Statistical Models of Video Structure for Content Analysis and Characterization

Nuno Vasconcelos, *Student Member, IEEE*, and Andrew Lippman, *Member, IEEE*

Abstract*— Content structure plays an important role in the understanding of video. In this paper, we argue that knowledge about structure can be used both as a means to improve the performance of content analysis and to extract features that convey semantic information about the content. We introduce statistical models for two important components of this structure, shot duration and activity, and demonstrate the usefulness of these models with two practical applications. First, we develop a Bayesian formulation for the shot segmentation problem that is shown to extend the standard thresholding model in an adaptive and intuitive way, leading to improved segmentation accuracy. Second, by applying the transformation into the shot duration/activity feature space to a database of movie clips, we also illustrate how the Bayesian model captures semantic properties of the content. We suggest ways in which these properties can be used as a basis for intuitive content-based access to movie libraries.*Content structure plays an important role in the understanding of video. In this paper, we argue that knowledge about structure can be used both as a means to improve the performance of content analysis and to extract features that convey semantic information about the content. We introduce statistical models for two important components of this structure, shot duration and activity, and demonstrate the usefulness of these models with two practical applications. First, we develop a Bayesian formulation for the shot segmentation problem that is shown to extend the standard thresholding model in an adaptive and intuitive way, leading to improved segmentation accuracy. Second, by applying the transformation into the shot duration/activity feature space to a database of movie clips, we also illustrate how the Bayesian model captures semantic properties of the content. We suggest ways in which these properties can be used as a basis for intuitive content-based access to movie libraries.**

I. Introduction

CURRENT video characterization and retrieval systems rely on image representations based on low-level visual primitives such as color, texture, and motion. While practical and computationally efficient, such characterization places on the interface to the retrieval system the burden of bridging the semantic gap between the low-level nature of the primitives and the high-level semantics that people rely on to perform the task. Because establishing this bridge is a difficult problem, there is interest in alternative representations built upon content descriptors at a higher semantic-level.

The most obvious of these alternatives is probably that of object-based representations. The ability to decompose a scene into the objects that compose it would allow semantic descriptions of arbitrary detail and enable intuitive video manipulation [8]. This type of scene decomposition can, however, only be achieved through sophisticated image segmentation. Despite significant recent progress in unsupervised segmentation [38], [48], [51], [52], [58], it does not seem that such algorithms, applicable across the different types of imagery that make up video databases and capable of producing semantic segmentations, will be achievable in the near future. On the other hand, supervised segmentation [9], [10] is not viable for video libraries, where very large volumes of content must be processed over relatively short periods of time.

This does not imply that region-based representations are not useful for content characterization and retrieval. In fact, several researchers have shown that they provide valuable extensions to the “query by example” paradigm [4], [11], [13], [40], [57]. The idea is to be able to search images by similar regions of coherent color, texture, shape, and motion instead of simpler feature representations such as the image histograms [42] or color layouts used by early retrieval systems [31], [33]. However, while regions may be better than simple features, the fact that they are not necessarily meaningful to people still poses some major difficulties. For example, it is not uncommon for an object to be segmented into several regions of different color or motion, one of these regions leading to an unintuitive match with a semantically unrelated object. Furthermore, not all regions have the same perceptual significance and it is not easy to decide how to weight each of them for the purpose of characterizing video.

Similarly to the earlier feature-based methods, one can always count on the user to reduce the gap between the machine representation and his/her own. In the context of region-based representations, this can be achieved by asking the user to interact with the machine in terms of regions. Some of the more sophisticated retrieval systems do just this: the user is allowed to formulate a query by specifying a few regions and their motion [11], or the internal region-based representation of the system is displayed to help the user figure out what went wrong during an unsuccessful query [4]. This type of interaction by reverse engineering of the machine representation and reasoning can however be unintuitive, at least for naive users.

The fact is that people characterize content according to high-level concepts, such as its amount of action, romance, or comedy that, most of the times, are not even related in a straightforward way to the visual attributes of the pixels that compose each image. Therefore, there is a need for new representations capable of capturing such high-level properties of the video. We believe that the most promising path for

The authors are with the Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.(e-mail: nuno@media.mit.edu; lip@media.mit.edu).

achieving this goal is to move away from the objective of understanding all the pixels in each image and concentrate instead on the higher level structure exhibited by the video. In fact, we argue that for most content classes that one would care to store in a video database, this structure is plentiful and plays a significant role in perceptual decoding of the video. The long-term goal of the research is to identify semantically relevant features and build computational models for their extraction and characterization. In this paper, we give the first steps towards this goal by developing statistical models for two important elements of the video structure: shot duration and activity.

