Learning and Classification of Semantic Concepts in Broadcast Video

John R. Smith, Murray Campbell, Milind Naphade, Apostol Natsev, Jelena Tesic
Intelligent Information Management Department
IBM T. J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532 USA
jsmith@us.ibm.com

Keywords: Imagery and Video Analysis, Search and Retrieval, Fusion, OSINT

Abstract
Broadcast video (news, documentaries, investigatory reports, etc.) is becoming an increasingly important medium for dissemination of information, sharing of knowledge and raising of public awareness on various issues. The ubiquity and open accessibility of broadcast video also gives it enormous potential as an important source of intelligence. However, technical challenges for automatically analyzing and understanding broadcast video content make it difficult to effectively tap this resource. In particular, manual processes for creating text-based metadata of broadcast video cannot keep up with the explosion of this content. Manual annotation is very costly and time consuming and often subjective, leading to incomplete and inconsistent annotations and poor system performance.

1. Introduction
Fueled by the rapid expansion of broadband Internet connectivity and increasing interest in online multimedia-rich applications, the growth of digital multimedia content has skyrocketed. This growth is driving new forms of interaction with images, speech, video and text. It is also compounding the need for more effective methods for indexing, searching, categorizing and organizing this information.

Unfortunately, manual processes for creating text-based metadata cannot keep up with the explosion of rich media content. Manual annotation is costly and inadequate. Manual labeling and cataloging is very costly and time consuming and often subjective, leading to incomplete and inconsistent annotations and poor system performance. New technologies are needed for reducing annotation costs. To address this challenge, we have developed a novel approach for machine learning of semantic concepts in broadcast video. In this paper, we describe the system and report on experimental results showing effective automatic detection of semantic concepts in the domain of broadcast news video.

2. Learning Video Semantics
We are developing a prototype multimedia analysis and retrieval system that uses multi-modal machine learning techniques for bridging the semantic gap for multimedia content analysis and retrieval. The objective is to automatically annotate broadcast video, making it possible to later search and retrieve content of interest. The system addresses both broadcast news video content analysis and search as follow:

- The broadcast news video content analysis engine applies machine learning techniques to model semantic concepts in video from automatically extracted audio, speech, visual content. It automatically assigns labels (with associated confidence scores) to new video data to reduce manual annotation load and improve searching and organizes semantic concepts using ontologies that exploit semantic relationships for improving detection performance.

- The broadcast news video search engine integrates multimedia semantics-based searching with other search techniques (speech, text, metadata, audio-visual features, etc.). It also combines content-based, model-based, and text-based searching for video searching.

2.1 Reducing Annotation Costs
Manual labeling of multimedia content is extremely human resource and cost intensive and typically requires in
excess of ten times greater time spent per unit time of
video, e.g., one hour of video requires ten hours of hu-
man effort for complete annotation. Furthermore, man-
ual labeling often results in incomplete and inconsistent
annotations. As shown in Figure 1, we are developing a
system that replaces manual processes requiring 100%
annotation with a semantics learning approach that in-
volves manual annotation of only 1-5% of video data.

- Manual labeling is extremely human resource
  and cost intensive (10x)
- Manual labeling also often results in incom-
  plete and inconsistent annotations
- Replace manual process requiring 100% annota-
  tion with a semantics learning approach that
  involves manual annotation of only 1-5% of
  video data
- Learning process builds models from labeled
  training data, adapts to annotation errors
- Automatically propagate labels to video
data, assign confidence score

A significant gain in annotation quality can be achieved
with modest levels of manual annotation & training of
statistical models from multimedia features. On the other
hand, manual annotation achieves high annotation quality
only with high completeness. The semantics learning
approach improves annotation quality at all levels of
completeness, and a significant gain in annotation quality
results from modest levels of training (1-5% impact).

2.2 Technical Approach
Recent advances in multimedia content analysis, feature
extraction and classification have improved capabilities
for effectively searching and filtering multimedia con-
tent. However, a gap remains between the low-level fea-
ture descriptions that can be automatically extracted from
the multimedia content, such as colors, textures, shapes,
motions, and so forth, and the semantic descriptions of
objects, events, scenes, people and concepts that are
meaningful to users of multimedia systems.

The system technology is unique in its approach for ana-
lyzing and fusing audio, visual and text information to
automatically annotate multimedia contents. The system
exploits relationships and correlations among annotated
labels to improve performance.

For example, detecting that a video scene shows an air-
plane is boosted by detection that the scene is outdoors as
opposed to indoors. The system integrates three different
search methods of multimedia content based on feature-
descriptors, concept models and text. This fusion ap-
proach provides users with resources for building queries
of multimedia repositories sequentially using multiple
individual search tools, and performs higher quality
searching.

2.3 Semantic Concept Detection
The system works by automatically labeling the multi-media contents using machine learning techniques that analyze the audio, visual and text components of multi-media data. As shown in Figure 2, the system uses text information such as speech, closed captions and transcript information along with automatically extract visual features, such as motion, texture, color, shapes for modeling semantic concepts. Using a machine learning approach a library of semantic models is created from training examples. Human interaction is required for the training process. However, this human input is required only for a small data set. Once the models are trained and validated, they are available for applying to large repository of unlabeled video content.

