

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

Nicolas James and Céline Hudelot

MAS Laboratory - École Centrale Paris

Grande Voie des Vignes, F-92295 Châtenay-Malabry Cedex, France

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

Abstract

This paper deals with the semantic enrichment of automatic annotations of images. Since it partially tackles the **Semantic Gap Problem**, semantic image annotation has received a large attention in the recent years. Nevertheless, the results of existing image annotation approaches are still not sufficient. We propose an original approach combining *a priori knowledge* (in our case, the WordNet lexical resource) and **visual knowledge** to build sense-tagged keywords-based annotation. First, a graph-based approach assigns a bag-of-keywords to a query image. Then, we propose to adapt a word sense disambiguation algorithm named SSI (Structural Semantic Interconnections), initially dedicated to text. We make two adaptations. First the grammar used in the SSI is modified to reflect the preponderance of semantic relations in image databases. Then, visual knowledge, including co-occurrence statistics in the visual domain and *visual cues*, is integrated. At last, a method to evaluate our approach is proposed.

1 Introduction

The development of the Web and the democratisation of information technologies have generated an explosion of digital images, requiring new effective methods to manage them. As *Text-Based Image Retrieval* and *Content-Based Image Retrieval* have shown their limits, **Content-Based Image Annotation (CBIA)**, has been widely studied in the literature [Li and Wang, 2003; Cusano *et al.*, 2003; Duygulu *et al.*, 2002; Blei and Jordan, 2003; Lavrenko *et al.*, 2003; Jeon *et al.*, 2003]. It consists in the automatic association of a set of semantic keywords, which depict their content, to images. A good classification of these different approaches can be found in [Wang *et al.*, 2007]. CBIA partially answers to the **Semantic Gap Problem** defined by [Smeulders *et al.*, 2000] as *the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation*.

However, the results of existing image annotation approaches are still not sufficient. First, resulting keywords

have no real semantics and as a consequence are prone to the **polysemy problem**. This is one of the image annotation challenge pointed in [Alm *et al.*, 2006]. Secondly, CBIA results in a **flat bag-of-keywords** without semantic nor structural relations between the different keywords. At last, resulting annotations still have **missing and irrelevant keywords**.

In Section 2, we propose a brief state-of-the-art on image annotation refinement and improvement. In this paper, we propose an original approach combining *a priori knowledge* (in our case, the WordNet lexical resource [Miller, 1995]) and **visual knowledge** to build sense-tagged keyword annotations. By visual knowledge, we mean both statistical information coming from word co-occurrences in image collections but also visual similarity between images. This approach is based on the adaptation of a word sense disambiguation algorithm named SSI, initially dedicated to text. We detail it in Section 3. At last, we propose a method to evaluate our approach and we discuss our future work.

2 Related Work

As illustrated in Figure 1, keyword annotations resulting from classical CBIA systems have to be refined. The image annotation refinement and improvement problem has been tackled in the literature with three main approaches:

- Approaches using *a priori knowledge* (or semantic knowledge), i.e. the semantic relatedness between keywords. They are commonly based on external lexical resources or ontologies, such as Wordnet. In [Khan, 2006; Jin *et al.*, 2005], a combination of semantic similarity measures is used to remove noisy keywords and a boosting method is used to add missing keywords. More recently, [Saenko and Darrell, 2008] have proposed an approach to address the visual polysemy problem, i.e. *the fact that a word has several dictionary senses that are visually distinct*, by using the text surrounding images and Wordnet. They use the Latent Dirichlet Allocation, or LDA, to extract hidden topics from text. Then, they build classifiers dedicated to a specific sense using web image search.
- Approaches using **visual knowledge**. Two kinds of visual knowledge can be used : the relationship between keywords in the visual domain (i.e. co-occurrence statistics of words in annotated images) and the vi-

visual relatedness (visual similarity) between the query image and selected images in the learning database. Co-occurrence statistics of words in images are exploited in [Bartolini and Ciaccia, 2007] using graph theory and in [Wang *et al.*, 2007] using random markov fields. [Wang *et al.*, 2007] also used visual relatedness. They use the visual similarity between the query image and images in which considered keywords co-occur. The same idea is proposed in [Kucuktunc *et al.*, 2008].

