SEMANTIC KNOWLEDGE CONSTRUCTION FROM ANNOTATED IMAGE COLLECTIONS

Ana B. Benitez and Shih-Fu Chang

Dept. of Electrical Engineering, Columbia University, New York, NY 10027
{ana, sfchang} @ ee.columbia.edu

ABSTRACT

This paper presents new methods for extracting semantic knowledge from collections of annotated images. The proposed methods include novel automatic techniques for extracting semantic concepts by disambiguating the senses of words in the annotations using the lexical database WordNet, and both the images and their annotations, and for discovering semantic relations among the detected concepts based on WordNet. Another contribution of this paper is the evaluation of several techniques for visual feature descriptor extraction and data clustering in the extraction of semantic concepts. Experiments show the potential of integrating the analysis of both images and annotations for improving the performance of the word-sense disambiguation process. In particular, the accuracy improves 4-15% with respect to the baselines systems for *nature* images.

1. INTRODUCTION

The important proliferation of digital multimedia content requires tools for extracting useful knowledge from the content to enable intelligent and efficient multimedia organization, filtering and retrieval. Knowledge is usually defined as facts about the world and is often represented as concepts and relationships among the concepts, i.e., semantic networks. Concepts are abstractions of objects, situations, or events in the world (e.g., color pattern and "car"); relationships represent interactions among concepts (e.g., color pattern 1 *visually similar to* color pattern 2 and "sedan" *specialization* of "car").

This paper focuses on the extraction of knowledge representing semantic information about the world depicted by, related to, or symbolized by an annotated image collection (e.g., concepts "animal" is a *generalization* of concept "human"). Semantic knowledge is the most powerful knowledge for intelligent multimedia applications because human communication often happens at this level. However, the construction of semantic knowledge is an open problem; current approaches are, at best, semi-automatic and very time consuming. As many images often have some text directly or indirectly describing their content (e.g., caption or web page of an image, respectively), both text and images can be used and integrated into the semantic knowledge extraction process.

Prior work on semantic knowledge construction includes word-sense disambiguation techniques for text documents [9][12][13]. Words in English may have more than one sense or meaning, for example "plant, industrial plant" and "plant, living organism" for the word “plant”. Word-sense disambiguation (WSD) is the process of finding the correct sense of a word within a document, which is a long-standing problem in Natural Language Processing. Although most English words have only one sense (80%), most words in text documents have more than one sense (80%) [12]. The two principles governing most word-sense disambiguation techniques are (1) that nearby words are semantically close or related and (2) that the sense of a word is often the same within a document [13]. In literature, there are unsupervised [9][13] and supervised [12] approaches that often use WordNet as the electronic word-sense lexicon. WordNet organizes English words into sets of synonyms (e.g., "rock, stone") and connects them with semantic relations (e.g., generalization) [10]. There are also image indexing approaches that try to disambiguate the senses of words in image annotations [1][11]. However, none of these approaches combine textual and image features during word-sense disambiguation. [1] integrates text and visual features only in the hierarchical image clustering.

This paper presents and evaluates new methods for automatically constructing semantic knowledge from annotated image collections including semantic concepts and relationships. The proposed approach for extracting semantic concepts consists of disambiguating the senses of words in image annotations using WordNet, and, in contrast with prior work, using both the images and the annotations. Semantic relationships are discovered among concepts based on relationships among word senses in WordNet. The input to this process is not only the annotated image collection but also perceptual knowledge extracted from the collection as described in [2]. The perceptual knowledge consists of a set of clusters grouping images based on visual and/or text feature descriptors, and relationships among the clusters. This paper evaluates several techniques for visual feature descriptor extraction and data clustering in the extraction of semantic concepts. In particular, for *nature* images, the accuracy improves 4-15% with respect to the baselines systems, i.e., text-based disambiguation, most frequent sense and random sense. For *news* images, the improvement is of 6-18% only with respect to text-based disambiguation and random sense.

These methods are developed and used within the IMKA system [3]. IMKA stands for "Intelligent Multimedia Knowledge Application". The objectives of the IMKA project are to develop methods for extracting knowledge from multimedia content and implementing intelligent applications that use that knowledge [3]. The multimedia knowledge is encoded using MediaNet, a knowledge representation framework that uses multimedia to represent both perceptual and semantic information about the world in terms of concepts and relationships among the concepts [4]. Methods for constructing perceptual knowledge from annotated image collections are presented in [2].

