In press : Proceedings of the 8th RIAO Conference on Large-Scale Semantic Access to Content,
Pittsburgh, May 30 to June 1, 2007

Cross-Media Entity Recognition in Nearly Parallel Visual and Textual Documents

Koen Deschacht, Marie-Francine Moens and Wouter Robeyns
Katholieke Universiteit Leuven - Legal Informatics and Information Retrieval
Tiensestraat 41, 3000 Leuven, Belgium
{koen.deschacht,marie-francine.moens}@law.kuleuven.be
wouter.robeyns@student.kuleuven.be

Abstract

We present a novel approach to automatically annotate images solely 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 900 image-text pairs of Yahoo! News.

1 Introduction

Our society deals with a growing bulk of unstructured, electronic information such as text, images and video, a situation witnessed in many domains (news, biomedical information, intelligence information, business documents, etc.). This growth creates an increasing demand for more effective tools to search the information and to summarize the results. 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 unstructured information would largely contribute to improved indexing, classification, filtering and interrogation. 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. Although images and associated texts never contain precisely the same information, in many situations texts offer valuable information that helps to interpret the image. Currently, object recognition in images is still a difficult task to perform and this has caused an increasing interest in using the accompanying textual descriptions as a weak annotation of image content.

Images (and video) can be accompanied by a variety of texts, such as captions, surrounding text on a web page or video transcripts. The content of these texts can match the content of the image very closely or only quite loosely. Often a kernel of parallel content is present, but the texts often contain other, complementary content which is not present in the image and vice versa.

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

In current approaches, the text that accompanies an image is often seen as a bag of words, ignoring that the text's discourse structure and semantics allow for a more fine-grained identification of what content might be present in the image. In this paper we test the feasibility of automatically annotating images by using textual information in nearly parallel image-text pairs, in which most of the content of the image corresponds to content of the text and vice versa. Recognized parallel patterns can then be used to better align content in comparable corpora or to annotate images even in the absence of text.

In this paper we focus on the relatively short texts accompanying images of news events. We will create an appearance model representing the content in the text that describe content in the image. Currently we only extract entities, i.e. persons and objects. Section 2 describes how we extract the entities from the text. The texts often describe many entities, of which only a fraction is present in the image. We want to be able to predict what entities are likely to appear in the image and what entities are not. Here for, we define two measures, *salience* and *visualness*. In section 3 we will define and compute the salience measure and in section 4 we define and compute the visualness. We will then combine these two measures in section 5 to compute the probability of an entity beeing present in the image. It is important to note that in none of these steps we perform an analysis of the image. We solely rely on the syntactic and semantic structure of the texts. This does not exclude that our methods can be used in combination with image recognition techniques. We have succesfully tested our technologies on 900 image-text pairs in section 6.

2 Entity detection

The first step in building our appearance model of the text is entity detection and classification. Here for we employ natural language processing, more specifically part-of-speech tagging, after which we classify the entities according to their WordNet synset or their proper name category (named entity recognition).

2.1 Natural Language Processing of the texts

We restrict the natural language processing of the texts to part-of-speech tagging (i.e., detecting the syntactic word class such as noun, verb, etc.) and sentence parsing (i.e., the detection of the dependency tree of a sentence). We have used LTPOS (Mikheev 1997) and the Charniak (2000) parser respectively. Part-of-speech tagging allows identifying nouns, which is needed for the recognition of entities. The parse tree of a sentence will contribute in the salience detection of an entity (see below). In order to more accurately detect the entities and their salience, the tool Lingpipe resolved the noun phrase coreferents that are in the form of pronouns[2]. Two entities are considered as coreferents when they both refer to the same noun phrase in the situation described by the text (e.g., in the sentences: "Dan Quayle met his wife in college. He married her shortly after he finished his studies", "his" and "he" corefer to "Dan Quayle", "her" corefers to "wife").

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

2.2 Entity classification

After determining the part-of-speech of every word in a sentence, and thus detecting entities, we want to classify every entity according to a certain ontology. We employ the WordNet (Fellbaum 1998) lexical database. This database 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. Usually a single word can be assigned to multiple synsets, each representing a different meaning of that word. For instance, the word “nail” is present in 3 synsets, representing the concepts “fingernail”, “piece of metal used in construction” and “former unit of length”. In order to correctly assign a noun in a text to its synset, i.e., to disambiguate the sense of this word, we use a classifier that was developed by the authors and which is described in (Deschacht and Moens 2006). However, this does not offer a satisfactory solution for proper names, since the amount of proper names is possible indefinite. To tag proper names we use a Named Entity Recognizer of Lingpipe. The Lingpipe package recognizes persons, locations and organizations. These labels allow us to assign the corresponding WordNet synset.

