SEMI-AUTOMATIC, DATA-DRIVEN CONSTRUCTION OF MULTIMEDIA ONTOLOGIES

Alejandro Jaimes * and John R. Smith Ψ

* Electrical Engineering Department
Columbia University
New York, NY 10027, USA

Ψ Pervasive Media Management
IBM TJ Watson Research Center
Hawthorne, NY 10532, USA

ABSTRACT
In this paper we investigate semi-automatic construction of multimedia ontologies using a data-driven approach. We start with a collection of videos for which we wish to build an ontology (an explicit specification of a domain). Each video is pre-processed: scene cut detection, automatic speech recognition (ASR), and metadata extraction are performed. In addition we automatically index the videos based on visual content by extracting syntactic (e.g., color, texture, etc.) and semantic features (e.g., face, landscape, etc.). We then combine standard tools for ontology engineering and tools in content-based retrieval to semi-automatically build ontologies. In the first stage we process the text information available with the videos (ASR, metadata, and annotations, if any). Stop words (e.g., a, on, the) are eliminated and statistics (e.g., frequency, TFIDF, and entropy) are computed for all terms. Based on this data we manually select concepts and relationships to include in the ontology. Then we use content-based retrieval tools to assign multimedia entities (e.g., shots, videos, collections of videos) to concepts, properties, or relationships in the ontology, and to select multimedia entities as concepts, relationships, or properties in the ontology. We explore this methodology to construct multimedia ontologies from 24 hours of educational films from the 1940s-1960s used in the TREC video retrieval benchmark and discuss the problems encountered and future directions.

1. INTRODUCTION
There have been many recent efforts to automatically analyze and index multimedia. At the same time, new applications in the context of the semantic web [11] have renewed interest in the creation of knowledge-based systems, and in particular on the exchange of information for different applications. As a result, new tools for creating and maintaining ontologies (formal, explicit specifications of a domain) have been developed [4], and there are many efforts to create domain-specific ontologies.

Ontologies have applications in many areas including natural language translation, medicine, standardization of product knowledge, electronic commerce, and geographic information systems, among others [6]. Without a doubt many of these applications use or will use multimedia data in the immediate future, making the creation of multimedia ontologies a crucial component.

Multimedia ontologies have applications in content visualization, indexing, annotation, knowledge sharing, learning, and reasoning, among others. Manual construction of ontologies, however, is a time consuming and therefore expensive task. Fully automatic construction of ontologies, on the other hand, is usually not feasible because many domain-specific decisions must be made to adequately specify the domain of interest.

In this paper, we investigate semi-automatic construction of multimedia ontologies using a data-driven approach. We process an input video collection by performing automatic scene cut detection, automatic speech recognition, parsing of metadata, and automatic content analysis [1]. Then we use standard text mining tools implemented in KAON [9][11] (e.g., stemming, term frequencies, etc.) to semi-automatically construct ontologies based on the textual data available in the collection. In the creation of these ontologies we also use an extensive content-based retrieval tool we built for the TREC video retrieval benchmark [1][14]. The textual ontology is manually augmented with multimedia entities (e.g., video shots, keyframes, regions, etc.), which are selected using our system. We present experiments using our approach, discuss the problems encountered and future directions.

1.1. Related work
In multimedia, ontologies have been used for photo annotation [10], audio structuring and retrieval [8], and image organization, browsing and retrieval [13][16], among others. In most cases the ontologies are built entirely automatically [2] or entirely by hand [10], and concepts are represented only by textual terms [16].

The work presented in this paper differs from previous approaches in two aspects: (1) we construct ontologies semi-automatically; (2) we use standard text mining tools and techniques in content-based retrieval to build data-driven ontologies. The work in [16] is interactive. However, we apply different text processing tools, more functions for content-based analysis, and use distinct kinds of metadata.

2. MULTIMEDIA ONTOLOGIES
An ontology is a formal, explicit specification of a domain. Typically, an ontology consists of concepts, concept properties, and relationships between concepts.

In a typical ontology concepts are represented by terms. In a multimedia ontology concepts might be represented by multimedia entities (images, graphics, video, audio, segments,
etc. as in [3]) or terms. In this paper, we are interested in textual and non-textual representation, and more importantly, in semi-automatic construction of ontologies from multimedia collections for multimedia applications.

Multimedia ontologies have many application areas, including the following:

- **Content visualization:** They can be used to create tables of content and used for browsing.
- **Content indexing:** They can be used to improve indexing consistency in manual annotation systems [10] (e.g., use the term apartment instead of flat), or in the propagation of labels in automatic indexing systems (e.g., a face detected implies a person was detected).
- **Knowledge sharing:** Annotated multimedia collections can be more easily shared if they use a common conceptual representation.
- **Learning:** Collections annotated by different individuals using common ontologies lead to annotation consistency which is of extreme importance in applying approaches based on learning techniques that use annotated collections for training.
- **Reasoning:** Information not explicit in the data may be obtained automatically with the help of an ontology.

Ontology construction is usually a manual, iterative process consisting of at least 3 steps: (1) selection of concepts to be included in the ontology; (2) establishment of properties for the concepts and relationships between concepts in the ontology; (3) maintenance of the ontology.