- Approaches combining semantic and visual knowledge such that [Barnard and Johnson, 2005; Barnard *et al.*, 2007] which propose a cross modal keyword disambiguation using both statistical visual information and knowledge from textual resources. In [Escalante *et al.*, 2007], the authors propose a re-ranking process based on both statistical visual information and statistical information about keyword usage extracted from an external collection of captions. More recently, interesting works have proposed the integration of ontologies directly into the automatic multilevel image annotation process [Fan *et al.*, 2008].

However, although all these approaches enable to improve the relevance of the annotation or to refine it, they do not allow to build semantic structured annotation and in particular sense-tagged keyword annotations (except [Barnard *et al.*, 2007]). Another weakness of the approaches based on semantic knowledge is the use of a very limited set of semantic relations : mainly hypernymy (*is-a* relation) and sometimes meronymy relations (*part-of*). This is a serious drawback in image annotation since relations between keywords describing an image are often more than hypernymy or meronymy relations. For instance complex spatial relations can be also of prime importance for image annotation [Millet *et al.*, 2005; Hollink *et al.*, 2004].

In this paper, we try to answer to the problem of image annotation disambiguation by combining both semantic and the two kinds of visual knowledge. The main requirements of our approach are : (1) Be able to take into account into a parameterized way, complex semantic relations; (2) Using relationships between keywords in the visual domain; (3) Using the visual relatedness in the disambiguation algorithm.

3 Automatic image annotation with keyword desambiguation

In this section, we describe our approach to build sense-tagged annotations. This process is composed of two steps. First, given I_q the image query to annotate, let $\mathcal{I} = \{I_0, \dots, I_m\}$ be the images in the learning database, and $\mathcal{L} = \{t_0, \dots, t_n\}$ be the lexicon used for the annotation. The content based annotation step results in the bag-of-keyword annotation, $A_{I_q} = \{(t_j, \mu(t_j)) \mid t_j \in \mathcal{L}, 1 \leq j \leq r\}$ 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 .

For this step, we use the approach proposed by [Pan *et al.*, 2004], based on a modeling of the learning dataset by a graph, called Mixed Multimedia Graph (MMG), described in Figure 2. A Random Walk With Restarts algorithm

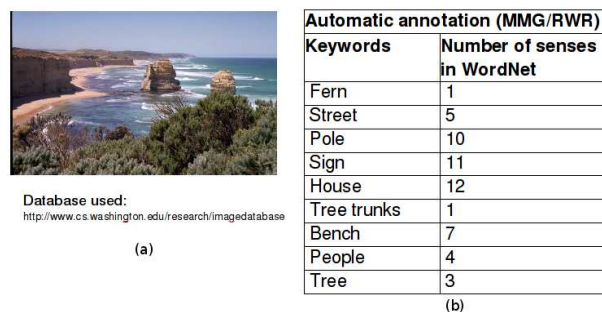


Figure 1: Content-based image annotation obtained with [Pan *et al.*, 2004]. (a) is the image query and (b) the computed keyword-based annotation. We can see that many keywords are polysemous. Some top-ranked keywords are irrelevant (e.g. *street*) and others are missing (e.g. *sea*). Moreover, there are no semantic relations between the obtained keywords.

