Appears in IEEE Conference Computer Vision and Pattern Recognition, Santa Barbara, California, 1998.

A Bayesian Framework for Semantic Content Characterization

Nuno Vasconcelos Andrew Lippman
MIT Media Laboratory
{nuno,lip}@media.mit.edu

Abstract

Current systems for content filtering, browsing, and retrieval rely on low-level image descriptors which are unintuitive for most users. In this paper, we propose an alternative framework that exploits the structured nature of most content sources to achieve semantic content characterization, and lead to much more meaningful user interaction. Computationally, this framework is based on the principles of Bayesian inference and can be implemented efficiently with Bayesian networks. As an illustration of its potential we apply it to the domain of movie databases.

1 Introduction

Given the massive amounts of imagery characteristic of modern multimedia applications, there has recently been a growing interest in the formulation of sophisticated algorithms for browsing, filtering, and retrieving content from image and video databases. Unfortunately, the majority of current content characterization frameworks rely on very low-level image descriptions (e.g. image color and texture), which limit their capability to support meaningful interaction with users that are not experts in the inner workings of the resulting systems.

The ability to infer semantic descriptions from images has, on the other hand, proven to be an elusive goal to attain. While significant progress has occurred, over the last decades, in the areas of segmentation, object recognition, and 3D scene understanding, it is relatively clear that approaches based on these principles will have limited success for the problems of information filtering, retrieval, classification, or summarization. The traditional approach of building a model of the world and/or imaging process, mapping the image observations into that model, and finally using it to infer the desired world or scene attributes is simply too complex for these problems, where there is no control over the scene or imaging set ups, and no model is generic enough to account for all the possible scene variability.

In this paper, we investigate an alternative path towards the semantic characterization of visual content where, instead of trying to interpret all the visual information, we use this information only as means to disambiguate conflicting scene interpretations. The fundamental assumption is that the process of content creation is not random but, instead, obeys a series of well established codes and conventions. These codes in turn impose a significant amount of structure on the content itself, originating characteristic visual patterns whose detection is sufficient to disambiguate the various possible semantic interpretations.

This leads us to formulate the problem of content characterization as one of sensor fusion, consisting of the design of 1) a set of sensors tuned to the relevant patterns, and 2) an architecture for the integration of all the sensory data and inference of the desired semantic attributes. Computationally, we pose sensor fusion as a problem of Bayesian inference and introduce a Bayesian framework, based on belief propagation with Bayesian networks, for content characterization. As an illustration of the potential of this framework, we apply it to the characterization of movies, a generic domain that exhibits a significant amount of structure.

2 Content structure

The fundamental assumption underlying our framework is that the bulk of the the content that one would care to store in, or retrieve from, video databases exhibits a significant amount of structure, that can be exploited for content characterization.

2.1 Natural modes

The assumption of structured content is an instantiation of the more general principles of *natural modes* [1, 4], and *non-accidental properties* [5]. The basic idea is that for a perceiver to develop the inferential leverage necessary to disambiguate among several conflicting configurations of the world, the world must behave regularly. In particular, it must evolve towards a discrete sampling along the dimensions that are important for the interaction of the diverse organisms with their environment, leading to the realization of *emergent* [1] or *modal properties* [4] that allow perceivers to make sense of it.

These principles are illustrated in [1] with examples from the biological world, where this clustering is due to two fundamental forces: physical constraints (e.g. it is physically impossible to build a flying elephant), and the pressures of evolution (e.g. if two organisms are only marginally different, and that difference is along a direction that is important to their survival, one will be fitter than the other and eliminate it).

2.2 Content production codes

Like biological processes, content production is subject to two fundamental forces: constraints in form and function, and evolutionary pressures. The constraints in form and function are imposed by physical limitations of the medium, limitations on the amount of cognitive resources that content consumers will devote to this task, and limitations on the amount of resources available for the content production itself. Evolutionary pressures are a consequence of the economics of content production.

It is well known that content consumption is mostly an informal activity, usually performed in parallel with other tasks that require a share of the consumer's cognitive resources (e.g. people customarily watch the news over breakfast, listen to the radio while driving, or watch movies while talking on the phone) [8]. In result, the message must be laid out in a way that minimizes the effort required for decoding it. Furthermore, because there is typically a limited amount of resources available for content production, it is important to standardize this process, so that the efficiency of the production is maximized.

