Empowering Video Storytellers:
Concept Discovery and Annotation for Large Audio-Video-Text Archives

A large amount of video content is available and stored by broadcasters, libraries and other enterprises. However, the lack of effective and efficient tools for searching and retrieving has prevented video becoming a major value asset. Much work has been done for retrieving video information from constrained structured domains, such as news and sports, but general solutions for unstructured audio-video-language intensive content are missing.

Our goal is to transform the multimedia archive into an organized and searchable asset so that human efforts can focus on a more important aspect - *storytelling*. This project involves a cross-disciplinary effort towards development of new techniques for organizing, summarizing, and question answering over audio-video-language intensive content, such as raw documentary footage. Such domains offer the full range of real-world situations and events, manifested with interaction of rich multimedia information.

The team includes investigators from Schools of Engineering and Applied Science, Journalism, and Arts. It has extensive experience and deep expertise in analysis of multimedia information - video, audio, and language - as well as broad experience with professional documentary film production and interactive media designs.

The project includes two major content partners, WITNESS and WNET, both providing unique video archives, well-defined application domains and user communities. In particular, WITNESS collects and manages a documentary video archive from activists all over the world to support the promotion of human rights. This archive has several important qualities: it is diverse (including interviews, documentation of events, and 'hidden camera' reporting), it is relatively 'raw' (shot by semi-professional operators), it has been shot for specific purposes (e.g. as the basis for a short documentary, or as evidence for a court case), and it has extensive hand-generated annotation logs for a portion of the existing database.

Our main research goals are three-fold. First, given the unique audio-video-text archive, we will develop automatic techniques for constructing a concept-level organization system, a *multimedia concept network* or *multimedia lexicon*, which captures predominant concepts, audio-visual classes, and rich concept-feature relations. Second, we will develop an automatic mechanism to mine for robust, accurate audio-video classifiers, which can be used to assign meanings to new audio-video data. Finally, we will develop a novel Question Answering engine and various multimedia navigation techniques to enable intuitive high-level access to large multimedia archives.

The outcome will include a set of automatic techniques for concept discovery and content annotation, and a variety of applications built on these techniques to provide personalized access to the WITNESS and WNET archives, variously tailored to the needs of documentary filmmakers and more general users such as students, researchers, and lawyers. The results will have impact on broad communities, such as public broadcasters, video producers, and eventually, consumers.

Through the collaboration with WITNESS and WNET, testbeds involving real content and users will be deployed in laboratories at Columbia University as well as at the operational facilities of the partners, for evaluating the new tools and shaping the technical directions. The project also involves the education of an interdisciplinary group of Ph.D. students, from engineering and journalism, efforts toward attracting women and minority students, and training workshops bringing trainers and mentors from the Schools of the Arts and Journalism together with user groups of WITNESS for experimenting with new technologies.

Empowering Video Storytellers:
Concept Discovery and Annotation for Large Audio-Video-Text Archives

1 Introduction

Ever since the inception of audio-video recording, managing archives of this material has been an ongoing problem. The profession with the longest history of video content management is film production, particularly for documentary films where the production process typically involves selecting and editing together material from a fixed but often extensive collection of relevant source material. Expert producers have developed various strategies to help them navigate and manage this material, involving highly specialized skills and very time-consuming processes.

With the advent of inexpensive desktop video production and personal multimedia archives, these issues will take on much broader relevance. It is but one step further to imagine consumers producing their own edited movies based on raw footage they have shot themselves. The impact and power of video compared to other modalities will ensure its primacy among archived content once the technical barriers to collection, storage and access are removed.

This proposal concerns access to large archives of audio-video material. Although this topic has received considerable attention in recent years, there are a number of aspects that make this project distinctive:

- We have partnered with WITNESS and WNET, both providing well-defined application domains and user communities. WITNESS collects and manages a documentary video archive from activists all over the world to support the promotion of human rights. This archive has several important qualities: it is diverse (including interviews, documentation of events, and 'hidden camera' reporting), it is relatively 'raw' (shot by semi-professional operators), it has been shot for specific purposes (e.g. as the basis for a short documentary, or as evidence for a court case), and it has extensive hand-generated annotation logs for a portion of the existing database.
- Our technical approach is centered around the idea of automatically-discovered concepts defined across audio-video content and textual annotations. That is, we propose not only to learn the relationships between concepts and object-related features derived from the soundtrack and image sequences, but the concepts themselves are automatically acquired through analysis of associated text materials. This fully automatic search will find the terms and concepts among a very large set that are both useful in the application as well as being practically derivable from the multimedia content features.
- Based on the derived network of concepts, access to the archive will occur through high-level navigation tools and a unique Question Answering engine, which analyzes natural-language questions and organizes raw multimedia material from the archive with the goal of providing a coherent overall response to the query, rather than simply returning an undigested list of 'hits'.

The outcome of this project will be a set of techniques for automatically discovering concepts relevant to an audio-video database, and systematically labeling new content with those concepts; an access mechanism based on those labels for retrieving items from an archive relevant to specific inquiries; and a variety of applications built on these techniques to provide personalized access to the WITNESS and WNET archives, variously tailored to the needs of documentary filmmakers and more general users such as students, researchers, and lawyers. In addition, these techniques will have impact in a wide range of more-or-less similar domains. We plan to share results with the Visual History Foundation's Shoah project (a very large archive of videos of Holocaust survivors telling their stories). Eventually, as suggested above, content organization techniques of this kind may become commonplace tools to use on the extensive personal archives each of us is likely to amass in the future.

2 The Content Archive and User Community

Our content partners, WITNESS and WNET, contribute two elements that give this project its focus and definition: an *archive* of primary video material, and a *user community* of nonfiction narrative filmmakers, defining the real applications that the research will address.

In particular, the WITNESS archive - over 800 hours of documentary video recorded by human rights activists from all over the world - is in many ways typical of large video archives held by content-production organizations such as broadcasters, documentary filmmakers, advertisers and video hobbyists. For our purposes, its important characteristics include:

- **Content diversity:** Many different kinds of real-world situations, ranging from interviews to demonstrations to exhumations, are portrayed.
- **Medium diversity:** The salient content in the WITNESS material is as likely to be found in the images, the nonspeech soundtrack, or in descriptive annotations, as in the dialog transcripts. None of these media alone gives an adequate description.
- **Raw, semi-professional footage:** Although WITNESS partners receive basic training in camera technique and shooting strategies, the bulk of the archive is raw footage of variable quality, with high relevance to future consumer applications.
- **Continual expansion:** The WITNESS archive is expanding exponentially as the network of partners grows, posing an urgent need for automated management tools.

Much of the activity at WITNESS and WNET concerns the production of short documentary films, a process of nonfiction video storytelling which involves winnowing down a huge pool of source material: a ten-minute piece might draw on 50 hours of original footage. Workflow typically involves manual **logging** of the raw footage, to generate concise textual descriptions of everything available; **select take** listing, in which the logs are used to identify potentially useful segments, organized within **narrative-relevant conceptual groupings**; then various stages of **editing** to reduce these elements to a finished piece, which may involve returning to the original logs to satisfy particular needs that arise during editing.

This project will produce new tools to support these operations. To this end, existing logs and select take lists available from WITNESS constitute invaluable **training data** for the development of our systems. A brief hint of the flavor of this data, which varies enormously in detail level, is provided in the excerpts presented below:

VIDEO LOG: Mental Disability Rights project: Tape 20, Psychiatric Facility

00:02:10	A man is walking around without pants, holding a quilt. An old man is turning his wheelchair. Talking noise in the background. The camera zooms in on the face of a man wearing a hat. Camera pans around and stops on what looks like a restraint.
00:03:01	As the camera pans fast, a few feet appear on the screen. People are walking by.
00:03:04	There are many people sitting or lying on the floor. They are naked or wrapping themselves in quilts. There is the sound of someone moaning. A man standing in front of the camera stretches his hand towards the screen. He looks surprised. …

SELECT TAKE LIST: Mental Disability Rights project

Exteriors	tape 4 06:00: sunset, pretty! tape 6 00:10: front door of facility with sign tape 6 26:20: exterior wall with barbed wire fence
Grim buildings	tape 1 57:10: concrete with tin roof tape 2 41:00: women behind fence tape 5 07:28: outside pen with barbed wire
Nothingness	tape 8 05:15: woman staring into space tape 8 06:40: two women sitting against wall in ward tape 8 13:30: women leaning against wall, one rocking head

…

3 Proposed Research

Our main research goals are three-fold. First, taking the unique audio-video-text corpus, we want to develop automatic techniques for constructing a concept-level organization system, called a *multimedia concept network* or *multimedia lexicon*, which captures predominant concepts, associated audio-visual classes, and rich concept-feature relations. Such automatic capability provides tremendous value to editors and public users in summarizing and accessing large collections of video content. Second, we propose to develop an automatic mechanism to mine for robust, accurate audio-video classifiers, which can be used to assign meaning to new audio-video data. Such utilities will greatly relieve the tedious manual labor of the current annotation process. Finally, to leverage applications of these automatic tools, we will develop a Question Answering system and new multimedia summarization interfaces allowing intuitive high-level access to large multimedia archives.

