The Multimedia Lexicon: Automatic object and structure discovery in audio-video-text content
Dan Ellis, Shih-Fu Chang & Kathleen McKeown - Columbia University

Project summary

We understand the real world in terms of objects: When we look at a picture or our surroundings, or when we close our eyes and listen, we organize the information arriving from the outside world in terms of attributes for perceived, independent objects. “Independence” can refer to physical separability (the pen on my desk), behavioral autonomy (the hissing steam valve on the radiator), or more conceptual distinctions, but it is something that human observers can largely agree upon.

For computers analyzing similar signals, this task is very challenging: Ongoing efforts to find robust algorithms to separate visual or auditory objects have yet to bear fruit. This absence is keenly felt; among the huge amount of audio and video material available, we can find particular events through manual annotations, or by reviewing the entire collection ourselves. Automatic systems with the ability to ‘watch’ the material and interpret it in human-relevant terms would help us with queries like “when was the speech in which President Clinton was applauded while waving his finger?”

The proposed project is concerned with building such a system. The aim is to develop the following technologies that are key to providing general-purpose object-based multimedia analysis:

- **Audio-visual object discovery**: Current object extraction is successful only for limited domains and single modalities. Integrating the best object-related feature extraction techniques from the complementary video and audio domains with machine learning algorithms to search for correlations in large unlabeled training corpora, will produce a new generation of genuinely scalable multimedia object extraction methods.

- **Automatic connection to lexical terms**: Unsupervised object discovery will result in a large number of self-defined candidate object classes, without any relation to verbal queries. To bridge this gap, the cross-modal clustering will be extended to cover terms derived, by parsing and similar techniques, from associated text material - such as descriptions, subtitles or speech-recognition transcripts. The outcome will be a *Multimedia Lexicon* relating textual terms and audio-visual feature descriptions.

- **Hierarchic topic and structure recovery**: Once the individual objects in the lexicon have been defined, their structure and regularity will be further analyzed to identify more abstract descriptions within the material. Relating these to higher-order structures will result in a lexicon including more conceptual terms such as ‘meeting’, ‘outdoors’, ‘Indonesia’ etc.

This research will result in a set of tools applicable to a complete movie, broadcast, or recording, to construct a hierarchic representation of its content. The project will focus on a *question-answering* application, returning excerpts from an audio-video database to best answer questions posed by users. This interaction will also be used to refine the lexicon. A collaboration with the Preservation Directorate at the Library of Congress will provide content and a user community.

Truly general object definition and extraction, made possible by the statistical correlation of audio, video and text, will result in a fundamental change in the ‘intelligence’ of automatic systems in a very wide range of applications. Machines will have the capacity to be ‘aware’ of events and structure in their environment, and, as our agents, to perform searching and monitoring tasks that at present require tedious and impractical manual review.
1. Introduction

The hallmark of intelligent perceptual systems is their ability to make sense of novel situations. This robust ability to adapt to new environments relies on two key skills: One is the compositional nature of perceptual organization: human observers do not rely on recognizing an entire scene, but can distinguish individual objects within the scene, relating them to previous experience with similar components in other environments. The second skill is the human capacity for perceptual learning, in which exposure to novel objects and situations can result in the construction of new categories with which to represent and understand future experiences.

Contrast this with computer ‘perception’. Despite the capacity to handle large-scale multimedia (audio-video) content, the most successful systems rely on highly constrained domains, such as network news videos or the voice of a known speaker, to limit the range of possible interpretations to permit a useful analysis. Outside these domains, intelligent processing rapidly falls apart.

This project proposes a radically different approach to the problem of computer interpretation of multimedia content: Rather than defining a specific task such as recovering the ground-truth of story boundaries in a news program, our interest, inspired by the characterization of human perception presented above, is in using unsupervised algorithms to discover self-defined classes within the signal content, and then relating these classes to meaningful categories through correlation with text material, and through interaction with human users. The principal lesson we derive from the example of human perception is the importance of scalability – both in terms of covering the greatest range of situations with a finite set of objects, and through the ability to extend that set as a result of exposure to qualitatively novel situations.