Thus, while there are clear incentives for innovation, content production evolves by building on previously developed formulas that have sustained the testing of both time and the market. In result, sophisticated content production codes or languages have evolved over time, becoming second nature to most of us. These languages are particularly evident in domains such as information delivery, where newscast present an impressive uniformity of structure across media, geographical locations, and even countries. Their scope is, however, much broader than these restricted domains and, even though not always obvious to the unatentive eye, they command the production of virtually all forms of content targeted to mass consumption. We next analyze in more detail the domain of film.

2.3 Structure in movies

It is well known in film theory that the stylistic elements of a movie are closely related to the message conveyed in its story. Historically, these stylistic elements have been grouped into two major categories: *montage* and *mise-en-scene* [7]. While montage refers to the aspects of film editing, mise-en-scene consists of the elements used in the composition of each shot: type of set, placement of the actors on the scene, lighting, camera angles, etc.

From the content characterization perspective, the important point is that while both elements of montage and mise-en-scene can be used to manipulate the emotions of the audience (this manipulation is, after all, the ultimate goal of the director), there are some very well established codes or rules to achieve this. For example, a director trying to put forth a text deeply rooted in the construction of character (e.g. a drama or a romance) will necessarily have to rely on a fair amount of facial close-ups, as close-ups are the most powerful tool for displaying emotion[1], an essential requirement to establish a strong bond between the audience and the characters in the story.

If, on the other hand, the same director is trying to put forth a text of the action or suspense genres, the elements of mise-en-scene become less relevant than the rhythmic patterns of montage. In action or suspense scenes it is imperative to rely on fast cutting, and manipulation of the cutting rate is the tool of choice for keeping the audience "at the edge of their seats". Directors who exhibit supreme mastery in the manipulation of the editing patterns are even referred to as *montage directors*[2].

Obviously, the structure due to the stylistic conventions is complemented by that due to the, more generic, production codes discussed above. For example, most Hollywood productions rely on a fundamental plot line which consists of 1) establishing main characters, 2) setting their goals, 3) introducing evil forces that pose a barrier to these goals, and 4) the "great finale" where good forces overcome evil forces [2]. Typically, these steps even occur at more or less standardized time intervals, allowing the attentive viewer to predict what will happen next.

There is, therefore, plenty of structure in most content to believe that it is possible for a machine to make semantic inferences about it, based on the analysis of the visual patterns and knowledge about the codes that determine its composition. We next introduce a computational framework to achieve that.

[1] The importance of close-ups is best summarized in the quote from the great Charles Chaplin: "Tragedy is a close-up, comedy a long shot".

[2] The best known example in this class is Alfred Hitchcock, who relied intensively on editing to create suspense in movies like "Psycho" or "Birds".

3 A Bayesian framework for semantic characterization

Our framework relies on the the principles of Bayesian inference through belief propagation. Computationally, this translates into the use of a *Bayesian Network* as the core of our content characterization architecture.

3.1 Bayesian networks

Given a set of random variables $\mathbf{X}$, a fundamental question in probabilistic inference is how 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 $P(\mathbf{U}|\mathbf{O} = \mathbf{o})$. While this computation is, theoretically, easy to perform using

$$P(\mathbf{U}|\mathbf{O} = \mathbf{o}) = \frac{\sum_{\mathbf{H}} P(\mathbf{U}, \mathbf{H}, \mathbf{O} = \mathbf{o})}{\sum_{\mathbf{O}} \sum_{\mathbf{H}} P(\mathbf{U}, \mathbf{H}, \mathbf{O})}, \qquad (1)$$

where $\mathbf{H} = \mathbf{X} - \{\mathbf{U} \cup \mathbf{O}\}$, and the summations are over all the possible configurations of the sets $\mathbf{H}$ and $\mathbf{O}$; in practice, the amount of computation involved in the evaluation of these summations makes the solution infeasible even for problems of relatively small size. A better alternative is to explore the relationships between the variables in the model to achieve more efficient inference procedures. This is the essence of Bayesian networks.

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. Together, the graph $\mathcal{G}$ and the set of probabilities $\mathcal{P}$ define the joint probability distribution for $\mathbf{X}$. Denoting the set of parents of the node associated with X_i by $\mathbf{pa}_i$, this joint distribution is

$$P(\mathbf{X}) = \prod_i P(X_i|\mathbf{pa}_i). \qquad (2)$$

The ability to decompose the joint density into a product of local conditional probabilities allows the construction of efficient algorithms where inference takes place by propagation of beliefs across the nodes in the network [9, 10].