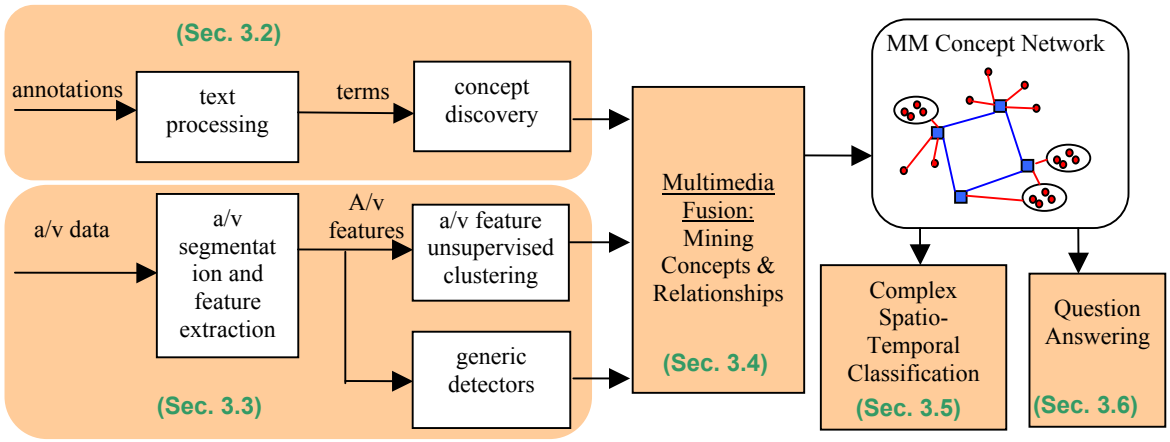


Figure 1. System architecture for Automatic Multimedia Concept Network construction and applications.

The structure of our research program is depicted in Figure 1. The concept discovery process (Section 3.2) takes associated textual information (e.g., video logs and speech transcripts) and uses natural language processing (NLP) techniques, to construct a baseline concept network. Our research will combine statistical and linguistic techniques to extract terms, group semantically related terms, and find concept labels. Audio-visual data (without annotations) are analyzed to extract various features, decompose continuous sequences into logical units, and construct object-based audio-video representations (Section 3.3). "Salient" patterns or classes of audio-video features are automatically discovered through a new method of guided unsupervised clustering and the use of manually selected generic detectors. Fusion of such mid-level audio-video representations with the text-based concepts provides the most interesting avenue of research for discovering new relationships at the semantic level (concept-to-concept), the feature level (class-to-class), and across levels (Section 3.4). The output of this process is a multimedia concept network (termed a concept network in the rest of the proposal), which consists of predominant concepts, audio-visual classes, and rich relations among them. In Section 3.5, spatial-temporal interactions among concepts and classes are explored, using graphical models to enhance the performance of automatic classifiers. Finally, in Section 3.6, a novel application of Question Answering, previously limited to text-based access, will be developed for accessing large multimedia archives.

3.1 Related Work and State of the Art

The Informedia project at CMU [WacHGC99] represents a long-running pioneer effort to index produced TV news videos by integrating technologies from speech recognition, video text

recognition, face recognition, and relatively basic processing of image and language. The Silver project [MyeCSD01] takes video indexes from Informedia system and develops hierarchical, multi-view synchronized editing interfaces. Other systems have also reported promising work in indexing news video [ShahrG95], video from WWW [SriPABP00], presentation video recordings [HeSGG99, UchFGB99], large image libraries [ForMFG96, MaM97], various domains of TV videos and films [DimiMADJ00, ArmaHC93, YeuYWL95, RuiHM98, KobDF97, ChenTABD99, FishLE95].

Several projects have been described in the area of content-based indexing by fusing multimedia - using caption information to annotate objects in news photographs [Sriha95], modeling the joint emission probabilities of text-visual features of images [BarnFor01], unsupervised clustering of raw 'ambulatory' audio and video [ClarkSP98, ClarkP99], and using multi-modality classifiers to detect high-level concepts [NapKFH98, PaekC00a].

Interesting relevant efforts in integrating multimedia information in knowledge representations include the following. The Multimedia Thesaurus project [TanBHLW00], the Mirror system [DooVre00], and Semantic Clustering system [SheCZ98] aim to capture the association between the semantic concept network and image templates or image concepts, obtained by image feature clustering or classification. However, selection of concept networks and concepts for building classifiers is manual. Information fusion is limited to partial text annotations and image features.

Our research differs from all the aforementioned by focusing on automatic concept network construction, automatic mining of concepts for building classifiers, the use of raw footage of documentary and non-fiction films, fusion of deep audio-video-text analysis, and content access through question answering and multimedia summarization.

In addition, we actively seek collaborations with other synergistic projects. We have initiated discussion about collaboration with the team at the Survivors of the Shoah Visual History Foundation, who recently have been granted an ITR project to index a large interview video collection with challenging issues in automatic speech recognition for the multilingual and highly accented speech data. Our work on integration of video, audio (with a distinctive focus on non-speech audio) and language is complementary and will provide significant mutual benefits (see the attached support letter from Dr. Sam Guston).

3.2 Concept Discovery from Text

In the process of developing a documentary, a director works from raw footage (i.e., large quantities of video from which the documentary will be constructed) and video logs, which provide detailed textual descriptions of the sceneries, the actions and dialog in the associated footage. From these two sources, a *select take list* is manually constructed; this associates concepts which the director may choose to illustrate in the film with multiple scenes from the raw footage which illustrate these concepts. In our work, we are proposing to develop tools for automated discovery of concepts (such as "Mexican exteriors," "hospital rooms") and the associated video that illustrates such concepts. We will also provide the facility for the director to specify more abstract concepts for illustration (e.g., "isolation", "depression") and will automatically find the links to such concepts in the video logs and transcripts. These tools will result in a concept network which contains the terms and phrases which refer to such concepts and the audio-visual properties which define them for the input domain.

In this section, we discuss the problem of deriving such concepts from the language in the video archive. This language includes the video logs associated with each video and the automatically derived transcripts of the raw footage. The language processing techniques we propose include clustering of audio-video-text into topically related segments, identification of terms, and discovery of clusters of semantically related terms which, as a group, imply some more general concept. The output of this processing will be a concept network consisting of semantically related terms and phrases which refer to general concepts. Later stages will determine the audio-

visual clusters that illustrate these general concepts and will use fusion of multiple media to refine the concept network.

Clustering by Topic

Rather than viewing the raw footage and associated video logs as one long, unordered collection of video, we will apply clustering technology at many levels to find groups of related video. At the highest level, we will separate the footage into sequences that are related by topic. To do this, we will first apply our previously developed segmentation tools [KanKM98, SundC00a, SundC00b] to divide the video into shorter sequences, each of which represents a single scene Some of these segments may be provided in the input data itself (e.g., segments which have been shot at distinct points in time and labeled with different time stamps).

Next, clustering techniques will be applied over the separate segments to group together segments that are related in topic. We will determine which segments cover the same events by extending technology that we have developed separately for topic detection and tracking [HatziGM00, http://www.cs.Columbia.edu/nlp/newsblaster] and for clustering of a partially annotated image collection [BenSC00]. In our earlier work, we experimented with a variety of features other than the typical bag of words approach. In particular, we did clustering based on terms, experimenting with varied weighting based on part-of-speech category (e.g., noun phrases vs. verb phrases) in order to capture the fact that discussions of different events each involve unique sets of people, organizations and places. We will further experiment with how we can identify unique features of different segments across multiple media in this work (Section 3.4).

NLP Techniques for Concept Discovery and Summarization

Given sets of related video segments and their video logs, we will apply three stages of discovery to the verbal portions of the data. We will first identify terms that are consistently used, followed by identification of semantically related groups of terms, and finally, label the clusters of semantically related terms with a concept name. This will constitute a more general concept of which each term in the semantically related set is an instance. We will stratify the concepts thus learned into those that were derived from dialog within the video, those that refer to what is visually depicted in video (e.g., showing rooms within a mental hospital), and those that label interpretations of the video (e.g. showing poor physical condition of the hospital, illustrating isolation of the patients). These concepts range from concrete to abstract.

We will use an approach to acquisition of terms that integrates syntactic processing, using a grammar of noun phrases, with statistical processing, using metrics that determine the level of cohesion of the words within the terms based on co-occurrence data [HatzMMJ99]. Our approach will have to be adapted depending on source (e.g., the grammar for transcripts will need to accommodate spontaneous speech).

We will experiment with a range of techniques for identifying semantically related terms which together may imply a more abstract concept. Part of our experimentation will measure whether results improve when we use related groups of topically clustered video to search for semantically related terms. Our techniques will include unsupervised techniques to identify semantically related groups of terms based on their function in relation to other words of the sentences in which they appear. For example, two adjectives (e.g., moderate, severe) that tend to modify the same group of nouns (e.g., problems, heart failure), fall into the same semantic scale [HatziMcK93, PereiraTL93]. Note here that the group of adjectives imply the more general concept of “severity”. We will extend this by looking at other linguistic relationships (e.g., noun-verb) for this data. We will also explore the use of linguistic resources such as WordNet [WordNet] to identify words that are synonyms, hypernyms or hyponyms of each other. We will apply and extend our work on paraphrase acquisition [BarzMcK01], which uses co-training techniques on parallel data, to identifying synonyms in topically related clusters of video. Here,

we will investigate learning types of semantic relations other than pure synonyms, which include general/specific, hypernym/hyponym, or semantic scales.

The final step is to label the clusters of learned, semantically related terms. While labeling of clusters is a hard, open problem, we hope to exploit the use of lexical resources such as WordNet to find hypernym relations between at least two members of a cluster and a term in WordNet which represents a more abstract concept. WordNet can also be used to help discriminate between the types of concepts learned. For example, words describing emotional state (e.g., isolation) will be categorized differently than words describing physical locations (e.g., place in a hospital). We will explore the use of other lexical resources over the course of the research, such as FrameNet [Baker98] or Levin's alternations [Levin93] to enhance both the labeling of clusters and discrimination between the types of abstracts concepts learned.

