

ENABLING VIDEO ANNOTATION USING A SEMANTIC DATABASE EXTENDED WITH VISUAL KNOWLEDGE

Gees C. Stein¹, Jens Rittscher and Anthony Hoogs
GE Global Research
One Research Circle
Niskayuna, NY 12303
USA

(gk@nycap.rr.com, hoogs@crd.ge.com, rittsc@crd.ge.com)

ABSTRACT

A semantic database has been extended with visual information to enable video annotation. This paper describes a lexical database, WordNet. We show its limitations with respect to describing visual characteristics, and describe an extension to WordNet that contains specific visual information. Having such a semantic database makes video annotation possible for Broadcast News: a domain that can cover any topic and involve a wide variety of events, objects and scenes. Combining basic visual analysis techniques and a semantic database containing visual descriptions avoids the problem developing large numbers of specific object and event detectors. Such a semantic database can be of great value for the analysis of multi-modal information. As far as we know, such a database has not been developed before.

1 INTRODUCTION

Although there is a growing demand for automated video annotation it is still a very difficult task. The annotation of broadcast video is particularly difficult due to the wide variety of topics and very short video sequences.

Currently, object and event recognition algorithms are not mature enough to scale up to such a large problem domain, as this would require enormous libraries of object and event representations and associated algorithms [6]. Furthermore, the semantic meaning of many events can only be recognized in its visual context. For example, the same hand motion may be a greeting, a signal for help, sign language, or even an insult.

We propose a novel approach that combines visual analysis with a semantic database describing events and objects. WordNet is a lexical database describing concepts both textually, and in terms of its relationships to other concepts. Our belief is that this approach will require a *limited* set of low-level visual detectors. Combining these low-level visual descriptions with a database of high-level object and events descriptions will enable the system to detect a wide variety of entities.

Below we will describe WordNet, discuss its limitations

with respect to visual information and explain how it was extended. We show how the extended semantic database enables improved Video Annotation and offer some suggestions how such a semantic database could be of use for other video analysis problems.

2 WORDNET

WordNet is a lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory [3]. English nouns, verbs, adjectives and adverbs are organized into synonym sets (synsets), each representing one underlying lexical concept. While in conventional dictionaries concepts are defined by a descriptive sentence and examples, resulting in a flat data structure, the meaning of a synset in WordNet is implied by its relationships to other synsets, thus creating a semantic network. In addition, every synset has a definition and example sentences. There are several relationships defined, depending on the part of speech. Nouns, for instance, have possible relationships synonymy, hyponymy (is an instance of), hypernymy (is a generalization of), holonymy (has parts) and meronymy (is a part of). Of these the hypernym-hyponym relationship for nouns and hypernym-troponym relationship for verbs are the most important, every noun and verb in WordNet are placed in the corresponding hierarchy.

Below an example from WordNet for *fighter*. Synonyms are given, a definition and its hypernyms (*fighter* is a kind of *airplane* is a kind of *heavier-than-aircraft* etc). *Fighter* has 10 links to its top node *entity*, *physical thing*. Note that concepts can have more than one hypernym, in this case a *fighter* is both an *airplane* and a *military vehicle*. Other relationships for *fighter* are *hyponym* and *meronym* (not shown). The noun 'is-a' hierarchy can be quite deep at places (see the *fighter* example). Overall, the verb network tends to be much shallower.

Currently WordNet contains about 110,000 synsets, with the noun group being by far the largest (about 75,000 synsets). WordNet is being added to daily, while parallel WordNets have been created for other languages, ranging from Dutch and Italian to Hindi. WordNet and its foreign counterparts have been used widely in the Natural Language Processing community for a variety of tasks such as

¹ At GE Global Research when this work was carried out. Currently not affiliated.

fighter, fighter aircraft, attack aircraft -- (a high-speed military or naval airplane designed to destroy enemy aircraft in the air)

- ⇒ airplane, aeroplane, plane -- (an aircraft that has a fixed wing and is powered by propellers or jets; "the flight was delayed due to trouble with the airplane")
 - ⇒ heavier-than-air craft -- (a non-buoyant aircraft that requires a source of power to hold it aloft and to propel it)
 - ⇒ aircraft -- (a vehicle that can fly)
 - ⇒ ...
 - ⇒ artifact, artefact -- (a man-made object taken as a whole)
- ⇒ military vehicle -- (vehicle used by the armed forces)
 - ⇒ ...
 - ⇒ artifact, artefact -- (a man-made object taken as a whole)

Machine Translation, Information Retrieval and Language Understanding. As far as we know, WordNet has never been used for Video Analysis on a large scale. In related work by [8], picture captions and other spatially related language were used to guide image analysis. Within this limited framework WordNet was extended with visual hierarchies to perform object recognition.