3.2 Bayesian content characterization

Figure 1 presents a Bayesian network that naturally encodes the content characterization problem. The set of nodes $\mathbf{X}$ is the union of two disjoint subsets: a set $\mathbf{S}$ of sensors containing all the leafs (nodes that do not have any children) of the graph, and a set $\mathbf{A}$ of semantic content attributes containing the remaining variables. The set of attributes is organized hierarchically, variables in a given layer representing higher level semantic attributes than those in the layers below.

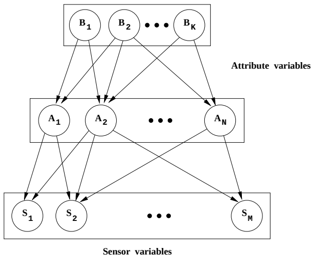

Figure 1: A generic Bayesian architecture for content characterization. Even though only three layers of variables are represented in the figure, the network could contain as many as desired.

The visual sensors are tuned to certain visual features deemed relevant for the semantic content characterization. The network infers the presence/absence of the semantic attributes given these sensor measurements, i.e. $P(\mathbf{a}|\mathbf{S})$, where $\mathbf{a} \subseteq \mathbf{A}$.

3.3 Semantic modeling

One of the strengths of the Bayesian framework is that the sensors in Figure 1 are not required to do a perfect job of identifying the desired image features. Reasons for this are the fact that 1) the model can account for the sensor precision, and 2) the network can integrate the sensor information to disambiguate conflicting hypothesis.

Consider, for example, the task of detecting sky in a sports database containing pictures of both skiing and sailing competitions. One way to achieve such goal would be to rely on a pair of sophisticated water and sky detectors to discriminate between water and sky. The underlying strategy is to interpret the images first and then decide on the characterization according to this interpretation.

While such strategy could be implemented under the Bayesian framework, a more efficient alternative would be to rely on the model of Figure 2. Here, the network consists of five semantic attributes and two simple sensors for large white and blue image patches. In the absence of any measurements, the variables *sailing* and *skiing* are independent. However, whenever the sensor of blue patches fires up, they do become dependent (or, in the Bayesian network lingo, *d-connected* [9]) and, the knowledge of the output of the sensor of white patches is sufficient to perform the desired inference.

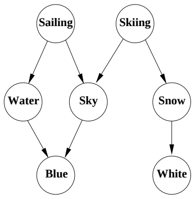

Figure 2: A simple Bayesian network for the classification of sports.

This effect is known as the "explaining away" capability of Bayesian networks [9]. Although there is no direct connection between the *white* sensor and the *sailing* variable, the observation of white reduces the likelihood of the sailing hypothesis and this, in turn, reduces the likelihood of the *water* hypothesis. So, if the *blue* sensor is active, the network will infer that this is a consequence of the presence of sky, even though we have not resorted to any sort of sophisticated sensors for image interpretation. I.e. the white sensor explains away the firing of the blue sensor.

This second strategy relies more in modeling the semantics and the relationships between them, than in the interpretation of the image itself. In fact, the image measurements are only used to discriminate between the different semantic interpretations. This has two practical advantages. First, a much smaller burden is placed on the sensors which, for example, do not have to know that "water is usually more textured than sky", and are therefore significantly easier to build. Second, as a side effect of the sky detection process, we obtain a semantic interpretation of the images which, in this case, is sufficient to classify them into one of the two classes in the database.

4 Characterizing movies

In this section we introduce a content characterization system, BMoViES[3], which exploits the structure inherent to the domain of film to achieve semantic characterization of movies.

BMoViES consists of two major modules. The first [11] relies on a Bayesian model of montage to segment the movie into its component shots. Each shot is then analyzed by the second module which infers the semantic content attributes. This module is an instantiation of the generic Bayesian framework of Figure 1.

4.1 The attributes

In the current implementation of BMoViES, the system recognizes four semantic shot attributes: the presence/absence of a close-up, the presence/absence of a crowd in the scene, the type of set (nature vs. urban), and if the shot contains a significant amount of action or not. These attributes can be seen as a minimalist characterization of mise-en-scene which, nevertheless, provides a basis for categorizing the video into relevant semantic categories such as "action vs dialog", "city vs country side", or combinations of these. Also, as the discussed in section 2.3, it captures the aspects of mise-en-scene that are essential for the inference of higher level semantic attributes such as suspense or drama.

4.2 The sensors

Currently, the sensor set consists of three sensors measuring the following properties: shot activity, texture energy, and amount of skin tones in the scene.

