Large-Scale Concept Ontology for Multimedia

Milind Naphade, John R. Smith, and Jelena Tesic
IBM T.J. Watson Research Center

Shih-Fu Chang, Winston Hsu, amd Lyndon Kennedy
Columbia University

Alexander Hauptmann
Carnegie Mellon University

Jon Curtis
Cyc Corp.

Multimedia content continues to grow rapidly. Bridging the semantic gap is essential to exploiting this growing data. Toward this goal, recent research has focused on automatically tagging multimedia content to support end-user interactions such as searching, filtering, mining, content-based routing, personalization, and summarization. However, to date, there's been limited progress on standardizing the set of semantics for machine tagging.

We describe a recent collaborative undertaking to develop the Large-Scale Concept Ontology for Multimedia. This effort is being led by IBM, Carnegie Mellon University, and Columbia University with participation from CyC corporation and various other research academic and industrial groups. (Some of the authors are leading this effort.) The Disruptive Technology Office (DTO) sponsored LSCOM as a series of workshops that brought together experts from multiple communities to create a taxonomy of 1,000 concepts for describing broadcast news video. The LSCOM taxonomy was designed to satisfy multiple criteria of utility, coverage, feasibility, and observability. Along with the taxonomy of 1,000 concepts, the LSCOM effort has produced a set of use cases and queries along with a large annotated data set of broadcast news video. The goal is to create a framework for ongoing research on semantic analysis of multimedia content.

Editor's Note

As increasingly powerful techniques emerge for machine tagging multimedia content, it becomes ever more important to standardize the underlying vocabularies. Doing so provides interoperability and lets the multimedia community focus ongoing research on a well-defined set of semantics. This column describes a recent collaborative effort of multimedia researchers, library scientists, and end users to develop a large standardized taxonomy for describing broadcast news video. The Large-Scale Concept Ontology for Multimedia (LSCOM) is the first of its kind designed to simultaneously optimize utility to facilitate end-user access, cover a large semantic space, make automated extraction feasible, and increase observability in diverse broadcast news video data sets.

—*John R. Smith*

Standardizing multimedia semantics

Several recent notable developments have greatly benefited multimedia retrieval research. The National Institute of Standards and Technology (NIST) TRECVID video retrieval evaluation (http://www-nlpir.nist.gov/projects/trecvid/) has been instrumental in advancing research on broadcast news video. New and powerful techniques have emerged for automatically tagging this video content using techniques such as statistical machine learning. These techniques are providing real benefits by greatly reducing human effort for cataloging large volumes of multimedia data. However, although TRECVID has been a catalyst for research, one open question remains: What set of semantic concepts should the community focus on as it explores new automated tagging techniques?

A scan of publications in the multimedia research community over the last 10 years reveals an ad hoc approach to research on multimedia semantics. Although a few recurring semantic concepts have emerged—such as indoors versus outdoors, cityscape versus landscape, people, faces, and so on—there's been no larger coordination of research. TRECVID created small sample sets of high-level features for its yearly evaluations, but the early TRECVID evaluations only provided limited coverage of the semantics interesting to users (see Table 1). The TRECVID experience highlights the first two critical requirements for any large-scale concept ontology for multimedia semantics, which is facilitating end-user access to multimedia repositories and covering the semantic space interesting to end users.

1070-986X/06/$20.00 © 2006 IEEE

Table 1. TRECVID high-level features.

Evaluation	No. of Semantic Concepts	Semantic Concepts
TRECVID-2002	10	Outdoors, indoors, face, people, cityscape, landscape, text overlay, speech, instrumental sound, and monologue
TRECVID-2003	17	Outdoors, news subject face, people, building, road, vegetation, animal, female speech, road vehicle, aircraft, news subject monologue, non-studio setting, sporting event, weather news, zoom in, physical violence, and person *x*
TRECVID-2004	10	Boat/ship, Madeleine Albright, Bill Clinton, train, beach, basket scored, airplane takeoff, people walking/running, physical violence, and road
TRECVID-2005 (LSCOM* lite)	39	People walking/running, explosion or fire, map, US flag, building exterior, waterscape/waterfront, mountain, prisoner, and sports car
TRECVID-2006 (LSCOM* base)	834	449 annotated concepts from base LSCOM, containing several broad categories such as objects, activities/events, scenes/locations, people, and graphics.

