Text Analysis for Automatic Image Annotation

Koen Deschacht and **Marie-Francine Moens**
Interdisciplinary Centre for Law & IT
Department of Computer Science
Katholieke Universiteit Leuven
Tiensestraat 41, 3000 Leuven, Belgium
{koen.deschacht,marie-france.moens}@law.kuleuven.ac.be

Abstract

We present a novel approach to automatically annotate images using associated text. We detect and classify all entities (persons and objects) in the text after which we determine the salience (the importance of an entity in a text) and visualness (the extent to which an entity can be perceived visually) of these entities. We combine these measures to compute the probability that an entity is present in the image. The suitability of our approach was successfully tested on 100 image-text pairs of Yahoo! News.

1 Introduction

Our society deals with a growing bulk of unstructured information such as text, images and video, a situation witnessed in many domains (news, biomedical information, intelligence information, business documents, etc.). This growth comes along with the demand for more effective tools to search and summarize this information. Moreover, there is the need to mine information from texts and images when they contribute to decision making by governments, businesses and other institutions. The capability to accurately recognize content in these sources would largely contribute to improved indexing, classification, filtering, mining and interrogation.

Algorithms and techniques for the disclosure of information from the different media have been developed for every medium independently during the last decennium, but only recently the interplay between these different media has become a topic of interest. One of the possible applications is to help analysis in one medium by employing information from another medium. In this paper we study text that is associated with an image, such as for instance image captions, video transcripts or surrounding text in a web page. We develop techniques that extract information from these texts to help with the difficult task of accurate object recognition in images. Although images and associated texts never contain precisely the same information, in many situations the associated text offers valuable information that helps to interpret the image.

The central objective of the CLASS project[1] is to develop advanced learning methods that allow images, video and associated text to be automatically analyzed and structured. In this paper we test the feasibility of automatically annotating images by using textual information in near-parallel image-text pairs, in which most of the content of the image corresponds to content of the text and vice versa. We will focus on entities such as persons and objects. We will hereby take into account the text's discourse structure and semantics, which allow a more fine-grained identification of what content might be present in the image, and will enrich our model with world knowledge that is not present in the text.

We will first discuss the corpus on which we apply and test our techniques in section 2, after which we outline what techniques we have developed: we start with a baseline system to annotate images with person names (section 3) and improve this by computing the importance of the persons in the text (section 4). We will then extend the model to include all

[1] http://class.inrialpes.fr/

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 1000–1007,
Prague, Czech Republic, June 2007.

Hiram Myers, of Edmond, Okla., walks across the fence, attempting to deliver what he called a 'people's indictment' of Halliburton CEO David Lesar, outside the site of the annual Halliburton shareholders meeting in Duncan, Okla., leading to his arrest, Wednesday, May 17, 2006.

Figure 1: Image-text pair with entity "Hiram Myers" appearing both in the text and in the image.

types of objects (section 5) and improve it by defining and computing the $visualness$ measure (section 6). Finally we will combine these different techniques in one probabilistic model in section 7.

2 The parallel corpus

We have created a parallel corpus consisting of 1700 image-text pairs, retrieved from the Yahoo! News website[2]. Every image has an accompanying text which describes the content of the image. This text will in general discuss one or more persons in the image, possibly one or more other objects, the location and the event for which the picture was taken. An example of an image-text pair is given in fig. 1. Not all persons or objects who are pictured in the images are necessarily described in the texts. The inverse is also true, i.e. content mentioned in the text may not be present in the image.

We have randomly selected 100 text-pairs from the corpus, and one annotator has labeled every image-text pair with the entities (i.e. persons and other objects) that appear both in the image and in the text. For example, the image-text pair shown in fig. 1 is annotated with one entity, "Hiram Myers", since this is the only entity that appears both in the text and in the image. On average these texts contain 15.04 entities, of which 2.58 appear in the image.

To build the appearance model of the text, we have combined different tools. We will evaluate every tool separately on 100 image-text pairs. This way we have a detailed view on the nature of the errors in the final model.

3 Automatically annotating person names

Given a text that is associated with an image, we want to compute a probabilistic *appearance model*, i.e. a collection of entities that are visible in the image. We will start with a model that holds the names of the persons that appear in the image, such as was done by (Satoh et al., 1999; Berg et al., 2004), and extend this model in section 5 to include all other objects.