2. SEMANTIC KNOWLEDGE EXTRACTION

The proposed approach for extracting semantic knowledge from an collection of annotated images, which has already been clustered based on text and/or visual feature descriptors as described in [2], consists of three steps, as shown in Figure 1: (1) the basic processing of tokenizing and chunking the textual annotations and tagging the words with their Part-Of-Speech (POS, e.g., "noun" and "verb"); (2) the extraction of semantic concepts by disambiguating the senses of the content words using WordNet and the image clusters; and (3) the discovery of relations and additional concepts from WordNet to relate the detected word senses.

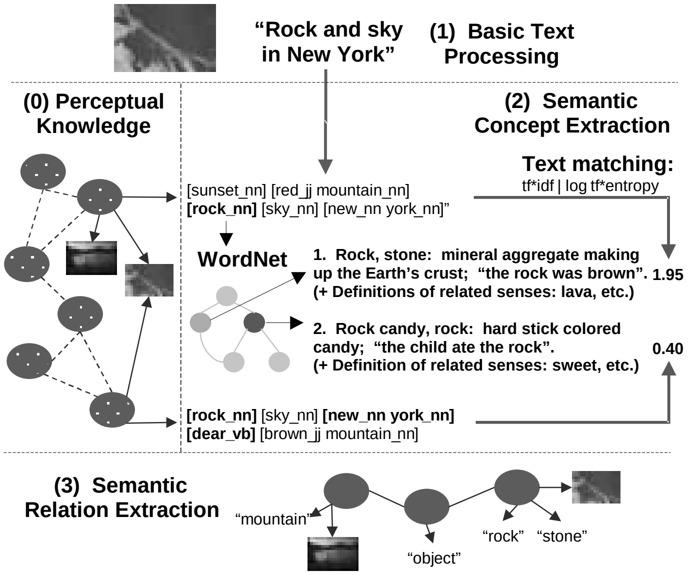

“Rock and sky in New York”
(1) Basic Text Processing
(0) Perceptual Knowledge
(2) Semantic Concept Extraction
Text matching:
tf*idf | log tf*entropy
[sunset_nn] [red_jj mountain_nn]
[rock_nn] [sky_nn] [new_nn york_nn]”
WordNet
1. Rock, stone: mineral aggregate making up the Earth’s crust; “the rock was brown”.
(+ Definitions of related senses: lava, etc.)
1.95
2. Rock candy, rock: hard stick colored candy; “the child ate the rock”.
(+ Definition of related senses: sweet, etc.)
0.40
[rock_nn] [sky_nn] [new_nn york_nn]
[dear_vb] [brown_jj mountain_nn]
(3) Semantic Relation Extraction
“mountain”
“object”
“rock”
“stone”

Figure 1: Semantic knowledge extraction process. Ellipses and dash lines in (0) represent perceptual concepts and relationships, respectively. Ellipses and plain lines in (3) represent semantic concepts and relationships, respectively. "_nn", "_vb", "_jj" and "_rb" are POS tags indicating previous words are nouns, verbs, adjectives or adverbs, respectively.

2.1. Basic text processing

During this step, textual annotations are tokenized and chunked into word phrases, and the words stemmed and tagged with their part-of-speech. Stopwords and non-content words are discarded.

The textual annotations of the images are first tokenized into words and punctuation signs. The words are tagged with their part-of-speech information (e.g., "noun" and "verb") and the annotations are chunked into noun and verb phrases (e.g., the sentence "The world loves New York" has two noun phrases "The world" and "New York", and one verb phrase "loves"). The IMKA system uses the LTG Chunker for chunking and POS tagging [8]. WordNet is also used in stemming words down to their base form (e.g., both "burns" and "burned" are reduced to "burn"), to correct some POS tagging errors (e.g., "dear" in Figure 1 can not be a verb), and to help group single words into compound words (e.g., "New York" is not two separate words, "New" and "York", but one compound word with one meaning in Figure 1). Finally, stopwords (i.e., frequent words with little information such as "be") and non-content words (i.e., words that not nouns, verbs, adjectives or adverbs) are discarded.

For the recognition of compound words, the IMKA system detects noun and verb phrases containing only nouns or verbs, respectively. Then, different combinations of the words, starting from the ones with more words and preserving word ordering, are searched in WordNet. If a word search is successful, the words are removed from the following word combinations until all the combinations have been searched. As an example, the noun phrase "New York" in Figure 1 will cause the following word searches: "New York ", "New" and "York"; the first search is successful so no additional searches are executed.

