Towards Semantic Image Annotation With Keyword Disambiguation Using Semantic And Visual Knowledge

James, N., Hudelot, C.

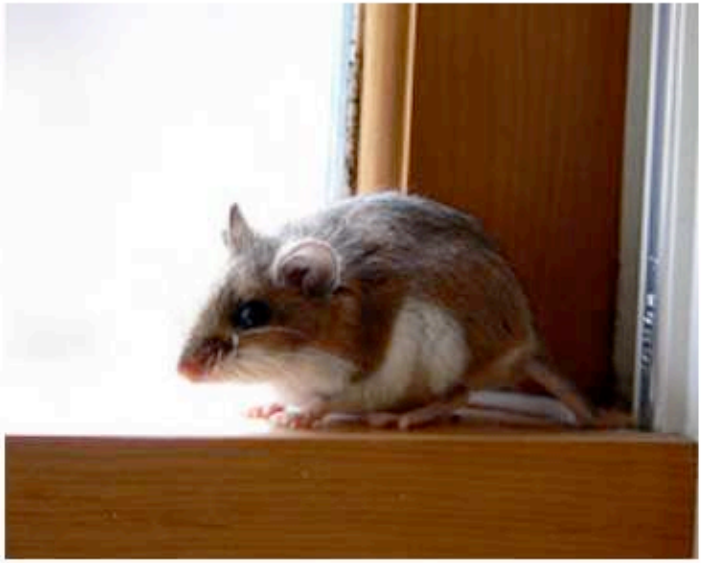
{nicolas.james,celine.hudelot}@ecp.fr

Applied Mathematics Laboratory - École Centrale Paris, France

July 13, CIAM 2009 Workshop







annotation={mouse,hood,window}

Which *mouse* ?



annotation={mouse,keyboard}

Which *mouse* ?

Outline

1. Introduction
   - Context
   - Related works / problems in automatic image annotation
   - Objectives
2. Disambiguating keywords
   - Approach overview
   - Semantic approach for disambiguating keywords
   - Adding visual context knowledge
   - Adding Visual knowledge
   - Combining the three information
3. Evaluation
   - The LabelMe dataset
   - Experimental protocol
4. Conclusion
   - Conclusion

Outline

1 Introduction
- Context
- Related works / problems in automatic image annotation
- Objectives

2 Disambiguating keywords
- Approach overview
- Semantic approach for disambiguating keywords
- Adding visual context knowledge
- Adding Visual knowledge
- Combining the three information

3 Evaluation
- The LabelMe dataset
- Experimental protocol

4 Conclusion
- Conclusion

Context

- Automatic image annotation: automatically associate text-annotation to an image.



(a)

annotation = {car, tree, window, pedestrian, asphalt, building, foliage, sky} (image from LabelMe)



(b)

annotation = {trees, beach, ocean, sky, cliffs} (image from UWDB)

Objectives

- Combining the best of text-based and content-based approaches.
- Build a text or text/visual index for searching from large collections.

Related works / problems in automatic image annotation

Related works

- Co-occurrence model [Mori & al., 99]
- Generative or probabilistic models [Barnard & al., 2003], [Jeon & al., 2003]
- Discriminative and classifier-based models [Jing & al., 2004], [Chang & al., 2003]
- Graph-based model [Pan & al., 2004]

Problems

- Irrelevant keywords and/or missing keywords in the annotation [Khan, 2006], [Jin et al., 2005]
- Many keywords are polysemous [Barnard and Johnson, 2005], [Saenko and Darell,2008]

Problems with Automatic Image Annotation

The keyword polysemy problem



(c)



(d)



(e)



(f)

Solving the problem of the polysemy

Objective

Computing keyword-based annotations exempt of polysemy !

