BAYESIAN MODELING OF VIDEO EDITING AND STRUCTURE: SEMANTIC FEATURES FOR VIDEO SUMMARIZATION AND BROWSING

Nuno Vasconcelos *Andrew Lippman*

MIT Media Laboratory, {nuno,lip}@media.mit.edu
http://www.media.mit.edu/˜nuno

ABSTRACT

The ability to model content semantics is an important step towards the development of intelligent interfaces to large image and video databases. While an extremely difficult problem in the abstract, semantic characterization is possible in domains where a significant amount of structure is exhibited by the content. Whenever this is the case, given their ability to integrate prior knowledge about this structure in the inferences to be made, Bayesian methods are a natural solution to the problem. In this paper, we present a Bayesian architecture for content characterization and analyze its potential as a tool for accessing and browsing through video databases on a semantic basis.

1. INTRODUCTION

The staggering possibilities for media-access inherent to the ubiquity of computing and connectivity that characterizes the modern information landscape, and the massive amounts of data originated by sophisticated multimedia applications has originated, in the recent past, an increased interest in the areas of content-based filtering [9], retrieval [3, 7, 10], browsing [2, 5], and summarization [14, 8]. A central problem for all these areas is that of content characterization, i.e. inferring content properties from the visual patterns exhibited by the imagery or video to characterize.

We are interested in characterizing content in terms of the semantic attributes that people rely on to perform the task themselves (e.g. action vs romance). For this, we exploit the fact that content production is usually governed by specific conventions and production codes that impose a significant amount of structure on the end product (e.g. romantic movies contain a lot more close-ups than action movies) and this structure provides a basis to recover the desired attributes. The challenge is, therefore, to set up a computational framework that can capture all the content structure and establish a map between semantic attributes and measurable image features. This leads to the idea of *semantic feature spaces* which allow the machine to reason in terms of concepts that are also intuitive to people, establishing a common language for interaction.

While we have been able to show [12, 11] that some apparently very high level concepts (such as classifying a movie according to the degree of action, violence, comedy, or romance) can be derived directly from relatively simple models of video structure, the task will, in general, require the ability to integrate information from diverse models and sensory measurements. To account for this, we have recently posed the problem as one of reasoning under uncertainty, and established a computational framework based on Bayesian principles to perform this integration [13]. This framework consists of 1) a set of extremely simple visual sensors trained to detect relevant visual features, and 2) a probabilistic (Bayesian) network that infers the state of a set of semantic content descriptors from all the sensory information. The goal is to rely as much as possible on prior knowledge about 1) the inter-relationships between the semantic attributes to be inferred, and 2) the relationships between these attributes and the observable sensory measurements, using the image features as a means to disambiguate conflicting semantic interpretations.

The BMoViES[1] system is the practical outcome of the application of this framework to the domain of movies, a domain which, although generic, is subject to a vast set of production codes that, in turn, translate into a significant amount of visual structure. In this system, the movie under analysis is first decomposed into shots according to a Bayesian model of the editing process and each shot is characterized according to a set of semantic attributes belonging to what film theorists refer to as the elements of mise-en-scene. In this paper, we give a brief presentation of the system, and analyze in some detail its potential as tool for characterizing and accessing video on a semantic basis. A more detailed description of the system as well as re-

[1] BMoViES stands for Bayesian Modeling of Video Editing and Structure.

sults in the related task of content-based retrieval can be found in [13].

2. BAYESIAN INFERENCE

Statistical inference consists of establishing a model of a process in the world and drawing conclusions about the parameters or other non-observed variables of the model given a set of observations resulting from the process. Bayesian inference considers all model parameters as non-observed random variables and allows and encourages the use of knowledge about the process to establish prior probabilities for these parameters that favor the configurations of the model which are *a priori* more likely.