Finally, there will be times when a director wants to specify a particular concept to be illustrated that is not represented in the concept network. To handle this case, we will develop tools that can identify links between user-specified concepts and actual terms used in the video log or video transcript. Such links result from some form of semantic relatedness between the concept and actual terms used in the text annotation. To discover which terms are semantically related to the concept, we will again use a combination of unsupervised statistical methods along with external lexical resources.

3.3 Audio-Video Analysis and Representation

This section briefly describes the initial processing of the audio and video signals that will form the basis of the concept network refinement and classifier construction in section 3.4. Robust automatic recognition of concepts from the audio-video signal depends upon a suitable analysis and representation of those signals. In particular, since many of the concepts we wish to detect will relate to particular objects within the scene, we need feature sets that relate to properties of individual objects (image segments or sound sources), in addition to global scene properties that can relate to more general concepts such as "outdoors", "babble" etc.

Multi-Level Video Decomposition and Feature Extraction

Video content can be characterized at multiple levels: program, scene, shot, object, and region. Much work can be found in the literature about automatic tools for decomposing video into separate units at each level. In this project, we will apply tools developed in our prior work and those available from other groups to process the video data and extract a comprehensive pool of features, which can be optimally selected or combined for the later tasks.

There is a large body of knowledge in literature about video shot and object segmentation. In [Zhong01], we developed a shot segmentation tool that detects shot transitions, dissolves, and flashes in video sequences. We achieved very good accuracy and real-time speed by integrating multiple features (color histogram and bi-directional motion correlation) and performing computations only in the minimally-decoded format. Given individual video shots, we performed automatic region segmentation and tracking [ZhongC97]. Segmentation uses adaptive color clustering, region merging, and edge-based boundary refinement. The tracking process is based on parametric motion projection and color/edge refinement. The output is a set of video regions each of which is associated with various spatial visual attributes (color, size, motion, speed, etc) and a coordinate trajectory over time. In [ZhongC01], we adopted an iterative background layer detection method and a long-term temporal decision rule to automatically detect foreground moving objects, which in turn consist of multiple underlying regions. Figure 2 shows the segmentation result for a dynamic object with rapid complex motion being tracked by panning camera motion over a multi-layer background. When applied to diverse complex scenes, this system achieves encouraging results, particularly in cases that include objects with heterogeneous internal motions and backgrounds with multiple layers. Continuing to improve the system in several areas (such as small objects and object collision), we will use such tools to obtain features

Figure 2. Example of automatic moving object segmentation. The image on the left shows the initial frame. The rest are the segmented objects and their underlying regions (indicated by different colors) at frame 1, 10, 20, and 30.

at the object level. Such a level is important as it typically corresponds the real-world entities (people, cars, etc.), although the extracted attributes may be approximate.

At a higher level, in [SundC00a] we developed a system to group shots into scenes based on domain-tuned production rules, audio-visual feature fusion, and simple models of the viewer's mental state. A video scene typically consists of multiple shots captured at the same location (but from different angles) or related to a consistent thematic concept (journey, party, etc.). The viewer model, similar to the visual coherence model used in [KenYeo00], is based on 'noticing' long-term audio-video coherence or style structure among shots that belong together in a single scene, and we have investigated several computational models for measuring such coherence, including pair-wise aggregative correlation, measures of the mutual information between shots, and models of common 'production styles' governing the ways in which shots are combined into scenes. Our system has shown promising performance in detecting 'true' scene structure in commercial movies; in the current project, these techniques will be adapted to segment raw footage into coherently related sequences, from which scene-level attributes will be extracted.

Audio Object-Based Feature Extraction

The corresponding analysis of audio into elements likely to relate to distinct objects is much less developed than the video methods described above. A key strength of our research group is a deep expertise in audio processing, which we will use to go beyond the basic cepstral coefficients (inherited from speech recognition) that are the most elaborate audio representation used in previous multimedia work e.g. [NapKFH98]. Our work on Computational Auditory Scene Analysis (CASA) – i.e. computational models of the ability of humans to organize sound mixtures [CookE01] – is based on the insight that bottom-up processes alone cannot resolve the ambiguity inherent in noisy, overlapping sounds. Instead, top-down constraints based on an internal 'world model' of hypothesized sound sources must be applied [Ellis99]. By modeling the cues to sound organization revealed by psychoacoustics (e.g. [Breg90]), and incorporating them into hypothesis-search frameworks borrowed from speech recognition, we will address the analysis of sound mixtures as mixtures, rather than trying to classify their aggregate properties with an inevitably inadequate set of composite labels (speech-against-noise, speech-against-babble, speech-against-music etc.). Integrating this source-based organization with missing-data speech recognition can significantly improve the performance of speech recognizers in noisy conditions too, as shown in our results on the Aurora noisy digits task [BarkCE01].

A good example of this source-separation approach to soundtrack analysis is our work on alarm sound detection, in which we compared two approaches, one based on traditional 'global' features (the cepstral coefficients from speech recognition), and the other using sinusoid models to describe prominent harmonics independent of the background noise [Ellis01]. While both systems performed about the same over a range of high noise conditions (0 dB SNR), the system based on global features made numerous false alarm errors in situations where the background noise differed from the training set, while the source-separation model was little affected by such changes, since detection was based only on the properties of the foreground sinusoids.

In contrast to previous work on joint audio-visual indexing [WacHGC99], speech recognition will not be a primary focus in this project. However, we will use ASR transcripts as an additional data source, likely derived from our work in broadcast news recognition [RobGE01]. Our experience

in the TREC Spoken Document Retrieval evaluations shows that even at the high word error rates that might be expected for this material, information retrieval can still perform quite effectively [AbbRRE00].

Other information to be explicitly detected in the soundtrack may include basic speech activity detection [WillE99], language identification (particularly since many videos feature interpreters who switch languages mid-sentence) [Ziss96], and speaker characteristics (such as fundamental frequency mean and variance, timing and pause rate, and average spectral characteristics, to support later speaker tracking within and between shots) [GenEM99]. We will adapt results from our current project into detecting singing segments within musical recordings, which is based on learning the particular patterns of output from a speech recognizer acoustic model [BerenE01]; although the model is quite inadequate to transcribe the speech, it is still sufficiently specialized to the particular properties of speech that singing segments can be detected with an error rate below 20% at the frame level.

3.4 Mining for New Relationships and Classifiers

The research goal at the center of this proposal is to find robust and accurate automatic mechanisms to attach conceptual, lexical annotations to audio-video (AV) content. This may be regarded as an effort to augment or duplicate the function performed by the human annotators at WITNESS who review each video to generate the logs introduced above. Although we cannot hope to derive logs of the accuracy or coherence produced by humans, through the combination of sophisticated object-related feature analysis for both audio and video domains, extensive training examples from WITNESS's existing logs, and identification of concepts in the logs as described in Section 3.2, this project has an unprecedented opportunity for significant and dramatic progress. Previous work in multimedia concept definition has typically been limited to a few hand-picked concepts for which automatic detectors were constructed [NapKFH98 , PaekC00a, SzumP98, VailJZ98]; the current project will be able to "mine" through a large set of candidate concepts and find the ones most successfully detectible via audio-video features. Given the equivocal results of earlier efforts to associate terms with features, we feel this exhaustive mining is the only way to improve our understanding of what can and cannot be detected by these techniques.

The results of the previous two sections will give us (a) a comprehensive collection of object-related and global features describing the AV content of each shot in the raw footage, and (b) a network of concepts derived from the annotations and transcripts, chosen for their importance to the material, and linked to each shot from which they are derived. The problem, then, is to sift through the many-to-many links resulting from these shot-level associations between terms and AV features to identify the strongly informative and robust relationships that can be used as the basis for automatic classifiers. We may also discover associations within the network of terms, and the space of AV features, that are only revealed via their connections to each other.

There are several approaches to the core classifier-construction problem. Using the terms from the NLP analysis as ground-truth labelings of the corresponding AV shots gives us the supervised learning problem of trying to find rules in terms of the content features that best discriminate between segments that do and do not have each term. We will investigate using decision trees [DudaHS01], which make a series of tests on specific feature dimensions, then make further tests on different dimensions depending on the earlier results. Greedy training schemes for growing decision trees find the single most 'informative' test to make at each stage by trying each possibility in turn, and choosing the one that gives the greatest increase in mutual information between tree node and ground-truth label across the training set.

Ideally, given an exhaustive set of audio-video features at various levels, this approach will find the most appropriate feature patterns. However, this ideal can be difficult to achieve, so we will augment it with opportunistic approaches. For instance, in previous work on visual scene

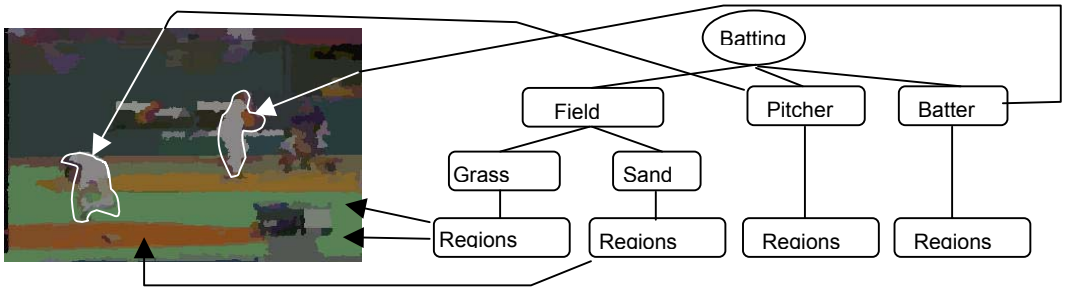


Figure 3. The Visual Apprentice uses manually-tagged image regions to construct individual classifiers as well as learning applicable spatial relationship constraints.

analysis we developed an interactive system called the Visual Apprentice [JaimC99] for users to interactively specify the spatial layout model of specific scenes, e.g., pitching scenes in baseball videos. During the training phase, the user tags the automatically segmented image regions to corresponding nodes in the scene model. Automatic selection of features and classifiers for each node is done by using a wrapper model [Koh95], which avoids exhaustive search in the joint feature-classifier space. We also adopted a best-first forward search method [RusNor95] to find the best subset of features for a given learning algorithm. Several different classifier-learning algorithms (e.g., ID3, Naive-Bayes, IB, MC4) are then compared based on the performance over the tagged training set. Figure 3 shows a segmented region map of baseball pitching image and the corresponding scene model. The best features and classifiers are automatically selected.