WordNet was selected because of its extensive concept coverage, and its hierarchical structure. The latter enables focused search in and immediate access to only relevant parts of WordNet's network. Also, WordNet is free (<http://www.cogsci.princeton.edu/~wn/>).

3 VISUAL INFORMATION AND WORDNET

The kind of visual descriptions created to search the semantic database varies from low semantic meaning ('parallel lines') to high semantic meaning ('face'). Roughly, one can distinguish between three semantic levels: Level 0 information consists of direct observables and forms the basis of the visual information hierarchy: edges, responses to filterbanks, segmented regions etc. Level 1 information is derived directly from image observables, and Level 2 information consists of objects and events, which are recognized using data from previous levels. Information recognized covers scene characterization (*man-made objects, vegetation, water/sea/ocean, rock and sky*), motion characteristics (*horizontal, up, down, smooth, erratic,...*) and a face detector (courtesy of CMU, [7]).

Ideally, the semantic database describes each object and event in visual terms that directly relate to the kind of descriptions video analysis is capable of generating. In WordNet concepts are described in terms of other related concepts (following their hyponym, hypernym etc. links) and contain free text; its definition and possible example sentences. Following links brings us to more definitions, the bottom line is that all knowledge in WordNet is given in free text.

When we look at WordNet's definition several limitations become apparent. First, many descriptions are functional, explaining the usage of an object, and not its visual features. In our *fighter* example above, we learn what the purpose is of a *fighter*. The definition of its hyponym *airplane* is more descriptive, (fixed wing, propellers), but again not sufficient. Second, there are many ways in language to describe the same visual features, which makes it harder to find visual descriptions if present. Finally, even

when known what to look for, say *rock* (the stone, a scene descriptor), one has to overcome the problem of polysemy: finding rock in a definition is not enough, rock can be a noun, or a verb and has more than one meaning. To fully understand text is a very hard problem, involving word sense disambiguation, parsing and semantic analysis; given the enormous size of WordNet this is not feasible.

On a positive note, when a WordNet concept does contain visual information it can be propagated down recursively to all its descendents following the hyponym links. For instance, we see that a fighter is a 'high-speed' plane, and that an aircraft 'can fly'. Thus we know that all types of aircraft (over 50 are defined in WN) can fly, and all fighters have high speed.

4 WORDNET ENHANCEMENTS

As shown in previous examples, only selected definitions contain useful visual information; therefore we decided to add explicit visual tags to WordNet.

The most basic 'tag' added describes *visibility*. WordNet contains many concepts that are not visible, such as music, ghost, hate, etc. We distinguish between *not_visible*, *visible* and *visualizable*. A concept is visualizable if one cannot only see it, but also draw it. All *visible* concepts that are not visualizable are labeled visible. Fruit, for instance, is labeled visible, since one cannot draw fruit; there are many different kinds of fruit. Banana, on the other hand, is visualizable. All visualizable concepts are visible, but not vice versa. This tag enables the search algorithm to discard all invisible concepts. We used the indices of two dictionaries ([1],[9]) to start with an initial list of words that are visible. For all the noun senses of a word the following was done: if it is monosemous, it was tagged as visualizable. If polysemous, for every sense the semantic category was found, depending on which a sense was labeled visualizable (for categories person, plant, animal etc.) or invisible (for categories cognition, feeling, etc.). For those phrases that were not found, the head noun was looked up: if found it was labeled visible. This way a little over 15,000 noun synsets were tagged, about 20% of the total of noun concepts in WordNet.

The choice of the additional visual aspects has been driven by low-level visual information provided by the visual analysis. The *motion type* tag describes basic motion characteristics (*no_motion* (house), *continuous* (car), *random* (fly)). If, for instance, motion is detected, the algorithm can discard entities known to be incapable of motion. (Note

that no motion detection does not contradict entities capable of motion.) The *motion texture* tag describes kind of motion (*articulate* (person), *non_articulate* (airplane)). A walking person has body parts that move in relation to each other, making this articulate motion. On the contrary, an airplane has no relative moving parts, making its motion non_articulate. When articulate motion is detected, non_articulate concepts can be ruled out, but not vice versa. The *motion speed* tag contains general information about potential maximum speed of an object (*high* (plane), *medium* (car), *low* (person)). When high-speed objects are detected, objects capable of only low or medium speed can be ignored. The *motion direction* gives the direction of events that describe motion: *up*, *down*, *horizontal*. Walking is a horizontal motion, while climbing involves upward motion. A tag related to scenery is the *inside/outside* tag which describes whether an object is typically found *outside* (e.g., tree) or *inside* (e.g., chair). One can envision additional tags (like *shape*) once the visual module can detect relevant discriminating data.