The ontology can be constructed using a concept-driven or a data-driven approach. The concept-driven approach does not require any data: the ontology is built from general or domain specific knowledge. A doctor, for example, could potentially create a domain specific ontology of diseases based on his expert knowledge. Or we could create a generic one about elements in the world. In the data-driven approach the ontology is constructed primarily from data, but domain knowledge is also used in manually constructing it.

Fully automatic construction of ontologies from data is desirable, and there are some recent efforts to achieve this goal. In general, however, fully automatic construction of ontologies is not possible because automatically selecting relevant concepts and relationships is hard. An alternative is to use semi-automatic ontology construction techniques, which aim at facilitating each of the steps above. Tools for semi-automatic ontology construction typically use text mining approaches to select and discover concepts and relationships between them. In addition, they include functionality to represent and manipulate ontologies (edit concepts, define relationships, etc.).

### 3. SEMI-AUTOMATIC CONSTRUCTION

In the data-driven approach we are exploring, we wish to semi-automatically construct a multimedia ontology that, using a video collection \( V \), models the concepts \( C \), their properties, and their relationships.

For each of the videos we apply the following steps: (1) automatic scene cut detection; (2) automatic speech recognition; (3) parsing of metadata; (4) automatic content analysis. Pre-processing consists of steps one through three, after which an ontology can be constructed based on textual content alone. Step four, described in section 3.3, forms the basis for manually adding multimedia components to the ontology.

#### 3.1. Pre-processing

Automatic scene cut detection is performed to break up the video into shots [1]. For automatic speech recognition (ASR) we use the system described in [7], which yields a time stamped list of recognized words and silence tags. Alignment is then performed to assign text to specific video shots (also in [7]). The result of this stage, then, is a video divided into shots which may have text associated with them. In addition, we parse metadata, if available with the videos, to extract relevant information (e.g., movie titles).

#### 3.2 Textual Content

We process the ASR and the metadata using standard text mining techniques implemented in KAON [11][9], a tool for semi-automatically building ontologies. This tool was chosen over other similar ones because it integrates the text mining techniques we describe here with an ontology management module.

![Figure 1. Example ontology built with KAON.](image-url)

The first step in processing the text consists of eliminating stop words (i.e., high frequency words which have little semantic content such as a, the, in, etc.). Once the stop words are removed, word stemming is performed (e.g., going replaced with go) to obtain terms and compound terms (e.g., social worker). For each term or compound term, the frequency, TFIDF (term frequency weighted by inverse document frequency), and entropy (logarithmic term frequency weighted by entropy of terms over the documents) scores are computed.

The ontology construction process outlined earlier then begins with the manual selection of relevant concepts, where the text pre-processing output is used to guide that selection. Terms with high TFIDF score, for example, might represent important concepts.

Once the user has selected concepts for the ontology, relationships between those concepts are discovered using KAON, which uses an algorithm for discovering generalized association rules. Relevant relationships are then manually selected by the user for inclusion in the ontology.
3.3. Visual Content

The construction of multimedia ontologies in the previous section is based only on text processing. While text can be used effectively, to create true multimedia ontologies we must integrate other aspects of multimedia.

As part of the TREC effort [1][14] we built a system that allows access to videos in multiple ways1. This includes textual search on metadata, ASR, and annotations, as well as several functions for searching and clustering based on content-based syntactic (e.g., color, texture, etc.), and semantic (i.e., concepts such as face, indoor, sky, music, monologue, etc.) features.

We use this system to support the construction of multimedia ontologies in the following ways:

- Assign visual entities to textual concepts, relationships, or properties in the ontology.
- Select visual entities as concepts, relationships, or properties of the ontology.

Semi-automatic ontology construction tools that use text and tools for content-based analysis can be integrated into a single framework [16] or used separately. In this paper we build the ontology using the method described in section 3.2, and using the TREC retrieval tool, which includes many retrieval functionalities not available in [2][16].

4. EXPERIMENTS

We used 24 hours of video from the NIST TREC 2002 Video Benchmark collection [1][14], which is a subset of the Internet Movie Archive [5]. The videos are of color and black and white educational and promotional films from the 1940s-1960s. Each of the videos was manually annotated1 shot by shot using VideoAnnex [15], in which keyframes are used to represent shots and users can label regions and scenes. For each shot, therefore, we have a set of region and/or scene labels. For each of the videos, therefore, we have metadata, manual annotations, and text from automatic speech recognition (ASR).

The metadata for each film consists of a short structured description with fields such as producer, and title, and a short sentence summary (e.g., dramatization of important military events and imagery of the places where they occurred).

For the entire 24 hour collection we generated a total of 22,815 annotations from 189 unique terms representing objects, scenes, audio (e.g., music), and audio-visual objects (e.g., monologue). From the metadata text we obtained approximately 5,000 terms in the entire collection. Speech recognition produced 391,945 terms for the videos in the collection, out of which only 11,234 were unique terms. For the 24 hours of video we obtained 11,744 keyframes representing the shots.