The shot activity [11] is computed by measuring the energy left after the frames in the shot are registered and subtracted. The registration is based on an affine transformation, making the measurements immune to most of the variation due to camera motion.

The texture energy sensor performs a 3-octave wavelet decomposition of each image, and measures the ratio of the total energy in the high-pass horizontal and vertical bands to the total energy in all the bands other than the DC. It produces a low output whenever there is a significant amount of vertical or horizontal structure in the images (as is the case in most man-made environments) and a high output when this is not the case (as is typically the case in natural settings).

Finally, the skin tones sensor identifies the regions of each image which contain colors consistent with human skin, measures the area of each of these regions and computes the entropy of the resulting vector (regarding each component as a probability). This sensor outputs a low value when there is a single region of skin and high values otherwise. The situation of complete absence of skin tones is also detected, the output of the sensor being set to one.

Sensor measurements are integrated across each shot by averaging the individual frame outputs. In order to quantize the sensor outputs, their range was thresholded into three equally sized bins. In this way, each sensor provides a ternary output corresponding to the states *no*, *yes*, or *maybe*.

4.3 The network

The Bayesian network implemented in BMoViES is presented in Figure 3. The parameters of this model

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

can either be learned from training data [3] or set according to expert knowledge. In the current implementation, we followed the latter approach. Both the structure and the probabilities in 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 overall performance of the system by tweaking the network probabilities.

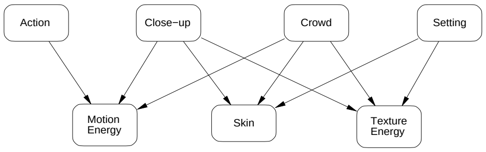

Figure 3: Bayesian network for content characterization.

To see how explaining away occurs in BMoViES, consider the observation of a significant amount of skin tones. Such observation can be synonymous with both a close-up or a scene of a crowd. If a crowd is present, though, there will also be a significant response by the texture sensor, while the opposite will happen if the shot consists of a close-up. Hence, the texture sensor "explains away" the skin tones observation, and rules out the close-up hypothesis, despite the fact that it is not a crowd detector.

5 Applications and results

Due to the fact that Bayesian networks do not have inputs and outputs, but only hidden and observed variables, a node which acts as an input for a given task can be used as an output for another. In result, the Bayesian framework works in both directions (i.e. given visual features, infer attributes that best classify them; or given attribute specifications, retrieve the data that best satisfies them), providing a unified solution to the problems of information filtering and retrieval, and allowing the construction of very flexible retrieval systems.

5.1 Classification

To evaluate the accuracy of the semantic classification of BMoViES, we applied the system to a database of about 100 video clips (total of about 3000 frames) from the movie "Circle of friends". The database is a subsampling of approximately 25 minutes of film, and contains a wide variety of scenes and high variation of imaging variables such as lighting, camera viewpoints, etc. To establish ground truth, the video clips were also manually classified.

Attribute	Action	Close-up	Crowd	Set
% Accuracy	90.7	88.2	85.5	86.8

Table 1: Classification accuracy of BMoViES.

Table 1 presents the classification accuracy achieved by BMoViES for each of the semantic attributes in the model. Overall the system achieved an accuracy of 88.7%. Given the simplicity of the sensors this is a very satisfying result. Some of the classification errors, which illustrate the difficulty of the task, are presented in Figure 4.

5.2 Retrieval

As a retrieval system, BMoViES supports two types of queries. The first is the standard query by example, where the user provides the system with a video clip and asks it to "find all the clips that look like this". BMoViES then classifies the example query and retrieves from the database the items that belong to the same category.

The interesting point is that the retrieval criteria, *semantic similarity*, is much more meaningful than the standard *visual similarity* criteria. In fact, whenever a user orders the machine to "search for a picture like this", the user is with high likelihood referring to pictures that are semantically similar to the query image (e.g. "pictures which also contain people") but which do not necessarily contain identical patterns of color and texture.

Figure 5 presents an example of retrieval by semantic similarity. The image in the top right is a key-frame of the clip submitted as a query by the user, and the remaining images are key frames of the clips returned by BMoViES. Notice that most of the suggestions made by the system are indeed semantically similar to the query, but very few are similar in terms of color and texture patterns.

