Building a Visual Ontology for Video Retrieval

L. Hollink
Vrije Universiteit Amsterdam
hollink@cs.vu.nl

M. Worring
University of Amsterdam
worring@science.uva.nl

A.Th.Schreiber
Vrije Universiteit Amsterdam
Schreiber@cs.vu.nl

ABSTRACT

To ensure access to growing video collections, annotation is becoming more and more important. Using background knowledge in the form of ontologies or thesauri is a way to facilitate annotation in a broad domain. Current ontologies are not suitable for (semi-) automatic annotation of visual resources as they contain little visual information about the concepts they describe. We investigate how an ontology that does contain visual information can facilitate annotation in a broad domain and identify requirements that a visual ontology has to meet. Based on these requirements, we create a visual ontology out of two existing knowledge corpora (WordNet and MPEG-7) by creating links between visual and general concepts. We test performance of the ontology on 40 shots of news video, and discuss the added value of each visual property.

1. INTRODUCTION

To ensure access to growing video collections, annotation is becoming more and more important. Ongoing research in video analysis has produced various concept detectors, which are used to detect the presence of a specific concept. This approach to video annotation works well within narrow domains, where the number of possible concepts is small. However, it gets difficult as soon as the collection gets broader. Domains like biology, art, family pictures and broadcast news are problematic, as it is infeasible to build detectors for all possible concepts.

One approach to this problem is to use background knowledge about the domain under consideration. There is structured background knowledge available about various topics, in the form of ontologies or thesauri. Examples are SnoMed, MESH and the Gene Ontology for health care, and AAT and IconClass for art. Also, non domain specific knowledge structures exist, such as WordNet and Cyc. These knowledge bases are currently used for manual annotation ([15, 10]). They are not suitable for automatic annotation since they contain little visual information about the concepts they describe. Ontologies are used in annotations for various reasons. If existing, well-established, ontologies are used, they provide a shared vocabulary. Not only the terms themselves are agreed upon, but also the meaning of the terms, since the meaning is captured in the (hierarchical) structure of the ontology. Polysemous terms can be disambiguated. Reasoning can be used to find relationships between classes [1].

Research has been done on using ontologies for retrieval of textual resources [4]. However, using ontologies for retrieval of visual resources is a relatively new area. Hauptmann [5] proposed to design an ontology of automatically detectable concepts that could provide a basis for annotation of broadcast video. Ongoing research will tell which concepts are suitable for inclusion in such an ontology. Mezaris et al. [14] combine a thesaurus with relevance feedback. They let users describe high-level keywords with terms from a small ontology of intermediate-level descriptors such as luminance and size. Descriptions of the keywords are compared to extracted features of regions in video footage and matching regions are returned. Relevance feedback is then used to refine the result set. Hoogs and Stein et al. [8, 18] extended WordNet with visual tags describing visibility, different aspects of motion, inside/outside, and frequency of occurrence. Their system first analyses video to detect general low-level features such as colour and motion, as well as a limited number of high-level features. Both are then used to search the extended WordNet for relevant annotations.

In this paper we describe a visual ontology that was built to aid video annotation in a broad domain, building upon the work by Hoogs and Stein et al.[8, 18]. The first question we address is: *What are the requirements for a visual thesaurus for video retrieval?* Following these requirements will make our visual ontology different from the extended WordNet by Hoogs and Stein. Our visual ontology was created out of two existing knowledge corpora, WordNet and MPEG-7. WordNet does contain concepts that are visible, like `material` and `color`, but this visual information is not structured in a way that makes it useable for annotation. The visual concepts are not linked to other classes; there are no statements saying that a `boat` is capable of `motion` or that `houses` are made of `brick` or `wood`. We created these links between visual concepts and general concepts, thus building a Visual Ontology (VO)[1]. The second question is:*How well can a visual ontology perform under ideal circumstances?* We test the VO on 40 shots of news video, and discuss the added value of each visual property.