The Visual Apprentice exploits the decomposability of visual scenes into elements with strong spatial (or temporal) relationships. The spatial relationships in this case are pitcher (indicated by color, approximate shape, and motion region attributes) located on the left, grass/ground (indicated by color and texture) located in the lower portion, and batter located on the upper right (or in the background). In our previous work, classes were chosen by hand to possess appropriate properties. In this project, we propose to explore an opportunistic approach to automatically mine for such classes. Analysis of the textual annotations can be used to detect references to spatial descriptions, such as "to the right of", or "in front of". We will explore integration of such multimedia cues to determine concepts that may have strong spatial-temporal patterns. Once we discover such concepts, we will apply the Visual Apprentice system to construct detectors, which can then be used for automatic annotation of new data.

Unsupervised Cluster Discovery

A second approach to discovering robustly-supported concept terms is a generalization of the Latent Semantic Analysis used for topic detection in text [DeerDFLH90]. In the multimedia domain, text documents are replaced by AV shots, and the feature vectors describe both the terms associated with that shot via log and transcript text, and classes or attributes derived from the AV features. Singular Value Decomposition of the document-by-feature matrix gives a set of basis vectors composed of terms and features that co-occur or are mutually dependent. This process clusters the annotation terms with their related content-derived attributes.

The identity of the content-based class dimensions is of course crucial to the success of this approach. Rather than being raw feature dimensions such as transform or histogram coefficients, they should be an intermediate or mid-level representation of the content in terms of a range of specific detectors or classes that represent a happy compromise between what can be reliably computed from the low-level features, and what is useful in the detection of higher level concepts [EllisR98]. One source of these attributes, drawing on previous work, is to use hand-built generic detectors, such as "outdoors", "face" or "alarm sound". Using only these categories, however, would limit the range of new classifiers that could be constructed.

A second source of intermediate class definitions is through unsupervised clustering of the content-derived features; the assumption is that if certain patterns of features occur in correlated and systematic patterns, these clusters may be a more useful abstract description of the signal

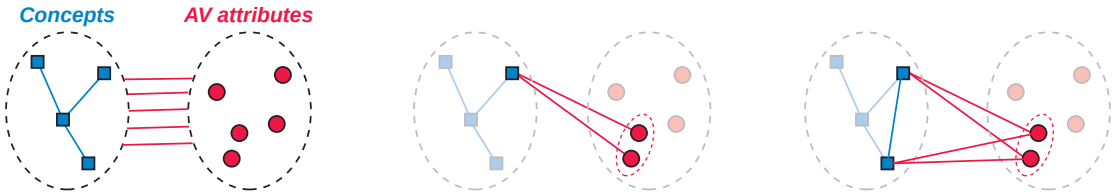
itself. K-means clustering and Gaussian mixture modeling over a combined audio-visual feature domain will be used to generate these data-derived classes.

In [PaekC00a], we used a similar approach to characterize images based on the occurrence frequency of intermediate classes derived from unsupervised clustering of different subsets of image features including color histograms, edge histograms, or textures. Occurrence counts of constituent image regions in each cluster give a frequency-like representation of each image. The weight of each cluster may be further modulated by the "rareness" of each cluster over the entire collection of images. We term such a representation, "object frequency-inverse image frequency" (OF/IIF) by analogy to the term frequency-inverse document frequency approach widely used in information retrieval. We used this representation for problems such as indoor/outdoor classification, and observed an encouraging performance statistically comparable to text-based approaches using annotations.

The danger with unsupervised clustering is that the resulting classes may have no particular relevance or utility in the later concept-definition task. To avoid this danger, we will guide our feature spaces and mid-level clustering methods towards useful classes by developing them to solve supervised problems. As an example, in our current project to cluster singers by perceived vocal similarity, we start with the supervised task of distinguishing between two or more known singers and optimize the features for that task, e.g. by linear techniques (LDA) or by nonlinear transformations (neural networks) [BerenE01]. The labels are then discarded, and unsupervised clustering is performed in this new space, which has been constructed to suppress irrelevant details, while retaining sufficient generality to reveal unanticipated clusterings.

We can use this idea to support the overall concept mining process. A large statistical clustering scheme cannot be guaranteed to find any highly discriminant definitions. If, however, we start with some example concepts for which somewhat reliable content-based detectors can be manually constructed, we can develop our unsupervised scheme until it is able to reproduce or better these token examples. Then, by having constructed a framework of which our manual examples are specific cases, we can expect that many other detectors of comparable or better performance will similarly be discovered. The kinds of concept detectors that we might start with to guide this development are things like "children" (based on vocal characteristics), "outdoors" (based on region structures and color histograms), and "shaky camera" (based on image sequence dynamics) to name a few.

The result of the relationship mining stage will be the identification of the most reliable associations between concepts and AV feature classes (see Figure 4(b)). The primary purpose of this is to generate sets of content-based feature attributes associated with one or a cluster of terms. These definitions can then be used for automatic annotation of new, unannotated video by associating terms whose corresponding feature attributes match the content of the new material. In this way, the reach of the concept-based access methods will extend to the entire video archive.



First-stage processing of audio-video content and text transcripts/ logs generates sets of concepts and AV attributes loosely related through association with the same shot.

(a)

"Link mining" will reveal the terms and supporting AV attributes that are most reliably associated with one another, defining new AV-feature-to-concept detectors.

(b)

By finding different concepts which are mutually dependent on similar sets of AV attributes, new relationships within the concept space may be established.

(c)

Figure 4. Link mining to generate new classifiers and new concept relationships.

As illustrated in figure 4(c), the process of relationship mining can also be used to discover associations between AV attributes that may (separately) correspond to particular concepts, and correspondingly finding new links between nodes in the concept network which are revealed only as common associations with particular AV features. In both these cases, the new associations can be used to define meta-classes, which can then be added to the attribute sets. Re-execution of the relationship mining scheme may then discover more abstract, higher-order relationships, both between concepts and AV features, or even simply within the concept network itself. New concept sets defined in this way can then be given a 'summary' term using the same techniques described in section 3.2.

3.5 Exploring Concept Interactions at Multiple Resolutions

The concepts present in an audio-video program shouldn't be treated as independent entities. Instead, very often there are significant interactions, which can be modeled and exploited using statistical tools, such as Bayesian Networks (BN), Hidden Markov Models (HMM), and Kalman Filters. Interactions among concepts may be static or temporal. Given audio-video data in a given time slice (e.g., an image, a shot, or a scene) and at a given data level (e.g., region, object, or image), there could be multiple related "salient" concepts; for example, a shot may show visuals of indoors and people, audio of speech, and annotations mentioning "meeting". Fusing observations and prior knowledge about related concepts provides great potential for improving the accuracy of individual detectors and the capability of inferring missing concepts. In our prior work [PaekSHJ99, PaekC00a], we investigated such issues in the consumer photograph domain and the news domain by using a BN-based approach. We compiled a test set of more than 3000 images, each with a textual annotation of variable length. The images are obtained from a realistic consumer study group organized by Eastman Kodak. We also included other images downloaded from online news sources. The BN models interactions among concepts, like "indoor/outdoor", "sky", "vegetation", and their classifiers based on visual features or textual features. A performance improvement was found in individual concept classification, particularly when other reliable relevant concepts were present.

Concept interactions exist at multiple resolutions, both spatial and temporal. Within a video scene that has multiple shots, the temporal structural patterns may characterize a specific concept. For example, concepts like dialogs, interviews, or scenes showing repetitive anchoring shots can be best characterized by their temporal patterns. In [SundC00a, SundC00b], we developed approaches based on topological modeling and cyclic self-correlation to detect dialogs and visual anchoring with promising performance (94% precision 86% recall). In [XieCDVS02], we used various HMM-based techniques to detect complex temporal concepts (such as multi-shot play segments in soccer video) that have strong temporal patterns.

In this project, we will further explore other statistical models such as Dynamic Bayesian Networks (DBN), with different variations such as tree-structured HMM, cross-product HMM, or switching state-space models [Gha98]. In these approaches, cross-dependent structures (similar to static BNs) are used to model the interactions among different concepts and multimedia feature observations within the same time slice (a shot or scene). The temporal (causal) interactions are modeled by the dependence links between time slices. Interesting work using similar approaches has been demonstrated in addressing problems such as detecting speaker in a kiosk setting [ChoRPV01] or speech recognition [ZweRus98]. We will expand such systems and our prior work mentioned above to address rich concept interactions in our audio-video-text content archive.

3.6 Question Answering

The result of the processing described above will be a Concept Network, relating acquired terms with their audio-visual counterparts. The question-answering application will critically depend on these results, helping us to evaluate their usefulness in a practical way.

Our approach will borrow from the paradigm of question answering that has been used in the text domain. Textual question answering systems attempt to find the relevant document passages that contain the answer to the user's query instead of returning the set of full documents that are related to the search terms. Such systems typically work by matching fragments of the user query against text contained within the documents returned by a standard retrieval system [TREC00], in some cases applying inference or thesauri to allow more than just exact matches to be found [HaraPM00]. However, question answering as used in the text domain cannot be directly applied here, since the video logs and scripts are sometimes incomplete or missing. Moreover, in many cases, these text resources do not contain a response to a user's query. Our research will determine how matching of query and answer can be done through exploitation of the different media and variations on query terms, both represented in the Concept Network. It will also have to address integration of different media and fragments from disparate portions of the video to form a response.