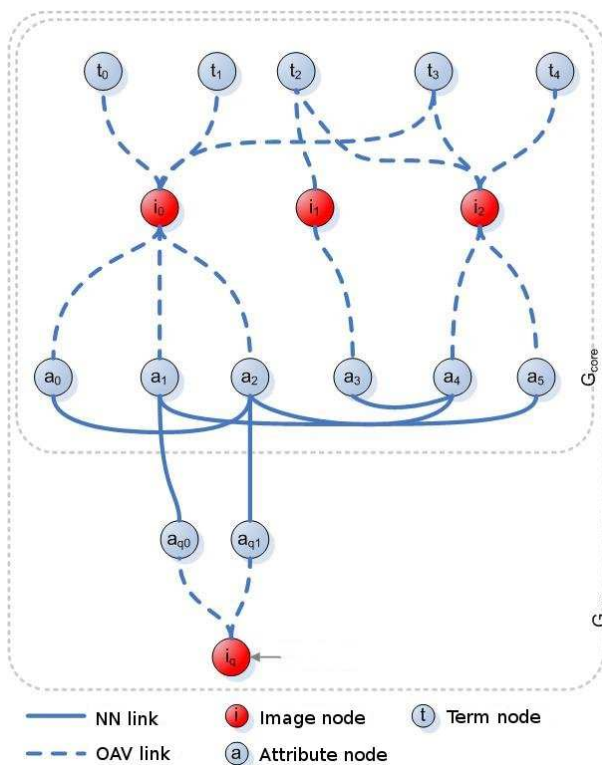


Figure 2: The MMG modeling the image dataset. Attribute nodes (a_i) represent image segment features. i_i are image nodes and t_i are term nodes. In our first experiments, each image is divided into parts and features are three mpeg-7 descriptors: ScalableColor, ColorLayout and EdgeHistogram.

(RWR) is applied to this graph for computing each $\mu(t_j)$ with $t_j \in \mathcal{L}$. $\mu(t_j)$ represents the steady state probability or the affinity of node t_j with I_q . The resulting bag-of-keywords is the set composed of the r terms having the highest affinities

with I_q .

Then, a second step consists in adding a disambiguation process to determine the sense of each keyword in A_{I_q} . Although Word Sense Disambiguation, or WSD, has been studied a lot for text analysis [Navigli, 2009], existing frameworks are not totally convenient for image annotation disambiguation. First, existing lexical resources or ontologies used in these algorithms are dedicated to text and built from textual corpus. As a consequence, the semantic relations between concepts do not really reflect the semantic relationships between concepts in the visual domain. We propose to adapt an existing WSD algorithm named SSI introduced by [Navigli and Velardi, 2005], which exploits the structure of a semantic concept network. One of the main interest of the SSI is to take into account only semantic interconnections which abide a context-free grammar. Our first adaptation, described in Section 3.2, consists in modifying this grammar for the image annotation problem. Then, to answer to requirements (2) and (3), we propose to integrate visual knowledge in the disambiguation process: term co-occurrences in the visual domain and visual relatedness between images in which the annotation context of a given term and the semantic neighbourhood of this term co-occur (see Section 3.3).

3.1 Knowledge-based disambiguation algorithm

For a good understanding of our approach, we give a brief overview of the SSI in the following. We define a **sense-tagged keyword** as a word linked to its relevant WordNet synset (i.e. *set of synonyms that together define a particular sense uniquely*). For example, the keyword *mouse* is referenced in WordNet by the following four synsets:

- mouse#n#1: {mouse} (with the gloss *any of numerous small rodents typically resembling diminutives rats,...*).
- mouse#n#2: {shiner, black eye, mouse} (with the gloss *a swollen bruise caused by a blow to the eye*)
- mouse#n#3: {mouse} (with the gloss *a swollen bruise caused by a blow to the eye*)
- mouse#n#4: {mouse, computer mouse} (with the gloss *a hand operated electronic device ...*).