One tag not directly related to visual information is the *topical* tag. If the context of the audio/visual input is known to be about a certain topic, the system can focus on entities known to belong to this topic.

Finally, a *frequency* tag was added. WordNet contains a lot of very specific information not relevant to our task. Either the concept is very rarely observed, or the concept is too specific describing a level of detail not of visual interest. An example of the first case is a sun eclipse, an example of the second case are all the hyponyms of grass (wheat grass, dog grass, beach grass, ...). One would like the search not to go below the more general concept of grass, which can be observed frequently in video. Since frequency data of events and objects observed in News broadcasts is not available, we used word frequency data from a large corpus: The British National Corpus (<http://www.hcu.ox.ac.uk/BNC/>). The underlying assumption is that frequently talked/written about concepts are also frequently observed. We chose the BNC because of its size (around 100 million words), the wide variety of styles and topics covered and the free availability of all words occurring more than 800 times

(<http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html>). Word frequency data did have a part of speech tag, but of course no WordNet sense number. A simple approximation was used for polysemous words, distributing frequency numbers across the different senses. These values were recursively summed upward the hierarchies to prevent high-level concepts, which are often unusual words such as 'entity', from getting low frequencies. The result is frequency data for concepts that are 'topic-neutral'; i.e., one can not easily tell what topic is discussed from the words with frequency > 0. This makes topic tags absolutely necessary: a topic label can prevent a low frequency concept from being discarded.

Currently all tags, except for the frequency tag and the visibility tag, were added by hand; using inheritance whole WordNet hyponym trees were tagged at a time. About 195,000 tags were added with relatively little effort. In the future we hope to develop methods to automatically tag more WordNet synsets.

5 USING A SEMANTIC DATABASE EXTENDED WITH VISUAL INFORMATION

The extended WordNet database described above was used in a system for annotating Broadcast News Video. The visual information recognized by the visual analysis, described earlier, translates into so-called elemental terms and pruning terms. An elemental term maps directly to a WordNet concept. *Man-made object*, for instance, maps directly to *artefact* and thus, using the hyponym relation, gives us access to all man-made objects. Visual features that were recognized and do not map directly to WordNet concepts become *pruning terms*, a number of possible textual descriptions for a visual characteristic. In short, a search starts at an *elemental term* and, using hyponym links for nouns and troponym links for verbs, *pruning terms* are used to guide the search. The search is breadth first, recursively expanding the set of candidates using the hyponym/troponym relationship and then pruning this set down using the pruning terms. For more detail see [4].

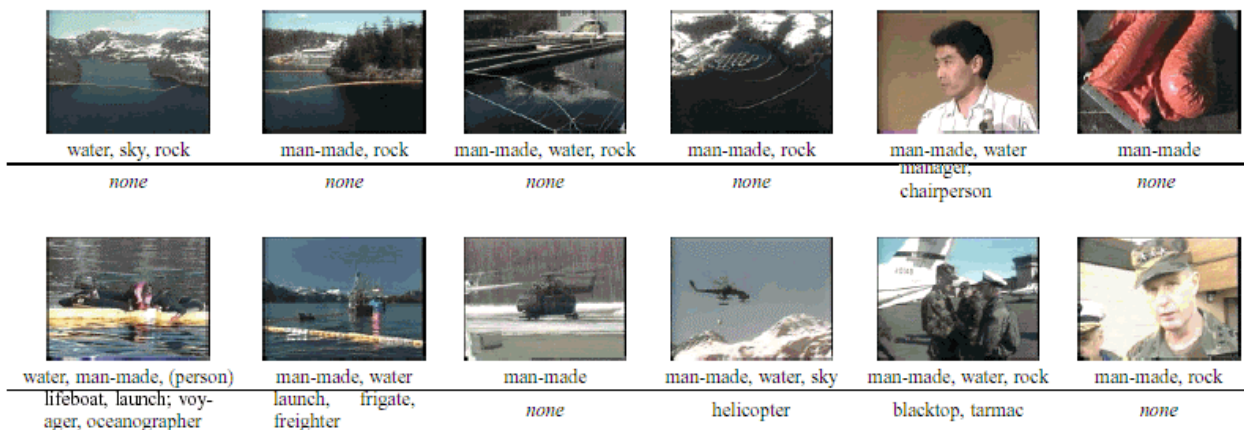


Figure 1 Representative images for 12 video clips with basic visual primitives and high-level semantic output.