5.3 Relevance feedback

The second query mode is even more intuitive than the first and consists of simply specifying to the system the desired semantic attributes. E.g. "show me all the action clips shot in a urban set". One of the problems with this retrieval paradigm is, however, that for a system with a large number of attributes it would be tedious to instantiate all of them whenever a query is formulated. Furthermore, it is not always the case that the user knows exactly what he/she is looking for. A more meaningful search strategy is, therefore, to rely on *relevance feedback* [6] mechanisms, where the

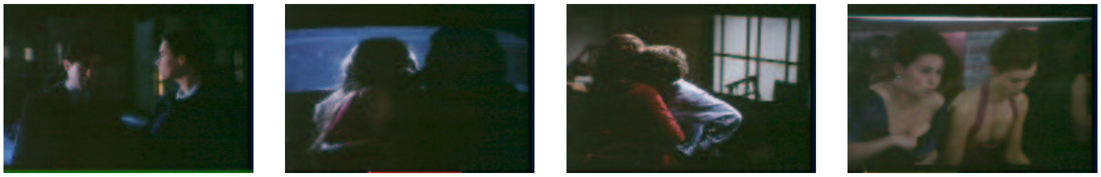

Figure 4: Classification errors in BMoViES. People were not detected in the left three clips, crowd was not recognized on the right.

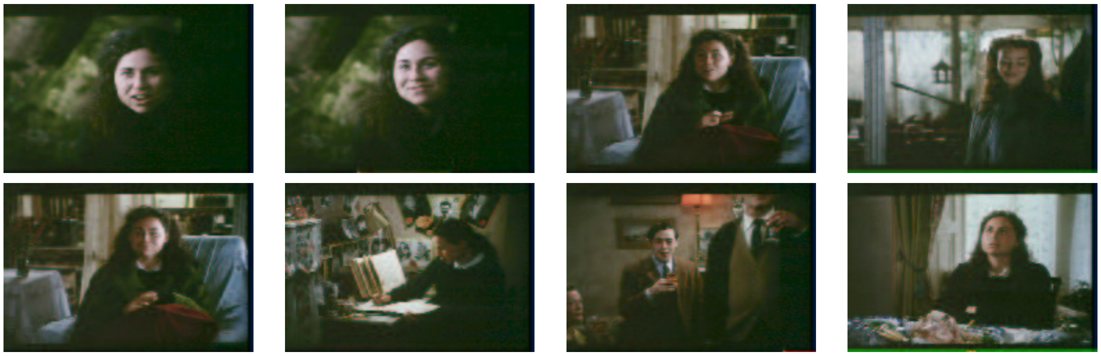

Figure 5: Example based retrieval in BMoViES. The top left image is a key frame of the clip submitted to the retrieval system. The remaining images are keyframes of the best seven matches found by the system.

user interacts with the system in order to accomplish his/her goals.

These type of incremental queries are very natural in the Bayesian setting, where the fact that the user provides more information at each iteration simply means that some of the attribute nodes change from the hidden to the observed state. The corresponding nodes of the network are then instantiated, and beliefs propagated to find out the mostly likely sensor configurations given those specifications. The video clips originating these configurations are then retrieved and presented to the user, which in response can refine the retrieval attributes.

Figure 6 illustrates the ability of the Bayesian framework to support meaningful user interaction. The top row of the figure presents the video clips retrieved in a response to a query where the action attribute was instantiated with *yes*, and the remaining attributes with *don't care*. The system suggests a shot of ballroom-dancing as the most likely to satisfy the query, followed by a clip containing some graphics and a clip of a rugby match.

In this example, the user was not interested in clips containing a lot of people. Specifying *no* for the crowd attribute, lead to the refinement shown in the second row of the figure. The ballroom shot is no longer among the top suggestions, which tend to include at most one or two people. At this point, the user specified that he was looking for scenes shot in a natural set, leading the system to suggest the clips shown in the third row of the figure. The clips that are most likely to satisfy the specification contain scenes of people running in a forest. Finally, the specification of *no* for the close-up attribute, lead to the suggestion of the bottom row of the figure, where the clips containing close-ups were replaced for clips where the set becomes predominant.

References

[1] A. Bobick and W. Richards. Classifying Objects from Visual Information. Technical report, MIT AI Lab Memo 879, June 1986.

[2] E. Dmytryk. *On Filmmaking.* Focal Press, 1986.

[3] D. Heckerman. A Tutorial on Learning with Bayesian Networks. Technical Report MSR-TR-96-06, Microsoft Research, March 1995.

[4] A. Jepson, W. Richards, and D. Knill. Modal Structure and Reliable Inference. In D. Knill and W. Richars, editors, *Perception as Bayesian Inference.* Cambridge Univ. Press, 1996.