Computationally, given a set of random variables $\mathbf{X}$, it allows us to infer the impact on a set of variables of interest $\mathbf{U} \subset \mathbf{X}$ of the observation of another (non-overlapping) set of variables $\mathbf{O} \subset \mathbf{X}$ in the model, i.e. the ability to compute $P(\mathbf{U}|\mathbf{O} = \mathbf{o})$ taking into account any prior beliefs about what that impact will be. In order to carry these computations efficiently, the underlying model is usually mapped into a graph, where inferences are achieved by *belief propagation* between adjacent nodes. Such graphs are known as *graphical models*, *Bayesian networks*, or *belief networks*, and encompass several of the most popular solutions to difficult image processing problems including Kalman filters, Hidden Markov Models, Markov Random Fields, and mixture models.

More specifically, a Bayesian network for a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ is a probabilistic model composed by 1) a graph $\mathcal{G}$, and 2) a set of local probabilistic relations $\mathcal{P}$. The graph consists of a set of nodes, each node corresponding to one of the variables in $\mathbf{X}$, and a set of links (or edges), each link expressing a probabilistic relationship between the variables in the nodes it connects. When the state of a node is observed the marginal probabilities for the states of the remaining nodes are updated by efficient algorithms that require only message passing between neighboring nodes in a structure (*junction tree*) derived from the graph [4].

3. CONTENT STRUCTURE

Due to its ability to incorporate prior knowledge about the problems at hand, the Bayesian setting is a natural one for building systems that are designed to operate on richly structured domains, where large amounts of domain structure can be translated into priors which constrain the resulting inferences to the appropriate regions of the probabilistic space. One such domain is that of feature film.

Without getting in much detail into film theory, we point out the fact that inumerous production codes and conventions influence the creation of a movie or television show. With regards to style, these factors are usually grouped into two major categories: *montage* and *mise-en-scene* [6]. While the elements of montage determine the manner in which camera shots are put together to compose the story, the elements of mise-en-scene are related to the visual composition of each shot and include aspects such as lighting, set, placement of the actors, camera angles and framing, etc.

From the content-characterization perspective the important fact is that there are very well established conventions relating the use of these stylistic elements and the message to be conveyed in the story. For example, a film maker directing a story based on the development of character (e.g. drama or romance) will need to rely on a significant number of close-ups, as close-ups are needed to show emotion, and displaying emotion is a requirement to establish a bond between audience and character. On the other hand, the director of a thriller or a text rooted in action will rely mostly on fast cutting as a means to manipulate the emotions of the audience. I.e., there exists a causal relationship between the story and the structure of the visual patterns that constitute a movie and there is, therefore, hope of learning about the content by analyzing these visual patterns.

4. THE BMOVIES SYSTEM

The BMoViES system relies on Bayesian principles to infer semantic content properties from the analysis of visual patterns in the video. It relies on a content model composed of two stages, the first one modeling the elements of montage and the second the elements of mise-en-scene. The movie to analyze is first decomposed into shots according to the Bayesian model of editing presented in [11]. Each shot is then analyzed with respect to the aspects of mise-en-scene.

In the current version, the system makes inferences about four semantic attributes: the *action* content of the scene, the type of *set* (man-made or nature) in which it occurs, and two aspects regarding the composition of the shot - if it consists or not of a facial *close-up* and if it contains or not a *crowd* (where crowd is defined as a group of four or more people). This is a minimalist characterization of mise-en-scene but 1) as illustrated by the discussion above, already provides sufficient information to allow discrimination between high-level semantic concepts such as drama vs suspense, and 2) as shown in section 5, provides a basis for video summarization and browsing based on commands such as "move ahead to the action scenes shot in the city". In the future, the system will be augmented with more attributes.

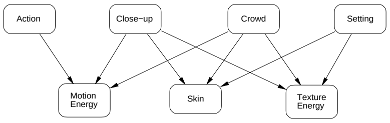

Figure 1: Bayesian network of the BMoViES system.