The multimedia analysis and retrieval research started with humble origins by modeling on the order of 5-6 semantic classes of images based on visual appearance -- things like "indoors" vs. "outdoors" and "nature" vs. "man-made". However, we soon discovered that it was possible to cover a much larger set of semantic classes for video content and improve accuracy by analyzing multiple channels of information -- namely, audio contents and speech in addition to the visual appearance -- and by using sophisticated statistical modeling techniques for "fusing" this multi-modal information. One of the early breakthroughs was the successful completion of a detector for "video clips of rocket launches" using multi-modal information that was accurate enough to give 90% precision in retrieval results.

This led to another breakthrough after expanding the vocabulary of semantic concepts to a large enough size (about 100 detectors) in creation of a "semantic basis" for higher-level video analysis and searching. By using the confidence scores generated from this larger set of detectors it is possible to perform more accurate searching and matching for queries based on a "model vector" technique. This led to another breakthrough of processing model vectors to exploit context by using confidence scores from some detectors to influence others. For example, if the computer assigned labels to a video clip indicating presence of "sky", "water", "sand", "people" then the confidence score for detecting "beach" are boosted. Furthermore, all of this boosting information could be learned automatically by the computer by extracting correlations and statistical information in a set of training examples.

2.4 Search
Given the ability to automatically extract descriptors from video and multimedia data, it is possible to search
for video content at a number of different levels, including features, models, and semantics. The shot boundary description defines the basic unit of matching and retrieval as a video shot or segment. Users may issue queries to the system in several ways: (1) feature-based – by selecting example key-frame images and video segments in which matches are found based on the MPEG-7 feature descriptions, (2) text-based – by issuing text query which is matched against the MPEG-7 textual annotations or speech transcriptions, (3) semantics-based – by issuing text queries or selecting from key-words that are part of MPEG-7 classification scheme, and (4) model-based – by selecting key-words. The distinction between model-based and semantics-based is that the confidence score for model-based search can be used for ranking and fusing results during the video searching process. The user-interface for the video search engine is shown in Figure 3.

3 Evaluation

We evaluated the semantic concept detection performance on the TRECVID evaluation benchmark. The TREC Video Retrieval Evaluation (TRECVID) was established in 2001 to evaluate and promote progress on research in automatic segmentation, indexing, and content-based retrieval of digital video. The TREC conference series is sponsored by the National Institute of Standards and Technology (NIST) with additional support from other U.S. government agencies. TREC provides an infrastructure for large-scale testing of retrieval technology realistic test collections uniform, appropriate scoring procedures a forum for the exchange of research ideas and for the discussion of research methodology.

TRECVID is a laboratory-style evaluation that attempts to model real world situations or significant component tasks involved in such situations. There are four main tasks in TRECVID with associated tests, as shown in Error! Reference source not found..

- Shot boundary determination
- Story segmentation
- High-level feature and semantic concept extraction
- Search

The TRECVID data set is comprised of approximately 170 hours of broadcast news video data from CNN Headline News, ABC World News Tonight and CSPAN. The system has achieved the top performance on the TRECVID semantic concept detection evaluation in 2003 and 2004. The concept detection task involves the automatic labelling of up to 1000 matching video clips for each of a predefined set of semantic concepts. Figure 4.
shows some of the automatic semantic concept detection results.

Overall, the semantic concepts defined in the TRECVID evaluation include:

- Indoors
- News subject face
- People
- Building
- Road
- Vegetation
- Animal
- Female speech
- Car/truck/bus
- Aircraft
- News subject monologue
- Non-studio setting
- Sporting event
- Weather news
- Physical violence

4 Scaling to 1,000 Concepts

While the TRECVID benchmark currently evaluates on the order to 10-20 semantic concepts, search and retrieval applications will require detection on the order of 1,000 or more semantic concepts. Therefore, scaling to the order of 1,000 is an important challenge in going from research technology to deployable system. We are currently addressing this challenge through a one-year workshop based effort called LSCOM (or Large Scale Concept Ontology for Multimedia), which is organized jointly by IBM T. J. Watson Research Center, Carnegie Mellon, Columbia University and Univ. California at Santa Barbara. The objective of this workshop is to create a collaborative activity in which representatives from three critical communities: broadcasters and users, library scientists, and systems and technology developers create a semantic ontology of 1,000 concepts for the broadcast news video domain. This ontology will provide a focal point for technical research on developing automated techniques for semantic concept detection. While LSCOM will help advance the field for automatic indexing of broadcast news video, it is hoped that it will serve as a model for creating concept ontologies for other content domains, such as sports, entertainment, music, and so on.

5 Conclusions

With the growing interest in digital video, efficient and effective systems are needed for video indexing and retrieval. In this paper, we described the development of a system for learning of semantic concepts in broadcast video, which reduces the required manual effort for creating content description metadata. The objective is to
better tap the enormous potential of broadcast video as an important source of intelligence. We described the integration of the video semantic analytics with video searching and reported on evaluation in the context of the NIST TRECVID evaluation.

6 References