4.1. Text Ontologies

As described in section 2, the first step in creating an ontology consists of manually selecting concepts. To do this, we examine the frequent terms (Table 1), and the terms with the highest TFIDF score (Table 2). The table shows the scores separately for each collection, but they could be computed jointly.

<table>
<thead>
<tr>
<th>Table 1. Terms with highest frequency score (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR</td>
</tr>
<tr>
<td>People</td>
</tr>
<tr>
<td>Time</td>
</tr>
<tr>
<td>Way</td>
</tr>
<tr>
<td>Day</td>
</tr>
<tr>
<td>Life</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Terms with highest TFIDF score (T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR</td>
</tr>
<tr>
<td>Prodigy</td>
</tr>
<tr>
<td>Social worker</td>
</tr>
<tr>
<td>Bike</td>
</tr>
<tr>
<td>Golden Gate</td>
</tr>
<tr>
<td>Cigarette</td>
</tr>
</tbody>
</table>

As can be seen from the tables, selecting the most important concepts for the ontology is by no means trivial. Although the individual scores are indicators of a particular term’s importance, the selection cannot be made based on the scores alone and general knowledge is necessary in making the selection. We discuss this further in section 4.1.

The next step is to automatically discover associations between concepts that are already in the ontology. Like terms, associations are automatically scored to help the user make the selection [9]. The user selects the relevant associations and defines the relationship type. For example, relationships may be hierarchical (concept x is related to concept y) or of property (concept x is a property of concept y). Table 3 shows some interesting associations found automatically. The right type of relationship (e.g., a face is part-of a person) is entered manually.

<table>
<thead>
<tr>
<th>Table 3. Example associations.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR</td>
</tr>
<tr>
<td>Therapy-diagnosis</td>
</tr>
<tr>
<td>Spaghetti-ketchup</td>
</tr>
<tr>
<td>Homework-bike</td>
</tr>
<tr>
<td>Germ-tuberculosis</td>
</tr>
</tbody>
</table>

4.1. Multimedia ontologies

We explore the construction of ontologies using the text data described above and the TREC content-based retrieval tool. First, however, there are several issues worth addressing.

First is the difference between text data sources. Issues to consider include whether a single ontology should be constructed or if several ontologies are required. An ontology derived from the metadata alone can be very useful for browsing based on general semantic categories. An ontology derived from annotations, on the other hand can be useful in getting a snapshot of the visual contents of the collection and on discovering relationships between such visual entities that were not immediately obvious. An ASR ontology can give us a detailed overview of the topics discussed in the videos.

The content-based retrieval tool can be used for tasks that include selection of representative multimedia entities for
concepts, properties, or relationships, and selection of multimedia entities as concepts, properties, or relationships. In other words, in some cases multimedia entities will themselves represent concepts, while in others they will be instances of concepts. Some examples of concepts that are better represented visually in our collection are shown in Figure 2.

![Figure 2. Example elements of a multimedia ontology built from the TREC data.](image)

In our experiments we found that constructing ontologies using textual information alone is not sufficient. This is partly due to the issues related to different text sources. The TREC retrieval tool was found to be helpful not only in searching for multimedia entities for inclusion in the ontology (e.g., Figure 2), but also in selecting concepts from the textual data. For example, the TREC tool clusters videos based on color information and type (photograph, graphic, etc.), and allows searches based on detected objects and scenes (e.g., face, sky, monologue, etc.) and visual content (e.g., color, texture, etc.). Since we have metadata associated with each video and most shots have ASR text associated with them, it is possible to browse, for instance, all shots associated with particular terms based on any of these criteria. We found this to be very helpful for constructing ontologies.

5. DISCUSSION

Constructing ontologies is difficult because different correct specifications of the same domain or collection are possible and many decisions have to be made which depend on the domain, the purpose of the ontology, the data, and the user’s knowledge.

We found that using standard tools for semi-automatic construction of ontologies is very helpful in building data-driven multimedia ontologies. The nature of multimedia data, however, makes it necessary to address issues that are not necessarily present in text-only collections. This includes the differences between multiple text sources and the ways in which content-analysis (i.e., visual, audio) techniques should be integrated in the construction of such ontologies.

6. CONCLUSIONS AND FUTURE WORK

In this paper we investigated semi-automatic construction of multimedia ontologies from multimedia collections. Our approach is based on combining existing tools for construction of ontologies from text, and tools for content-based retrieval.

We outlined the possible applications of multimedia ontologies and applied our methodology to a large set of educational films from the 1940s-1960s used in the TREC video retrieval benchmark.

Future work includes developing systematic methodologies for the construction of multimedia ontologies, development of evaluation metrics, and further work on investigating different ways to integrate text based approaches and content-based analysis.

ACKNOWLEDGMENTS

The authors would like to thank all of the members of the IBM 2002 TREC team for annotating the TREC collection and for their many technical contributions. We are also very grateful to Harriet Nock for providing us with the text from speech transcriptions and to Bell Tseng for aligning those transcriptions with individual shots.

REFERENCES

[5] Internet Movie Archive (http://www.moviearchive.org)