

# Multimedia Content Analysis: The Next Wave

Nevenka Dimitrova

Philips Research, 345 Scarborough Rd.,  
Briarcliff Manor, NY 10510, USA  
Nevenka.Dimitrova@Philips.com

**Abstract.** We have witnessed a decade of exploding research interest in multimedia content analysis. The goal of content analysis has been to derive automatic methods for high-level description and annotation. In this paper we will summarize the main research topics in this area and state some assumptions that we have been using all along. We will also postulate the main future trends including usage of long term memory, context, dynamic processing, evolvable generalized detectors and user aspects.

## 1 Introduction

After a decade of exploding interest in the multimedia content analysis and retrieval [1,6,8,15], there has been enough research momentum generated that we can finally reflect on the overall progress. The goal has been to develop automatic analysis techniques for deriving high level descriptions and annotations, as well as coming up with realistic applications in pursuit of the killer application. Meanwhile the MPEG-7 has standardized the description of metadata – data describing information in the content at various levels. Applications range from home media library organization that contains volumes of personal video, audio and images, multimedia lectures archive, content navigation for broadcast TV and video on demand content. The tools have emerged from traditional image processing and computer vision, audio analysis and processing, and information retrieval.

In this paper we will first present an overview of the active research areas in video content analysis in Section 2. Next, we will make high level observations about the current practices in Section 3. Section 4 will attempt to provide future directions. Section 5 will conclude the paper.

## 2 Active Research Areas

In Figure 1 we show a conceptual pyramid where the sides represent visual, audio, auxiliary (e.g. data provided by content creator) data and textual processing. The features computed range from low level – closer to the base of the pyramid to the high level semantics – closer to the top. Although there is variability in terminology,

we have seen that the algorithms can be largely categorized in “detectors,” intermediate descriptors such as genre, structure, event, and affective descriptors and high level “abstractors.” Detectors in turn can be very basic, for example face detection, video-text detection as well as complex: for example anchor detector based on face detector and shot classifier. The high level abstractors reveal the essence of the underlying content. For example, a summary will contain the essential content elements in a condensed form with less data. Examples of these features will be given in the following subsections.

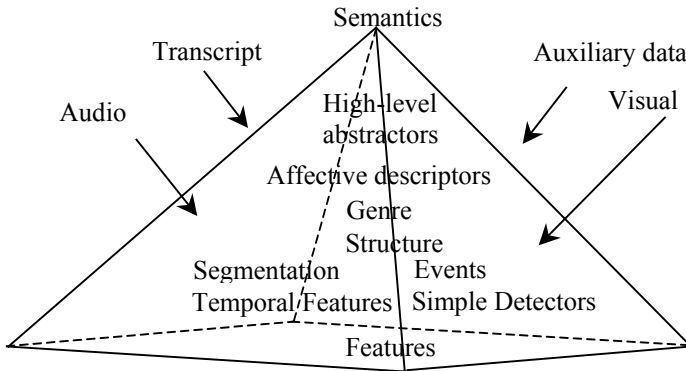
**2.1 Object Detection Algorithms**

These detectors bring important semantic information in the video content analysis and indexing. Basic object detectors in video include videotext detection and face detection. At the generic level both have to first delineate the desired object from the “background” via simple image processing operators such as color, edge, and shape extractors and then apply area filters in order to focus on finding the desired shape.

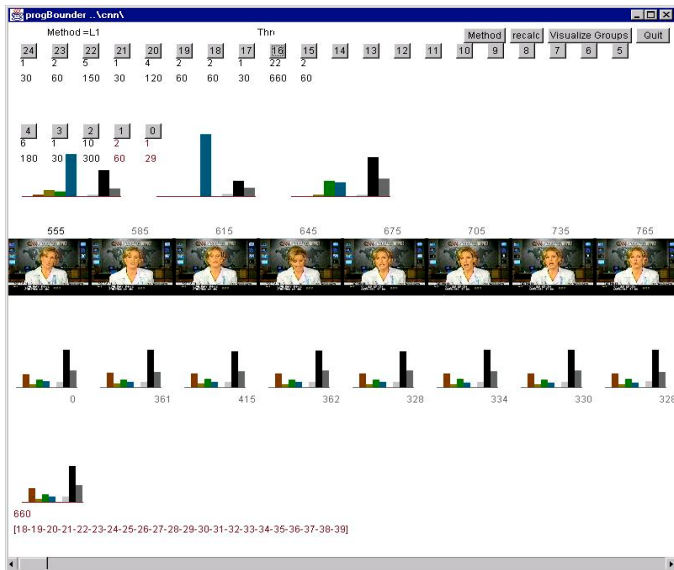
Textual information brings important semantic clues in video content analysis such as name plates, beginning and ending credits, reporter names, etc. We investigated a method for detection and representation of text in video segments. The method consists of seven steps: Channel Separation, Image Enhancement, Edge Detection, Edge Filtering, Character Detection, Text Box Detection, and Text Line Detection [1]. Our results show that this method can be applied to English as well as non-English text (such as Korean) with precision and recall of 85%.