While building good models for video is an interesting exercise, the ultimate measure of success of a given model is the efficiency of the solutions for practical and objective tasks that may be derived from it. In this context, once the models are built, we concentrate on two such tasks: temporal video segmentation and semantic characterization. Because knowledge of the video structure is a form of prior knowledge, Bayesian procedures [18] become a natural solution for these tasks. We therefore introduce a Bayesian framework for segmentation and characterization that is shown to significantly outperform currently used approaches and to capture relevant semantic properties of the content.

The paper is organized as follows. In section II, we argue that high-level structure is prevalent in most video domains and has a direct impact in our ability to understand the video. We identify two important semantic features, shot duration and activity, and introduce statistical models for these features in sections III and IV. The remainder of the paper is devoted to practical applications of these models. In section V, we introduce a Bayesian solution to the problem of temporal video segmentation. A detailed experimental analysis of its performance is carried out in section VI. Section VII then illustrates the semantic content characterization ability of the shot-duration/scene-activity feature-space on a large database of movie clips. Finally, section VIII presents some conclusions and pointers for future work. Preliminary reports on this work were previously presented in [46], and [49].

II. VIDEO STRUCTURE

The most obvious example of a domain where structure plays a significant role in the characterization of video is that of television newscasts. A newscast contains a significant amount of both spatial and temporal structure. Examples of spatial structure are the standardized spatial layouts to which the shots of the anchor-person and the various separators between different sections of the newscast usually obey. Examples of temporal structure are the regularly spaced appearances of the anchor-person, usually indicating the start of a new story, or the standardized temporal intervals in which the different news sections are covered (e.g. the sports section always appears "x" minutes after the start of the newscast). As a consequence of all this structure, it is relatively easy to give a machine the understanding of commands such as "skip ahead to the sports section" without the need to understand all the pixels in every image of the video sequence. This has made the topic of analyzing newscasts a very popular one in the video databases literature [3], [20], [26], [55].

The ability to infer semantic information is directly related to the amount of structure exhibited by the content. While newscasts are at the highly structured end of the spectrum and constitute one of the easiest classes to analyze, the raw output of a personal camcorder exhibits almost no structure and is typically too difficult to characterize semantically [25]. Between these two extrema, there are various types of content for which the characterization has a varying level of difficulty. Our interests are mostly in domains that are more generic than newscasts, but still follow enough content production codes to exhibit a significant amount of structure. A good example of such a domain is that of feature films.

A. Structure in movies

While the analysis of all the elements that contribute to the structure of a movie would require significantly more space than what it is available here, it is worth to point out that there are some well known principles in film theory that can be exploited for the purpose of semantic characterization. In particular, it is well known that the stylistic elements of a movie are closely related to the message conveyed in its story [6], [29], [36]. Historically, these stylistic elements have been grouped into two major categories: the *elements of montage* and the *elements of mise-en-scene*. Montage refers to the temporal structure, namely the aspects of film editing or the way in which the different shots are composed to form the scenes in the movie. On the other hand, mise-en-scene deals with spatial structure, i.e. the composition of each image, and includes variables such as the type of set in which the scene develops, the placement of the actors on the scene, aspects of lighting, focus, camera angles, and so on.

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].

[1] The importance of close-ups is best summarized in the quote from 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" [6].

It is therefore to be expected that the analysis of the elements of montage and mise-en-scene can lead to feature spaces where the content is laid out in a way that would allow a machine to characterize it on a semantic basis. While there is a fundamental element of montage, the shot duration, it is harder to identify a single defining characteristic of mise-en-scene. It is, nevertheless, clear that among the elements of mise-en-scene, one important property for the semantic characterization of a movie is the amount of activity in its shots: while action packed movies typically contain a strong component of active shots, character based movies are mainly composed of scenes with shots (e.g. dialogues) of smaller activity. Furthermore, important events for the semantic classification of action movies, such as explosions, car chases, or fights are typically associated with highly active shots. Finally, the amount of activity is usually correlated with the amount of violence in the content (at least that of a gratuitous nature) and can provide clues for its detection. For these reasons, in this work, we develop computational models for the shot duration and activity.