[1]Available on www.cs.vu.nl/~laurah/VO/VO.html

2. REQUIREMENTS

The following requirements for a VO have been identified:

1. Visuality A VO needs to contain classes and properties that describe visual information, such as color and shape of objects, in order to support automatic annotation of visual resources. These properties needs to be **visually perceptible**; characteristics like mass, smell and status are thus disregarded. Moreover, the property needs to be **detectable** from the visual data. Finally, visual properties need to be **discriminating**. In at least some cases the value of a visual property needs to tell something about the class an object belongs to. If too many objects have the same value for a visual property, it is not discriminating and therefore not useful for a VO.

2. Generality For our purposes of annotation in a broad domain, a VO needs to comprise terms from a broad domain.

3. General-Visual Relations It needs to contain relations between visual and general concepts, such as the statement that a `wheel` is `round`.

4. Interoperability The ontology as well as the resulting annotations need to be usable and reusable by various applications. Using existing knowledge corpora and standards instead of introducing ones own terms increases interoperability. The need for interoperability of data has been widely recognised in the semantic web and digital library communities (see for example [11]). Martinez et al. state that "one of the current trends is that the content is created only once, but it should be accessible via any access network and client device" ([13], p.1).

Hoogs en Stein extended wordnet with visual terms [8, 18]. The use of WordNet ensured that they met the generality requirement, while the added visual properties fulfilled the Relations requirements. We seek to build an ontology that meets all four requirements.

3. DESIGN OF A VISUAL ONTOLOGY

3.1 Using existing ontologies

We chose MPEG-7 to describe the visual information. MPEG-7 is a standard for describing multimedia content published by the Moving Picture Experts Group (MPEG) [12]. It is aimed at a broad range of applications. The MPEG-7 OWL[2] ontology as published by Hunter [9] contains low-level visual properties like color, shape and motion.

Following the choice of Hoogs and Stein, we used WordNet as a general ontology. WordNet is a widely used lexical database in which nouns, verbs, adjectives and adverbs are organised into synonym sets, each representing one underlying lexical concept [2]. Containing over 100,000 synsets, its broadness makes it desirable for annotation in broad domains. We used the RDF(Schema) translation of WordNet by van Assem et al [19].

The broadness of WordNet combined with the visual information in Mpeg-7 ensure compliance with the first two requirements of Generality and Visuality. By using an ISO standard like MPEG-7 and a widely used lexicon like WordNet we seek to fulfill the fourth requirement of Interoperability. We use RDF(S) as the language for our visual thesaurus, to ensure interoperability also on the syntactic level. In order to link general concepts to visual concepts and meet the third requirement, we add statements of the form

```
<general concept> <visual property> <visual value>
 wordnet:subway -  VO:environment -  wordnet:indoor
 wordnet:car    -  mpeg7:motion   -  wordnet:rigid
```

3.2 Selecting visual properties

As identified in Section 2, a visual property needs to be visually perceptible, detectible and discriminating to be useful in a VO. We use two sources of visual properties. Properties from the Mpeg7 ontology [9] were used for the most low-level descriptors. From 'Mpeg7 Visual' we used `color`, `shape` and `motion`. Mpeg7 provides one more descriptor that is both visual, detectable and discriminating: `texture`. Since values for a texture descriptor are not easily described in words, we did not incorporate this in our VO. Texture is used in detecting materials by for example Geusebroek et al. [3].

While MPEG-7 offers low-level visual properties, WordNet contains more high-level visual properties. The hierarchy under `property/attribute` contains the following concepts that meet the requirements for visual properties: `visibility`, `naturalness`, `environment` and `material`. Visibility can have the values invisible or visible. Visible can be further refined into visualisable and viewable. A visualisable concept was defined by Stein et al. as a concept that "one can not only see [..], but also draw"[18]. All instances of a visualisable class must have visual characteristics in common like shape and colour. We consider a concept viewable if it can be seen in today's daily life without instruments. Microbacteria, intestines and galleons are not viewable.