Our starting point will be questions requesting the system to find and/or summarize certain parts of the video. This will require a study of the kinds of questions that people want answered through access to videos. There are many types of questions that are beyond the scope of the system (e.g., ``At what point did the patients receive enough medication?''), and we will provide a classification of the kinds of questions that we expect to handle.

We will then determine how query words can be matched against video logs, select take lists, scripts, and corresponding entries in the Concept Network. For example, given a query such as "Show me the part of the footage where a nurse is filling a bottle from a faucet", we may not find any direct reference to a "faucet" in the video log, either because it isn't mentioned or because the only reference to this concept does not use the same terminology (e.g. "water tap"). We may, however, be able to find a segment with similarities to video and audio features of "faucet" in the Concept Network. Additionally, we can expand our prior work on concept discovery by identifying terms that have a similar semantic meaning, allowing us, for example, to link ``faucet'' and ``tap'' together.

Select take lists will also be used to improve the performance of our question answering application, since they can provide us with concepts that would typically be difficult to discover without a deep understanding of the scene. If the user looks for a scene "showing evidence of lack of personal space in the hospital", we can infer from the select take list that the scene which is described in the video log as a "man changing clothes in an open cupboard" corresponds to the idea of lacking personal space. Answering this kind of query will exercise the results of our work on concept discovery.

Our approach will be to avoid any real understanding of the answer, instead finding and integrating video logs, scripts, and select take lists along with the multiple and relevant portions of the video that match terms in the query. In addition to finding matching portions of the data, a significant portion of the research will go into integrating these portions to formulate an appropriate response. During the process of integration, we will develop techniques that can identify when information in different segments complements each other and should be combined, when it repeats and should be eliminated, where syntactically cohesive structures exist they should be retained, and where the different media can be used in complementary ways to respond. These issues are all part of the general problem of information fusion and will draw on our work in summarization of text [McKeKHB99, BarzME99] and syntax based video skimming [SundC00a].

4 Applications

The motivation for this research is to enable the creation of a variety of new applications for the users of the video archive. We will develop applications in the following areas:

4.1 Concept-Based Storytelling

As introduced in Section 2, the process of video storytelling essentially involves extracting and arranging key story elements from a pool of source material vastly larger than the finished product. Editors conventionally work with the select take lists which arrange segments identified in the raw footage according to the relevant conceptual groupings which they plan to use in creating their story. Currently, select take lists and the underlying video logs are generated manually, and only organizations with the resources to undertake these huge and laborious tasks can contemplate this kind of production. At the same time, archives with sparse relevance are often ignored even though they may contain a few unique and valuable elements.

The concept-based analysis, organization and retrieval of raw footage described above is the foundation for a system that can support the generation of select take sets from raw material regardless of whether it has been manually logged. Automatic analysis can suggest a set of key concepts (particularly for logged material), and the editor can interact with the query engine to retrieve and refine elements from the raw footage applicable to his or her story.

The system should also be able to browse footage based on associative methods because storytelling is a fluid process, characterized by experimentation and serendipity. After seeing the results for initial concept lists, the producer may augment the concept set, add specialized conditions, or search for 'similar' material. Making overt the previously subjective process of take selection will allow for a richer collaborative environment and process insights.

4.2 User Interfaces and Workflow Management

Critical to the success of the project is the "translation" between engineering capabilities and user needs. On the practical level, these two sides meet in the implementation of the user interface and its implications for the workflow process. Since these technologies impact the core of documentary production, this process will be a delicate compromise between benefits of the new techniques, and the disruption to existing work patterns. In particular, editors who have not reviewed the entire archive during logging need alternative ways such as browsing and summarization to get a 'feel' for the entirety of material available. Our user interface design will respect and accommodate the wide range of working styles and modes of real editors.

Information captured via the workflow user interface may have additional uses. Simply tagging how often and in what contexts a particular segment has been viewed can move towards collaborative-filtering ideas such as identifying the most salient segments of a recording, as well as alerting editors to tangential connections through previous projects.

4.3 Access Mechanisms for Broader Communities

In addition to the filmmakers who are the primary consumers of the archive, WITNESS also has a mandate to make their factual records available for use by other activists and interested parties for many purposes other than film production. An alternative user interface will be developed for this class of users, for instance journalists or human rights advocates engaged in fact-finding. This interface could be used proactively to generate automatic hyperlinks within and between archives for applications in online learning environments, improving and enhancing the value and utility of existing resources.

5 Testbed and Evaluation

It is critical that the development of the technology proceed with strong iterative feedback from the content partners, in order for the tools to have maximum value for producers and editors. The key will be paying attention to how the tools integrate into, and help evolve, workflow processes. To this end, we will deploy realistic testbeds involving the actual use of large video collections and different types of users. The deployment will follow a phased approach, first in the research labs, then in the prototyping environment at the Interactive Design Lab (IDL) at the Journalism School, and finally in the operational facilities of the content partners (WITNESS and WNET).

The Digital Video Lab (headed by PI Chang) has extensive experience in designing and prototyping large scale video testbeds, including an online object-based video search engine for a 30-hour professional video footage archive (http://www.ee.columbia.edu/videoq) [ChaCMS97] and an on-going effort to index 750 tapes of echocardiogram videos representing important cardiologic diseases (http://www.ctr.columbia.edu/dvmm/research.htm#pcdl) [EbadCW01].

For evaluation, on the technical side, we will start with standard, rigorous testing methodologies: recognition and discovery accuracy under various content conditions, speed and complexity of various algorithms, ability to reproduce or outperform hand-designed concept definitions, and the ability to scale up to various types of content. On the user side, the group at IDL will manage the extensive process through careful user studies including:

- Detailed mapping of current workflow processes: Initial work will track representative projects at the content partners, creating detailed models of workflow including team structure, task analysis and measurements of time allocation for personnel in each phase.
- Definition of user requirements: close observations and interviews of system users will identify key criteria for defining successful concepts for development and refinement. This process will track several projects that are representative of the categories of work ongoing at WITNESS and WNET, in order to be able to identify requirements that are generic to all projects and those that are specific to particular domains of content or types of storytelling (e.g. concepts relevant to interviews will be different than those relevant to outdoor action scenes).
- User interface and system integration design and assessment: Effective tool development requires a user-centered approach to interface design and system integration. Close observation of users and user interviews will reveal key requirements and form the basis for assessment and refinement. Interface analysis will focus on "robustness" (correlation of tool's capabilities to superset of task requirements), "efficiency" (time, interaction, and complexity required for task fulfillment), and user satisfaction (qualitative).
- Modified workflow mapping: It is likely that new user behaviors will alter the "ecology" of task management, resulting in significant modification of workflow processes. Mapping these modifications will point to features of "next-generation" workflow processes for video storytelling.

IDL, WITNESS, and WNET will select a diverse sample of projects for analysis. These will be chosen to represent between them a broad range of creative styles and essential content domains of mainstream production at these institutions. Testbeds will be set up at WITNESS and WNET, using prototype video databases loaded with source material for the sample projects.

Field observation and interviews will be conducted on-site by IDL's dedicated Ph.D. student, selected for extensive prior experience with professional documentary video production. Regular "milestone" meetings between content partners and Columbia researchers will review findings and set the agenda for system refinement.

6 Education

Our proposed research will involve education of an interdisciplinary group of Ph.D. students, including four students from electrical engineering, two from computer science and one from journalism. For journalism students, this project will provide new opportunities for research, specifically around the nexus of cutting edge media technologies and innovative storytelling methods and production techniques. For engineering students, this project will provide opportunities for feedback from end users and domain experts in early stages of development to shape the system and in later stages for meaningful evaluation.

We will also provide opportunities for involvement of undergraduates in the research process through independent research for course credit using the organized research liaison program in the Department of Computer Science and the research experience for undergraduate (REU) program in the Department of Electrical Engineering. We will attempt to attract female and minority students both through this program and as Ph.D. students. We feel that the topic material of the project, involving humanitarian efforts and film documentaries relating to these efforts, is more likely to appeal to non-traditional students than more engineering oriented projects.

Finally, due to the participation of WITNESS and IDL, the development of this project will promote reciprocal benefits with educational activities under way at Columbia. Last year, IDL initiated a joint training workshop with WITNESS, bringing trainers and mentors from the School of the Arts' graduate film program together with WITNESS staff and partner human rights Non-Government Organizations (NGOs) for a week-long intensive program in documentary video production and digital editing (see http://www.columbia.edu/cu/news/01/08/witness.html). This program can be expanded to include training and user testing of the developed tools. This would provide a solid next step for tool development and refinement, moving beyond the in-house WITNESS and WNET testbeds. In collaboration with WNET, it will also provide opportunities for training and experimentation with professional documentary production communities and Public Television professionals.

7 Results from Prior NSF Support

In Section 3.3, we have described technologies from prior work of ourselves and others that will be used and extended in this project. In the following we summarize the research in prior funded projects and their relations with this project. In the project, "Generation of Coherent Summaries of Online Documents: Combining Statistical and Symbolic Techniques (IRI-96-18797)," we focus on developing a multi-document summarization system to automatically generate a concise summary by identifying similarities and differences across a set of related documents. Results of this system will be extended for summarizing the multimedia responses from the Question Answering component proposed here.