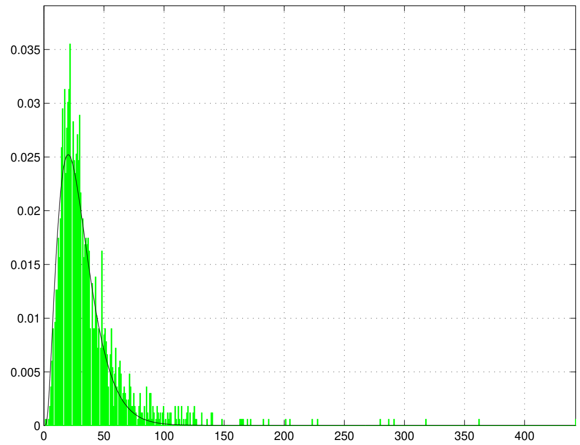

Fig. 1. Shot duration histogram, and maximum likelihood fit obtained with the Erlang distribution.

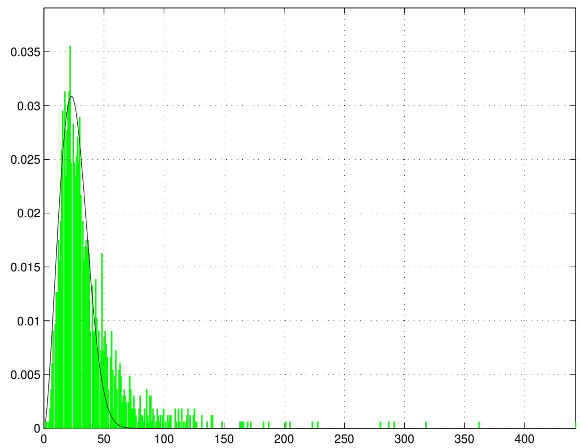

Fig. 2. Shot duration histogram, and maximum likelihood fit obtained with the Weibull distribution.

III. Modeling shot duration

Several probabilistic models can be used for shot duration. Because shot boundaries can be seen as arrivals over discrete, non-overlapping temporal intervals, a Poisson process seems an appropriate prior for the task of boundary detection [14]. However, events generated by Poisson processes have inter-arrival times characterized by the exponential density which is a monotonically decreasing function of time. This is clearly not the case for the shot duration, as can be seen from the histogram of Figures 1 and 2. In this work, we consider two more generic models, the Erlang and Weibull distributions, which provide more realistic models for shot duration.

A. The Erlang model

The first model that we consider for the time elapsed since the previous shot boundary is the Erlang distribution [14]. Denoting by τ the time since the previous boundary, the Erlang distribution is described by

$$\epsilon_{r,\lambda}(\tau) = \frac{\lambda^r \tau^{r-1} e^{-\lambda\tau}}{(r-1)!}. \tag{1}$$

It is a generalization of the exponential density, characterized by two parameters: the order r, and the expected inter-arrival time $(1/\lambda)$ of the underlying Poisson process. When $r = 1$, the Erlang distribution becomes the exponential distribution. For larger values of r, it characterizes the time between the r^{th} order inter-arrival time of the Poisson process. This leads to an intuitive explanation for the use of the Erlang distribution as a model of shot duration: for a given order r, the shot is modeled as a sequence of r events which are themselves the outcomes of Poisson processes. Such events may reflect properties of the shot content, such as "setting the context" through a wide angle view followed by "zooming in on the details" when $r = 2$, or "emotional buildup" followed by "action" and "action outcome" when $r = 3$.

Our experiments show that the Erlang distribution provides a good model for shot duration. Figure 1 presents a shot duration histogram, obtained from the training set to be described in section VI, and its maximum likelihood (ML) Erlang fit obtained according to the procedure described in Appendix . It can be seen that the Erlang fit is a good approximation to the empirical density.