We denote $Senses(t)$, the set of synsets which represents the possible senses of the keyword t in WordNet. We say that t is sense-tagged when it has the predicted sense s , where s is a synset. We denote it $s = \hat{sense}_t$. To compute it, we use the semantic context of t . Our objective is to build the semantic interpretation of A_{I_q} , i.e. $\hat{S}_{A_{I_q}} = \{\hat{sense}_{t_1}, \dots, \hat{sense}_{t_r}\}$. Given:

- an ambiguous term $t_i \in A_{I_q}$,
- the annotation context of t_i denoted σ_{t_i} which is the set of all monosemous or ever disambiguated term $\in A_{I_q} \setminus t_i$,
- a lexical knowledge resource \mathcal{W} (WordNet) allowing to compute semantic interconnections between two given concepts.

The idea of the SSI algorithm is the following. It is an iterative algorithm. It selects for t_i the sense that maximize the degree of mutual interconnection of t_i with all terms in the

context σ_{t_i} : $\hat{sense}_{t_i} = \underset{s \in Senses(t_i)}{\operatorname{argmax}} f(s, \sigma_{t_i})$ where the function f estimates the weight of semantic interconnections. In the original article of the SSI, the f function is a sum function. The algorithm is described in pseudo-code in Algorithm 1.

Input: A_{I_q} : a bag-of-keyword, $\hat{S}_{A_{I_q}}$: a list of sense-tagged keywords (initially empty)

Output: $\hat{S}_{A_{I_q}}$: a list of sense-tagged keywords of A_{I_q} (an array)

```

foreach  $t \in A_{I_q}$  do
  if  $t$  is monosemic then
     $\hat{S}_{A_{I_q}}[t] = \text{sense}(t, 1)$ 
  end
end
 $P = \{t \in A_{I_q} \mid \hat{S}_{A_{I_q}}[t] = \text{null}\}$ 
while  $P \neq \emptyset$  do
  foreach  $t \in P$  do
     $\hat{sense}_t = \text{null}, \text{maxValue} = 0$ 
    foreach sense  $s$  of  $t$  do
       $f[s] = 0$ 
      foreach sense  $s'$  of  $\hat{S}_{A_{I_q}}$  do
         $\phi = \emptyset$ 
        foreach semantic path between  $s$  and  $s'$  do
           $\phi = \phi + \text{weight}(s, s')$ 
        end
         $f[s] = f[s] + \phi$ 
      end
      if  $f[s] > \text{maxValue}$  then
         $\text{maxValue} = f[s]$ 
         $\hat{sense}_t = s$ 
      end
    end
    if  $\text{maxValue} > 0$  then
       $\hat{S}_{A_{I_q}}[t] = \hat{sense}_t$ 
       $P = P \setminus \{t\}$ 
    end
  end
end
return  $\hat{S}_{A_{I_q}}$ 

```

Algorithm 1: The SSI algorithm. A represents $S_{A_{I_q}}$, $\hat{S}_{A_{I_q}}$ represents σ . $\text{sense}(t, i)$ represents the i -ith sense of t . P is the set of terms to disambiguate. $\text{weight}(s, s')$ is a function that estimate the weight of a semantic path.

3.2 Modifying the grammar of the SSI to image annotation

An important point in the SSI is that among semantic interconnections found in the WordNet database, only those which abide a context-free grammar are taken in consideration. This grammar is manually defined and encodes the relevant semantic patterns between two concepts. To adapt it to image annotation disambiguation, we use an interesting study done in [Hollink *et al.*, 2007]. This study was dedicated to query expansion to improve image retrieval. It proposes a set

of semantic relation patterns that seems to be useful to improve the retrieval. In the following, we will investigate the learning of these semantic relation patterns and their associated weights in large annotated image collections such as LabelMe¹ or Flickr².