**2.2 Computing Scenes: Micro vs. Macro Boundaries**

Temporal boundary detection was initially a very active area of research [9]. The temporal segmentation referred to shot boundary detection. However, boundaries are also detected at the scene level as well as the program structure level. We can think of these boundaries as: micro (e.g. shots), macro (e.g. scene) and mega (program structure) boundaries.



**Fig. 1.** The video research pyramid



**Fig. 2.** Superhistogram representation of a news video program

Micro boundaries are associated to the smallest video units -- video microunits -- for which a given attribute is constant or slowly varying. The attribute can be any feature in the visual, audio, or text domain. Macro boundaries delineate collections of video micro segments that are clearly identifiable, organic part of an event defining a structural (action) or thematic (story) unit. Mega boundaries delineate collections of macro segments which exhibit a structural and feature (e.g. audio-visual) consistency.

Note that although in the literature it is well accepted that scenes comprise of one or more shots, there are complete movies or long sections of movies that defy this rule. Instead, a single shot consists of multiple scenes. Complete movies such as Hitchcock's "Rope" and consumer home video comprise of seemingly a single shot. Each scene boundary within the movie would represent a micro-segment -- while multiple scenes can comprise a macro-segment.

We investigated different methods for computing super-histograms for color representation of micro and macro segments (see Figure 2). We build cumulative histograms for video shots and scenes. A video segment can be represented with the color histograms of its most dominant scenes. The superhistograms representing episodes of the same sitcom look strikingly similar, while TV news superhistograms are not similar to the ones from sitcoms. This method can be used for video classification and retrieval in studio archival, digital libraries, authoring tools, and web crawling.

### 2.3 Structure and Classification

The first notable step in structure and classification is to detect the non-program segments such as commercials and future program announcements. Furthermore, after the commercial segments have been isolated, the inner structure of the program can

be recovered. Video programs such as news, talk shows, game shows, sports programs have an internal structure and detectable well-defined format. This provides transparency of the program content and gives users direct overview and access to meaningful modules of the program. To this end, we developed a multimodal analysis system, called Video Scout, for processing video, extracting and analyzing transcript, audio and visual aspects, determining the boundaries of program segments and commercial breaks and extracting a program summary from a complete broadcast [12].

## 2.4 Genre Detection

In absence of electronic program guide or metadata describing the video content, we need to use automatic methods for genre detection. Video content classification is a necessary tool in the current merging of entertainment and information media. Systems that help in content management have to discern between different categories of video in order to provide for fast retrieval. We developed a method for video classification based on face and text trajectories [10] based on the observation that in different TV categories there are different face and text trajectory patterns. Face and text tracking is applied to arbitrary video clips to extract faces and text trajectories. We used Hidden Markov Models (HMM) to classify a given video clip into predefined categories, e.g., commercial, news, sitcom and soap. Our results show classification accuracy of over 80% for HMM method on short video clips.

## 2.5 Multimedia Summary

Video summarization is the process of condensing the content into a shorter descriptive form of the original content. There is a variety of flavors that have been considered under the topic of summarization: video skimming, highlights, and various types of multimedia summaries. Next, we distinguish between local summaries for part of a program (e.g. for a scene), global summaries for the entire program, and meta-level summaries of a collection of programs.