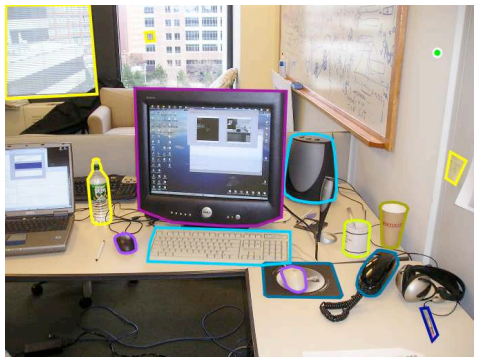


Figure: A sample image in our LabelMe sub-dataset, annotated with the ground truth: {light, cpu, keyboard, mouse, can, speaker, telephone, mousepad, mug, bottle, post-it, window, pen}

Using a knowledge resource like WordNet [Miller, 1995]:

```
<annotation>
<sense-tagged-keyword>
<keyword>mouse</keyword>
<wn-synset>mouse#n#4</wn-synset>
</sense-tagged-keyword>
...
</annotation>
```

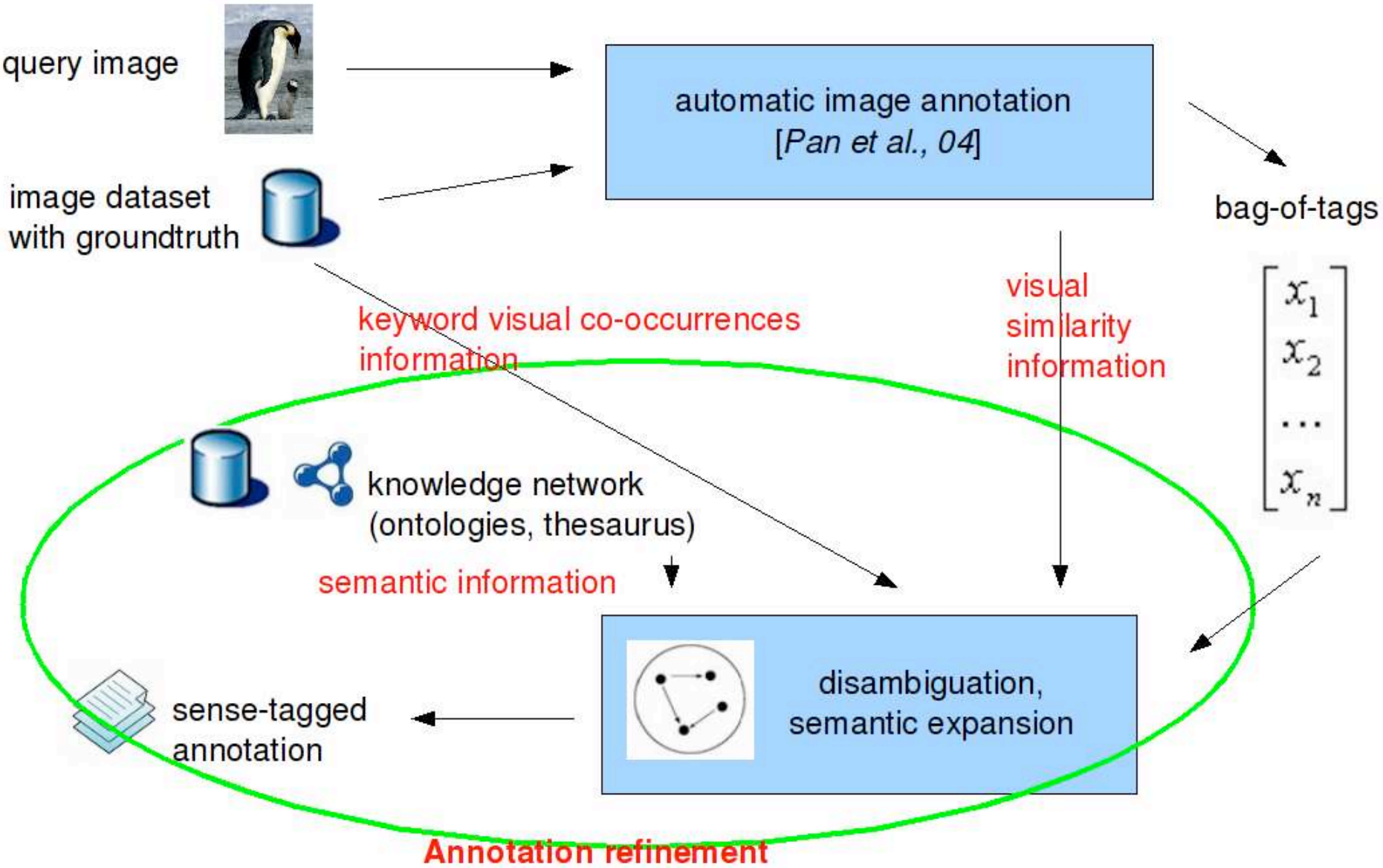
Outline

1 Introduction
- Context
- Related works / problems in automatic image annotation
- Objectives

2 Disambiguating keywords
- Approach overview
- Semantic approach for disambiguating keywords
- Adding visual context knowledge
- Adding Visual knowledge
- Combining the three information