The RDFS graph of the visual thesaurus is depicted in Fig. 1, showing that the top-level WordNet concept `entity` can have values for the seven visual properties. SubClass relations are denoted by arrows, property names are in italics and class names are in normal font. The ranges of values that the visual properties can take were all taken from WordNet. The set of possible values is large for some properties, such as for `material`, which can have values from all WordNet subclasses of cloth, building material and matter. This was done not to restrict the usability of the thesaurus to a particular domain or use case. Also, an annotation is never too specific as in a thesaurus a specific concept can always be traced back to its more general parent. The classes motion, colour and shape are modeled as subclasses of corresponding MPEG-7 classes, in order to make sure that the MPEG-7 classes are not changed themselves [19].

3.3 Populating the VO with instances

As a proof of concept, we implemented part of the visual thesaurus: all classes in the WordNet hierarchy under the class `conveyance` were extended with visual properties using the schema described above. The hierarchy under `conveyance` contains 564 classes. This number makes it feasible to manually assign the values of the visual properties, while it is still large enough to demonstrate annotation in a broad domain. We made use of the subclass-of relations already present in WordNet; if a class has a certain value for a property, all its subclasses have the same value. In many cases no value can be assigned to a property, such as to the `shape` property of the class `public transport`. In total, 548 property-value pairs were assigned, excluding the pairs that can be deduced from the subClass hierarchy.

[2]The Web Ontology Language OWL

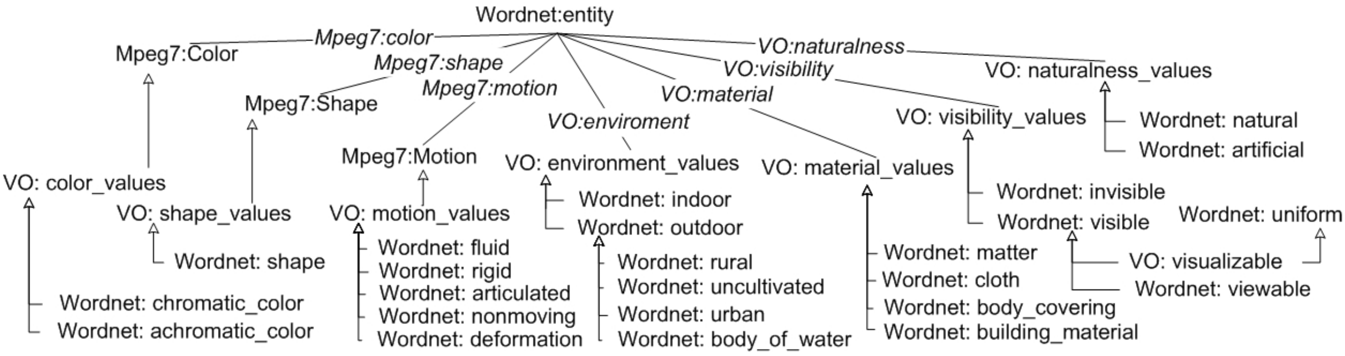

Figure 1: RDFS graph of the visual properties from MPEG-7 and WordNet and their values from Wordnet.

3.4 Use of the Visual Ontology

The aim of the VO is to facilitate search as well as quick semantic annotation. Annotation with the VO can be done by detecting color, shape, motion, material, naturalness, and/or environment of a region in a video, and then searching the VO for concepts with matching properties. In this way, the list of possible annotations is deduced to a manageable size from which relevant annotations can be selected.

Search with the VO can be done by looking for regions within a video whose visual characteristics match the visual properties of a query concept taken from the VO. This will reduce the video collection to a smaller set of shots, through which the searcher can browse to find relevant items.

4. A THOUGHT EXPERIMENT

4.1 setup

An evaluation of the search capabilities would require the detection of visual characteristics of all video's in a collection, which is beyond the scope of this paper. Instead, we evaluate the quality of *annotations* made with the VO. To decouple detector quality and the power of the VO, we make the assumption that we have perfect detectors for the six detectable properties and their categorisation into symbolic values. Currently, detectors exist for each property, and detector quality is improving fast, as can be seen from [17]. This makes it plausible that good detectors will be available in the near future.

