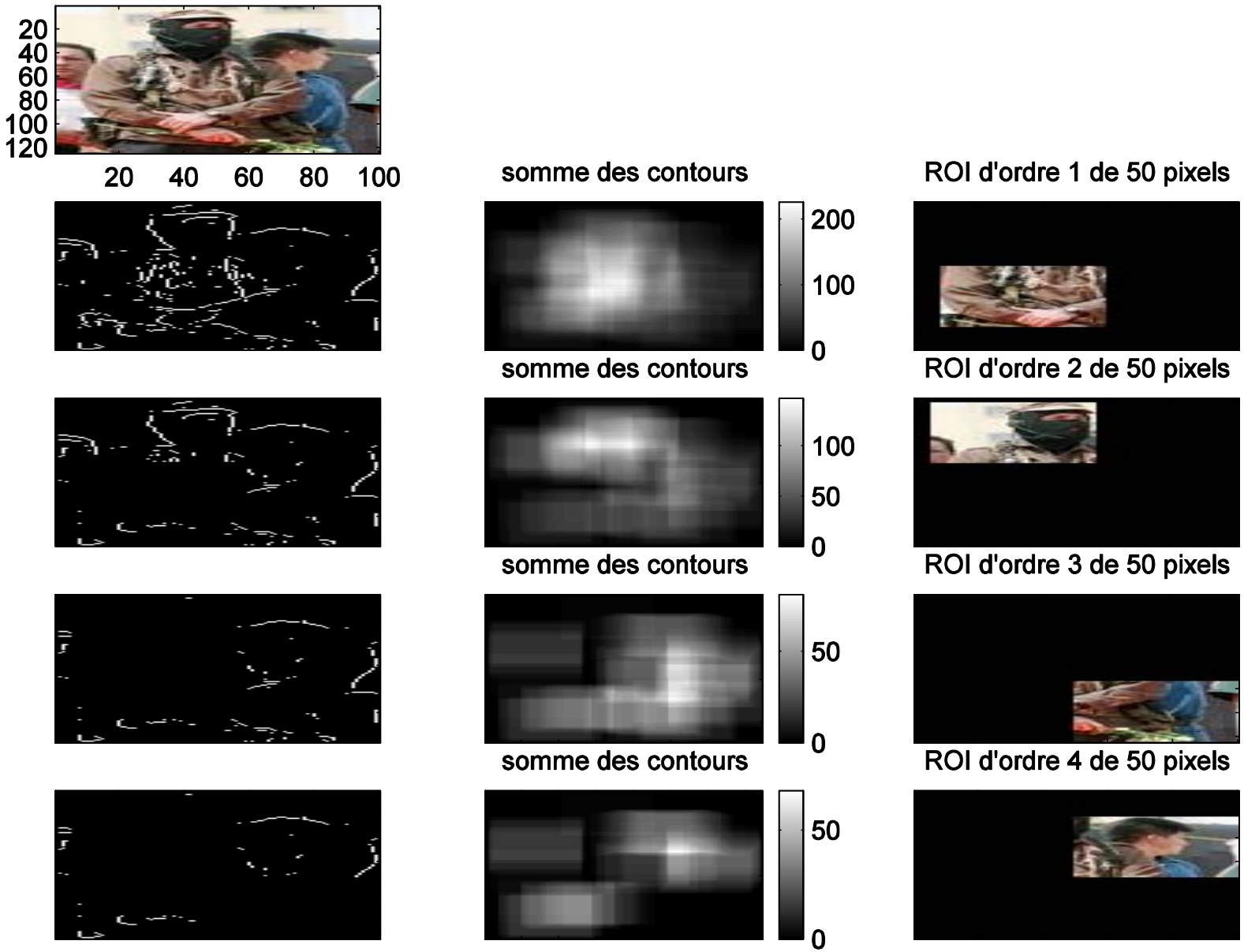
Source : *Statistique Canada*

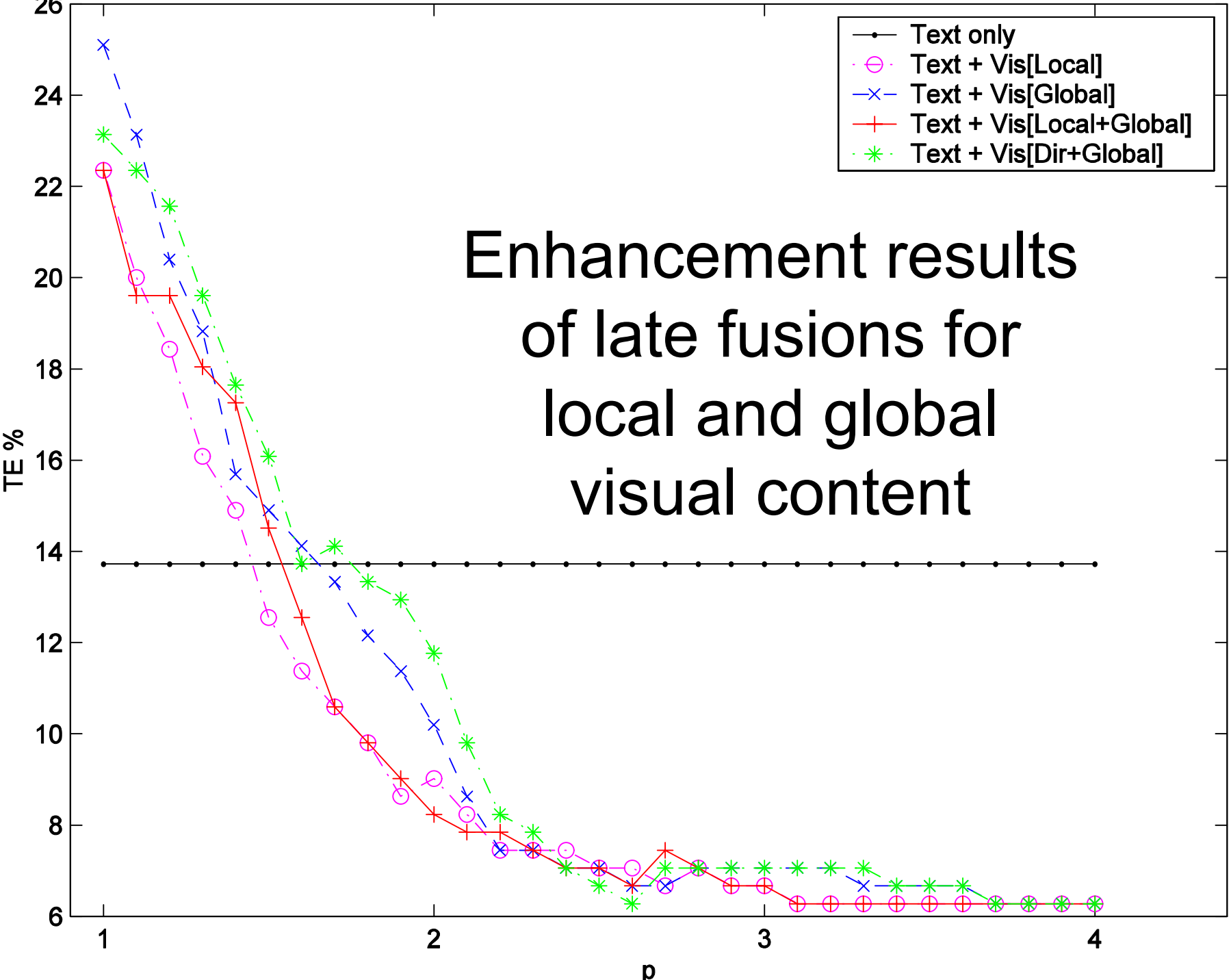
Automatic indexing :

- economy
- dollars
- ~~family~~ ?
- ~~people~~

Thank you for your attention

Local visual content and Region of interest




Text only
Text + Vis[Local]
Text + Vis[Global]
Text + Vis[Local+Global]
Text + Vis[Dir+Global]
Enhancement results of late fusions for local and global visual content
TE %
p
26
24
22
20
18
16
14
12
10
8
6
1
2
3
4