1.1 Overview

The goal of this project is to build a system that will take an audio-video signal such as a movie, and analyze it into a representation in terms of abstract classes that are understandable to people. An archive of such representations will then be used to support a question-answering system that selects excerpts from the archive to answer questions posed by users.

The key to this system is the acquisition of objects that will be recognized. In contrast to the approach of starting with a given set of ‘relevant’ objects, the system will define its own categories based on machine-learning applied to the signal content, then learn lexical labels for these classes through statistical correlation with words from associated text or dialog. This will result in a Multimedia Lexicon of terms associated with classifiers defined on the audio-video content.

The Multimedia Lexicon has broad applications in enhancing human's interaction with audio-video material, such as constructing term-based indexes analogous to the those in books. The question-answering application will analyze user-posed queries into the terms of the lexicon and various basic relationships such as conjunctions and sequences.

Combining the modalities of text, sound and vision provides a huge advantage for object extraction. Objects have very different manifestations in these domains, and correlated activity is highly informative. Indistinct sounds can be disambiguated by the accompanying video information, and vice versa; theories of human perceptual bootstrapping typically invoke this principle of cross-
modal integration. Unsupervised machine learning will result in unpredictable, sometimes meaningless, categories, but statistical analysis of the co-occurrence of base-level objects and their correlation with more abstract terms and concepts derived from the analysis of associated text will prune the results and support the construction of higher-order categories such as “party”.

We have selected movies as our exploratory domain. Scalability is a major goal of this work, and thus we deviate from the usual practice of limiting the domain to videos with explicit structure. Movies can contain the full range of real-world situations and events; they also offer a wide range of associated text material, such as screenplays and commentaries from which lexical labels can be derived for the content-derived objects. Subsequent analysis can be applied to material without such text, by reusing the lexical classes already defined.

1.2 Related work

From the large body of work on multimedia content analysis, we highlight the following: In [ClarkSP98, ClarkP99], unsupervised clustering was applied to raw ‘ambulatory’ audio and video, but without an object orientation nor any lexical connection. [NapKFH98] have defined ‘multijects’ as detectors for high-level concepts in audio-video signals; the classes of the Multimedia Lexicon are like self-defined multijects without needing training examples. The Informedia project [WacHGC99] indexes TV news segments with terms from speech recognition, closed captions etc., but does not address general object recognition in multimedia content.

2. Proposed research

The research breaks down into the development of the underlying audio and video features, the construction of the Multimedia Lexicon through category formation including text-derived terms, and the use of the lexicon through the question answering application and other user interactions.

2.1 Object-based feature analysis

Audio features

Auditory organization has been extensively investigated in experimental psychoacoustics, informing us of several key principles that lead to the integration of acoustic energy across time-frequency into separately-perceived objects [Breg90, CookE01]. We will continue our research into computer modeling of these principles, known as computational auditory scene analysis (CASA). Our ‘prediction-driven’ approach [Ellis99] searches for an explanation of the entire sound scene in terms of a limited vocabulary of generic elements.

We will also develop audio processing specific to particular tasks in the object analysis. Statistical modeling of segment characteristics can support the discovery of speaker turn boundaries and other segments [FerrE00]; segments can be broadly classified (e.g. speech vs. music) from model output statistics [WillE99]. Speech will be subject to automatic transcription [RobGE01] and speaker tracking [GenEM99], with music being analyzed along the lines of [TzanC00].

Video features

In video object analysis, there is clear trade-off between accuracy, level of representation, and domain generality. Our prior work has demonstrated algorithms and tools for segmenting video objects, including fully automatic unrestricted segmentation of regions with consistent color, motion and shape features [ChaCMS97], semi-automatic interactive systems for segmenting whole-body video objects (e.g. people, cars) [ZhongC99], and detection of specific objects (e.g., faces [WangC97]). In this project, we will extend these automatic and semi-automatic tools to
generate a rich set of video objects and their associated features which will be used as the foundation for discovery of higher level structures.