3 Detection of the Salience of an Entity

The salience of an entity gives a measure of the importance of this entity in the text. Typically a *tf* (term frequency) x *idf* (inverse document frequency) is computed for the terms that represent an entity. For certain types of tasks this calculation yields acceptable results. For short texts, as the ones we use in our experiments below, the majority of the entities are only mentioned once, making it impossible to use this measure. To reliable discriminate the salience of entities, we resort to an in depth salience analysis of the discourse and sentences. We present here a method for computing the salience score of an entity based on the analysis of the discourse and an analysis of the individual sentences.

3.1 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 the 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 the 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.

3.2 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. Compared to the main topic of a sentence, we can determine the relative importance of the other entities in a sentence 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 (also see section 6) 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.

4 Detection of the Visualness of an Entity

In sections 2 and 3 we have created a appearance model of the text. This appearance model represents the entities in the text together with an importance score for every entity. When trying to predict what entities are visible in the image, we experienced that some entities are far more probably to appear than others. Take for example the entity "thought". This entity will never (or only indirectly) appear in the image. To capture this kind of information, we define the measure *visualness*, which is defined as the extent to which an entity can be perceived visually. To determine this visualness, we rely on the external resource WordNet. The computation of the visualness for a given synset is independent of the text that synset is used in.

4.1 WordNet similarity

We determine the visualness for every synset in a text 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.

4.2 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 occurences 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

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

where $C(S_i)$ is the number of occurences 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.

4.3 Seed synsets

We have manually selected 25 seed synsets in WordNet, trying to cover the topics we were likely to encounter in the test corpus (see section 6). We have selected the visual (*person, vehicle, animal, house, organism, desk, coat, book, weapon, body, physical entity*) and the not visual seed synsets (*power, danger, thought, knowledge, air, sleep, test, country, organization, location, picture, contest, abstract entity, thing*) by hand and set their visualness 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:

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

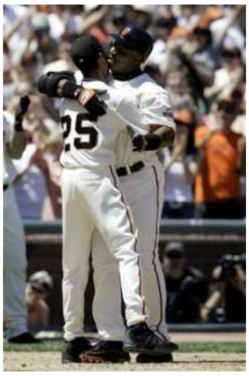
5 Cross-media entity recognition

We align content (e.g., persons, objects) found in a text t with the accompanying image based on the assumption that content of a text that can be visualized is present in the image, and that the probability of the occurrence of an entity e_{im} in the image, given a text t, $P(e_{im}|t)$, is proportional with 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. $P(e_{im}|t)$ defines a ranking of the text's entities.

6 Experiments and results

We carry out experiments on a data set of images and accompanying texts, retrieved from the Yahoo! News website[3]. Every image has an accompanying text which describes the content of the

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



San Francisco Giants' Barry Bonds hugs his son Nikolai after hitting a two-run home run off of Colorado Rockies' Byung-Hyun Kim, of South Korea, in the fourth inning of their baseball game in San Francisco, Sunday, May 28, 2006. It was Bonds' career home run number 715, surpassing Babe Ruth on the all time home run list.

Figure 1: Example image-text pair

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. Not all persons or objects who are pictured in a photograph are necessarily described in the text. The inverse is also true, i.e. content mentioned in the text may not be present in the image. On average the texts have a length of 40.98 words, containing 21.10 words that refer to entities of which 4.29 refer to entities that are present in the image.

An example of an image-text pair is given in fig. 1. Here the entities "Barry Bonds" and "Nikolai" are discussed in the text and appear on the image. The entities "San Francisco", "Giants", "home run", "Colorado Rockies", "Byung-Hyun Kim", "inning", "baseball game", "Sunday", "May", "career", "number","Babe Ruth" and "list" are discussed in the text but are not pictured in the image.

In the framework of the CLASS project we evaluate several aspects of our appearance model of the text. First we evaluate the visualness score that we assign to proper and common names. Secondly, we evaluate the ranking of the entities in the text with the entities in the image. Except for the first evaluation where we separately evaluate visualness, all other performance measures evaluate the integrated technologies for salience and visualness detection. We do not separately evaluate our technology for salience detection as this technology was already extensively evaluated in the past (Moens et al. 2005; Moens 2006). We do not evaluate part-of-speech tagging and sentence parsing on this corpus since we assume that the accuracy for these tasks is similar to values reported in literature (Mikheev 1997; Charniak 2000).