The main limitation of the Erlang model is its dependence on the constant arrival rate assumption [21] inherent to the underlying Poisson process. Because λ is a constant, the expected rate of occurrence of a new shot boundary in the next frame interval is the same if 10 seconds or 1 hour have elapsed since the occurrence of the previous one. This type of behavior is clearly inappropriate for most physical processes and several alternative models have been proposed in the statistics literature to handle the problem. We next consider one such alternative: the *Weibull* distribution [21].

B. The Weibull model

The Weibull distribution generalizes the exponential distribution by considering an expected rate of arrival of new events that is a function of time τ

$$\lambda(\tau) = \frac{\alpha \tau^{\alpha-1}}{\beta^\alpha},$$

and of the parameters α and β, leading to a probability density of the form

$$w_{\alpha,\beta}(\tau) = \frac{\alpha \tau^{\alpha-1}}{\beta^\alpha} \exp\left[-\left(\frac{\tau}{\beta}\right)^\alpha\right]. \quad (2)$$

For most values of β, the distribution does not have heavy tails, and robust estimation procedures are required to avoid high sensitivity to outliers. Figure 2 presents the robust ML Weibull fit, obtained according to the procedure described in Appendix , to the shot duration histogram of Figure 1. Once again we obtain a good approximation to the empirical density estimate provided by the histogram.

IV. Modeling shot activity

Given a sequence of images, a considerable number of methods can be used to obtain an estimate of the amount of activity in the underlying scene. In this work, we consider two metrics that can be derived from low-level image measurements: the *color histogram distance* and the *tangent distance* between successive images in the sequence.

A. Color histogram distance

The color histogram distance has been widely used as a measure of (dis)similarity between images for the purposes of object recognition [16], [42], content-based retrieval [23], [27], [31], [32], [40], and temporal video segmentation [7], [17], [30], [41], [53], [56]. A histogram is first computed for each image in the sequence and the distance between successive histograms is used as a measure of local activity. Among the metrics proposed in the literature, we use the L_1 norm of the histogram difference,

$$\mathcal{D}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{B} |a_i - b_i|, \quad (3)$$

where $\mathbf{a}$ and $\mathbf{b}$ are histograms of successive frames, and B the number of histogram bins. This metric has been shown to perform well for temporal video segmentation [7] and is equivalent (for normalized histograms) to the *histogram intersection* metric [42] commonly used in the retrieval literature.

Since histograms are invariant to most types of motion either by the camera or the objects in the scene, they perform very well for tasks where invariance is a big plus, e.g. video segmentation and retrieval. It is however not clear that invariance is a major advantage for a metric of scene activity, since part of what has to be measured is precisely the amount of motion in the scene. In fact, the metric should only be invariant with respect to camera motions, e.g. pans and zooms, which do not necessarily correlate with activity[3]. The tangent distance [39] between successive images has this property.

[3]For example, pans are prevalent on scenic videos that depict scenes of very low activity.

B. Tangent distance

The key idea behind the tangent distance is that, when subject to spatial transformations, images span manifolds embedded in a much higher dimensional Euclidean space, and a metric invariant to those transformations should measure the distance between those manifolds instead of the distance between other properties of (or features extracted from) the images themselves. However, because the manifolds can be very complex, minimizing the distance between them is a hard optimization problem. The problem can be made tractable by considering instead the minimization of the distance between the tangent planes to the manifolds.

Given two images $M(\mathbf{x})$ and $N(\mathbf{x})$, and a transformation $T_\mathbf{q}$ parameterized by the vector $\mathbf{q}$, the distance between the associated manifolds is

$$\mathcal{D}(M, N) = \min_{\mathbf{p},\mathbf{q}} ||T_\mathbf{q}[M(\mathbf{x})] - T_\mathbf{p}[N(\mathbf{x})]||^2. \quad (4)$$

Assuming, for simplicity, that one of the images (M) is fixed, and replacing $T_\mathbf{p}[N(\mathbf{x})]$ by the tangent plane at the point $N(\mathbf{x})$, we obtain the (one-sided) tangent distance

$$\mathcal{D}(M, N) = \min_{\mathbf{p}} ||M(\mathbf{x}) - N(\mathbf{x}) - (\mathbf{p} - \mathbf{I})^T \nabla_\mathbf{p} T_\mathbf{p}[N(\mathbf{x})]||^2. \quad (5)$$