**Segmentation**

Dividing the source material into coherent segments is a necessary subtask of this work in order to obtain suitable long-span regions for classification, and to limit the scope of feature clustering and associated text correspondence. As throughout the project, combining information across modalities will afford huge advantages. On the audio-video side, we have been developing the concept of ‘computable scenes’ [SundC00a, SundC00b] which groups shots in a movie on the basis of consistency in visual and acoustic properties. Based on a causal, finite-memory viewer model, we segment the audio and video by determining the local minima of the correlation of the most recent data in the memory with the past, then optimally align the two modalities.

Text from automatic or manual transcripts can be used to produce a topical segmentation. In previous work [KanKM98], we developed a method for discovering a segmental discourse structure of a document, using extraction of noun phrases, pronominal forms and lexical chains of the noun phrases. We also used term distribution to aid in identifying segment function, labelling segments as **summary** (central to the document), **anecdotal** (extraneous to the central point), and **supporting** (providing details). We will extend our approach to work on spoken language, exploring different methods for learning the optimal combination of this text-derived information with the signal-based boundary hypotheses derived from video and audio features.

**2.2 Defining the Multimedia Lexicon**

The Multimedia Lexicon is a set of natural language terms associated with statistical signal classifiers defined in terms of the features of the previous section. It is formed by unsupervised clustering of the object-based features, structure discovery in the clustered domain, correlation of structures with text-derived terms, and a search for further structure among the new definitions.

**Clustering and structure discovery**

We will use standard clustering algorithms such as vector quantization and Gaussian mixture modeling on the feature elements to find robust patterns that recur over a wide range of training material. We will also extend the ideas from a novel histogram-based approach called object frequency–inverse image frequency (of•iif, similar to tf•idf for text), which computes the statistics of dominant objects and their coincidence relationships [PaekSHJ99]. We anticipate large benefits from clustering across the complementary audio and video domains.

While the raw clustering will be applied to simultaneously-present features, the next stage of processing will search for regularities in temporal structure. This includes recognizing mutually-informative fixed-interval correlations (using the tools of [EllisB00]), complex inferences via Bayesian Networks [PaekC00], and hidden Markov models of sequential dependence.

**Lexical correlation**

We will correlate the anonymous self-defined classes with collections of terms derived via natural language processing, in order to find the categories that ‘make sense’. Text sources will include descriptive annotation such as stage directions in a screenplay, narration, screenplay dialog, closed-caption subtitles, and speech recognizer output. We will base term identification on our work to date which uses a combination of statistical metrics to identify words that commonly appear together, along with noun phrase grammars to filter statistical results [SmadjM90,
HatzMMJ99, FungM97]. The problem will be to develop techniques that can identify meaningful sets of terms, filtered linguistically, that correspond to audio-video objects. The result of this stage will be a pruning away of signal-based classes for which no systematic lexical correlates could be identified, leaving the Multimedia Lexicon as a collection of feature-space classes associated with particular natural language terms.

Hierarchic structure

Further processing of the entities defined in the lexicon will search for higher-order structure among this new set of object classes, expanding on our prior work using Belief Networks and probabilistic reasoning for integrating image and text-based classes and classifiers [PaekC00]. Interactions between the various types of classification can be used to improve accuracy of individual classifiers, to infer new categories, to produce different types of categories and to focus the kind of processing needed within an individual classifier. For example, if we know that a video contains people, is indoor, contains question and answer sequences, and occurred as part of a news clip, we can infer that a likely category is a press conference. In our work on text-based image categorization [SableH99], we found that different features supported different types of categories; categorization based on the text of an article yielded a high level topical category (e.g., politics, disaster), while categorization based on the caption of an accompanying image yielded a more visual category (e.g., indoor/outdoor; people/no people). We will explore how high level topical categorizations based on movie transcripts and associated scripts can provide guidance to low level classifiers in finding the kinds of categories that humans find useful.