*Large-Scale Concept Ontology for Multimedia

Alternatively, there are several standard controlled vocabularies and classification schemes for multimedia. For example, MPEG-7 has standardized more than 140 classification schemes that describe properties of multimedia content. Similarly, TGM-I provides a large thesaurus for cataloging graphical material. (See the "Multimedia Controlled Vocabularies" sidebar for more examples and details.)

However, these standard schemes have received little attention from the multimedia research community, mostly because many of the terms in these schemes aren't suitable for automated tagging. For example, the MPEG-7 Genre Classification Scheme (urn:mpeg:mpeg7:cs:GenreCS:2001, which is a genre classification scheme defined by the MPEG-7 standard), which is used to classify programs based on their con-

Multimedia Controlled Vocabularies

We examined several multimedia controlled vocabularies:

- MPEG-7 Multimedia Description Schemes (http://www.chiariglione.org/mpeg/) standardizes classification schemes comprising 754 controlled terms.

- TV-Anytime (http://www.tv-anytime.org/) standardizes 954 terms for broadcast TV content.

- Escort 2.4 defines 115 terms.

- Thesaurus of Graphical Material (TGM-I) defines index terms for cataloging graphical (image and video) content.

- SMPTE Metadata Registry from Society of Motion Picture and Television Engineers (SMPTE; http://smpte.org/). This metadata dictionary defines a registry of metadata element descriptions for association with essence or other metadata. A full explanation is contained in SMPTE 335M.

- IPTC Core Metadata, NewsML, SportsML, and ProgramGuideML from International Press Telecommunications Council (IPTC, http://www.iptc.org). The versatile News Markup Language for global news exchange. NewsML 1 is designed to provide a media-independent, structural framework for multimedia news.

- P/Meta Metadata Scheme from European Broadcasting Union (EBU; http://www.ebu.ch/) an audio–visual metadata schema.

- Military Lexica provides several hundred concepts covering buildings, weapons, transportation routes, vehicles, and situations.

- Comstock offers 3,000 lexical terms for Comstock image categories.

- Time Life gives 368 categories for classifying images.

- SLA News Division provides 108 terms from the Special Libraries Association's News Division Web site, http://www.ibiblio.org/slanews/.

In all, we reviewed more than 10,000 terms and found that most other schemes don't adequately consider observability in video data sets, feasibility, or automating extraction.

tent or subject matter, defines terms such as "special events" and "remarkable people." The terms might be useful for classifying multimedia content but don't lend themselves well to automated extraction. Such subjective concepts also make it difficult for two annotators to completely agree, which further complicates this issue. This highlights the third critical requirement for the multimedia concept ontology: the feasibility of automated extraction.

To advance multimedia retrieval research, one final ingredient—the availability of large annotated data sets—is necessary. Researchers have addressed similar requirements for other related problem domains such as speech recognition, which builds on large phoneme-level annotated speech corpora and optical character recognition (OCR). In particular, statistical machine-learning techniques use these data sets to build models and classifiers. A lack of sufficient training data severely limits progress on automating extractions. This completes the final requirement for the large multimedia concept ontology: the development of large annotated video data sets that support the observability of the semantic concepts.

LSCOM's Design

According to these needs, we identified the following requirements for LSCOM:

- *utility*, or a high practical relevance in supporting genuine use cases and queries;
- *coverage*, or a high coverage of the overall semantic space of interest to end users within the target domain;
- *feasibility*, or a high likelihood of automated extraction considering a five-year technology horizon; and
- *observability*, or a high frequency of occurrence within video data sets from the target domain.