*Video skim* is a temporally condensed form of the video stream that preferably preserves the most important information. A method for generating visual skims based on scene analysis and using the grammar of film language is presented in [17]. Ma et al. proposed an attention model that includes visual, audio, and text modalities for summarization of videos [13].

*Video highlights* is a form of summary that aims at including the most important events in the video. Various methods have been introduced for extracting highlights from specific subgenre of sports programs: goals in soccer video [7], hits in tennis video, touch down in baseball, important events in car racing video [13] and others.

*Multimedia video summary* is a collection of audio, visual, and text segments that preserve the essence and the structure of the underlying video (e.g. pictorial summary, story boards, surface summary). Uchihashi et al., present methods for automatically creating pictorial summaries of videos [18] using image and audio analysis to find relative importance of segments. The output consists of static images linked to the video and the users can interact with it. Surface level summarization takes into account the structure of the video. For example, Agnihotri et al. present a surface

summarization method for talk shows that includes representative elements for the host portion and each of the guests [2]. The system consists of: transcript extractor, program type classifier, cue extractor, knowledge database, temporal database, and inference engine. Aner et al. introduce mosaic-based scene representation for clustering of scenes into physical settings [5].

Meta-level summaries provide an overview of a whole cluster of related videos. For example, meta summary of all available news items from Web and TV sources is provided by the MyInfo system [11] (see Figure 3). Summary of the news items is extracted by the reportage analysis and presented according to a personal profile.



Fig. 3. Summary of Web and TV news items in MyInfo

### 3 Current Practices and Assumptions

We rarely pause to reflect on the well accepted practices and assumptions that we use in multimedia content analysis. Here we will make observations based on our own work and the papers in the recent literature.

#### 3.1 Short Memory

Short” of course is a relative term. In audio processing usually 20ms window is used in order to make local assessments which are further used for audio segmentation and classification. In video, for cut detection usually the window is two frames, for soft scene transitions (fade, dissolve) the window can be half a minute. In all these cases the length of the temporal window is dictated by the detection task. Information about the wider scope of the signal is usually not used.

This kind of short-term memory that we are using for specialized tasks is reminiscent of a special type of medical condition. The main hero in the movie “Memento,” Leonard, is suffering from a condition called anterograde amnesia, which means that

he cannot create new long-term memories. His attention span is about 15 minutes and the current memories cannot be permanently implanted in his brain. He operates by using notes, Polaroid snapshots and tattoos (externalized substitute for long term memory). How does this analogy translate to multimedia processing? Short-term buffers are used and most of the information that could be deemed useful in the long term is thrown away. However, we think that the short term memory processing has two consequences: a) loss of accuracy and b) brittleness. In the face detection example: the experience shows that the changing lighting conditions usually mean that we get false negatives although the face is consistently present in the whole shot.

### **3.2 Focused Processing**

Currently in content analysis the processing is specifically directed to find an object, behavior, or event. As an example, in face detection the algorithm starts with skin-tone detection, followed by shape filtering, and in some cases, post-analysis that tries to reduce the number of false alarms and missed faces. In the process, the algorithm would miss what is “obvious” to us due to variations in color, position and occlusion. The focus of the algorithms is on the features that describe the face and not on the anatomical features or the physical laws – since the face does not appear and disappear within a split second.

In image face detection there is usually limited additional information. However, in video face detection the motion information is also available and this can be exploited as well. Instead of focused processing on a few frames, the algorithms can take into account the physical laws and the cinematographic practices.

### **3.3 Utilizing Available Features**

Visual, auditory and text features have been used to extract content descriptors. In most cases, the assumption has been that we can use color, motion, shape, and text features to define objects and events of interest. In this past decade we did not question whether these features are the most representative of the underlying content for the task at hand. Features that were available were used and re-used in order to generate more detectors. Feature selection has only recently come to be the focus of attention.

### **3.4 Inherent Production Syntax**

Produced video such as TV programs and movies follow cinematographic principles. The language syntax is present in the final produced movie or TV program. In movie production, the main syntactic elements include camera angles, continuity, cutting (multiple perspectives of the current and introduction of new scenes), close-ups, and composition. Content analysis area devises methods for recognition of structure, where structure represents the syntactic level composition of the video content. In specific domains, high-level syntactic structures may correspond well to distinctive semantic events. For example, we rely on the video news to have anchor shots and reportage shots in order to convey the full background of the story and on-the-scene

information. We also rely on the production syntax being consistent without frequent changes.