Many transformations can be used in this equation. Because we are mostly interested in invariance against camera motion, we consider the set of affine transformations $T_\mathbf{p}[N(\mathbf{x})] = N(\psi(\mathbf{x}, \mathbf{p}))$, with

$$\psi(\mathbf{x}, \mathbf{p}) = \begin{bmatrix} x & y & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x & y & 1 \end{bmatrix} \mathbf{p} = \mathbf{\Phi}(\mathbf{x})\mathbf{p}, \quad (6)$$

capable of compensating for translation (panning), scaling (zooming), in-plane rotation, and shearing. The cost function of equation (5) can be minimized using a multiresolution variant of Newton's method [5], leading to the following algorithm [2], [47]. For a given level l of the multiresolution decomposition:

1) Compute $N'(\mathbf{x})$ by warping image $N(\mathbf{x})$ according to the best current estimate of $\mathbf{p}$, and compute its spatial gradient $\nabla_\mathbf{x} N'(\mathbf{x})$.
2) Update the estimate of $\mathbf{p}_l$ according to

$$\mathbf{p}_l^{n+1} = \mathbf{p}_l^n + \alpha \left[\sum_\mathbf{x} \mathbf{\Phi}(\mathbf{x})^T \nabla_\mathbf{x} N'(\mathbf{x}) \nabla_\mathbf{x}^T N'(\mathbf{x}) \mathbf{\Phi}(\mathbf{x})^T\right]^{-1} \times \left[\sum_\mathbf{x} [M(\mathbf{x}) - N'(\mathbf{x})]\, \mathbf{\Phi}(\mathbf{x})^T \nabla_\mathbf{x} N'(\mathbf{x})\right].$$

3) Stop if convergence, otherwise go to 1.

Once the final $\mathbf{p}_l$ is obtained, it is passed to the multiresolution level below, by simply doubling the translation parameters. The rescaled vector is then used as initial estimate at the level $l + 1$, and the above process repeated. Once this iterative procedure has converged for all levels of the multiresolution decomposition, the tangent distance between the images is computed through equation (5), using the optimal parameter vector $\mathbf{p}$.

C. Statistical models of activity

Given a set of activity features, the second step of the modeling consists of finding good statistical representations for them. Here, it is important to realize that there may not be a universal model applicable all of the time but different models may be best suited for the different states through which the video progresses. For example, the principle of *continuity in editing* states that, in order not to confuse the viewer, the first frames after a shot transition should be significantly different than the last frames before it [36]. Thus, during a typical shot transition, any activity metric is likely to take values that are significantly different from those observed within each shot. For simplicity, in this work we restrict ourselves to a video model composed of two states: "regular frames" and "shot transitions". The fundamental principles are however applicable to more complex models with more states.

1) Modeling state densities: We start by defining two states: $\mathcal{S} = 0$ associated with regular frames, and $\mathcal{S} = 1$ associated with shot boundaries. These states are not directly measurable from the video, but can only be inferred from the activity features described above. We designate these features by $\mathcal{D}$. Once the states are identified, the goal is to design good conditional density models for the observed activity features given each state. This is not as easy as in the case of the shot duration since there do not seem to exist simple densities that can provide a good fit to the empirical observations. In order to overcome this limitation, we rely on generic mixture models.

A mixture density [44] is defined as

$$P(\mathcal{D}) = \sum_{i=1}^{C} \pi_i p(\mathcal{D}|\phi_i),$$

where C is the number of mixture components, π_i the probability of the i^{th} class, $p(\mathcal{D}|\phi_i)$ the likelihood of the observed data for this class, and ϕ_i the corresponding parameter vector. For example, in the case of a Gaussian mixture component $\phi_i = \{\mu_i, \sigma_i\}$ and

$$p(\mathcal{D}|\phi_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(\mathcal{D}-\mu_i)^2}{2\sigma_i^2}},$$

while for a Erlang component $\phi_i = \{r_i, \lambda_i\}$ and $p(\mathcal{D}|\phi_i) = \epsilon_{r_i,\lambda_i}(\mathcal{D})$ (as defined by equation (1)), and for a uniform component $\phi_i = \{a_i, b_i\}$ and $p(\mathcal{D}|\phi_i) = 1/(b_i - a_i)$, where a_i and b_i are the extrema of the uniform density. The basic idea behind this model is that each observation of $\mathcal{D}$ is generated in two steps: first a class is selected according to the π_i, and the observation is then drawn from the corresponding class likelihood.