3.3 Introducing visual knowledge in the SSI algorithm

As emphasized in [Wang *et al.*, 2007], using only external *a priori* knowledge to improve keyword annotations is not sufficient. Indeed, this external knowledge often reflects the semantic relatedness of keywords in the textual domain which can be different in the visual domain. For instance, in [Wu *et al.*, 2008], the importance of a *concurrency* relation which represents the co-occurrence of concepts or background coherence in visual domain is underlined. The authors propose to build a visual ConceptNet using Flickr and a Visual Language Model, and they applied it to image management tasks as image annotation and image clustering. To answer to our second requirement, in addition to semantic knowledge used in the SSI, we also integrate visual knowledge. Semantic knowledge refers to semantic correlations between keywords that we can extract in lexical resources. Visual knowledge refers to the co-occurrences statistics between keywords and images in the learning dataset. We also integrate visual relatedness between images in which the annotation context of a given term and its semantic neighbourhood co-occur. This information is made available by the first step, i.e. the random walk with restarts over the MMG graph. It is the probability $\mu(I, I_q)$ where $I \in \mathcal{I}$.

Adding visual knowledge

A strong hypothesis of our method is to suppose that there is only one occurrence of a given concept in the query image. Moreover, we suppose that if the keyword corresponds to a visual concept, the keyword is present in the image under only one meaning.

Let $A_{I_q} = \{t_1, \dots, t_r\}$ be the keyword-based annotation to disambiguate. Given, t_k , a polysemous term to disambiguate, $A_{I_q} \setminus t_k$ is the annotation context of t_k . We build the set $Rel_{sense(t_k, i)}$ which is a set of terms extracted from the semantic neighbourhood of t_k in its i -th sense. To build this set, we concatenate the hypernymy, meronymy and gloss relations into a more general *related-to* relation. More precisely, for a given term t_k under its i -th sense, *related-to*($sense(t_k, i)$) is built by taking :

- direct hypernyms,
- meronyms,
- WordNet nouns found in the gloss.

For instance, with the term *street* and its first sense in WordNet, $related-to(street\#n\#1) = \{thoroughfare, pavement, paving, sidewalks, buildings\}$.

Then $Rel_{sense(t_k, i)} = \mathcal{L} \cap related-to(sense(t_k, i))$. The intersection with \mathcal{L} ensures that we only keep the term in the annotated learning image database.

To compute the semantic relatedness of terms in the visual domain, we compute, for a given polysemous term t_k , and its i -th sense, the co-occurrence matrix $W_{sense(t_k, i)}$ between the terms in $A_{I_q} \setminus t_k$, the annotation context of t_k and $Rel_{sense(t_k, i)}$ the semantic neighbourhood of t_k . Finally to take into account the third requirement, we weight the co-occurrence number using the visual similarity with the query image (as in [Wang *et al.*, 2007; Kucuktunc *et al.*, 2008]). As a consequence, each element of the matrix is defined by :

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

where $\mu(I, I_q)$ is the steady state probability of the image I given the image query I_q , that can be assimilated to a visual similarity between I and I_q . It is given by the RWR. N_{t_i} is the number of images annotated by the term t_i in the learning dataset, and N_{t_i, t_j} is the number of images annotated by the terms t_i and t_j .

Integration in the SSI

Our objective is to disambiguate bag-of-keyword annotations using both semantic knowledge and visual knowledge. The contribution of the semantic knowledge can be estimated by computing the semantic relatedness between a term and its disambiguated annotation context (σ) using semantic interconnections. This quantity can be estimated as :

$$\frac{f[s]}{\sum_{s \in sense(t_k)} f[s]} \quad (\text{where } f[s] \text{ is the variable in Algorithm 1}).$$