6.1 Evaluating visualness

To evaluate the visualness measure, one human annotator has classified all nouns 30 texts of the Yahoo! News corpus as either visual (220) or non-visual (377). A noun is considered visual if it represents an entity which can be perceived visually. Note that for this experiment it is not required that the entity does effectively appear in the image (as opposed to the experiments in section 6.2). Since our algorithm for determining the visualness returns a floating number instead of a binary classification, we use the following evaluation measure for comparing the machine classification with the manual classification:

$$\text{Eval1} = \frac{1}{N} \sum_{r} (1 - |VIS(e_r) - vis(e_r)|) \tag{1}$$

	Eval1
Baseline	36.03%
Visualness algorithm	89.37%

Table 1: Evaluation of visualness on 597 noun phrases using *Eval1* (eq. 1).

Here N is the number of entities e_r in the 30 annotated texts, $vis(e_r)$ is a number between 0 and 1 which denotes the visualness of the entity s_r as determined by our algorithm, and $VIS(e_r)$ is either 0 or 1, denoting the visualness as determined by the annotator. We compare our algorithm with a baseline approach where all nouns are considered visible, consistent with the common bag-of-words approach where all nouns in a text annotate the image. The results are shown in table 1. We see that our algorithm performs significantly better than the baseline approach.

6.2 Evaluating alignment

To test the ranking generated by our system we have annotated 900 image-text pairs of the Yahoo! News dataset. For every text-image pair one human annotator has selected the entities that appear both in the text and in the image (3430 entities) and sorted these based on their perceived importance in the image. If two or more entities were considered equally important in the image by the annotator, they were ranked on the same position of salience. We call the resulting ranking the expert ranking E. For example, in fig. 1 two entities are present in the image, "Barry Bonds" and "Nikolai" where both have equal importance. The entities "Byung-Hyun Kim" and "Babe Ruth" (and others) are present in the text but are not present in the image and are thus not annotated. The ranking that was automatically generated from the text and ranked according to the probability of presence in the image is called the machine ranking or M.

For our first evaluation we take a simple approach where all entities above a certain cut-off position in the machine list M are considered to appear in the image and all the entities below this cut-off position are considered not to appear in the image. We compute the average recall, precision and accuracy at different cut-off positions in M. Recall is the percentage of annotated entities in the image that have been correctly predicted to appear in the image by our algorithm, precision gives the percentage of predicted entities that are actually present in the image, and accuracy computes the percentage of correctly classified entities as being present or not present in the image. We refer to this evaluation as the *static cut-off*, of which the results are shown in fig. 2.

We see that precision is high using a small cut-off value, and drops when the cut-off value increases. This confirms the hypothesis that entities with a high combined salience and visualness have a large probability of appearing in the image. As we increase the cut-off value, entities with a low combined visualness and salience are also included in the appearance model. These entities are less likely to appear in the image, hence the decrease of precision. Analogically, the accuracy drops with a growing cut-off value as more entities are erroneously predicted to appear in the image. At cut-off position 15 all entities in all texts are predicted to appear in the corresponding image and recall is 97.26% (a small amount of entities were not detected due to errors in part-of-speech tagging) and precision is 22.96%, which is equal to the ratio of entities in the text that are present in the image.

We also consider a different approach where the cut-off position is dynamically determined per image-text pair by the number of objects annotated for that particular image. This statistic is valuable because we assume that it is possible to automatically determine the number of (salient) entities that are recognized in the image, which gives us an indication of a suitable cut-off value in

	Accuracy
Baseline	79.20%
Combined visualness and salience	89.56%

Table 2: Average accuracy over 900 image-text pairs using a dynamic cut-off value in the machine ranking.

M. For example, when trying to recognize faces (i.e. to detect a face and then assign the correct name to it) one could use a face detector which detects the number of faces present in the image, which can already be performed with an accuracy of above 90% for frontal faces (Viola and Jones 2001). Given the number of faces, we can then extract the most probable names which belong to these faces from the text.

We have compared our algorithm with a baseline approach where all nouns are considered to be entities which are present in the image, and are ranked according to occurence position in the text (i.e. the first entity in the text is given the highest ranking etc.). The results of these evaluations are given in table 2. We see here that the performance of the baseline approach is comparable to the performance of our algorithm, although our algorithm achieves statistically significantly better results. The reason for this is that the short texts in our corpus very often describe the most important entities first, then the second most important entities etc., which is the reason that this simple heuristic already achieves good results.