To address these requirements, the teams embarked on a series of workshops. The aim was to ensure impact through focused collaboration of multimedia researchers, library scientists, and end users to achieve a balance when designing the large-scale concept ontology for multimedia. In particular, the specific tasks were to

- identify the target domain (broadcast news video),
- solicit input from end-user communities and capture use cases,
- survey related metadata standards and cataloging practices,
- organize large video data sets,
- analyze technical capabilities for concept modeling and detection,
- develop a draft taxonomy and validate in TRECVID evaluation,
- annotate large video data sets,
- expand to 1,000 concepts and construct an ontology, and
- evaluate and conduct gap analysis and identify outstanding research challenges.

To date, the LSCOM effort has produced an ontology of 1,000 concepts that's proving to be a valuable resource for the multimedia research community. The teams have used approximately 450 concepts to manually annotate a large corpus of 80 hours of broadcast news video. The experiments evaluate the utility in supporting 268 queries taken from 39 use cases developed in the LSCOM workshop series.

Use case design

With help from practitioners and end users, the teams designed a set of 39 use cases capturing needs for accessing large news video data sets. The use cases cover both generic and more transient current topics. They include stories that cover natural calamities, crime, breaking news, and so on. Some cover topics that have long duration such as "The War on Terror" or "The Oil Crisis." Others cover recurring events such as government elections.

To evaluate the ontology using these use cases, we expanded each use case into a set of queries (typically somewhere between six to 37 queries per use case). Figure 1 gives an example use case for "The War on Terror" in Afghanistan.

Video domain and data sets

The LSCOM effort selected broadcast news video as its target domain due to the significant interest in news video as an important multimedia source of information. This choice also

Named entities relevant to use case: Afghanistan, Kandahar, Kabul, Pakistan, Hindu Khush, Taliban, Hamid Karzai.

Expanded queries:

- Battles/violence in mountains
- Landmines exploding in barren landscapes
- Masked gunmen
- Camps with masked gunmen without uniforms
- Armored vehicles driving through barren landscapes
- Mountainous scenes with openings of caves visible
- People wearing turbans with missile launchers
- Group of people with a pile of weapons
- Refugee camps with women and children visible
- Political Leaders making speeches or meeting with people
- Predator drone flying over mountainous landscape
- Munitions being dropped from aircrafts over landscape
- Munitions being dropped from aircrafts in mountains
- Dead people and injured people
- Bearded men speaking on satellite phones in mountainous landscape
- Convoy of several vehicles on makeshift roads
- Empty streets with buildings in state of dilapidation
- Groups of people commenting on terrorism
- Map of Afghanistan with Kandahar and Kabul shown
- Afghan flag atop building
- Scenes from the meetings of political leaders
- Militia with guns firing across mountains
- Insurgents
- Men in black Afghan dresses with weapons exercising with bunkers in the background
- Military personnel watching battlefield with binoculars
- Series of explosions in hilly terrain
- Man firing at soldier
- Fired missile in the air
- Incarcerated people in makeshift jail
- Funeral procession of young victims of bombing
- Afghan warlords with weapon-carrying bodyguards in a village meeting discussing strategy and tactics

Figure 1. An example Large-Scale Concept Ontology for Multimedia (LSCOM) use case for news footage of Afghan battles, demobilization, and disarmament.

proved synergistic with the TRECVID evaluation, which similarly focuses on broadcast news video.

In addition, our choice to use broadcast news video as the target domain ensured a variety of available video data sets. The LSCOM effort used the development data set from the TRECVID 2005 corpus, which consists of broadcast news videos captured during October and November 2004 from 11 different sources. The 80-hour news video corpus was segmented into 61,901 shots, and each shot was represented by a single keyframe.

Concept modeling

To analyze the feasibility of tagging the news video content automatically, we drew on the expertise of the multimedia analysis researchers and experience of the NIST TRECVID video retrieval evaluation. We chose a five-year technology horizon for assessing feasibility. In determining feasibility, we assessed the following characteristics of each concept:

- the class of concepts to which it belongs (objects, scenes, roles, and so on),
- the frequency of occurrence in the TRECVID corpus,
- the specificity or generality of the concept, and
- the semantic gap as a function of these characteristics that tells us how easy or hard it is to automatically model this concept when tagging automatically.