Inference of the semantic attributes is based on the Bayesian network of Figure 1. The bottom layer is composed by a set of visual sensors, trained to detect features that are deemed relevant for the characterization. Currently there are three sensors: one for *activity*, one for *large connected blobs of skin tones*, and one for *texture energy*. All the sensors perform very simple operations: the activity sensor measures the residual energy after the frames in the shot are aligned by an affine transformation; the skin tones sensor detects pixels whose color lies on a small predefined region of the color spectrum which is consistent with human flesh, groups them into connected regions, and computes a metric that is highest if there is a large single connected region, and small if there are several small regions; and the texture energy sensor performs a wavelet decomposition of the frames in the shot, measures the ratio between the energy in the diagonal bands and that in the horizontal and vertical ones, outputting a large value when the ratio is large (indicating natural scenes) and a small value when it is small (man made environments). The sensor outputs are then quantized into three uniform bins that signal an answer of *no*, *maybe*, or *yes* to the presence of the feature in the video. See [13, 12] for a more detailed explanation of the sensor implementation.

The prior and conditional probabilities of the model were hand-coded using common-sense (e.g. the output of the skin tones sensor will be *yes* with probability 0.9 for a scene of a crowd in a man-made set). No effort was made to optimize the performance of the system by tweaking with network parameters. All the probabilities could also have been learned from training data but, given the relatively small size of the network, we have so far felt no need for that.

One of the most interesting properties of Bayesian networks is the capability to integrate information from all the observed nodes during inference. This phenomena, commonly referred to as *explaining away* in the literature, is visible in the network of BMoViES. Consider, for example, the observation of skin tones which can be a sign of both the scene being composed by a close-up or a crowd. While if a crowd is present there will also be a significant response of the texture sensor, the opposite will happen in the case of a close-up. Hence, when its output is high (low), the texture sensor "explains away" the observation of skin tones and rules out the close-up (crowd) hypothesis, even though it is not a close-up or crowd detector.

5. SEMANTIC USER INTERACTION

In [13], we show experimental evidence of the fact that semantic characterization can be achieved with rates as high as 90% on real movie databases. In this section, we illustrate the large potential for user interaction inherent to this type of characterization by considering the tasks of video summarization and browsing.

5.1. Video summarization

For a system capable of inferring content semantics, summarization is a simple outcome of the characterization process. Because system and user understand the same language, all that is required from the system is that it can display the inferred semantic attributes in a way that does not overwhelm the user. The user can then use his/her own cognitive resources to extrapolate from these semantic attributes to other attributes, usually of higher semantic level, that may be required for a coarse understanding of the content.

5.2. Semantic time-lines

In BMoViES, this graphical summarization is attained in the form of a time-line that displays the evolution of the state of the semantic attributes throughout the movie. Figure 2 presents the time-lines resulting from the analysis of the promotional trailers of the movies "Circle of friends" (COF), and "The river wild" (TRW). Each line in the time-line corresponds to the semantic attribute identified by the letter on the left margin - "A" for action, "D" for close-up, "C" for crowd, and "S" for natural set - and each interval between small tick marks displays the state of the attribute in one shot of the trailer - filled (empty) intervals mean that the attribute is active (not present). The shots are represented in the order by which they appear in the trailer.

By simply looking at these time-lines, the user can quickly extract a significant amount of information about the content of the two movies. Namely, he/she will understand right away that while COF contains very few action scenes, consists mostly of dialogue, and is for the most part shot in man-made sets; TRW is mostly about