3 Evaluation
- The LabelMe dataset
- Experimental protocol

4 Conclusion
- Conclusion


query image
automatic image annotation
[Pan et al., 04]
image dataset
with groundtruth
bag-of-tags
keyword visual co-occurrences
information
visual
similarity
information
knowledge network
(ontologies, thesaurus)
semantic information
disambiguation,
semantic expansion
sense-tagged
annotation
Annotation refinement

Three types of information

- a semantic information
- a visual knowledge
- a visual context knowledge

Some notations, be:

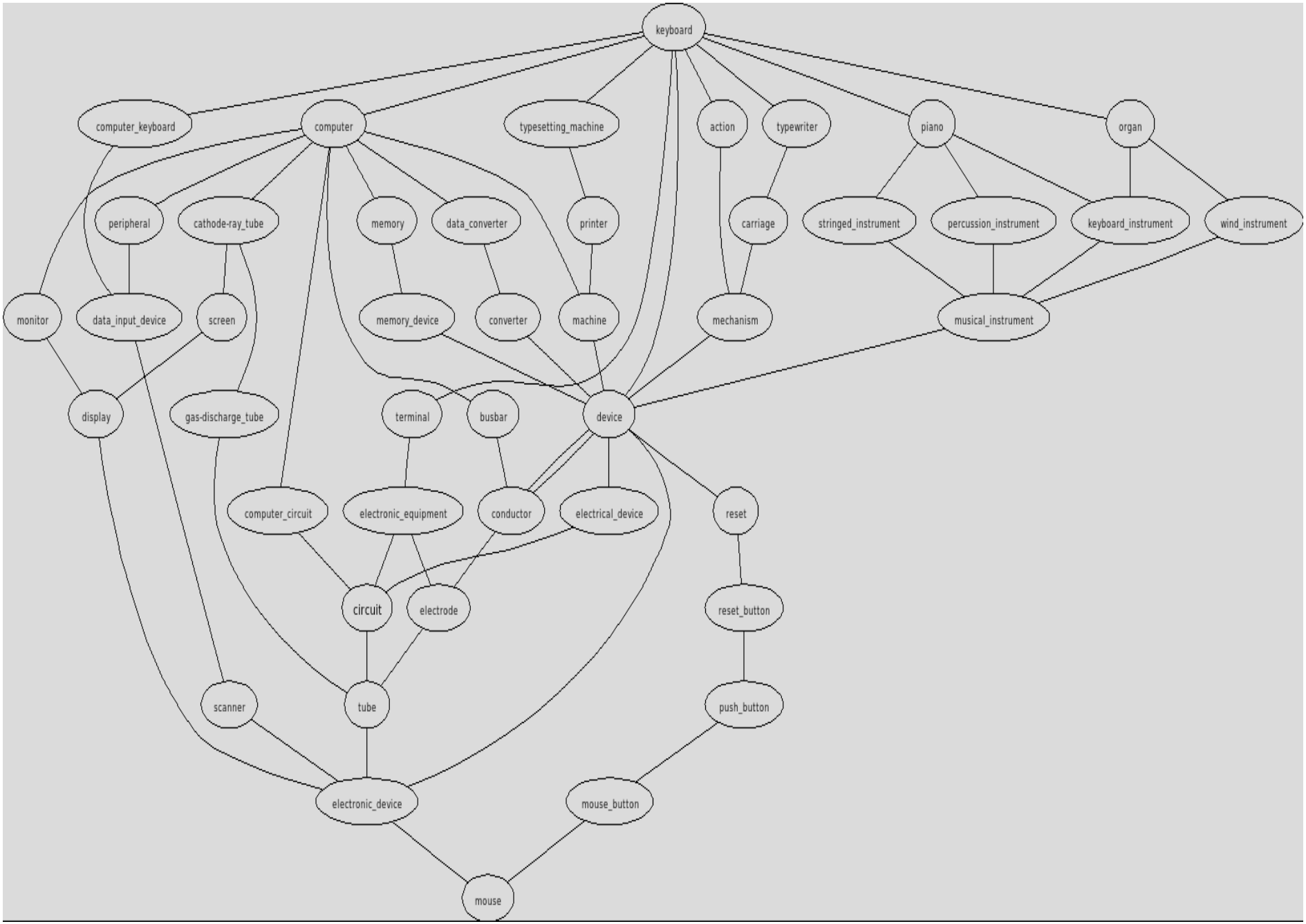
- $\mathscr{L}$, the annotation vocabulary
- A_{I_q}, the computed annotation for the image I_q
- σ_{t_i}, the annotation context for the term t_i
- $\mathscr{W}$, a lexical knowledge resource
- $\hat{S}_{A_{I_q}}$, the sense-tagged annotation
- $Senses(t_i)$, the set of the different senses of the term t_i