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

If we make the strong assumption that the contribution of the semantic knowledge and the contribution of the visual knowledge are relatively independent, we can adapt the SSI algorithm by the following:

```

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 2: Adapted SSI algorithm with visual cues

4 Evaluation

To evaluate our approach, we propose to apply the disambiguation method to the LabelMe dataset.

¹<http://labelme.csail.mit.edu/>

²<http://www.flickr.com/>

4.1 The LabelMe dataset

The LabelMe dataset is build around a prototype [Russell *et al.*, 2008] for building an image dataset using Web users to annotate images. Downloaded from the experiment website on the 9 nov. 2008, the dataset is composed of 46302 images for 282838 annotated objects. Annotated objects are drawn polygons, and annotations are attached to each polygon. Among these annotated objects, we can find 9698 classes, or keywords. Experiments have ever been made to apply disambiguation methods with WordNet on LabelMe [Russell *et al.*, 2008]. However, the disambiguation is made only on the classes independently of the real annotated object in the images. Moreover they assumes that each polysemous class is present under only one meaning. At last, as the association between annotation classes and WordNet synsets are made manually in their method, it could not be applied for large semantic image databases (i.e. image databases with a large number of polysemous classes, as for instance Flickr).

4.2 Ground truth

To evaluate our method, we have to build a ground truth database from LabelMe. Indeed, LabelMe is usually considered as an image annotation ground truth but due to free-text annotations, the quality is very unpredictable.

So, in our first experiments, we have removed from annotations terms those that are not found in WordNet 3.0. In our final dataset, we have aligned 32% of all the annotations in LabelMe with WordNet synsets, i.e. the annotation is composed of words found in one or several WordNet synsets, be 3118 classes. And among these annotations, 1735 are polysems using WordNet, be 55% of all classes.



Figure 3: A sample image in our LabelMe sub-dataset, annotated with the ground truth: {light, cpu, keyboard, mouse, can, speaker, telephone, mouse, mousepad, mouse, mug, bottle, post-it, window, pen}

4.3 Evaluation protocol

To evaluate the method, we propose to compare our approach to the one using the most common sense in WordNet. For

instance, considering the word *mouse* used for annotating the image of Figure 3, the most common sense in WordNet is *mouse#n#1* whose the gloss is *any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails* which is not the sense of the annotated object in the figure 3. In WordNet, the word *mouse* has the following senses:

- sense 1: {mouse} – (any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails)
- sense 2: {shiner, black eye, mouse} – (a swollen bruise caused by a blow to the eye)
- sense 3: {mouse} – (person who is quiet or timid)
- sense 4: {mouse, computer mouse} – (a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad; "a mouse takes much more room than a trackball")

Considering this example, we can feel the benefits of our method. Our first experiments on a small set of polysemous terms are promising. Nevertheless, we have not enough quantitative results to validate this method. An important next step of our works will be to challenge our approach on existing corpus or benchmarks.

5 Conclusion and future works

To sum up, in this article we propose a method to tackle the visual polysemy problem in the context of automatic image annotation. Compared to other methods, we propose an original approach which enables to take into account *a priori* knowledge (WordNet), visual cues (by taking the visual information on the query image I_q) and also term cooccurrences in the image learning dataset. Our method allows to build an image annotation composed of sense-tagged keywords. Our first experiments take a modified LabelMe dataset as our evaluation dataset and are still in running. Future work is logically to quantify these experiments well and to validate it on other databases (e.g. Flickr).

References

- [Alm *et al.*, 2006] Cecilia Ovesdotter Alm, Nicolas Loeff, and David A. Forsyth. Challenges for annotating images for sense disambiguation. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pages 1–4, Sydney, Australia, July 2006. Association for Computational Linguistics.