For example, concepts we observed in the corpus (such as airplane flying, protests, riots, parade, and aircraft carrier) we deemed feasible for automated tagging. On the other hand, we deemed more subjective concepts (such as discovery, challenge, and happiness) unobservable and infeasible. Overall, we filtered the 10,000 candidate concepts down to 834 based on feasibility and utility. This list was then further filtered based on observability to produce a base set of 449 concepts.

Potentially, given a larger corpus, we believe we would find a high degree of observability for most of the 834 concepts. To reach 1,000 concepts we are working with the CyC Knowledge Base to identify gaps. In addition, we're continuing to evaluate the taxonomy's coverage in supporting queries, which will identify additional candidates to be included in the taxonomy.

The taxonomy design organized the 834 concepts into six categories on a top level: objects, activities/events, scenes/locations, people, graphics, and program categories. We then refined these categories, such as by subdividing objects into buildings, ground vehicles, flying objects,

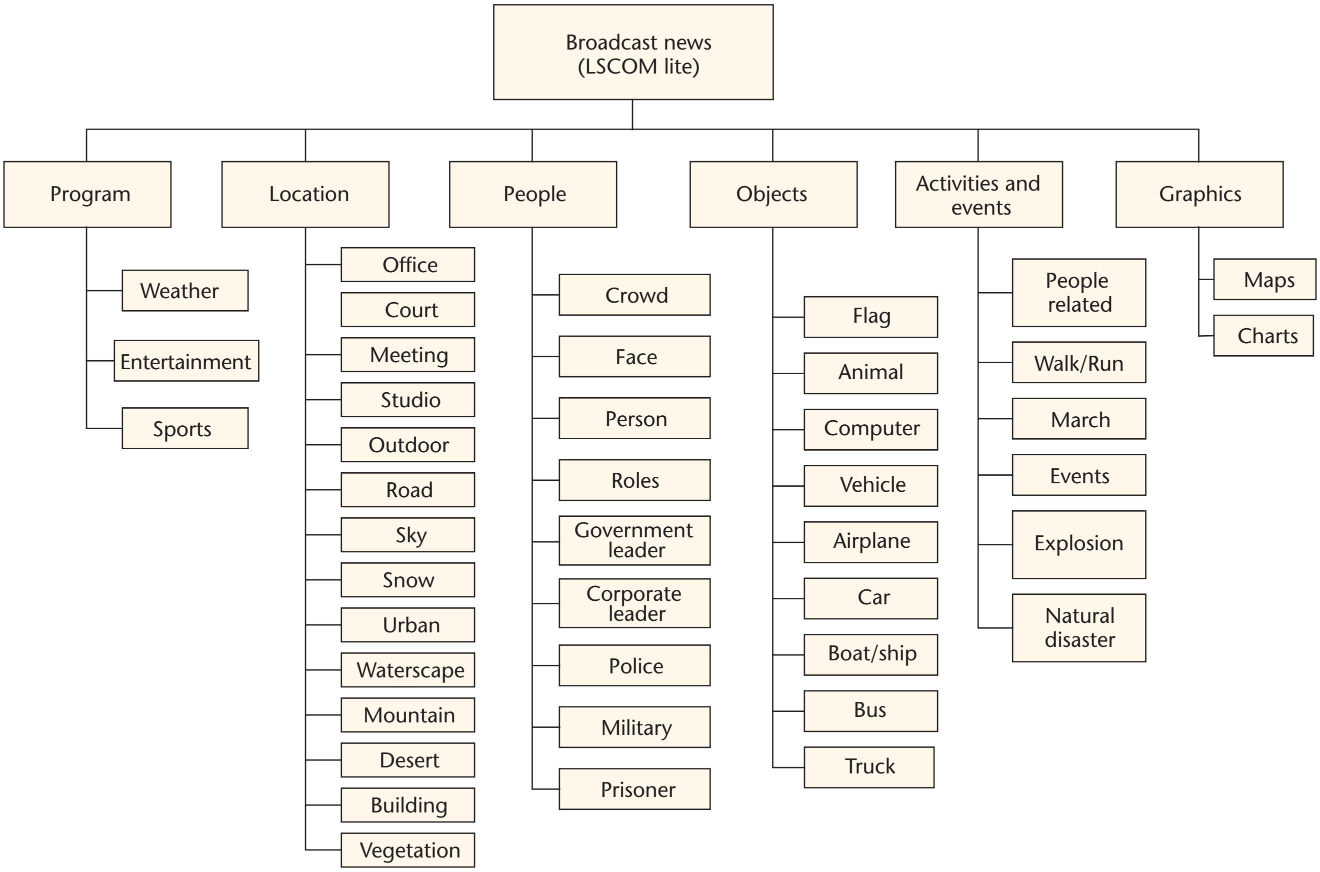

Figure 2. A light version of the LSCOM taxonomy as applied to the high-level feature detection task of the National Institute of Standards and Technology's TRECVID 2005 and 2006.

and so forth. Figure 2 shows the top few levels of the LSCOM taxonomy.

To create a fully annotated video data set, the LSCOM team labeled the presence or absence of each of the 449 concepts in 61,901 shots of broadcast news video. This required more than 28 million human judgments. We believe the annotated data set will prove to be of enormous value for research. It's also being explored for the TRECVID-2006 video retrieval evaluation.

Ontology design

To further enrich the base LSCOM taxonomy toward the goal of identifying 1,000 concepts, we're currently mapping it into the Cyc Knowledge Base, which is a repository of more than 300,000 concepts and 2 million assertions (rules and ground assertions). This mapping is intended to help ensure the following:

- The LSCOM ontology will be more than a taxonomy (simple hierarchy) of concepts. Cyc will provide a knowledge-rich representation, complete with rules and additional background knowledge needed to support a previously unattained level of semantic video and image annotation and retrieval.
- The LSCOM ontology will have sufficient broadness by growing the concepts to include peer and other related nodes identified in Cyc.