40 Shots from the TRECVID 2003 collection that contain a form of conveyance were selected for testing. The shots display boats, trains, cars and planes, with 10 shots in each category. First, keyframes of all shots were segmented [6]. Second, property values were assigned manually to one region in each keyframe. The value for the 'environment' property was determined by looking at the values of the neighbouring regions. Finally, the VO was searched for viewable concepts matching the list of property values of a region.

4.2 Results

We evaluated the results with two measures, precision and reliability. Precision is defined as the number of relevant annotations found, divided by the total number of annotations found. Reliability is the percentage of shots for which at least one relevant annotation is found. Relevance of an annotation was decided manually, based on the free text descriptions that WordNet provides for all concepts. Fig. 2 shows examples of retrieved annotations. At least one relevant annotation was found for 93% of all shots (Table 1). Precision of the annotations has a mean of 8%. On average, the 564 concepts of the conveyance hierarchy have been reduced to 57. This is a manageable list from which relevant annotations can be picked manually or automatically.

	Train	Car	Boat	Plane	Total
Retrieved	61.0	37.7	37.8	91.0	56.9
Retrieved relevant	1.9	3.8	2.3	9.2	4.3
Precision	0.03	0.10	0.06	0.10	0.08
Reliability	0.90	1.00	0.80	1.00	0.93

Table 1: Precision and reliability of annotations

Retrieval relied mostly on only three properties: material, motion and environment. Colour and shape were detected in the shots, but were not present in the visual thesaurus. This is because the majority of the WordNet concepts are not visualisable and do not have a fixed shape or colour for every instance. In the domain of conveyance, almost all concepts are viewable, with the exception of some historic or fantastic vehicles like `galleon`, which are visible but not viewable. All WordNet classes in this domain have the value artificial, except for `pirogue`, which is a canoe made out of a whole tree.

	Correct	**– Incorrect descriptions**
	taxicab	electric-car, berlin, limousine, minicab, gypsy cab.
	passenger train, commuter train	freight train, bullet train, tandem trailer, articulated lorry, trailer-truck, helicopter, skyhook, freight-liner, ladder truck, single rotor helicopter, shuttle helicopter, cargo helicopter.

Fig. 2: Derived descriptions

5. DISCUSSION AND CONCLUSION

In this paper we have described our experiences during creation and use of a VO in a collection of video. We propose that a VO for this purpose needs to contain broad general concepts, visual descriptions and links between those two. Furthermore it needs to comply to existing standards. A combination of two existing ontologies, Mpeg-7 and WordNet, meets these requirements. Visual properties that are represented in the visual thesaurus need to be visually perceptible, detectable and discriminating. In a study on the conveyance domain, it appeared that the visual properties `visibility` and `naturalness` were not discriminating in this domain: too many classes had the same value for these properties. We believe, however, that in other domains, such as landscapes or politics, these properties could be important discriminating attributes. Therefore, we did not remove them from the visual thesaurus.

Colour and shape were not useful as visual properties, since most of the WordNet conveyance concepts can not be described in terms of colour and shape: they are not visualisable. Further study is needed to determine if colour and shape are useful in other domains, such as botanical or food domains. In this study, motion, material and environment were both detectable and discriminating, and therefore the most important properties for retrieving relevant annotations.

Using existing knowledge bases ensures interoperability. However, it also means that one has to use resources who's design may not be ideal for the current purpose [7]. WordNet, is not ideal for visual descriptions, since the hierarchy is more functionally orientated then visually, as was pointed out by Stein et al. [18]. This means that all members of a class have functional properties in common, but not necessarily visual properties. In addition, WordNet is originally not a subclass hierarchy, but a hierarchy of hyponyms. Synsets represent lexical concepts. As language is not always consistent, WordNet is not always consistent. Because of this, visual properties of a concept do not always propagate to all its hyponyms, which complicates the process of assigning visual properties to classes.

One of the questions we asked was "How well can a VO perform in ideal circumstances?" We successfully reduced the list of possible annotations to a list of manageable size from which relevant annotations can be picked. Going through the result can be done by a human annotator. Another option is to combine the results from the VO with text associated with a shot. After detecting the visual properties of a region, a VO can tell which of the words in the description or transcription of the video shot are relevant for the region.