Using semantic knowledge

SSI, Structural Semantic Interconnection, [Navigli, 2005]

- from WSD approaches
- for a given bag-of-tags A_{I_q}, computes $\hat{S}_{A_{I_q}} = \{\hat{sense}_{t_1}, \ldots, \hat{sense}_{t_r}\}$
- use a lexical knowledge resource $\mathscr{W}$ (WordNet) allowing to compute semantic interconnections between two given concepts
- iterative algorithm
- for t_i, the sense computed is the one that maximizes the degree of mutual interconnections of t_i with all terms in the context σ_{t_i}:
  $\hat{sense}_{t_i} = argmax_{s \in Senses(t_i)} f(s, \sigma_{t_i})$

Example of semantic interconnection graph



Adding visual context knowledge

query image 

annotation $[t_1, t_2, \ldots, t_i = 'jaguar', \ldots, t_n]$

Defining a semantic relation *related-to(Synset s)*: represent closed synsets according to severals semantic relations: (hyper|hypo|mero|holo)nymy, part-of etc.

For example:
related-to(jaguar#n#1) = { cat, big cat, feline, america, leopard, classification, genus felis }

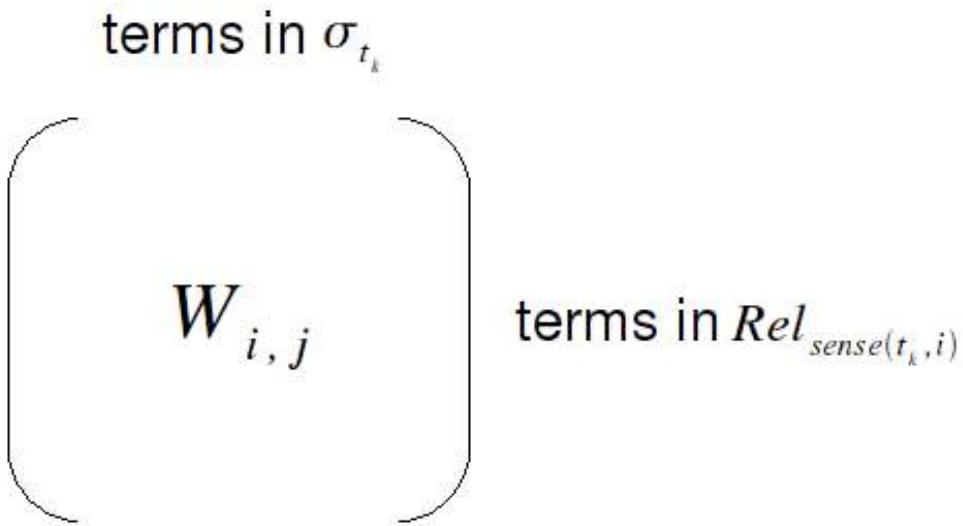


Image dataset with groundtruth

keywords	jaguar, vegetal, rock	jaguar, cat	car, jaguar, sea, sydney, opera house
images			