2.2. Semantic concept extraction

The second step in the semantic knowledge extraction process is to disambiguate the senses of content words in the annotations using WordNet and images clusters. Each detected sense is considered a semantic concept. The image clusters group similar images based on one or more visual feature descriptors (e.g., color histogram and Tamura texture) and/or text feature descriptors (e.g., tf * idf, term frequency weighted by inverse document frequency) and using a variety of clustering techniques (e.g., k-means algorithm and self-organizing maps) [2].

The intuition behind the proposed approach is that the images that belong to the same cluster are likely to share some semantics although these semantics may be very general (e.g., images of animals and flowers surrounded by vegetation are likely to be clustered together based on global color; they share semantics such as "nature" and "vegetation"). The proposed techniques also observe the two principles for word-sense disambiguation: consistent sense for a word and semantically relatedness of nearby words within clusters’ annotations, in this case.

The word-sense disambiguation procedure consists of two basic steps (see Figure 1). First, the different senses of words annotating the images in a cluster are ranked based on their relevance (the more relevant the concept, the higher the rank). An image can belong to several clusters; the second step is then to add the ranks for the same word and the same image obtained for the different clusters to which the image belongs. The sense of each word is considered to be the highest ranked sense.

The IMKA system ranks the different senses of a word for an image in a cluster by matching the definitions of the senses obtained from WordNet with the annotations of all the images in the cluster using standard word weighting schemes in Information Retrieval. The IMKA system implements two of the most popular schemes: tf*idf, term frequency weighted by inverse document frequency; and log tf*entropy, logarithmic term frequency weighted by Shannon entropy of the terms over the documents. The latter has been proven to outperform the former in Information Retrieval [6]. In this process, the extended definition of each possible sense for a word is considered to be a document; the document collection is basically the extended definitions of all the possible senses of a word; and the query keywords are the aggregated textual annotations of the images in the cluster. The task of word-sense disambiguation is to choose the most relevant sense out of all possible senses given the textual annotations aggregated through image clusters (by both visual and textual feature descriptors). Latent Semantic Indexing (LSI) can be optionally used before text matching to reduce the dimensionality of the text feature vectors [5].

The extended definition of a sense (e.g., "rock, stone" in Figure 1) is constructed from the synonym set (e.g., "rock, stone"), the actual definition (e.g., "mineral aggregate making up the Earth's crust") and the usage examples of the sense (e.g., "the work was brown") together with the synonym set, the actual definition and usage examples of directly or indirectly related senses (e.g., sense "lava", which is a *specialization* of "rock, stone") provided by WordNet. Different weights can be assigned to the synonym set, the actual definition, and the usage examples of a sense. As an example, higher weights should be assigned to the words in the synonym set (e.g., 1.0 for "rock" and "stone") compared to the usage examples (e.g., 0.5 for "rock" and "brown"). In the same way, weights assigned to the words in the synonym set, the definition and the usage examples of directly and indirectly related senses are made to decrease based on the type and the number of relationships to the original sense (e.g., 1.0 for words in definition of "rock, stone", 0.5 for words in definition of "lava"). The relation weights basically determine the kinds and levels of relationships used in sense definitions.

2.3. Semantic relationship extraction

The third step is to discover semantic relationships among the semantic concepts based on the relationships between the corresponding senses in WordNet.

Relationships among the detected senses, or semantic concepts, are taken from WordNet together with additional senses, the necessary to connect every pair of detected senses. Table 1 lists the semantic relations in WordNet together with definitions and examples. For example, if the senses "mountain" and "rock, stone" have been detected during the word-sense disambiguation process, both concepts will be connected at this stage through the concept "object", their common ancestor, and *generalization* relationships among them (see Figure 1).

Relationship	Definition	Example
Synonymy	Similar	rock ↔ stone
Antonymy	Opposite	white ↔ black
Hypernymy Hyponymy	Generalize Specialize	animal → dog rose → flower
Meronymy Holonymy	Component of Whole of	ship → fleet martini → gin
Troponymy	Manner of	whisper → speak
Entailment	Cause or necessity	divorce → marry

Table 1: Relations in WordNet with definitions and examples.

During this process, the IMKA system finds all the paths connecting every pair of detected senses in WordNet, either with a direct relationship or through relationships to intermediate senses. All the semantic relationships and the intermediate senses on these paths are also added to the extracted semantic knowledge. Therefore, the constructed knowledge will not be restricted to the detected senses but it will also contain intermediate senses among them. In other words, a subset of WordNet is selected in the semantic knowledge extraction process that includes the detected senses and all the paths between them.