[5] D. Lowe. *Perceptual Organization and Visual Recognition.* Klewer, 1985.

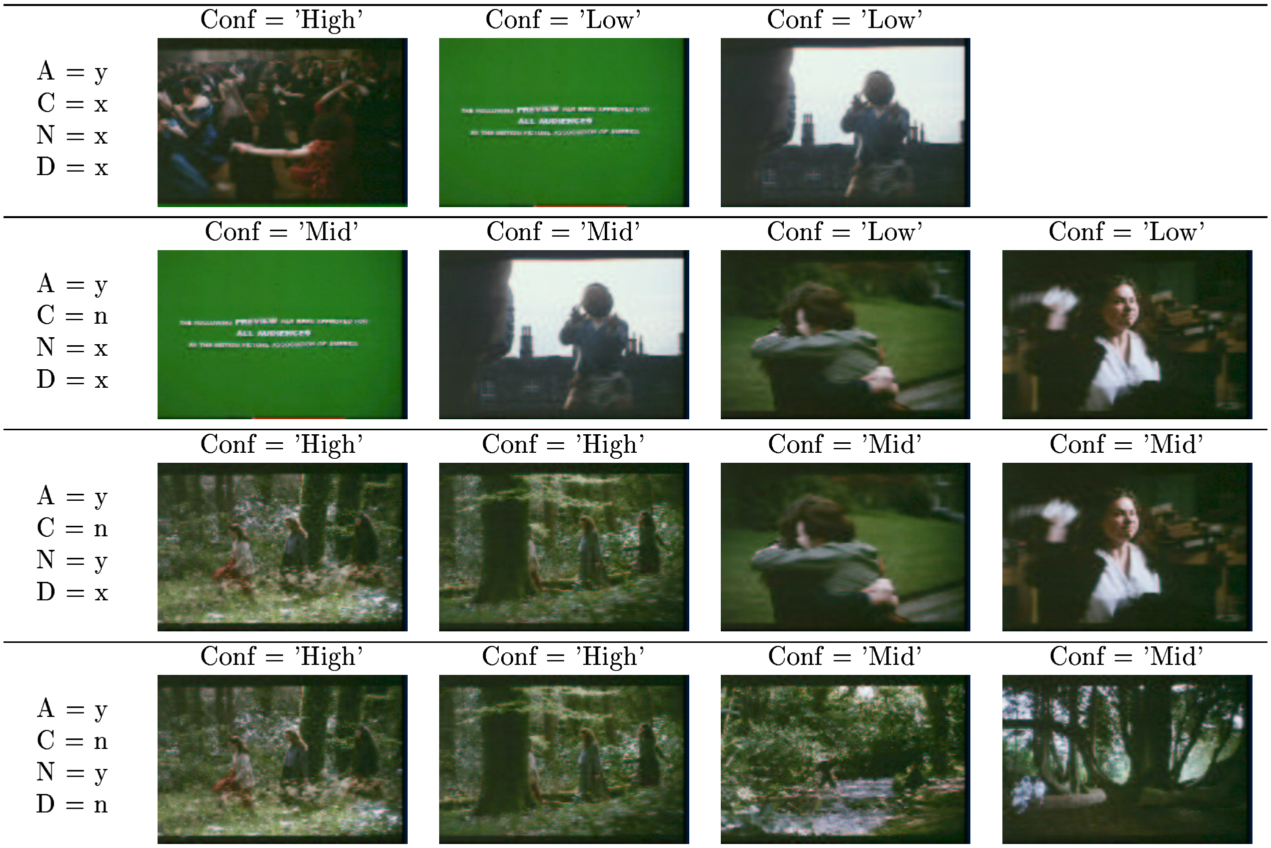

Figure 6: Relevance feedback in the Bayesian framework. Each row presents the response of the system to the query on the left. The action (A), crowd (C), natural set (N), and close-up (D) attributes are instantiated with *yes* (y), *no* (n), or *don't care* (x). The confidence of the system on each of the retrieved clips is shown on top of the corresponding key frame.

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

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

[8] W. R. Neuman. *The Future of the Mass Audience.* NY: Cambridge University Press, 1991.

[9] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.

[10] D. Spiegelhalter, A. Dawid, S. Lauritzen, and R. Cowell. Bayesian Analysis in Expert Systems. *Statistical Science*, 8(3):219–283, 1993.

[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.