In the project, "STIMULATE: An Environment for Illustrated Briefing and Follow-up Search over Live Multimedia Information (IRI-96-19124)," we studied approaches to fusing image and textual features for image classification. We studied interactions among manually identified concepts based on a basic Bayesian Network framework, image classifiers using texts of different lengths, with different levels of tagging and parsing, and image classifiers using image features.

In the project, "A Patient Care Digital Library: Personalized Retrieval and Summarization of Multimedia Information (IIS-98-17374)," our research focuses on tailoring search, presentation, and summarization of online medical literature and consumer health information, and multimedia medical data to the end user, whether patient or healthcare provider using the secure online patient records available at Columbia Presbyterian Medical Center (CPMC).

8. References

[AbbRRE00] D. Abberley, S. Renals, T. Robinson, D. Ellis (2000). "The THISL SDR System At TREC-8," *Proc. Eighth Text Retrieval Conference TREC-8*. http://trec.nist.gov/pubs/trec8/papers/shef-proc-trec8.pdf

[ArmaHC93] Farshid Arman, Arding Hsu, and Ming-Yee Chiu., "Feature management for large video databases.," *Proc. SPIE Storage and Retrieval for Image and Video Databases*, 1993.

[Baker98] C.F. Baker, C.J. Fillmore, and J.B. Lowe, "The Berkeley FrameNet Project", in the *Proceedings of COLING98*, Montreal, 1998.

[BarkCE01] J. Barker, M. Cooke and D.P.W. Ellis (2001). "Integrating bottom-up and top-down constraints to achieve robust ASR: The multisource decoder," *Proc. Workshop on Consistent and Reliable Acoustic Cues CRAC-01*, Aalborg, Denmark.

[BarnFor01] K. Barnard and D.A. Forsyth, "Learning the semantics of words and pictures," *Int. Conf. Computer Vision*, 2001

[BarzMcK01] R. Barzilay and K.R. McKeown, Extracting paraphrases from a parallel corpus, in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, July 2001.

[BarzME99] R. Barzilay, K.R. McKeown & M. Elhadad (1999). "Information fusion in the context of multi-document summarization," *Proc. 37th Ann. Mtg. of Assoc. for Comp. Linguistics*, College Park, pp. 550-557.

[BenSC00] A. B. Benitez, J. R. Smith, and S.-F. Chang, "MediaNet: A Multimedia Information Network for Knowledge Representation," *Proceedings of the SPIE 2000 Conference on Internet Multimedia Management Systems* (IS&T/SPIE-2000), Vol. 4210, Boston, MA, Nov 6-8, 2000.

[BerenE01] A. Berenzweig & D. Ellis (2001). "Locating Singing Voice Segments within Music Signals," *Proc. IEEE Workshop on Apps. of Sig. Proc. to Acous. and Audio*, Mohonk.

[Breg90] A.S. Bregman (1990). *Auditory Scene Analysis: the perceptual organization of sound*, MIT Press.

[ChaCMS97] S.-F. Chang, W. Chen, H.J. Meng, H. Sundaram & D. Zhong (1997). "VideoQ - An Automatic Content-Based Video Search System Using Visual Cues," *ACM Multimedia Conference*, Seattle, WA.

[ChenTABD99] J.-Y. Chen, C. Taskiran, A. Albiol, C. A. Bouman, and E. J. Delp, "Vibe: A video indexing and browsing environment," Proceedings of the SPIE Conference on Multimedia Storage and Archiving Systems IV, vol. 3846, September 1999, Boston, MA, pp. 148--164.

[ChoRPV01] Choudhury, T., Rehg, J., Pavlovic, V., and Pentland, A. "The Role of Boosting and Structure Learning in Dynamic Bayesian Networks for Multi-Modal Speaker Detection", Submission to International Conf. on Computer Vision and Pattern Recognition (CVPR01), May 2001.

[ClarkSP98] B. Clarkson, N. Sawhney & A. Pentland (1998). "Auditory Context Awareness via Wearable Computing," *Proc. Perceptual User Interfaces Workshop*, San Francisco.

[ClarkP99] B. Clarkson & A. Pentland (1999). "Unsupervised Clustering of Ambulatory Audio and Video," *Proc. IEEE Int. Conf. on Acous., Speech & Sig. Proc.*,

Phoenix.

[CookE01] M. Cooke, D. Ellis (2001). "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication* (accepted for publication).

[DeerDFLH90] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, "Indexing by Latent Semantic Analysis," *J. Am. Soc. Inf. Sci*. 41(6), pp. 391-407. 1990.

[DimiMADJ00] N. Dimitrova, T. McGee, L. Agnihotri, S. Dagtas, R. Jasinschi, "On Selecting Video Content Analysis and Filtering," Proc. SPIE on Image and Video Databases, San Jose, January 2000.

[DooVre00] M. G. L. M. van Doorn and A. P. de Vries, "The Psychology of Multimedia Databases", *Proceedings of the 5th ACM Conference on Digital Libraries*, San Antonio, TX, USA, June 2-7, 2000.

[DudaHS01] R.O. Duda, P.E. Hart and D.G. Stork, *Pattern Classification (2nd ed.),* Wiley Interscience, 2001.

[EbadCW01] Shahram Ebadollahi, Shih-Fu Chang, Henry Wu, Echocardiogram Video Summarization, SPIE Medical Imaging, San Diego, CA, Feb. 2001.

[Ellis99] D.P.W. Ellis (1999). "Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis, and its application to speech/nonspeech mixtures," *Speech Communications* 27. http://www.icsi.berkeley.edu/~dpwe/research/spcomcasa98/ spcomcasa98.pdf

[Ellis01] D.P.W. Ellis (2001). "Detecting Alarm Sounds," *Proc. Workshop on Consistent and Reliable Acoustic Cues CRAC-01*, Aalborg, Denmark.

[EllisR98] D.P.W. Ellis & D.F. Rosenthal, "Mid-Level representations for Computational Auditory Scene Analysis," in: D.F. Rosenthal & H.G. Okuno (eds.) *Computational Auditory Scene Analysis* (Lawrence Erlbaum, Mahwah), pp. 257-272, 1998.

[FerrE00] J. Ferreiros & D.P.W. Ellis (2000). "Using acoustic condition clustering to improve acoustic change detection on Broadcast News," *Proc. ICSLP-2000*, Beijing. ftp://ftp.icsi.berkeley.edu/pub/speech/papers/icslp00-acd.pdf

[FishLE95] Stephan Fischer, Rainer Lienhart and Wolfgang Effelsberg, Automatic Recogition of Film Genres, Proc. ACM Multimedia 95, San Francisco, CA, Nov. 1995, pp. 295-304.

[ForMFG96] D. Forsyth, J. Malik, M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson and C. Bregler, "Finding Pictures of Objects in Large Collections of Images" Int. Workshop on Object Recognition for Computer Vision,April 13-14 1996, Cambridge, England.

[GenEM99] D. Genoud, D. Ellis & N. Morgan (1999). "Combined Speech And Speaker Recognition With Speaker-adapted Connectionist Models," *Proc. Int. Works. on Auto. Speech Recog. and Underst. ASRU'99*, pp. 177-180. ftp://ftp.icsi.berkeley.edu/pub/speech/papers/asru99-spsp.pdf

[Gha98] Ghahramani, Z. "Learning Dynamic Bayesian Networks " In C.L. Giles and M. Gori (eds.), Adaptive Processing of Sequences and Data Structures . Lecture Notes in Artificial Intelligence, 168-197. (1998) Berlin: Springer-Verlag.

[HaraPM00] S. Harabagiu, M. Pasca and S. Maiorano, "Experiments with open-domain textual question answering," *Proc. COLING-2000*, Saabruken, August 2000.

[HatziGM00] V. Hatzivassiloglou, L. Gravano, and A. Maganti, An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-00)*, Athens, Greece, July 2000.

[HatziMcK93] V. Hatzivassiloglou and K.R. McKeown, Towards the automatic identification of adjectival scales: clustering of adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Oh, June 1993.

[HatzMMJ99] V. Hatzivassiloglou, O. Merport, K.R. McKeown, D.A. Jordan (1999). "Extracting patient profiles from patient records and online literature," *Proc. AMIA 1999 Annual Symp*., Washington.

[HeSGG99] Liwei He, Elizabeth Sanocki, Anoop Gupta, and Jonathan Grudin, "Auto-Summarization of Audio-Video Presentations," ACM Multimedia Conference, Orlando, FL, Nov. 1999.

[JaimC99] A. Jaimes & S.-F. Chang (1999). "Model Based Image Classification for Content-Based Retrieval," *SPIE Conference on Storage and Retrieval for Image and Video Databases*, San Jose.

[KanKM98] M.-Y. Kan, J.L. Klavens & K. McKeown (1998). "Linear segmentation and segment relevance," *Proc. 6th Int. Workshop of Very Large Corpora WVLC-6*, Montreal, pp. 197-205.

[KobDF97] Kobla, V., Doermann, D., and Faloutsos, D., Video Trails: Representing and Visualizing Structure in Video Sequences, ACM Multimedia 97, Seattle, WA, Nov. 1997.

[Koh95] R. Kohavi, "Feature Subset Selection Using the Wrapper Model: Overfitting and Dynamic Search Space Topology," in *proceedings of First International Conference on Knowledge Discovery and Data Mining*, pages 192-197, 1995.

[Levin93] B. Levin, *English Verb Classes and Alternations*, University of Chicago Press, Chicago, 1993.

[MaM97] W. Y. Ma and B. S. Manjunath, "NETRA: A toolbox for navigating large image databases," IEEE International Conference on Image Processing, Santa Barbara, California, October1997.

[McKeKHB99] K.R. McKeown, J.L. Klavens, V. Hatzivassiloglou, R. Barzilay & E. Eskin (1999). "Towards multidocument summarization by reformulation: Progress and prospects," *Proc. 17th Nat. Conf. on Artif. Intel. AAAI-99*, Orlando, pp. 453-460.