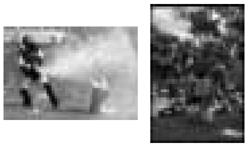
3. EVALUATION

Semantic knowledge was constructed for a collection of images with associated category labels, textual annotations, and image clusters. The image clusters were constructed from the images and the annotations using the techniques described in [2]. Accuracy in disambiguating the senses of the words in the textual annotations of 10% of the images in the collection was used to evaluate the semantic concept extraction process.

3.1. Experiment setup

The test set was a diverse collection of 3,624 *nature* and *news* images from the Berkeley's CalPhotos collection (http://elib.cs.berkeley.edu/photos/) and the ClariNet current news newsgroups (http://www.clari.net/), respectively. The images in CalPhotos were already labeled as *plants* (857), *animals* (818), *landscapes* (660) or *people* (371). The *news* images from ClariNet were categorized into *struggle* (236), *politics* (257), *disaster* (174), *crime* (84) and *other* (67) by researchers at Columbia University. The category labels were not used during the word-sense disambiguation. The *nature* and *news* images had short annotations in the form of keywords or well-formed phrases, respectively (see Figure 2).

Caption: **South Korea's police fire tear gas May 10 in front of a Seoul University.**



What: **People, culture, Zulu warrior**
Where: **Africa**
When: **1975-10-01**
Creator: **R. Thomas**

Figure 2: Example of a *news* image (left) and a *nature* image (right) with corresponding textual annotations.

During the perceptual knowledge extraction process, the images were clustered using different algorithms -k-means and SOM – and different visual feature descriptors -color histogram and Tamura texture - into different number of clusters. During the semantic knowledge extraction process, different image clusters were compared using different visual feature descriptors, clustering techniques and numbers of clusters. The extended sense definitions of senses were generated assigning different weights to the synonym set with respect to the actual definition and usage examples of a sense, and to the definitions of directly and indirectly related senses. Lof tf * entropy was used to match sense definitions and cluster annotations using the cosine metric.

The criterion to evaluate the word-sense disambiguation process was the percentage of words correctly disambiguated, in other words, the word-sense disambiguation accuracy. The first author of this paper generated the ground truth for the annotations of 10% of randomly selected images in the collection; no training was needed for that. The accuracy of the proposed approach was compared to three baseline approaches: (1) selecting a random sense for each word, (2) selecting the most frequent sense for each word, and (3) considering a cluster per image, i.e., only the text associated with each individual image is used during word-sense disambiguation.

3.2. Experiment results

Table 2 shows the accuracy results for best image clusters (BI), worst image clusters (WI), cluster-per-image (TT), most frequent senses (MF), and random senses (RD). The accuracy results are provided separately for the *nature* and the *news* images, and for nouns, verbs, adjectives, adverbs and all the content words.

Interesting conclusions can be drawn from Table 2. Consistently for both sets of images, best image clusters outperformed cluster-per-image and random senses. For *nature* images, best image clusters provided much better results than most frequent senses for all kinds of content words; even worst image clusters had similar results to most frequent senses, taking into account that most of the words in the annotations were nouns. The results for the *news* images were quite different: most frequent senses outperformed even best image clusters.

Nature Images

	BI	**WI**	**TT**	**MF**	**RD**
Nouns	91.32	87.7	82.92	85.92	74.44
Verbs	62.96	44.44	59.26	44.44	44.44
Adjectives	56.64	37.59	40.85	55.71	44.29
Adverbs	100.00	37.50	37.50	100.00	75.00
All words	**88.73**	**84.64**	**84.72**	**83.80**	**72.42**

News Images

	BI	**WI**	**TT**	**MF**	**RD**
Nouns	63.50	56.06	57.88	68.59	45.86
Verbs	46.44	35.63	39.07	58.48	24.08
Adjectives	69.50	53.50	54.68	72.00	46.77
Adverbs	71.88	50.00	62.50	74.19	45.16
All words	**58.85**	**51.32**	**52.33**	**66.58**	**40.52**

Table 2: Word-sense disambiguation accuracy for best image clusters (BI), worst image clusters (WI), image-per-cluster (TT), most frequent senses (MF), and random senses (RD) for the *nature* and the *news* images.