Another assumption is linear progressive passage of time – which is the common sense model from our own experience in the real world. However, while this is true of many TV programs, in movie making, flashbacks are also used in order to fill in gaps in the present story. In the movie “Memento” if we assign letters to the backward color scenes and numbers to the monochrome scenes, then what the director Christopher Nolan presents is the following sequence of scenes: opening credits, 1, V, 2, U, 3, T, 4, S, 5, R, 6, Q ... all the way to 20, C, 21, B, and, finally, a scene Klein calls 22/A [4]. A skimming method can operate by making assumptions of the forward and backward passage of time. However, if this assumption is not verified, we might cut off the last portion of the shot – which in the backward case means the most important part of the shot.

## 4 Next Wave

The new trends are to get down to earth with the recognized assumptions and develop beyond the areas from which we originated and learned.

### 4.1 Memory

Memory is important aiding factor in content analysis with long-term goals. In this respect our methods are designed to just keep very localized information about the current computations. However, in multimedia processing we need to keep more information for longer periods of time, such as full programs, episodes and genres. Here we refer to long-term behavior of features in not only a single shot/scene but the whole (TV) program or even the whole series or genre. A director chooses to use a particular editing style, color scheme, that is consistent throughout the movie (e.g. in sitcoms: limited number of background sets, regular cast, and theme). The “long term” behavior of features can be thought of as “priors” in probabilistic terms and used for both high level processing and improving results of the low level detectors [12].

### 4.2 Multimedia Context

Context is the larger environmental knowledge that includes the laws of biology, physics and common sense. In philosophical terms, we have been using what can be termed the “Hume” model of signal processing where the only things that exist in the present frame are real, and we should transcend to the “Kant” model where there is a representation which accounts for contextual knowledge and assumptions about the “apriory models” - expected behavior of the entities that are sought for.

### 4.3 Dynamic Processing for the Evolving Production Process

As observed earlier, multimedia content analysis can rely on the inherent syntactic structure in order to devise methods for structure analysis of video. However, the main issue is that an ever evolving media – both TV programs and films strive to break the old rules and introduce novelty. In film it is the introduction of novel camera techniques, in news it is the Web page-like appearance showing simultaneously multiple sources of information that are orthogonal to the main news story (e.g. weather, stock information, breaking news highlights at the bottom of the screen). This issue presents great challenge in the long term, because most of our methods for content analysis require training and assume that the production rules are not going to change fast.

### 4.4 Domain Specific Processing: Specific vs. General Detectors

We need to be able to generate new detectors and methods in order to learn new concepts that are evolving all the time. Repetitiveness is one of the most important aspects of the objects and events in both spatial and temporal domain and reason for applying learning methods and statistical pattern recognition [16,19].

We have made domain (genre) specific methods which have targeted focus. We have impressive results especially in news analysis and retrieval, sports highlights detection. The question is which of these methods can be generalized with little effort to the other domains and which ones would perform better on a certain domain – better than any general detector.

### 4.5 User Input and Feedback

We explore research topics under the assumption that people are going to need the results. However, user needs analysis studies are necessary to see what are the important algorithms, topics and their relevance. Also, testing the final results and surveying the usefulness of the system aspects will provide insights into applications that can eventually have impact in our everyday life.

## 5 Conclusions

In creating video databases we travel the round trip: from brains to bits and back. In film production we started with an idea expressed in a script, and then followed by production and capture of this idea into bits. Accessing this information in a video database requires enabling to travel from the bits back to consumption and playback. The applications are in enabling to travel this path from bits back to brains in the enterprise, home environment and accessing public information. We should look at the generators of the content, not only the 3500 movies that Hollywood and its equivalents around the world produce every year, but also all the camera devices in surveillance, mobile communication, live event streaming, conferencing and personal home video archives.

In this paper we summarized the global trends of multimedia content processing and we presented a view outlining the future research directions. In this endeavor I believe that we have to expand our algorithms and theory to include context, memory, dynamic processing, general evolvable detectors and user aspects in order to tackle the wide array of applications.