Given an observed sample $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$, the ML estimates for the mixture parameters $\{\pi_i, \phi_i\}, i = 1, \ldots, C$, can be obtained with the expectation-maximization (EM) algorithm [12], [35]. EM is an iterative procedure that iterates between an expectation (E) and a maximization (M) step. The E-step computes the posterior probability, h_{ij}, that a sample point j was drawn from mixture component i

$$h_{ij} = P(class_i|\mathcal{D}_j) = \frac{p(\mathcal{D}_j|\phi_i)\pi_i}{\sum_{k=1}^{C} p(\mathcal{D}_j|\phi_k)\pi_k}, \forall i, j.$$

The M-step then finds the set of parameters that maximize the weighted log-likelihood of the sample $\mathcal{D}$ given the class assignments computed in the E-step

$$\{\pi_i^{new}, \phi_i^{new}\} = \arg \max_{\{\pi_i, \phi_i\}} \sum_j h_{ij}\{\log p(\mathcal{D}_j|\phi_i) + \log \pi_i\},$$

under the constraint $\sum_i \pi_i^{new} = 1$.

The algorithm is guaranteed to converge to at least a local maximum of the likelihood of the sample $\mathcal{D}$ [12]. Furthermore, it can be shown [34], [35] using Lagrange multipliers that, independently of the conditional densities $p(\mathcal{D}_j|\phi_i)$ considered in the model, the optimal values for the π_i updates are

$$\pi_i^{new} = \frac{\sum_j h_{ij}}{\sum_{ij} h_{ij}}.$$

The M-step updates for the remaining parameters are similar to standard ML estimates, with the difference that we now have to take into account the weights h_{ij}. Usually this leads to a slight modification of the expressions obtained by ML. For example, in the Gaussian case [34], [35],

$$\mu_i^{new} = \frac{\sum_j h_{ij} x_j}{\sum_j h_{ij}}, \quad \sigma_i^{new} = \frac{\sum_j h_{ij}(x_j - \mu_i^{new})^2}{\sum_j h_{ij}}, \tag{7}$$

while in the Erlang case r_i, and λ_i can be found by a procedure similar to the one described in Appendix using a slightly modified expression for the optimal value of λ

$$\lambda_i^{new} = \frac{r_i \sum_j h_{ij}}{\sum_j h_{ij}\mathcal{D}_j}.$$

2) Regular frames: Figure 3 presents the histogram of the activity features for the "regular frames" state for both the histogram and tangent distance metrics. It is clear that the distributions are very similar: asymmetric about their mean, always positive and concentrated near zero. This suggests that a mixture of Erlang distributions is an appropriate model for this state, a suggestion that is confirmed by the fits obtained with EM, that are also depicted in the figure. In both cases, we have used a mixture of three Erlang components and a uniform component. The uniform component accounts for the tails of the distribution and, due to its small amplitude, is not perceptible in the figures.

3) Shot transitions: Figure 4 presents the histogram of the activity features for the shot transition state together with the fit obtained with a mixture of a Gaussian and a uniform densities. Once again, the uniform density accounts for the tails of the distribution. When the shot transition state is considered, a simple Gaussian model appears to provide a reasonable approximation to the empirical density. This model has indeed been previously used as a basis for setting shot detection thresholds [56].

The fact that the density estimates depicted in Figures 1 to 4 approximate well the empirical observations indicates that the models now introduced are good representations for shot duration and activity. However, because the ultimate measure of effectiveness of any given model is the degree to which it leads to effective solutions for objective tasks, in the remainder of the paper we concentrate on two such tasks: shot segmentation and semantic content characterization.