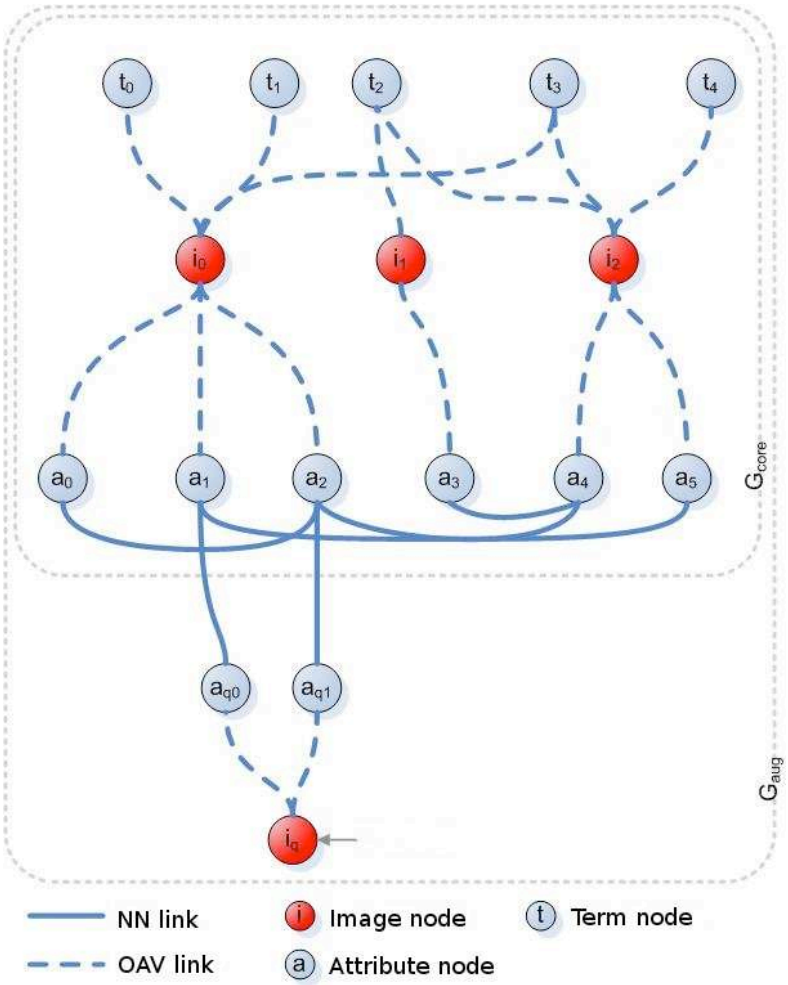
Adding visual context knowledge

- we compute a co-occurence matrix $W_{sense(t_k,i)}$ between the terms in σ_{t_k}, the annotation context of t_k and $Rel_{sense(t_k,i)}$ the semantic neighborhood of the i-ieth sense of t_k



Where $W_{i,j}$ is the number of images where t_i co-occurs with t_j.

Adding Visual knowledge



Be:

- I_q the image to annotate
- $\mathscr{I} = \{I_0, \ldots, I_m\}$ the images in the learning database
- $\mathscr{L} = \{t_0, \ldots, t_n\}$ the lexicon used for the annotation

Mixed Multimedia Graph

- modeling the image dataset
- attribute nodes represent image segment features
- terms nodes are the annotation vocabulary

Adding Visual knowledge

Mixed Multimedia Graph - Random Walks with Restarts

Using [Pan & al., 2004] for the computing of the bag-of-tags:

- $A_{I_q} = \{(t_j, \mu(t_j)) \mid t_j \in \mathscr{L}, 1 \leq j \leq r\}$ a set of r couples $(t_j, \mu(t_j))$ where $\mu(t_j)$ represents the probability that t_j is relevant to annotate I_q.

Adding Visual knowledge

image query



annotation $[t_1, t_2, \ldots, t_i = 'jaguar', \ldots, t_n]$

$Rel_{sense(jaguar, i)}$





Image dataset with groundtruth

keywords	jaguar, vegetal, rock	jaguar, cat	car, jaguar, sea, sydney, opera house
images			

Adding the visual knowledge

- $W_{sense(t_k,i)}$ is a co-occurrence matrix between the terms in σ_{t_k} and the terms in $Rel_{sense(t_k,i)}$.
- for each co-occurrence, we weighting the co-occurence number using the visual similarity of the image implied in the co-occurrence with the query image.

$$w_{t_i,t_j} = \frac{\sum_{I \in \mathscr{I}} \begin{cases} 1 * \mu(I, I_q) & \text{if terms } t_i \text{ and } t_j \text{ co-occurr} \\ 0 & \text{otherwise} \end{cases}}{(max_{I \in \mathscr{I}} \mu(I, I_q)) * (N_{t_i} + N_{t_j} - N_{t_i,t_j})}$$

where:

- $\mu(I, I_q)$ is the visual similarity between I and I_q
- N_{t_i} is the number of images annotated by the term t_i in the learning dataset
- N_{t_i,t_j} is the number of images annotated by the terms t_i and t_j

Combining the three information

The contribution of the visual knowledge may be estimated using the previous co-occurrence matrix : $\frac{||W_s||}{\sum_{s \in sense(t_k)} ||W_s||}$.