There are several factors that can explain the performance differences between *nature* and *news* images. First, WordNet has a more comprehensive coverage of nature concepts (e.g., animals and plants) than the one of news concepts because several animal and plant thesauri where used in its construction (i.e., WordNet has a bias for nature concepts compared to news concepts). Second, the textual annotations of *news* images are well-formed phrases so there are more words in the annotations that can potentially confuse the word-sense disambiguator compared to mostly keywords annotating the *nature* images. A third factor that may contribute to these results are that *news* images are much more diverse visually than *nature* images and, therefore, their clusters may not be as "meaningful" as those of *nature* images. Another possible explanation is that for *nature* images the annotations describe concepts that have a high correlation to visual features; whereas, the gap is larger between concepts and the visual features for *news* images.

Other important results not reflected in Table 2 due to space limitations follow. Extensive experiments were run for clustering the images into a number of clusters in the range from 1 to 1000. The best accuracy results for both *nature* and *news* images were obtained for numbers of clusters in the range from 8 to 30. These results are surprising because the larger the number of clusters, the purer the clusters were in terms of the categories of the images assigned to the clusters. This reinforces the fact that visual clusters are useful for word-sense disambiguation. The use of different visual feature descriptors or clustering algorithms had no obvious impact in the results shown in Table 2.

4. CONCLUSIONS

This paper proposes novel techniques for automatically extracting semantic knowledge from annotated image collections. The evaluation of the proposed word-sense disambiguation approach for extracting semantic concepts has shown that perceptual knowledge in the form of clusters generated from visual feature descriptors has the potential to improve performance compare to most frequent sense and purely text-based word-sense disambiguation (5% for *nature* images).

Our current work is focused on extending the evaluation to clusters using text feature descriptors, which have a higher correlation with semantic categories [2]. We are also working on automatic ways to evaluate arbitrary knowledge and to discover interactions among knowledge at different abstraction levels. For example, how to interrelate the semantic knowledge discovered in this paper and the perceptual knowledge discovered in [2], and use such interrelations for knowledge summarization, image classification, and automated concept illustration.

5. ACKNOWLEDGMENTS

This research is partly supported by a Kodak fellowship awarded to the first author of the paper.

6. REFERENCES

[1] Barnard, K., P. Duygulu, and D. Forsyth, "Clustering Art", *CVPR-2001*, Hawaii, USA, Dec. 9-14, 2001.

[2] Benitez, A.B., and S.-F. Chang, "Perceptual Knowledge Construction From Annotated Image Collections", *ICME-2002*, Lausanne, Switzerland, Aug 26-29, 2002; also Columbia University ADVENT Technical Report #001, 2002.

[3] Benitez, A.B., S.-F. Chang, and J.R. Smith, "IMKA: A Multimedia Organization System Combining Perceptual and Semantic Knowledge", *ACM MM-2001*, Ottawa, Sept. 2001.

[4] Benitez, A.B., J.R. Smith, and S.-F. Chang, "MediaNet: A Multimedia Information Network for Knowledge Representation", *IS&T/SPIE-2000*, Vol. 4210, Boston, MA, Nov 6-8, 2000.

[5] Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by Latent Semantic Indexing", *JASIS*, Vol. 41, No. 6, pp. 391-407, 1990.

[6] Dumais, S.T., "Improving the retrieval of information from external sources", *Behavior Research Methods, Instruments and Computers*, Vol. 23, No. 2, pp. 229-236, 1991.

[7] Jain, A.K., M.N. Murty, and P.J. Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, Vol. 31, No. 3, pp.264-323, 1999.

[8] McKelvie, D., C. Brew and H. Thompson, "Using SGML as a basis for data-intensive NLP", *Applied Natural Language Processing*, Washington, USA, April 1997.

[9] Mihalcea, R., and D.I. Moldovan, "A Method for Word-sense disambiguation of Unrestricted Text", *ACL-1999*, June 1999.

[10] Miller, G.A., “WordNet: A Lexical Database for English”, *Comm. of the ACM*, Vol. 38, No. 11, pp. 39-41, Nov. 1995.

[11] Rowe, N.C., "Precise and Efficient Retrieval of Captioned Images: The MARIE Project", *Library Trends*, Fall 1999.

[12] Stetina, J., S. Kurohashi, M. Nagao, "General Word Sense method based on a full sentential context", *COLING-ACL Workshop*, Montreal, CA, July 1998.

[13] Yarowsky, D., "Unsupervised Word-sense disambiguation Rivaling Supervised Methods", *Association of Computational Linguistics*, 1995.