Extending the VO from only the hierarchy under conveyance to the complete WordNet hierarchy will increase the size of the result list: it will occur more often that concepts in different places in the hierarchy have the same set of property values. One way to overcome this is to use more visual properties in the VO, as the range of available detectors increases. Hoogs et al. [8] deal with this problem by employing a limited set of high-level concept detectors as a starting point for a search through their visually extended WordNet. In this way classes in different parts of the hierarchy with equal visual properties can be disambiguated. Detectors for this purpose should not be too specific, as they are meant to start a search rather than find a specific concept. In addition, they should correspond to one or more concepts in WordNet. Wordnet concepts that meet these criteria, and for which detectors are available (see e.g. [17]) are `animal`, `human` and `vegetation`.

6. REFERENCES

[1] G. Antoniou and F. van Harmelen. *A Semantic Web Primer*. MIT Press, April 2004.

[2] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[3] J.M. Geusebroek and A.W.M. Smeulders. A six-stimulus theory for stochastic texture. *Int. Journal of Computer Vision*, 62(1/2):7–16, 2005.

[4] S. Handschuh and S. Staab. *Annotation for the Semantic Web*. IOS Press, 2003.

[5] A.G. Hauptmann. Towards a large scale concept ontology for broadcast video. In *Proc. of the 3rd int conf on Image and Video Retrieval (CIVR'04)*, 2004.

[6] M.A. Hoang, J. Geusebroek, and A.W.M. Smeulders. Color texture measurement and segmentation. In *Proceedings of the 2nd international workshop on Texture Analysis and Synthesis*, 2002.

[7] L. Hollink, A.Th. Schreiber, J. Wielemaker, and B. Wielinga. Semantic annotation of image collections. In *Proc. of the K-CAP Semannot Workshop*, 2003.

[8] A. Hoogs, J. Rittscher, G. Stein, and J. Schmiederer. Video content annotation using visual analysis and a large semantic knowledgebase. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*, 2003.

[9] J. Hunter. Adding multimedia to the semantic web - building an mpeg-7 ontology. In *International Semantic Web Working Symposium*, 2001.

[10] E. Hyvonen, S. Saarela, K. Viljanen, E. Mkel, A. Valo, M. Salminen, S. Kettula, and M. Junnila. A cultural community portal for publishing museum collections on the semantic web. In *Proc. of ECAI Workshop on Appl. of Semantic Web Technologies to Web*, 2004.

[11] E. Lee. Building interoperability for united kingdom environment information resources. In *Proc. of the European Conference on Digital Libraries*, 2004.

[12] J.M. Martinez. Overview of the mpeg-7 standard. Technical Report 5.0, ISO/IEC, Singapore, 2001.

[13] J. M. Martnez, C. Gonzlez, O. Fernndez, C. Garca, and J. de Ramn. Towards universal access to content using mpeg-7. In *Proc. of ACM MM*, 2002.

[14] V. Mezaris, I.Kompatsiaris, and M.G.Strintzis. Region-based image retrieval using an object ontology and relevance feedback. *J. on Appl. Signal Proc.*, 2004.

[15] A.Th. Schreiber, B. Dubbeldam, J. Wielemaker, and B.J Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, 16(3):66–74, May/June 2001.

[16] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12), December 2000.

[17] C.G.M. Snoek, M. Worring, J.M. Geusebroek, D.C. Koelma, and F.J. Seinstra. The mediamill trecvid 2004 semantic video search engine. In *TREC Video Retrieval Evaluation Online Proceedings*, 2004.

[18] G.C. Stein, J. Rittscher, and A. Hoogs. Enabling video annotatation using a semantic database extended with visual knowledge. In *Proceedings of ICME*, 2003.

[19] M. van Assem, M.R. Menken, A.Th. Schreiber, J. Wielemaker, and B. Wielinga. A method for converting thesauri to rdf/owl. In *Proc. of the Third International Semantic Web Conference*, 2004.