3.1 Named Entity Recognition

A logical first step to detect person names is Named Entity Recognition (NER). We use the OpenNLP package[3], which detects noun phrase chunks in the sentences that represent persons, locations, organizations and dates. To improve the recognition of person names, we use a dictionary of names, which we have extracted from the Wikipedia[4] website. We have manually evaluated performance of NER on our test corpus and found that performance was satisfying: we obtained a precision of 93.37% and a recall of 97.69%. Precision is the percentage of identified person names by the system that corresponds to correct person names, and recall is the percentage of person names in the text that have been correctly identified by the system.

The texts contain a small number of noun phrase coreferents that are in the form of pronouns, we have resolved these using the LingPipe[5] package.

3.2 Baseline system

We want to annotate an image using the associated text. We try to find the names of persons which are

[2] http://news.yahoo.com/

[3] http://opennlp.sourceforge.net/

[4] http://en.wikipedia.org/

[5] http://www.alias-i.com/lingpipe/

both described in the text *and* visible in the image, and we want to do so by relying *only* on an analysis of the text. In some cases, such as the following example, the text states explicitly whether a person is (not) visible in the image:

> President Bush [...] with Danish Prime Minister Anders Fogh Rasmussen, not pictured, at Camp David [...].

Developing a system that could extract this information is not trivial, and even if we could do so, only a very small percentage of the texts in our corpus contain this kind of information. In the next section we will look into a method that is applicable to a wide range of (descriptive) texts and that does not rely on specific information within the text.

To evaluate the performance of this system, we will compare it with a simple baseline system. The baseline system assumes that all persons in the text are visible in the image, which results in a precision of 71.27% and a recall of 95.56%. The (low) precision can be explained by the fact that the texts often discuss people which are not present in the image.

4 Detection of the salience of a person

Not all persons discussed in a text are equally important. We would like to discover what persons are in the focus of a text and what persons are only mentioned briefly, because we presume that more important persons in the text have a larger probability of appearing in the image than less important persons. Because of the short lengths of the documents in our corpus, an analysis of lexical cohesion between terms in the text will not be sufficient for distinguishing between important and less important entities. We define a measure, *salience*, which is a number between 0 and 1 that represents the importance of an entity in a text. We present here a method for computing this score based on an in depth analysis of the discourse of the text and of the syntactic structure of the individual sentences.

4.1 Discourse segmentation

The discourse segmentation module, which we developed in earlier research, hierarchically and sequentially segments the discourse in different topics and subtopics resulting in a table of contents of a text (Moens, 2006). The table shows the main entities and the related subtopic entities in a tree-like structure that also indicates the segments (by means of character pointers) to which an entity applies. The algorithm detects patterns of thematic progression in texts and can thus recognize the main topic of a sentence (i.e., about whom or what the sentence speaks) and the hierarchical and sequential relationships between individual topics. A mixture model, taking into account different discourse features, is trained with the Expectation Maximization algorithm on an annotated DUC-2003 corpus. We use the resulting discourse segmentation to define the salience of individual entities that are recognized as topics of a sentence. We compute for each noun entity e_r in the discourse its salience ($Sal1$) in the discourse tree, which is proportional with the depth of the entity in the discourse tree -hereby assuming that deeper in this tree more detailed topics of a text are described- and normalize this value to be between zero and one. When an entity occurs in different subtrees, its maximum score is chosen.

4.2 Refinement with sentence parse information

Because not all entities of the text are captured in the discourse tree, we implement an additional refinement of the computation of the salience of an entity which is inspired by (Moens et al., 2006). The segmentation module already determines the main topic of a sentence. Since the syntactic structure is often indicative of the information distribution in a sentence, we can determine the relative importance of the other entities in a sentence by relying on the relationships between entities as signaled by the parse tree. When determining the salience of an entity, we take into account the level of the entity mention in the parse tree ($Sal2$), and the number of children for the entity in this structure ($Sal3$), where the normalized score is respectively inversely proportional with the depth of the parse tree where the entity occurs, and proportional with the number of children.

We combine the three salience values ($Sal1$, $Sal2$ and $Sal3$) by using a linear weighting. We have experimentally determined reasonable coefficients for these three values, which are respectively 0.8, 0.1 and 0.1. Eventually, we could learn these coefficients from a training corpus (e.g., with the

	Precision	Recall	F-measure
NER	71.27%	95.56%	81.65%
NER+DYN	97.66%	92.59%	95.06%

Table 1: Comparison of methods to predict what persons described in the text will appear in the image, using Named Entity Recognition (NER), and the salience measure with dynamic cut-off (DYN).