The result is a set of ranked hypotheses, where each hypothesis is a WordNet concept (with implied tree of descendants) together with visual features that were satisfied. Note that although the number of potential objects and events output can still be quite large, it is a very focused and specific set compared to ‘the whole world’ contained in WordNet and possibly shown in a Broadcast Video.

Figure 1 shows representative images for 12 news clips. Above the line we see the basic visual features recognized for the clip. Below the line we see correct information found in the top 10 objects that were output by the WordNet search. Some of them are high-level concepts for which it would be very hard to write general detectors.

Having a semantic database that contains visual descriptions on such a large scale creates many opportunities for object and event recognition.

For instance, Duygulu et al. annotate images taken from the Corel data base [2]. Their approach uses a large training set of annotated images and a fixed set of low-level features. The approach is probabilistic, and creates a lexicon containing word –feature-vector combinations. Since the images in the Corel database have a descriptive text associated with them, this approach could benefit from the additional explicit visual information found in the enhanced WordNet: given basic visual features and WordNet, words from the captions could either be linked to specific regions in the image or ruled out. For instance, the training data includes abstract notions (i.e., not visible), the enhanced WordNet would know these cannot be part of the image. Duygulua et al. work on still images and motion characteristics are thus non-relevant. It would be interesting to try their approach on representative images of video sequences, and in addition use WordNet’s motion features.



Figure 2. “Ruben refueled jets on the Enterprise deck.”

Another interesting applications is annotation and word sense disambiguation of video transcripts. Look at Figure 2: a clip from Broadcast News that shows a jet landing on an aircraft carrier, with transcript. Doing some simple text analysis on the transcript gives us as possible visible objects: *jets* and *Enterprise deck*. Video analysis gives us: motion, and man-made object. When we look up jet in WordNet we find several senses for the noun meaning:

1. **jet**, jet plane, jet-propelled plane -- (an airplane powered by one or more jet engines)
2. **jet**, squirt, spurt, spirt -- (the occurrence of a sudden discharge (as of liquid))
3. **jet** -- (a hard black form of lignite that takes a brilliant polish and is used in jewellery or ornamentation)
4. fountain, **jet** -- (an artificially produced flow of water)

Only sense 1 for jet is known to be both man-made and capable of motion. In other words, we have disambiguated the word jet. In addition, we can build up glosses for word meanings using this approach: the transcript could be added to jet (sense 1) as an additional example sentence. One might even do some simple analysis on the sentences and learn that jets can be refueled. Word sense disambiguation is a very important problem in Natural Language Processing. Adding glosses and even more specific information to a lexicon is also of great usage to Machine Translation and Text Understanding, to name a few.

6 CONCLUSION AND FUTURE WORK

We have described a semantic lexical database, WordNet, and how it was extended with visual information specifically to be used to aid visual analysis in a general domain (News Broadcast). We have shown how this database enhances visual analysis capabilities while only involving a *limited* set of visual features. In addition we suggest some other useful applications for this kind of database.

We would like to enhance WordNet by better usage of analyzing the existing definitions, but also by mining other sources such as large corpora and other existing semantic nets (such as Cyc [5]). One non-trivial issue is that all information will have to be mapped to WordNet concepts. We also hope to make better use of the broadcast transcripts to direct the semantic search. More experiments will be carried out to establish the validity of the approach.

The work described in this paper was supported in full by ARDA.

References

- [1] Corbeil, Jean Claude (ed.) and Ariane Archambault. 1995. *The MacMillan Visual Dictionary*. Hungry Minds, Inc.
- [2] Duygulu, P., K. Barnard, J.F.G. de Freitas and D.A. Forsyth. 2002. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. *European Conference on Computer Vision*, 97-112.
- [3] Fellbaum, C. (ed). 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- [4] Hoogs, A., J. Rittscher, G. Stein (forthcoming). Video Content Extraction Using Visual Analysis and a Large Semantic Knowledgebase. *CVPR 2003*.
- [5] Lenat, D. and R. Guha. 1990. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley: Reading, MA.
- [6] Nelson, R. and A. Selinger. 2000. Learning 3D recognition models for general objects from unlabeled imagery: An experiment in intelligent brute force. In *Proceedings of ICPR*, volume I, 1-8.
- [7] Rowley, H.A., S. Baluja and T. Kanade. 1998. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (1): 23-38.
- [8] Srihari, R., and Z. Zhang. 2000. Show & tell: a semi-automated image annotation system. *IEEE Multimedia*, 7(3): 61-71, July 2000.
- [9] *The Ultimate Visual Dictionary 2001*. Dorling Kindersley Publishing.