A third evaluation of the alignment takes into account the machine ranking and will especially focus on the suitability of this ranking and the accurateness of the probability scores that combine salience and visualness. First, we perform a very strict evaluation of the ranking of the entities:

$$score(e_r) = \begin{cases} 1 & \text{iff } p_E(e_r) == p_M(e_r) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where e_r is the text entity to be judged, $p_E(e_r)$ is the position of this entity in the expert ranking E and $p_M(e_r)$ is the position of this entity in the machine ranking M. Here, a score of one represents that the relevant entity is found in the text in the same ranking position as classified based on the

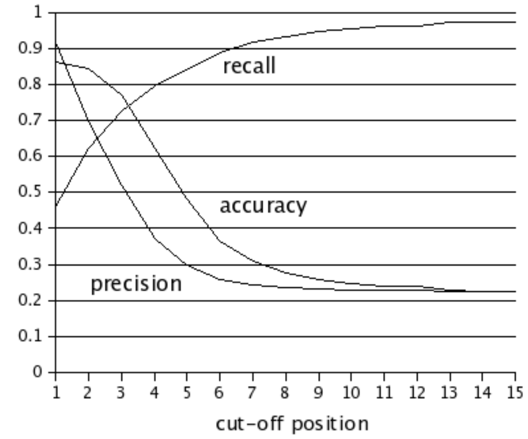


Figure 2: Average recall, precision and accuracy over 900 image-text pairs using different static cut-off values in the machine ranking.

	Macro		Micro	
	Strict	**Penalty**	**Strict**	**Penalty**
Baseline	3.12%	58.30%	2.49%	56.90%
Combined visualness and salience	69.41%	86.82%	67.56%	85.92%

Table 3: Macro (eq. 4) and micro (eq. 5) average accuracy of the classification of text entities according to image salience taken over 900 image-text pairs, using a strict (eq. 2) and penalty-based (eq. 3) evaluation measure.

image content. In all other cases the score is zero. For each image we have a number of visual classes or positions in E, augmented by one non-visual class, the 'not present in image'-category, which is always ranked last in E. Our evaluation determines whether these classes are correctly attributed to the text entities.

We also want to take into account the differences in ranking or differences in probability scores in the evaluation because not all errors count as equally important. Here for, we use a penalty-based score where an entity which has received a wrong position, receives a penalty which is proportional with the distance from the expert position:

$$score(e_r) = 1 - \frac{|p_E(e_r) - p_M(e_r)|}{P} \tag{3}$$

where P is the number of positions in the expert list E. Note that the class "not present in image" occupies the last position in E. We could also work with absolute differences in probabilities, if sensible probabilities are given by the experts. Note that in all the above schemes, several entities can occupy the same position in the ranking, both for E and M.

We now compute an average of these scores for all texts. We do this in two ways, using micro and macro averages. For the macro average, we evaluate every text separately, after which we average the evaluations of all texts. This weights all texts equally, regardless of the number of entities which are present in the text.

$$Macro = \frac{1}{T}\sum_t \frac{1}{N(t)}\sum_r score(e_r) \tag{4}$$

Here T is the number of texts, $N(t)$ is the number of entities in text t and r ranges over all entities in text t. We also compute the micro average. This evaluation weights every (error of) entity ranking equally, regardless of what text that entity belongs to.

$$Micro = \frac{1}{N}\sum_t \sum_r score(e_r) \tag{5}$$

Here N is the total sum of all entities of all texts.

We have computed both the micro and macro average for the strict (eq. 2) and the penalty-based (eq. 3) evaluation measure. The results are shown in table 3. We compare our approach with a baseline approach where all entities in a text are considered visible and where the entities are ranked according to the position in the text. We see that the baseline performs very badly for the strict classification. Although entities in first positions of the short texts give already a reasonable indication of which entities are present in the image, the baseline cannot correctly discriminate these entities according to salience. Also for the more lenient evaluation where the penalty of misclassification is proportional with the distance in ranking position, we see that our algorithm quite correctly ranks the entities according to their importance in the image. Although, an entity might not be classified in the correct position, the distance to its correct position is small. As in the foregoing evaluation, our algorithm performs significantly better than the baseline approach.