Expectation Maximization algorithm).

We do not separately evaluate our technology for salience detection as this technology was already extensively evaluated in the past (Moens, 2006).

4.3 Evaluating the improved system

The salience measure defines a ranking of all the persons in a text. We will use this ranking to improve our baseline system. We assume that it is possible to automatically determine the number of faces that are recognized in the image, which gives us an indication of a suitable cut-off value. This approach is reasonable since face detection (determine whether a face is present in the image) is significant easier than face recognition (determine which person is present in the image). In the improved model we assume that persons which are ranked higher than, or equal to, the cut-off value appear in the image. For example, if 4 faces appear in the image, we assume that only the 4 persons of which the names in the text have been assigned the highest salience appear in the image. We see from table 1 that the precision (97.66%) has improved drastically, while the recall remained high (92.59%). This confirms the hypothesis that determining the focus of a text helps in determining the persons that appear in the image.

5 Automatically annotating persons and objects

After having developed a reasonable successful system to detect what persons will appear in the image, we turn to a more difficult case : Detecting persons *and* all other objects that are described in the text.

5.1 Entity detection

We will first detect what words in the text refer to an entity. For this, we perform part-of-speech tagging (i.e., detecting the syntactic word class such as noun, verb, etc.). We take that every noun in the text represents an entity. We have used LTPOS (Mikheev, 1997), which performed the task almost errorless (precision of 98.144% and recall of 97.36% on the nouns in the test corpus). Person names which were segmented using the NER package are also marked as entities.

5.2 Baseline system

We want to detect the objects and the names of persons which are both visible in the image and described in the text. We start with a simple baseline system, in which we assume that every entity in the text appears in the image. As can be expected, this results in a high recall (91.08%), and a very low precision (15.62%). We see that the problem here is far more difficult compared to detecting only person names. This can be explained by the fact that many entities (such as for example *August*, *idea* and *history*) will never (or only indirectly) appear in an image. In the next section we will try to determine what types of entities are more likely to appear in the image.

6 Detection of the visualness of an entity

The assumption that every entity in the text appears in the image is rather crude. We will enrich our model with external world knowledge to find entities which are not likely to appear in an image. We define a measure called *visualness*, which is defined as the extent to which an entity can be perceived visually.

6.1 Entity classification

After we have performed entity detection, we want to classify every entity according to a certain semantic database. We use the WordNet (Fellbaum, 1998) database, which organizes English nouns, verbs, adjectives and adverbs in synsets. A synset is a collection of words that have a close meaning and that represent an underlying concept. An example of such a synset is “person, individual, someone, somebody, mortal, soul”. All these words refer to a human being. In order to correctly assign a noun in a text to its synset, i.e., to disambiguate the sense of this word, we use an efficient Word Sense Disambiguation (WSD) system that was developed by the authors and which is described in (Deschacht and Moens, 2006). Proper names are labeled by the Named Entity Recognizer, which recognizes persons, locations and organizations. These labels in turn allow us to assign the corresponding WordNet synset.

The combination of the WSD system and the NER package achieved a 75.97% accuracy in classifying the entities. Apart from errors that resulted from erroneous entity detection (32.32%), errors were mainly due to the WSD system (60.56%) and in a smaller amount to the NER package (8.12%).

6.2 WordNet similarity

We determine the visualness for every synset using a method that was inspired by Kamps and Marx (2002). Kamps and Marx use a distance measure defined on the adjectives of the WordNet database together with two seed adjectives to determine the emotive or affective meaning of any given adjective. They compute the relative distance of the adjective to the seed synsets "good" and "bad" and use this distance to define a measure of affective meaning.

We take a similar approach to determine the visualness of a given synset. We first define a similarity measure between synsets in the WordNet database. Then we select a set of seed synsets, i.e. synsets with a predefined visualness, and use the similarity of a given synset to the seed synsets to determine the visualness.

6.3 Distance measure

The WordNet database defines different relations between its synsets. An important relation for nouns is the hypernym/hyponym relation. A noun X is a hypernym of a noun Y if Y is a subtype or instance of X. For example, "bird" is a hypernym of "penguin" (and "penguin" is a hyponym of "bird"). A synset in WordNet can have one or more hypernyms. This relation organizes the synsets in a hierarchical tree (Hayes, 1999).