Finally, we can adapt the SSI algorithm :

Input: A_{I_q} bag-of-keywords

Output: $\hat{S}_{A_{I_q}}$ set a sense-tagged keywords

while *P is not empty* **do**

  foreach $t \in P$ **do**

    Compute $\hat{S}_{A_{I_q}}[t] = argmax_{s \in sense(t)} \left(\frac{f[s]}{\sum_{s \in sense(t)} f[s]} * \frac{||W_s||}{\sum_{s \in sense(t)} ||W_s||} \right)$

  end

  Remove t_k from P

end

return $\hat{S}_{A_{I_q}}$

Algorithm 1: Adapted SSI algorithm with visual cues

Outline

1 Introduction
- Context
- Related works / problems in automatic image annotation
- Objectives

2 Disambiguating keywords
- Approach overview
- Semantic approach for disambiguating keywords
- Adding visual context knowledge
- Adding Visual knowledge
- Combining the three information

3 Evaluation
- The LabelMe dataset
- Experimental protocol

4 Conclusion
- Conclusion

LabelMe statistics (in nov. 2008)

- 46,302 documents
- 282,719 annotated objects
- 9,666 annotation classes and 282,719 instances
- 1,731 keywords are polysemous using WordNet 3.0.

Experimental protocol

- building a ground truth database from an extended version of LabelMe (using Flickr CC at the moment).
- comparing our approach to the one using the most common sense in WordNet.
- the implementation of a prototype is in course, thus we don't have explicit results for the moment.

An example on the LabelMe dataset for the word `mouse` (visual cooccurrences matrices):

terms in $Rel_{mouse\#N\#1}$

terms in σ_{I_k}

	rodent	gnawer	rat	snout	ear	body	slender	tail
light	0	0	0	0	0	0	0	0
cpu	0	0	0	0	0	0	0	0
keywoard	0	0	0	0	0	0	0	0
can	0	0	0	0	0	0	0	0
speaker	0	0	0	0	0	0	0	0
telephone	0	0	0	0	0	0	0	0
mousepad	0	0	0	0	0	0	0	0
mug	0	0	0	0	0	0	0	0

terms in $Rel_{mouse\#N\#4}$

terms in σ_{I_k}

	device	cursor	computer	screen	pad	ball
light	0	0	6	11	0	0
cpu	0	0	1	3	0	0
keyboard	0	0	0	0	0	0
can	0	0	19	23	1	0
speaker	0	0	4	11	0	0
telephone	0	0	5	9	0	0
mousepad	0	0	2	6	1	0
mug	0	0	0	0	0	0

terms in related-to($mouse\#N\#i$) but not in $\mathcal{L}$

Outline

1 Introduction
- Context
- Related works / problems in automatic image annotation
- Objectives

2 Disambiguating keywords
- Approach overview
- Semantic approach for disambiguating keywords
- Adding visual context knowledge
- Adding Visual knowledge
- Combining the three information

3 Evaluation
- The LabelMe dataset
- Experimental protocol

4 Conclusion
- Conclusion

Conclusion

To conclude

- We propose a method to tackle the polysemy problem combining three information types: a semantic information, a visual knowledge based on visual similarity, a visual co-occurrence information.
- Validating the method allows us to tackle the other problem of large annotation vocabulary.
- The method could also be used for making a sense-tagging of the dataset (e.g. create a sense-tagged version of the LabelMe dataset).

Future work

- a robust and allowing large vocabulary prototype implementation
- introducing other knowledge in the disambiguation step (spatial knowledge on visual objects)

This work is funded by the POPS, Peta Operation Per Second, project and the Fame2 project from the SYSTEM@TIC PARIS-REGION Cluster.