Mapping into Cyc has revealed some side benefits, such as ontology repair, whereby nodes in the LSCOM taxonomy that had been originally misplaced in the taxonomy were moved to more appropriate locations. For example, US_Flag originally appeared directly under Object and was moved under Flag.

The mapping will also produce a Web Ontology Language (OWL) export of the relevant Cyc subset and will contain that portion of the Cyc content for the LSCOM concepts that OWL supports. This includes most binary relations known (short of theorem proving) to hold among target concepts, as well as higher relations (so-called *rule macros*) that stand in for rules by relating target concepts to binary relations. This also includes many explicit rules (universally quantified if/then statements), so long as they can be represented in the Semantic Web Rule Language (SWRL), which is the intersection of OWL and RuleML. Consequently, the releasable version can support some level of theorem proving in video- and image-retrieval software.

Evaluation

We're currently working to quantify the benefits of the LSCOM taxonomy in supporting more than 250 queries corresponding to 39 use cases. In the first step, we're using the LSCOM taxonomy to manually and automatically expand the 250 queries. This will help measure taxonomy coverage in terms of broadness and depth.

In the second step, we're using the concepts for querying video based on the 250 queries. We're comparing the LSCOM concept-based query results to those obtained using baseline methods such as text retrieval on speech transcripts. We will also combine the LSCOM-based retrieval with text-based retrieval to assess improvement in retrieval effectiveness.

Future directions

We plan to make the results of the LSCOM workshop available to the community at large. Already to date, a lite version of LSCOM (see Figure 2) was explored by hundreds of researchers worldwide for the TRECVID-2005 video retrieval evaluation (see http://www-nlpir.nist.gov/projects/tv2005/tv2005.html). The plan is to use a larger base taxonomy of 449 concepts for the TRECVID-2006 evaluation.

We believe many of the outputs from this effort, such as the standardized LSCOM ontology, use cases, queries, and large annotated broadcast news video corpus will serve as an important catalyst for further work in this field.

More information about LSCOM is available at http://www.ee.columbia.edu/dvmm/lscom/. The site also contains the current draft of the LSCOM taxonomy, which is available for download. **MM**

Acknowledgments

This material is based on work funded by the US government. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the US government.

Readers may contact the authors at naphade@us.ibm.com.