## References

- [1] Lalitha Agnihotri and Nevenka Dimitrova, Text Detection in Video Segments, IEEE Workshop on Content Based Access to Image and Video Libraries, June 1999.
- [2] Lalitha Agnihotri, Kavitha Devara, Thomas McGee, and Nevenka Dimitrova, "Summarization of Video Programs Based on Closed Captioning", SPIE Conf. on Storage and Retrieval in Media Databases, San Jose, CA, January 2001, pp. 599-607.
- [3] Aigrain P., Zhang H.J., Petkovic D., "Content-based Representation and Retrieval of Visual Media: A State-of-the-Art Review", International Journal of Multimedia Tools and Applications, Kluwer Academic Publishers, Vol.3, No.3, 1996.
- [4] A. Klein, Everything you wanted to know about "Memento", Salon.com, 6/28/2001. [http://archive.salon.com/ent/movies/feature/2001/06/28/memento\\_analysis/index.html](http://archive.salon.com/ent/movies/feature/2001/06/28/memento_analysis/index.html)
- [5] A. Aner and J. R. Kender, "Video Summaries through Mosaic-Based Shot and Scene Clustering", In Proc. European Conf. on Computer Vision, Denmark, May 2002.
- [6] Shih-Fu Chang, What is the Holy Grail of Content-Based Media Analysis?, IEEE Multimedia, Spring 2002.
- [7] S. Dagtas, T. McGee, and M. Abdel-Mottaleb, "Smart Watch: An automated video event finder", ACM Multimedia'2000, LA California, October 2000.
- [8] Dimitrova N., Sethi I., Rui Y., "Media Content Management", in Design and Management of Multimedia Information Systems: Opportunities and Challenges, edited by Mahbubur Rahman Syed, Idea Publishing Group, 2000.
- [9] Dimitrova, N., Agnihotri, L., and Jainschi, R., Temporal Video Boundaries, in Video Mining, A. Rosenfeld, D. Doermann, D. Dementhon eds., Kluwer, 2003, pp. 63-92.
- [10] N. Dimitrova, L. Agnihotri, and G. Wei, Video Classification based on HMM using Text and Faces, EUSIPCO 2000.
- [11] N. Haas, R. Bolle, N. Dimitrova, A. Janevski, and J. Zimmerman, Personalized News Through Content Augmentation and Profiling, IEEE ICIP, September 22-25, 2002.
- [12] R.S. Jasinschi, N. Dimitrova, T. McGee, L. Agnihotri, J. Zimmerman, D. Li, "Integrated Multimedia Processing for Topic Segmentation and Classification", In the Proceedings of IEEE Intl. Conf. on Image Processing (ICIP), Greece, 2001.

- [13] Yu-Fei Ma; Lie Lu; Hong-Jiang Zhang; Mingjing Li, “A User Attention Model for Video Summarization,” *ACM Multimedia 2002*, Juan Les Pin, December 1-5, 2002.
- [14] M. Petkovic, V. Mihajlovic, W. Jonker, “Multi-Modal Extraction of Highlights from TV Formula 1 Programs”, *IEEE Conf. on Multimedia and Expo*, Lausanne, 2002.
- [15] Smeulders A.W.M., Worring M., Santini S., Gupta A., and Jain R. Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000.
- [16] J. R. Smith, C-Y. Lin, M. Naphade, A. Natsev and B Tseng, Statistical Techniques for Video Analysis and Searching, in *Video Mining*, A. Rosenfeld, D. Doermann, D. Dementhon eds., Kluwer Academic Publishers, 2003, 259-284.
- [17] H. Sundaram; L. Xie; S-F Chang, A Utility Framework for the Automatic Generation of Audio-Visual Skims , *ACM Multimedia 2002*, Juan Les Pin, December 1-5, 2002.
- [18] S. Uchihashi, J. Foote, A. Girgensohn and J. Boreczky, Video Manga: Generating semantically meaningful video summaries, *ACM Multimedia 1999*, pp 383-392.
- [19] L. Xie, S-F Chang, A. Divakaran, H. Sun, Mining Statistical Video Structures, *Video Mining*, A. Rosenfeld, D. Doermann, D. Dementhon eds, Kluwer 2003, 285- 324.