2.3 Application, interaction and evaluation

Question answering

The result of the processing described above will be a Multimedia Lexicon, realized as a set of audio-video classifiers that can be applied to multimedia content to represent it using the lexicon’s terms. The question-answering application will help us to motivate, evaluate and interact with this analysis process. We will expand our prior work [ChenC00] using a Bayesian relevance feedback framework to facilitate rapid generation of multiple views and templates based on the initial description from users and subsequent interactions. We will borrow from the paradigm of question answering that has been used in the text domain, where systems answer fact-based questions by matching query fragments against text documents [TREC00]. For the proposed project, we must address how matching can be done based on the different media, and the integration of different media and fragments from disparate portions of the movie.

Our starting point will be questions requesting the system to find and/or summarize certain parts of the movie. This will require firstly identifying questions that are beyond the scope of the system (“did the movie have a happy ending?”), then determining how query words can be matched against the movie transcript or script, and corresponding entries in the lexicon. For example, given a query such as “show me the part of the movie where the boy falls into the river,” we may not find any spoken dialogue referring directly to the fall. We can, however, find a segment with similarities to audio and video features of “fall” and “river” in the Multimedia Lexicon. Our approach would be to avoid any real understanding of the answer, instead finding and integrating text from the transcript along with the multiple, relevant portions of the movie. The integration of the portions is crucial in providing an appropriate response and we will draw on our work in summarization of text [McKeKHB99, BarzME99] and video [ChangSBB97].
Iterative refinement

Contact between the automatically-derived Multimedia Lexicon and real users provides a valuable opportunity to improve the lexicon through the incorporation of subjective judgments. By collecting overall ratings of the usefulness or success of each query response, the system can ‘back-propagate’ blame or credit to each of the lexicon definitions used as a basis for that result. Explicit judgments can be sought by presenting the alternative results of competing candidate definitions within the lexicon. This information can be used to tune the existing lexicon, or as a basis for a rebuild that moves the class definitions towards subjective salience.

With the candidate classes and clusters obtained above, we will develop a model-based framework which allows users to take a more active role in defining multi-level hierarchical models for classes in specific domains. In [JaimC99], we have demonstrated a system called Visual Apprentice, which allows users to define hierarchical models for visual scenes (e.g., handshake) and interactively train the model to find optimal features and classifiers for each node in the hierarchy.

Evaluation

The novelty of this work makes developing suitable evaluation metrics a challenging and subtle research problem. Once the lexicon begins to take shape, we can measure, on a disjoint test set, the accuracy of detection for specific terms that it claims to have learned. Since a major part of system ‘performance’ relates to the nature and range of the lexicon terms it has acquired, we will develop metrics for the utility of the lexicon based on term discrimination (how much information each term provides when it is applied to a scene) and a measure of actual utility (does detecting this category actually do anything useful for our anticipated applications?)

We will also participate in developing common standards for multimedia content analysis. The NIST TREC (Text Retrieval Conferences) have this year launched a video documents track, and we will participate in these evaluations (continuing our previous involvement with the spoken documents track [AbbRRE00]). We are also actively involved in the MPEG7 activity [MPEG98].

We have established a partnership with the Preservation Directorate of the Library of Congress to collaborate on this project. This will provide us with access to a very large body of material, as well as providing specific, real application areas and test user groups.

3. Impact

The majority of the project funds will go to supporting six graduate students through their Ph.D. training. Operating as a team that includes the faculty PIs, every participant will be exposed to the full range of topics across the disciplines of signal processing, machine learning, natural language processing and human-computer interaction. We expect to offer one or more related graduate seminar courses (possibly involving our contacts at the Columbia Film School) leading to the development of strong, ongoing links between our three laboratories (in image and video processing, speech and audio recognition, and natural language processing).

The ability to identify and extract the kinds of objects perceived by humans in audio and video signals would enable a whole new generation of automatic systems, ranging from smart searching for audio-video archives, through autonomous monitoring and alarm systems, to a new breed of interactive robots having a true ‘awareness’ of their physical environments, and some understanding of how those environments affect their human interlocutors. While the ultimate consequences of such developments lie far in the future, we are convinced that the development of scalable, cross-modal, signal-to-language Multimedia Lexicon is the best avenue towards their realization.
4. References


Tech. for Digital Libraries.