7 Related Research

Using text that accompanies the image for annotating images and for training image recognition is not new. The earliest work that only considers person names is by Satoh et al. (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 alams. 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 an appearance 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 et al. (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). (Berg et al. 2004) have also processed the nearly parallel image-text pairs found in the Yahoo! news corpus. They perform an analysis of both the text (NER recognition) and the image (face detection) to link faces in the image with names in the text. They do not consider other objects. By considering all possible pairs of person names (text) and faces (image) and using clustering with the expectation maximization algorithm to find the faces belonging to the same 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 iteration. They obtained an accuracy of 84% on person face recognition, thus confirming our findings about the importance of the texts that are associated with an image.

8 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 nearly 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 nearly 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 gives the probability that the entity is present in the accompanying image. The suitability of our approach was succesfully tested on 900 image-text pairs of the Yahoo! News collection. We were able to detect the persons and objects in the text that are also present in the image with an accuracy of more than 89%, where the cut-off position is determined by the number of persons and objects in the image. In addition, the accuracy of the ranking according to salience and visualness approaches 86%. These values were substantially better than a baseline approach that considers all nouns of the text and ranks them according to their position in the text. Even if we cannot resolve all ambiguity, keeping the most confident hypotheses generated by our textual hypotheses will greatly assist in indexing the image sources.

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. We will therefor work together with our partners in the CLASS project, which are specialized in image recognition. We want to combine evidence from many images and accompanying texts to improve content recognition and disambiguation in both media. We expect the results to improve if we consider many good correlations in a large data set of image-text pairs. Also, the approach of Berg et al. can be augmented with our features, namely salience and visualness. On the other hand our approach is also valuable when there are few image-text pairs that picture a certain person or object. In addition, we will investigate how we can build more refined appearance models that incorporate attributes and actions of entities.

Acknowledgements

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 data set.

References

A. Amir, J. Argillander, M. Campbell, A. Haubold, G. Iyengar, S. Ebadollahi, F. Kang, M. R. Naphade, A. Natsev, J. R. Smith, J. Tešió, and T. Volkmer. IBM Research TRECVID-2005 Video Retrieval System. In *Proceedings of TRECVID 2005*, Gaithersburg, MD, 2005.

S. Ayache, G. M. Qunot, J. Gensel, and S. Satoh. CLIPS-LRS-NII Experiments at TRECVID 2005. In *Proceedings of TRECVID 2005*, Gaithersburg, MD, 2005.

K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D.M. Blei, and M.I. Jordan. Matching Words and Pictures. *Journal of Machine Learning Research*, 3(6):1107–1135, 2003.

T.L. Berg, A.C. Berg, J. Edwards, and D.A. Forsyth. Who's in the Picture? In *Proceedings of the 18th Annual Conference on Neural Information Processing Systems*, pages 137–144, 2004.

E. Charniak. A Maximum-Entropy-Inspired Parser. In *Proceedings of the First Conference on North American chapter of the Association for Computational Linguistics*, pages 132–139. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2000.

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

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

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

J. Kamps and M. Marx. Words with Attitude. In *Proceedings of the 1st International Conference on Global WordNet*, pages 332–341, Mysore, IN, 2002.

S. Landes, C. Leacock, and R.I. Tengi. Building Semantic Concordances. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database.* The MIT Press, 1998.

D. Lin. An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th International Conf. on Machine Learning*, 1998.

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

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

M.-F. Moens, R. Angheluta, and J. Dumortier. Generic Technologies for Single- and Multi-Document Summarization. *Information Processing and Management*, 41(3):569–586, 2005.

M.-F. Moens, P. Jeuniaux, R. Angheluta, and R. Mitra. Measuring Aboutness of an Entity in a Text. In *Proceedings of HLT-NAACL 2006 TextGraphs: Graph-based Algorithms for Natural Language Processing*, East Stroudsburg, 2006. ACL.

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

T. Pedersen, S. Patwardhan, and J. Michelizzi. 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 2004.

S. Satoh, Y. Nakamura, and T. Kanade. Name-It: Naming and Detecting Faces in News Videos. *IEEE MultiMedia*, 6(1):22–35, January-March 1999.

P. Viola and M. Jones. Rapid Object Detection Using a Boosted Cascade of Simple Features. *Proc. CVPR*, 1:511–518, 2001.

T. Westerveld. Image Retrieval: Content versus Context. In *Proceedings of the RIAO 2000 conference : Content-Based Multimedia Information Access*, pages 276–284, April 2000. ISBN 2-905450-07-X.

T. Westerveld, J.C. van Gemert, R. Cornacchia, D. Hiemstra, and A. de Vries. An Integrated Approach to Text and Image Retrieval. In *Proceedings of TRECVID 2005*, Gaithersburg, MD, 2005.