The similarity measure defined by Lin (1998) uses the hypernym/hyponym relation to compute a semantic similarity between two WordNet synsets S_1 and S_2. First it finds the most specific (lowest in the tree) synset S_p that is a parent of both S_1 and S_2. Then it computes the similarity of S_1 and S_2 as

$$sim(S_1, S_2) = \frac{2logP(S_p)}{logP(S_1) + logP(S_2)}$$

Here the probability $P(S_i)$ is the probability of labeling any word in a text with synset S_i or with one of the descendants of S_i in the WordNet hierarchy. We estimate these probabilities by counting the number of occurrences of a synset in the Semcor corpus (Fellbaum, 1998; Landes et al., 1998), where all noun chunks are labeled with their WordNet synset. The probability $P(S_i)$ is computed as

$$P(S_i) = \frac{C(S_i)}{\sum_{n=1}^{N} C(S_n)} + \sum_{k=1}^{K} P(S_k)$$

where $C(S_i)$ is the number of occurrences of S_i, N is the total number of synsets in WordNet and K is the number of children of S_i. The WordNet::Similarity package (Pedersen et al., 2004) implements this distance measure and was used by the authors.

6.4 Seed synsets

We have manually selected 25 seed synsets in WordNet, where we tried to cover the wide range of topics we were likely to encounter in the test corpus. We have set the visualness of these seed synsets to either 1 (visual) or 0 (not visual). We determine the visualness of all other synsets using these seed synsets. A synset that is close to a visual seed synset gets a high visualness and vice versa. We choose a linear weighting:

$$vis(s) = \sum_i vis(s_i) \frac{sim(s, s_i)}{C(s)}$$

where $vis(s)$ returns a number between 0 and 1 denoting the visualness of a synset s, s_i are the seed synsets, $sim(s,t)$ returns a number between 0 and 1 denoting the similarity between synsets s and t and $C(s)$ is constant given a synset s:

$$C(s) = \sum_i sim(s, s_i)$$

6.5 Evaluation of the visualness computation

To determine the visualness, we first assign the correct WordNet synset to every entity, after which we compute a visualness score for these synsets. Since these scores are floating point numbers, they are hard to evaluate manually. During evaluation, we make the simplifying assumption that all entities with a visualness below a certain threshold are not visual, and all entities above this threshold are visual. We choose this threshold to be 0.5. This results in an accuracy of 79.56%. Errors are mainly caused by erroneous entity detection and classification (63.10%) but also because of an incorrect assignment of the visualness (36.90%) by the method described above.

7 Creating an appearance model using salience and visualness

In the previous section we have created a method to calculate a visualness score for every entity, because we stated that removing the entities which can never be perceived visually will improve the performance of our baseline system. An experiment proves that this is exactly the case. If we assume that only the entities that have a visualness above a 0.5 threshold are visible and will appear in the image, we get a precision of 48.81% and a recall of 87.98%. We see from table 2 that this is already a significant improvement over the baseline system.

In section 4 we have seen that the salience measure helps in determining what persons are visible in the image. We have used the fact that face detection in images is relatively easily and can thus supply a cut-off value for the ranked person names. In the present state-of-the-art, we are not able to exploit a similar fact when detecting all types of entities. We will thus use the salience measure in a different way. We compute the salience of every entity, and we assume that only the entities with a salience score above a threshold of 0.5 will appear in the image. We see that this method drastically improves precision to 66.03%, but also lowers recall until 54.26%.

We now create a last model where we combine both the visualness and the salience measures. We want to calculate the probability of the occurrence of an entity e_{im} in the image, given a text t, $P(e_{im}|t)$. We assume that this probability is proportional with

	Precision	**Recall**	**F-measure**
Ent	15.62%	91.08%	26.66%
Ent+Vis	48.81%	87.98%	62.78%
Ent+Sal	66.03%	54.26%	59.56%
Ent+Vis+Sal	70.56%	67.82%	69.39%

Table 2: Comparison of methods to predict the entities that appear in the image, using entity detection (Ent), and the visualness (Vis) and salience (Sal) measures.

the degree of visualness and salience of e_{im} in t. In our framework, $P(e_{im}|t)$ is computed as the product of the salience of the entity e_{im} and its visualness score, as we assume both scores to be independent.