[MyeCSD01] Brad A. Myers, Juan P. Casares, Scott Stevens, Laura Dabbish, Dan Yocum, Albert Corbett, "A Multi-View Intelligent Editor for Digital Video Libraries." The First ACM+IEEE Joint Conference on Digital Libraries, JCDL'01, June 24-28, 2001, Roanoke, VA. pp. 106-115.

[NapKFH98] M.R. Naphade, T.T. Kristjansson, B.J. Frey, and T.S. Huang (1998). "Probabilistic Multimedia Objects (Multijects): A Novel Approach to Video Indexing and Retrieval in Multimedia Systems," *Proc. IEEE Int. Conf. on Image Proc. ICIP-98*, Chicago.
http://www.ifp.uiuc.edu/~trausti/Papers/2029.ps

[PaekC00a] S. Paek & S.-F. Chang (2000), "A knowledge engineering approach for image classification based on probabilistic reasoning systems," *IEEE Int. Conf. on Multimedia and Expo*, New York.

[PaekSHJ99] S.Paek, C.L. Sable, V. Hatzivassiloglou, A. Jaimes, B.H. Shiffman, S.-F. Chang, K.R. McKeown (1999). "Integration of Visual and Text-Based Approaches for the Content Labeling and Classification of Photographs," Proc. ACM-SIGIR'99, Berkeley CA. http://www.cs.columbia.edu/~sable/research/sigir99.ps

[PereiraTL93] F. Pereira, N. Tishby and L. Lee, Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Oh, June 1993.

[RobGE01] A.J. Robinson, G.D. Cook, D.P.W. Ellis, E. Fosler-Lussier, S.J. Renals & D.A.G. Williams (2001). "Connectionist Speech Recognition of Broadcast News." *Speech Communication* (accepted for publication).

[RuiHM98] Yong Rui, Thomas S. Huang and Sharad Mehrotra, "Constructing Table of Contents for Videos," ACM J. of Multimedia Systems, 1998.

[RusNor95] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall series in artificial intelligence, Prentice Hall, Englewood Cliffs, N.J., 1995.

[ShahrG95] B. Shahraray and D. C. Gibbon, "Automatic Generation of Pictorial Transcript of Video Programs," SPIE Vol. 2417, pp.512-518, 1995.

[SheCZ98] G. Sheikholeslami, W. Chang, and A. Zhang, "Semantic Clustering and Querying on Heterogeneous Features for Visual Data, " *Proceeding of the ACM International Multimedia Conference (ACMM-1996)*, Bristol, MA, USA, Sep. 12-16 1998.

[Sriha95] R. K. Srihari, "Automatic Indexing and Content-Based Retrieval of Captioned Images", IEEE Computer Magazine, Sep. 1995, Vol 28, No 9, pp. 49-58.

[SriPABP00] S. Srinivasan, D. Ponceleon, A. Amir, B. Blanchard, D. Petkovic, "Engineering the Web for Multimedia", in Web Engineering workshop (WEBE), WWW-9, Amsterdam, May 2000.

[SundC00a] H. Sundaram & S.-F. Chang (2000). "Video Scene Segmentation Using Video and Audio Features," *IEEE Int. Conf. on Multimedia and Expo*, New York.

[SundC00b] H. Sundaram & S.-F. Chang (2000). "Computable Scene and Program Level Structures in Films using Audio-Visual Memory Models," *ACM Multimedia Conference*, Los Angeles.

[SzumP98] M. Szummer and R. Picard, "Indoor-Outdoor Image Classification," IEEE International Workshop on Content-Based Access of Image and Video Databases CAIVD '98, Bombay, India, Jan. 1998.

[TanBHLW00] R. Tansley, C. Bird, W. Hall, P. Lewis, and M. Weal, "Automatic Linking of Content and Concept", *Proceeding of the ACM International Multimedia Conference and Exhibition (ACM-2000)*, Los Angeles, CA, USA, Oct./Nov. 30-4, 2000.

[TREC00] TREC Editors (2000). "Proceedings of the eighth text retrieval conference (TREC-8)," National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec8/t8_proceedings.html

[UchFGB99] Shingo Uchihashi, Jonathan Foote, Andreas Girgensohn, John Boreczky, "Video Manga: Generating Semantically Meaningful Video Summaries," ACM Multimedia Conference, Orlando, FL, Nov. 1999.

[VailJZ98] A. Vailaya, A. Jain and H.J. Zhang, "On Image Classification: City vs. Landscape", IEEE Workshop on Content-Based Access of Image and Video Libraries, June 21, 1998, Santa Barbara, California.

[WacHGC99] H. Wactlar, A. Hauptmann, Y. Gong & M. Christel (1999). "Lessons learned from the creation and deployment of a terabyte digital video library," *IEEE Computer* 32(2), pp. 66-73. http://www.informedia.cs.cmu.edu/

[WillE99] G. Williams & D. Ellis (1999). "Speech/music discrimination based on posterior probability features", *Proc. Eurospeech-99*, Budapest.

[WordNet] C. Fellbaum, G.A. Miller et al., "WordNet: A lexical database for the English language," http://www.cogsci.princeton.edu/~wn/ .

[XieCDVS02] L. Xie, S.-F. Chang, A. Divakaran, A. Vetro, H. Sun, "Structure Analysis of Soccer Video with Hidden Markov Models," to appear in IEEE ICASSP, Orlando, FL, May 2002.

[YeuYWL95] M.M. Yeung, B.-L. Yeo, W. Wolf, and Bede Liu, "Video Browsing using Clustering and Scene Transitions on Compressed Sequences," IS&T/SPIE Symposium Proceedings, Feb. 1995, San Jose, California. Vol. 2417, pp. 399-413.

[ZhongC97] D. Zhong and S.-F. Chang, Video Object Model and Segmentation for Content-Based Video Indexing, IEEE International Symposium on Circuits and Systems (ISCAS'97), Hong Kong, June 1997, Special Session on Networked Multimedia Technology and Application.

[ZhongC99] D. Zhong & S.-F. Chang (1999). "An Integrated System for Content-Based Video Object Segmentation and Retrieval," *IEEE Tr. Circuits and Systems for Video Technology* 9(8), pp.1259-1268.

[ZhongC01] D. Zhong and S.-F. Chang, Long-Term Moving Object Segmentation and Tracking Using Spatio-Temporal Consistency, IEEE International Conference on Image Processing, Greece, Oct. 2001.

[Zhong01] Di Zhong, Segmentation, Index and Summarization of Digital Video Content, Ph.D. Thesis, Columbia University, 2001.

[Ziss96] M. Zissman (1996). "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Transactions on Speech and Audio Processing* , 4(1).

[ZweRus98] G. Zweig and S. Russell, "Speech Recognition with Dynamic Bayesian Networks," In Proc. AAAI-98, Madison, Wisconsin: AAAI Press, 1998.

Management Plan

This project represents an exciting and ambitious collaboration among four research laboratories in different schools at Columbia University and two major content organizations –

- Digital Video/Multimedia (DVMM) Laboratory (http://www.ee.columbia.edu/dvmm),
- Laboratory for Recognition and Organization of Speech and Audio (LabROSA) (http://labrosa.ee.columbia.edu/),
- Natural Language Processing (NLP) Laboratory (http://www.cs.columbia.edu/nlp),
- Interactive Design (IDL) Laboratory (http://www.columbia.edu/idl),
- WITNESS – a well-known human rights promotion organization (http://ww.witness.com), and
- Thirteen/WNET New York- the flagship public broadcaster in New York, New Jersey, and Connecticut metro area.

Successful management of the project is ensured by the long-standing fruitful collaboration among the groups, well-defined responsibilities, and careful plans for different phases of the project.

Long-Standing Collaboration Among All Partners

Located in the same research building, the DVMM, LabROSA and NLP labs have a long tradition of fruitful collaboration, including joint projects, seminar series, and study groups. The IDL team has had extensive collaboration with the technology research counterparts through several new media initiatives at Columbia University. The relationships with the content partners have been very strong also. In 2000, WITNESS and IDL held a successful user training workshop, bringing trainers and mentors from the School of the Arts' graduate film program together with WITNESS staff and partner human rights NGOs for a week-long intensive program in documentary video production and digital editing. We will capitalize on these existing multi-link relationships to undertake various tasks in different phases of this project.

Responsibilities:

The groups in DVMM, LabROSA, and NLP will be responsible for research and development of new theories, algorithms, and tools for new technologies presented in the proposal description. These three groups provide comprehensive expertise in tackling challenging issues in multimedia analysis, recognition, summarization, and access. Two Ph.D. students will be dedicated to research tasks in each of the three areas: video, audio, and language.

The IDL lab, founded in 1999, is a joint initiative of Columbia's Graduate School of Journalism and School of the Arts for advanced research and development of interactive media content. The group will play a major role in managing the extensive processes of user studies and system integration designs. They will also play the role of "liaison" between the technology research groups and the content partners at WITNESS and WNET. The co-directors of the lab and one dedicated Ph.D. student will be responsible for tasks in this area.

Multi-Phase Milestones:

The initial phase of the project will focus on analysis of diverse samples of content, users, and workflow processes, through construction and testing of a manually selected system. This involves the following.

- Acquire and analyze a sample content set (about 30 hours of video) including a diverse sample of content and projects
- Map and understand current user workflow processes
- Manually construct a concept network and map the initial content to the concept network
- Conduct user studies to assess user's interaction with the manually constructed system

The 2nd phase will take the feedback and experience from Phase 1 and focus on solving the major research issues in (1) automatic construction of concept network, (2) mining of concepts for automatic annotation, and (3) evaluation of the automatic system. The work will involve

- development of techniques for concept discovery and summarization
- development of new audio-video annotation tools
- systematic evaluation of the new tools and systems

The outcome in Phase 2 will include a baseline system using fully automatic processes and performance evaluation results of such system and tools.