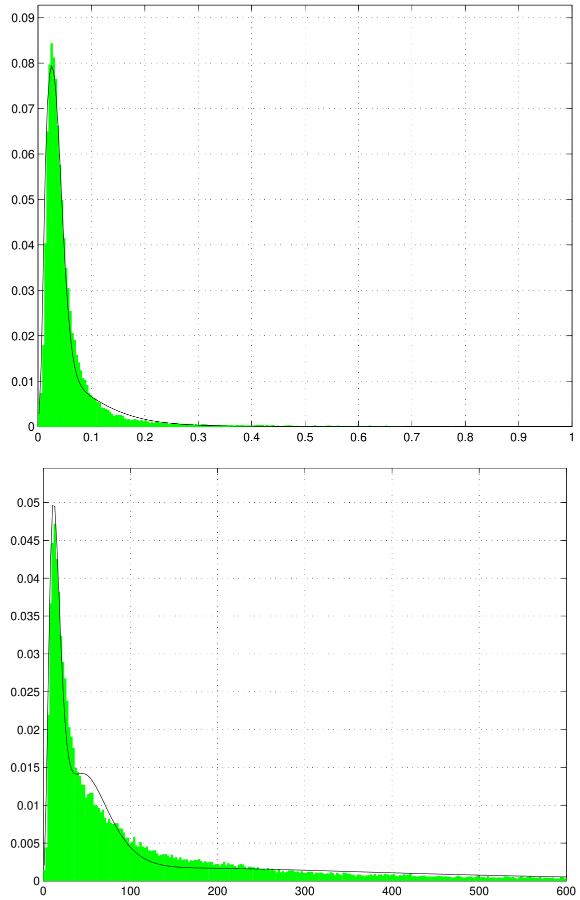

Fig. 3. Conditional activity histogram given that there are no shot changes, and best fit by a mixture with three Erlang components and a uniform component. Top: histogram distance. Bottom: tangent distance.

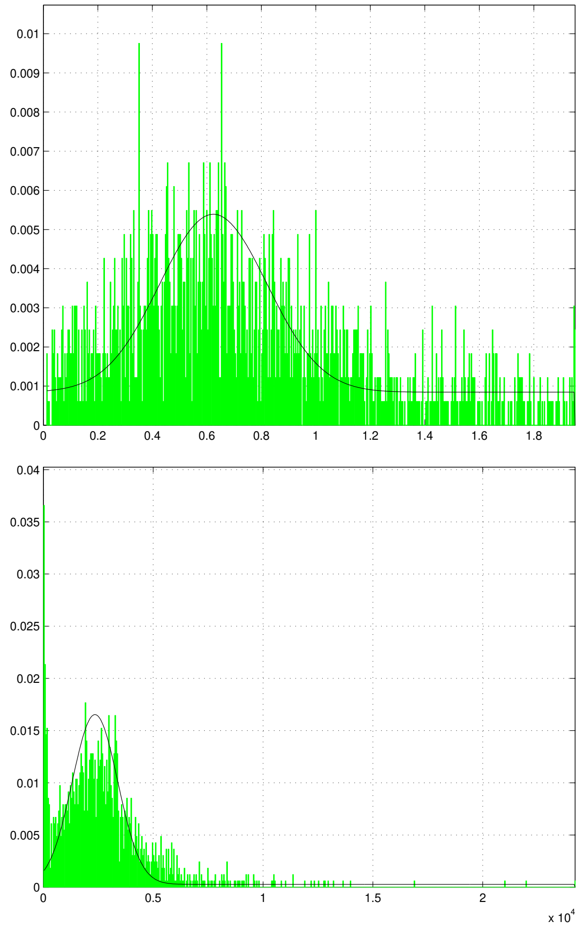

Fig. 4. Conditional activity histogram for shot transitions, and best fit by a mixture with a Gaussian and a uniform component. Top: histogram distance. Bottom: tangent distance.

V. A Bayesian framework for shot segmentation

Because shot segmentation is a pre-requisite for virtually any task involving the understanding, parsing, indexing, characterization, or categorization of video, the grouping of video frames into shots has been an active topic of research in the area of content-based retrieval [7], [17], [19], [30], [41], [53], [56]. Extensive evaluation of various approaches has shown that simple thresholding of histogram distances performs surprisingly well and is difficult to beat [7], [17]. In this work, we consider an alternative formulation that regards the problem as one of statistical inference between two hypothesis:

- $\mathcal{H}_0$: no shot boundary occurs between the two frames under analysis ($\mathcal{S} = 0$),
- $\mathcal{H}_1$: a shot boundary occurs between the two frames ($\mathcal{S} = 1$).