Again, for evaluation sake, we choose a threshold of 0.4 to transform this continuous ranking into a binary classification. This results in a precision of 70.56% and a recall of 67.82%. This model is the best of the 4 models for entity annotation which have been evaluated.

8 Related Research

Using text that accompanies the image for annotating images and for training image recognition is not new. The earliest work (only on person names) is by Satoh (1999) and this research can be considered as the closest to our work. The authors make a distinction between proper names, common nouns and other words, and detect entities based on a thesaurus list of persons, social groups and other words, thus exploiting already simple semantics. Also a rudimentary approach to discourse analysis is followed by taking into account the position of words in a text. The results were not satisfactory: 752 words were extracted from video as candidates for being in the accompanying images, but only 94 were correct where 658 were false positives. Mori et al. (2000) learn textual descriptions of images from surrounding texts. These authors filter nouns and adjectives from the surrounding texts when they occur above a certain frequency and obtain a maximum hit rate of top 3 words that is situated between 30% and 40%. Other approaches consider both the textual and image features when building a content model of the image. For instance, some content is selected from the text (such as person names) and from the image (such as faces) and both contribute in describing the content of a document. This approach was followed by Barnard (2003).

Westerveld (2000) combines image features and words from collateral text into one semantic space. This author uses Latent Semantic Indexing for representing the image/text pair content. Ayache et al. (2005) classify video data into different topical concepts. The results of these approaches are often disappointing. The methods here represent the text as a bag of words possibly augmented with a *tf* (term frequency) x *idf* (inverse document frequency) weight of the words (Amir et al., 2005). In exceptional cases, the hierarchical XML structure of a text document (which was manually annotated) is taken into account (Westerveld et al., 2005). The most interesting work here to mention is the work of Berg et al. (2004) who also process the nearly parallel image-text pairs found in the Yahoo! news corpus. They link faces in the image with names in the text (recognized with named entity recognition), but do not consider other objects. They consider pairs of person names (text) and faces (image) and use clustering with the Expectation Maximization algorithm to find all faces belonging to a certain person. In their model they consider the probability that an entity is pictured given the textual context (i.e., the part-of-speech tags immediately prior and after the name, the location of the name in the text and the distance to particular symbols such as "(R)"), which is learned with a probabilistic classifier in each step of the EM iteration. They obtained an accuracy of 84% on person face recognition.

In the CLASS project we work together with groups specialized in image recognition. In future work we will combine face and object recognition with text analysis techniques. We expect the recognition and disambiguation of faces to improve if many image-text pairs that treat the same person are used. On the other hand our approach is also valuable when there are few image-text pairs that picture a certain person or object. The approach of Berg et al. could be augmented with the typical features that we use, namely salience and visualness. In Deschacht et al. (2007) we have evaluated the ranking of persons and objects by the method we have described here and we have shown that this ranking correlates with the importance of persons and objects in the picture.

None of the above state-of-the-art approaches consider salience and visualness as discriminating factors in the entity recognition, although these aspects could advance the state-of-the-art.

9 Conclusion

Our society in the 21st century produces gigantic amounts of data, which are a mixture of different media. Our repositories contain texts interwoven with images, audio and video and we need automated ways to automatically index these data and to automatically find interrelationships between the various media contents. This is not an easy task. However, if we succeed in recognizing and aligning content in near-parallel image-text pairs, we might be able to use this acquired knowledge in indexing comparable image-text pairs (e.g., in video) by aligning content in these media.

In the experiment described above, we analyze the discourse and semantics of texts of near-parallel image-text pairs in order to compute the probability that an entity mentioned in the text is also present in the accompanying image. First, we have developed an approach for computing the salience of each entity mentioned in the text. Secondly, we have used the WordNet classification in order to detect the visualness of an entity, which is translated into a visualness probability. The combined salience and visualness provide a score that signals the probability that the entity is present in the accompanying image.

We extensively evaluated all the different modules of our system, pinpointing weak points that could be improved and exposing the potential of our work in cross-media exploitation of content.

We were able to detect the persons in the text that are also present in the image with a (evenly weighted) F-measure of more than 95%, and in addition were able to detect the entities that are present in the image with a F-measure of more than 69%. These results have been obtained by relying only on an analysis of the text and were substantially better than the baseline approach. Even if we can not resolve all ambiguity, keeping the most confident hypotheses generated by our textual hypotheses will greatly assist in analyzing images.