In Phase 3, the baseline system will undergo iterative processes of refinement and be tested against a larger set of content (about 300 hours). The refined system will also be deployed in the IDL lab for actual user studies and integration in a modified workflow process for video production. In addition, we will expand the system with other advanced features, such as question answering functionality, concept navigation and search interfaces, and multimedia summarization interfaces.

The last phase (Phase 4) involves integration and testing of the system at WITNESS and WNET, using video databases loaded with source material for the sample production projects. We will also conduct users training workshops, similar to the one we did in 2000 (for more info, see http://www.columbia.edu/cu/news/01/08/witness.html), to bring trainers and mentors from the School of the Arts' graduate film program together with WITNESS staff and partner human rights groups for an intensive program in next-generation video storytelling. This would provide a solid next step for tool development and refinement, moving beyond the in-house WITNESS and WNET testbeds.

Over the course of the project, regular "milestone" meetings between content partners and Columbia researchers will review findings and set the agenda for technology refinement, system integration, and user study experiments.

List of all Personnel Associated with the Proposal

Shih-Fu Chang, Columbia University
Daniel P Ellis, Columbia University
Kathleen R McKeown, Columbia University
Andrew Lih, Columbia University
John Kelly, Columbia University
Gillian Caldwell, WITNESS
Anthony Chapman, Thirteen/WNET New York

WITNESS
353 Broadway
New York NY 10013

www.witness.org
witness@witness.org
Tel (212) 274 1664
Fax (212) 274 1262

November 5, 2001

Prof. Shih-Fu Chang
Assistant Professor, Electrical Engineering
Columbia University
530 West 120th Street
New York, NY 10027

Dear Prof Shih-Fu Chang:

This letter is to express our enthusiastic support for a collaboration between WITNESS and Columbia University to research and develop a next-generation search and retrieval system for large video archives. We are very excited about your proposal, "Empowering video storytellers: Concept discovery and annotation for large-video-text archives," for funding under the ITR program of NSF. We are honored to have been approached by the very well-renowned team at Columbia, and see this as a tremendously important opportunity for our organization.

As you know, WITNESS (www.witness.org) advances human rights advocacy through the use of video and communications technology. In partnership with non-governmental organizations and activists, WITNESS strengthens grassroots movements for change by providing video technology and assisting its partners to use video as evidence before courts and the United Nations, as a tool for public education, and as a deterrent to further abuse. WITNESS also gives local groups a global voice by distributing their video to the media and on the Internet, and by helping to educate and activate an international audience around their causes. We want to empower people use digital video record injustice as it happens and tell stories about human rights issues to a wide audience.

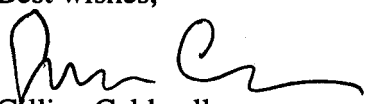
Since our inception in 1992, WITNESS has gathered a unique archive of over 800 hours of videotape, shot by human rights defenders in over 50 countries around the world. At the moment, all our videotape is stored on VHS analog tapes, and logged or transcribed on word documents. This method of functioning severely limits the effectiveness of our work. Producers have to painstakingly go through the large archive to find usable footage and select clips when formulating and editing the story. A searchable and digitized archive will be more accessible to expert users and, - through the Internet - to our NGO partners who could download and edit imagery from remote locations. This would accelerate and increase access to footage for partner groups in remote locations, and once remote uplinks were created, it would allow WITNESS to receive material from partner groups faster. It could also enable a broader public easier and less expensive access to archival materials. We currently receive and respond to numerous requests for archival material every week from student, press, and other human rights organizations and our process for doing so is encumbered by our inability to direct people to an online

database of digitized imagery. It also requires that institutional memory be developed within individuals, rather than in a searchable system that will outlast an individual's tenure at our organization.

Additionally, in the long term we need to safeguard our archive by creating digital backups that can be secured in an offsite server. Much of our older footage, from as long as nine years ago, is –time-sensitive, degradable stock. The older material was recorded on Hi-8 tape that becomes increasingly fragile with time; it is delicate and irreplaceable. If we can create digital backup for this material, we can be confident that it will be safe and preserved for the historical record. The World Trade Center crisis and the ruin in downtown Manhattan, only blocks away from our office, has reinforced and accelerated our concern to make digitized back-ups immediately, to be stored off-site.

We very much look forward to collaborating with Columbia University to develop these much needed new technologies, which will advance our work and the work of many organizations around the world who are still managing their video archives manually and without the benefit that new technologies can offer. We already have close collaboration with the Interactive Design Lab at the Journalism School of Columbia University. We are very excited about expanding this fruitful relationship. We will offer our full cooperation to this important endeavor.

Best wishes,



Gillian Caldwell
Executive Director

Enc.

thirteen
WNET NEW YORK

450 West 33rd Street
New York NY 10001
tel: 212.560.1313
fax: 212.560.1314
www.thirteen.org

November 9, 2001

Professor Shih-Fu Chang
Department of Electrical Engineering
Columbia University

Dear Professor Chang

We at Thirteen/WNET New York offer our strong support of your research initiative for developing advanced search and retrieval methods for video databases. Indeed, we think that realizing the vision outlined in your proposal, "Empowering video storytellers: Concept discovery and annotation for large-video-text archives," would prove extremely valuable for us, and indeed for the Public Television community in general. We are eager to collaborate with you as your work progresses, providing feedback that will help you ensure the tool's value to content creators.

For nearly four decades, Thirteen/WNET New York has been dedicated to the idea that television can be a consistently positive force in people's lives. A not-for-profit public service institution, Thirteen is a leader in television production, broadcast, and interactive media. One of the key program providers for the Public Broadcasting Service (PBS), Thirteen brings such acclaimed and award-winning series as NATURE, GREAT PERFORMANCES, AMERICAN MASTERS, RELIGION & ETHICS NEWSWEEKLY, WORLD NEWS FOR PUBLIC TELEVISION, and CHARLIE ROSE -- as well as the work of Bill Moyers -- to audiences nationwide. Our Interactive & Broadband group produces first-rate interactive experiences that enhance the value of our on-air programming, and our commitment to using new technologies to further the objectives of Public Television is strong.

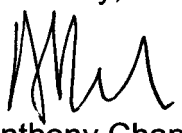
There are a number of ways in which advanced tools for searching video databases would benefit our mission. As a center for producers of television content, including numerous high-quality documentaries such as Ric Burns' NEW YORK, we can attest that an enormous amount of work often goes into combing through vast archives of film and video. Hours upon hours of time are spent finding minutes or seconds of footage. This process is at best frustrating and expensive, and at worst, prohibits the exploitation of culturally valuable material because the cost of finding it is simply too high. A tool that could help producers "see" into hundreds of hours of poorly catalogued video would be of inestimable

value, allowing the production of many compelling shows that simply would not be produced otherwise.

We also see great potential in the long run for such tools to make our storehouse of finished programming valuable in new ways, and to new audiences. The ability to make a vast collection of historical documentaries into a rich reference resource, or for an opera student to find and view multiple performances of an aria, could give programming a life beyond broadcast that would be of extraordinary value to the public.

We believe the efforts of your research group can help bring these tantalizing visions of the future closer to reality, and we heartily endorse and look forward to cooperating in this project.

Sincerely,



Anthony Chapman
Director, Interactive & Broadband
Thirteen/WNET New York

Dear Professor Chang

I am writing to express my support of your proposed project, "Empowering Video Storytellers", to develop automatic annotation and access tools for the WITNESS human rights video archive. I am very encouraged that you and your colleagues will be contributing new ideas and techniques to make large video archives more useful and broadly accessible through technology. I am also pleased that you are collaborating with WITNESS, an organization that we have worked with too, and whose goals and achievements I particularly admire.

As you know, the Survivors of the Shoah Visual History Foundation was set up by Steven Spielberg in 1994 to collect first-hand accounts of the Holocaust from survivors and others. By 1998, we had collected testimonies from more than 50,000 eyewitnesses comprising around 120,000 hours of recorded video. Since our mission is to make this material available for educational purposes globally, we have been very active in investigating new technologies such as broadband distribution, and computer support for our cataloguers who perform the critical step of indexing each recorded interview to allow later retrieval.

At the Visual History Foundation, we have been fortunate enough to receive a grant from the National Science Foundation to support a research collaboration for advanced automatic indexing techniques. This project is focused on automatic speech recognition for the multilingual and often highly accented speech in our archive, as well as various other archive management tools. We will be happy to share parts of this work with you where they have applicability to your task (barring intellectual property issues with our partners).

I am excited to see that your project is centered around areas that we were unable to address, such as analyzing the visual and nonspeech content of the archives. The WITNESS collection is ideal for this kind of work, being a far more visually active format than the Foundation's oral testimonies which have a person sitting in a chair for most of the interview. WITNESS's active involvement in documentary film production also gives your project a somewhat different and very specific set of user needs to work with.

Although VHF is having human cataloguers describe our entire archive, there will always be second-tier material where the expense of manual annotation cannot be justified. There are visual aspects to our archive that will not be described by the Automatic Speech Recognition or human catalogers, that we hope you develop.

Given the impressive caliber and reputation of you and your collaborators, I am confident that you will generate real, practical results for your partners and WITNESS. I am also very pleased and grateful that you have agreed to share your applicable results with us, and impatiently anticipate the new abilities you will bring to documentary archive management technology!

Sam Gustman
Executive Director of Technology
Survivors of the Shoah Visual History Foundation