- [Barnard and Johnson, 2005] Kobus Barnard and Matthew Johnson. Word sense disambiguation with pictures. volume 167, pages 13–30, Essex, UK, 2005. Elsevier Science Publishers Ltd.
- [Barnard *et al.*, 2007] K. Barnard, K. Yanai, M. Johnson, and P. Gabbur. *Cross Modal Disambiguation*, pages 238–257. Springer Berlin / Heidelberg, 2007.

- [Bartolini and Ciaccia, 2007] I. Bartolini and P. Ciaccia. Imagination: Accurate image annotation using link-analysis techniques. *5th International Workshop on Adaptive Multimedia Retrieval (AMR 2007)*, Paris, 2007.
- [Blei and Jordan, 2003] David M. Blei and Michael I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 127–134, New York, NY, USA, 2003. ACM.
- [Cusano et al., 2003] C. Cusano, G. Ciocca, and R. Schettini. Image annotation using SVM. In S. Santini and R. Schettini, editors, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 5304 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 330–338, December 2003.
- [Duygulu et al., 2002] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *Proceedings of the 7th European Conference on Computer Vision-Part IV*, pages 97–112, 2002.
- [Escalante et al., 2007] H.J. Escalante, M. Montes y Gomez, and L.E. Sucar. Word co-occurrence and markov random fields for improving automatic image annotation. pages xx–yy, 2007.
- [Fan et al., 2008] J. Fan, Y. Gao, H. Luo, and S. Satoh. New Approach for Hierarchical Classifier Training and Multi-level Image Annotation. *LECTURE NOTES IN COMPUTER SCIENCE*, 4903:45, 2008.
- [Hollink et al., 2004] Laura Hollink, Giang Nguyen, Guus Schreiber, Jan Wielemaker, and Bob Wielinga. Adding spatial semantics to image annotations. In *4th International Workshop on Knowledge Markup and Semantic Annotation at ISWC04*, pages 31–40, 2004.
- [Hollink et al., 2007] Laura Hollink, Guus Schreiber, and Bob Wielinga. Patterns of semantic relations to improve image content search. *Web Semant.*, 5(3):195–203, 2007.
- [Jeon et al., 2003] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 119–126, 2003.
- [Jin et al., 2005] Y. Jin, L. Khan, L. Wang, and M. Awad. Image annotations by combining multiple evidence & wordnet. *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 706–715, 2005.
- [Khan, 2006] L. Khan. Improving image annotations using fuzzy pruning and association rule mining. *MDM/KDD'06*, 2006.
- [Kucuktunc et al., 2008] Onur Kucuktunc, Sare Gul Sevil, A. Burak Tosun, Hilal Zitouni, Pinar Duygulu, and Fazli Can. Tag suggestr: Automatic photo tag expansion using visual information for photo sharing websites. In David J. Duke, Lynda Hardman, Alexander G. Hauptmann, Dietrich Paulus, and Steffen Staab, editors, *SAMT*, volume 5392 of *Lecture Notes in Computer Science*, pages 61–73. Springer, 2008.
- [Lavrenko et al., 2003] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Proceedings of Advance in Neural Information Processing*, 2003.
- [Li and Wang, 2003] Jia Li and James Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1075–1088, 2003.
- [Miller, 1995] George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.
- [Millet et al., 2005] Christophe Millet, Isabelle Bloch, I. Bloch, Patrick Hde, and Pierre-Alain Mollic. Using relative spatial relationships to improve individual region recognition. In *In Proceedings of 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, pages 119–126, 2005.
- [Navigli and Velardi, 2005] Roberto Navigli and Paola Velardi. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(7):1075–1086, 2005.
- [Navigli, 2009] Roberto Navigli. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):1–69, 2009.
- [Pan et al., 2004] J.Y. Pan, H.J. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 653–658, 2004.
- [Russell et al., 2008] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77(1-3):157–173, 2008.
- [Saenko and Darrell, 2008] K. Saenko and T. Darrell. Unsupervised learning of visual sense models for polysemous words. In *Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems (NIPS), Vancouver, Canada, to appear*, 2008.
- [Smeulders et al., 2000] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. 2000.
- [Wang et al., 2007] Changhu Wang, Feng Jing, Lei Zhang, and Hong-Jiang Zhang. Content-based image annotation refinement. *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.
- [Wu et al., 2008] Lei Wu, Xian-Sheng Hua, Nenghai Yu, Wei-Ying Ma, and Shipeng Li. Flickr distance. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, pages 31–40, New York, NY, USA, 2008. ACM.