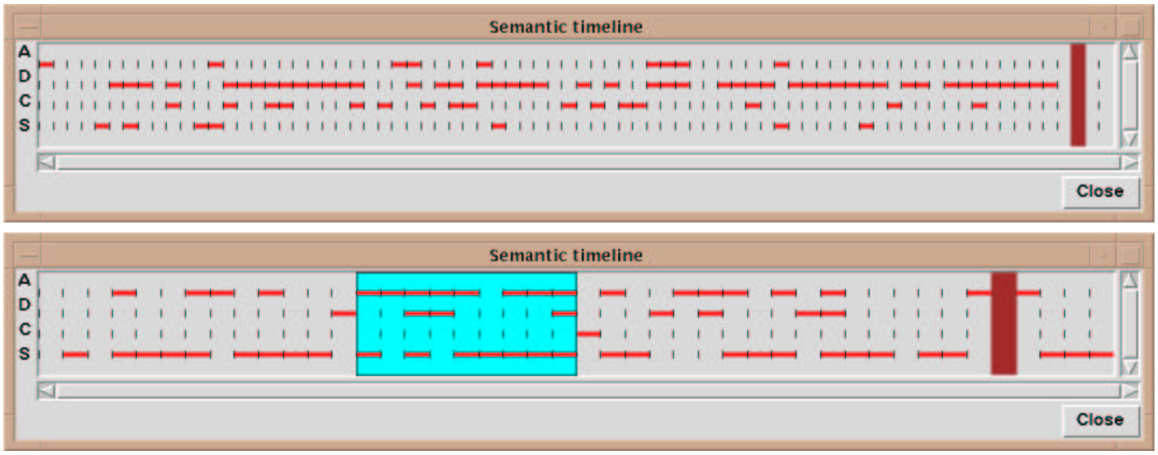

Figure 2: Semantic time-lines for the trailers of the movies "Circle of friends" (top) and "The river wild" (bottom).

action, contains few dialogue, and is shot in the wilderness. When faced with such descriptions, few users looking for a romance would consider TRW worth of further inspection, and few users looking for a thriller would give COF further consideration.

In fact, we believe that given a written summary of the two movies few people would have doubts in establishing the correspondence between summaries and movies based on the information provided by the semantic time-lines alone. To verify this consider yourself the summaries provided for these movies by the Internet Movie Database [1]:

Circle of Friends:

A story about the lives, loves and betrayals of three Irish girls, Bennie, Eve, and Nan as they go to Trinity College, Dublin. Bennie soon seems to have found her ideal man in Jack, but events conspire to ruin their happiness.

The River Wild:

Gail, an expert at white water rafting, takes her family on a trip down the river to their family's house. Along the way, the family encounters two men who are unexperienced rafters that need to find their friends down river. Later, the family finds out that the pair of men are armed robbers. The men then physically force the family to take them down the river to meet their accomplices. The rafting trip for the family is definitely ruined, but most importantly, their lives are at stake.

We are currently designing experiments with human subjects that will allow us to achieve a more objective assessment of the benefits of semantic summarization.

5.3. Semantic content-access and browsing

It can obviously be argued that the example above does not fully stretch the capabilities of the semantic characterization, i.e. that the movies belong to such different genres that the roughest of the semantic characterizations would allow a smart user to find the desired movie. What if instead of distinguishing COF from TRW, we would like to differentiate TRW from "Ghost and the Darkness" (GAD)? GAD is summarized as follows:

Ghost and the Darkness:

Set in 1898, this movie is based on the true story of two lions in Africa that killed 130 people over a nine month period, while a bridge engineer and an experienced old hunter tried to kill them.

After all, one would also expect GAD to contain lots of action, few dialog, and be shot in the wilderness. How would the semantic characterization help here?

There are two answers to this question. The first is that it would not because the characterization is not *fine* enough to distinguish between TRW and GAD. The solution would be to augment the system with finer semantic attributes, e.g. to subdivide the natural set attribute into classes like "river", "forest", "savannah", "desert", etc. The second, significantly simpler, is to say that while simply looking at the time-lines would not help, interacting with them would.

Consider the action scenes in the two movies. While in TRW we would expect to see a river, a woman, and good and bad guys, in GAD we would expect to see savannah, lions and hunters. Thus, the action scenes would probably be the place to look first. Consider next, the TRW time-line in the bottom of Figure 2. The high concentration of action shots in the high-