In the future we hope to extrinsically evaluate the proposed technologies, e.g., by testing whether the recognized content in the text, improves image recognition, retrieval of multimedia sources, mining of these sources, and cross-media retrieval. In addition, we will investigate how we can build more refined appearance models that incorporate attributes and actions of entities.

Acknowledgments

The work reported in this paper was supported by the EU-IST project CLASS (Cognitive-Level Annotation using Latent Statistical Structure, IST-027978). We acknowledge the CLASS consortium partners for their valuable comments and we are especially grateful to Yves Gufflet from the INRIA research team (Grenoble, France) for collecting the Yahoo! News dataset.

References

Arnon Amir, Janne Argillander, Murray Campbell, Alexander Haubold, Giridharan Iyengar, Shahram Ebadollahi, Feng Kang, Milind R. Naphade, Apostol Natsev, John R. Smith, Jelena Tešió, and Timo Volkmer. 2005. IBM Research TRECVID-2005 Video Retrieval System. In *Proceedings of TRECVID 2005*, Gaithersburg, MD.

Stéphane Ayache, Gearges M. Qunot, Jrme Gensel, and Shin'Ichi Satoh. 2005. CLIPS-LRS-NII Experiments at TRECVID 2005. In *Proceedings of TRECVID 2005*, Gaithersburg, MD.

Kobus Barnard, Pinar Duygulu, Nando de Freitas, David Forsyth, David Blei, and Michael I. Jordan. 2003. Matching Words and Pictures. *Journal of Machine Learning Research*, 3(6):1107–1135.

Tamara L. Berg, Alexander C. Berg, Jaety Edwards, and D.A. Forsyth. 2004. Who's in the Picture? In *Neural Information Processing Systems*, pages 137–144.

Koen Deschacht and Marie-Francine Moens. 2006. Efficient Hierarchical Entity Classification Using Conditional Random Fields. In *Proceedings of the 2nd Workshop on Ontology Learning and Population*, pages 33–40, Sydney, July.

Koen Deschacht, Marie-Francine Moens, and W Robeyns. 2007. Cross-media entity recognition in nearly parallel visual and textual documents. In *Proceedings of the 8th RIAO Conference on Large-Scale Semantic Access to Content (Text, Image, Video and Sound)*. Cmu. (in press).

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Brian Hayes. 1999. The Web of Words. *American Scientist*, 87(2):108–112, March-April.

Jaap Kamps and Maarten Marx. 2002. Words with Attitude. In *Proceedings of the 1st International Conference on Global WordNet*, pages 332–341, India.

Shari Landes, Claudia Leacock, and Randee I. Tengi. 1998. Building Semantic Concordances. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press.

Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. In *Proc. 15th International Conf. on Machine Learning*.

Andrei Mikheev. 1997. Automatic Rule Induction for Unknown-Word Guessing. *Computational Linguistics*, 23(3):405–423.

Marie-Francine Moens, Patrick Jeuniaux, Roxana Angheluta, and Rudradeb Mitra. 2006. Measuring Aboutness of an Entity in a Text. In *Proceedings of HLT-NAACL 2006 TextGraphs: Graph-based Algorithms for Natural Language Processing*, East Stroudsburg. ACL.

Marie-Francine Moens. 2006. Using Patterns of Thematic Progression for Building a Table of Content of a Text. *Journal of Natural Language Engineering*, 12(3):1–28.

Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka. 2000. Automatic Word Assignment to Images Based on Image Division and Vector Quantization. In *RIAO-2000 Content-Based Multimedia Information Access*, Paris, April 12-14.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. In *The Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04)*, Boston, May.

Shin'ichi Satoh, Yuichi Nakamura, and Takeo Kanade. 1999. Name-It: Naming and Detecting Faces in News Videos. *IEEE MultiMedia*, 6(1):22–35, January-March.

Thijs Westerveld, Jan C. van Gemert, Roberto Cornacchia, Djoerd Hiemstra, and Arjen de Vries. 2005. An Integrated Approach to Text and Image Retrieval. In *Proceedings of TRECVID 2005*, Gaithersburg, MD.

Thijs Westerveld. 2000. Image Retrieval: Content versus Context. In *Content-Based Multimedia Information Access, RIAO 2000 Conference Proceedings*, pages 276–284, April.