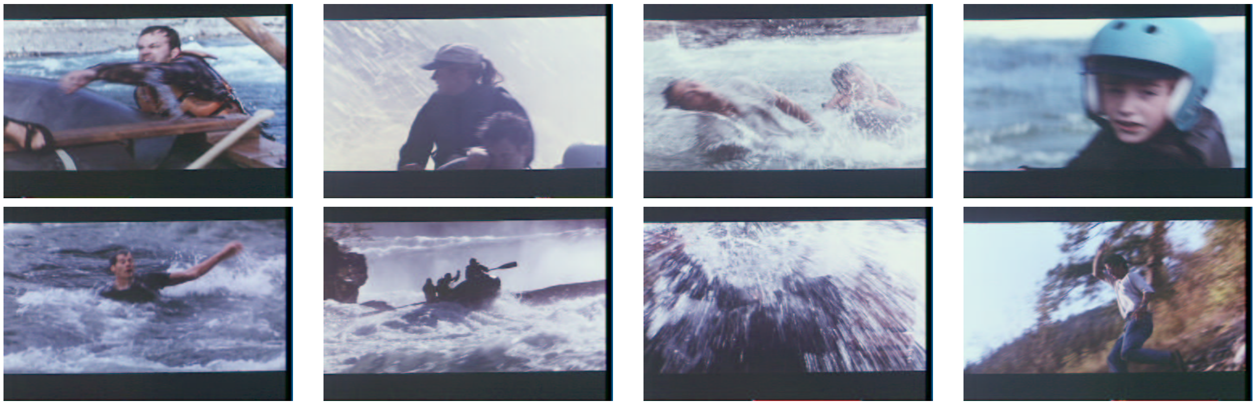

Figure 3: Key-frames of the shots in the high-lighted area of the time-line in the bottom of Figure 2. The shot (correctly) classified as not containing action depicts a promotional message and is not included.

lighted area indicates that this is likely to be the best area to look for action. This is confirmed by Figure 3, which presents key frames for each of the shots in the area. By actually viewing the shots represented the figure, it is clear that the action occurs in a river, that there are good and bad guys (the first, and third shots depict a fight), and there are a woman and a child in the boat. I.e. even when the information contained in them is not enough to completely disambiguate the content, the semantic attributes provide a way to quickly *access* the relevant portions of the video stream. Semantic-based access is an important feature on its own for browsing as it allows users to quickly move on to the portions of the video in which they are really interested.

6. REFERENCES

[1] *Internet Movie Database.* http://us.imdb.com/.

[2] I. Cox, M. Miller, S. Omohundro, and P. Yianilos. PicHunter: Bayesian Relevance Feedback for Image Retrieval. In *Int. Conf. on Pattern Recognition*, Vienna, Austria, 1996.

[3] W. Niblack et al. The QBIC project: Querying images by content using color, texture, and shape. In *Storage and Retrieval for Image and Video Databases*, pages 173–181, SPIE, Feb. 1993, San Jose, CA.

[4] F. Jensen. *An Introduction to Bayesian Networks.* Springer-Verlag, 1996.

[5] T. Minka and R. Picard. Interactive learning using a "society of models". Technical Report 349, MIT Media Lab, 1995.

[6] J. Monaco. *How to Read a Movie.* Oxford University Press, 1981.

[7] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Content-based Manipulation of Image Databases. In *SPIE Storage and Retrieval for Image and Video Databases II*, number 2185, Feb. 1994, San Jose, CA.

[8] H. Sawhney and S. Ayer. Compact Representations of Videos Through Dominant and Multiple Motion Estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 18, August 1996.

[9] J. Smith and S. Chang. Visually Searching the Web for Content. *IEEE Multimedia*, 4(3):12–20, July-September 1997.

[10] S. Smoliar and H. Zhang. Video Indexing and Retrieval. In B. Furth, editor, *Multimedia Systems and Techniques.* KAP, 1996.

[11] N. Vasconcelos and A. Lippman. A Bayesian Video Modeling Framework for Shot Segmentation and Content Characterization. In *Proc. IEEE Workshop on Content-based Access to Image and Video Libraries*, CVPR97, San Juan, Puerto Rico, 1997.

[12] N. Vasconcelos and A. Lippman. Towards Semantically Meaningful Feature Spaces for the Characterization of Video Content. In *Proc. Int. Conf. Image Processing*, Santa Barbara, California, 1997.

[13] N. Vasconcelos and A. Lippman. A Bayesian Framework for Semantic Content Characterization. In *Proc. IEEE Computer Vision and Pat-*

tern Recognition Conf., Santa Barbara, California, 1998.

[14] B. Yeo and B. Liu. Rapid Scene Analysis on Compressed Video. *IEEE Trans. on Circuits and Systems for Video Technology*, 5(6):533–544, December 1995.