In this setting, the optimal decision is provided by a likelihood ratio test where $\mathcal{H}_1$ is chosen if

$$\mathcal{L} = \log \frac{P(\mathcal{D}|\mathcal{S} = 1)}{P(\mathcal{D}|\mathcal{S} = 0)} > 0, \tag{8}$$

and $\mathcal{H}_0$ is chosen otherwise. It can be shown that standard thresholding is a particular case of this statistical formulation, in which both conditional densities are assumed to be Gaussians with the same covariance. From the discussion in the previous section, we know that this does not hold for real video. One further limitation of the thresholding model is that it does not take into account the fact that the likelihood of a new shot transition is dependent on how much time has elapsed since the previous one. On the other hand, the statistical formulation can easily incorporate the shot duration models developed in section III. For this, we start by making some conventions with regards to notation.

A. Notation

Because video is a discrete process, characterized by a given frame rate, shot boundaries are not instantaneous, but last for one frame period. To account for this, states are defined over time intervals, i.e. instead of $\mathcal{S}_t = 0$ or $\mathcal{S}_t = 1$, we have $\mathcal{S}_{t,t+\delta} = 0$ or $\mathcal{S}_{t,t+\delta} = 1$, where t is the start of a time interval, and δ its duration. We designate the features observed during the interval $[t, t+\delta]$ by $\mathcal{D}_{t,t+\delta}$.

To simplify the notation, we reserve t for the temporal instant at which the last shot boundary has occurred and make all temporal indexes relative to this instant. I.e. instead of $\mathcal{S}_{t+\tau,t+\tau+\delta}$ we write $\mathcal{S}_{\tau,\tau+\delta}$, or simply $\mathcal{S}_\delta$ if $\tau = 0$. Furthermore, we reserve the symbol δ for the duration of the interval between successive frames (inverse of the frame rate), and use the same notation for a simple frame interval and a vector of frame intervals (the temporal indexes being themselves enough to avoid ambiguity). I.e., while $\mathcal{S}_{\tau,\tau+\delta} = 0$ indicates that no shot boundary is present in the interval $[t+\tau, t+\tau+\delta]$, $\mathcal{S}_{\tau+\delta} = \mathbf{0}$ indicates that no shot boundary has occurred in any of the frames between t and $t+\tau+\delta$. Similarly, $\mathcal{D}_{\tau+\delta}$ represents the vector of observations in $[t, t+\tau+\delta]$.

B. Bayesian formulation

Given that there is a shot boundary at time t and no boundaries occur in the interval $[t, t+\tau]$, the posterior probability that the next shot change happens during the interval $[t+\tau, t+\tau+\delta]$ is, using Bayes rule,

$$\begin{aligned} &P(\mathcal{S}_{\tau,\tau+\delta} = 1|\mathcal{S}_\tau = \mathbf{0}, \mathcal{D}_{\tau+\delta}) \\ &\quad = \gamma P(\mathcal{D}_{\tau+\delta}|\mathcal{S}_\tau = \mathbf{0}, \mathcal{S}_{\tau,\tau+\delta} = 1) P(\mathcal{S}_{\tau,\tau+\delta} = 1|\mathcal{S}_\tau = \mathbf{0}), \end{aligned}$$

where γ is a normalizing constant (also known as partition function). Similarly, the probability that there is no change in $[t+\tau, t+\tau+\delta]$ is

$$\begin{aligned} &P(\mathcal{S}_{\tau,\tau+\delta} = 0|\mathcal{S}_\tau = \mathbf{0}, \mathcal{D}_{\tau+\delta}) = \\ &\quad = \gamma P(\mathcal{D}_{\tau+\delta}|\mathcal{S}_{\tau+\delta} = \mathbf{0}) P(\mathcal{S}_{\tau,\tau+\delta} = 0|\mathcal{S}_\tau = \mathbf{0}), \end{aligned}$$

and the *posterior odds ratio* [18] between the two hypothesis is

$$\begin{aligned} &\frac{P(\mathcal{S}_{\tau,\tau+\delta} = 1|\mathcal{S}_\tau = \mathbf{0}, \mathcal{D}_{\tau+\delta})}{P(\mathcal{S}_{\tau,\tau+\delta} = 0|\mathcal{S}_\tau = \mathbf{0}, \mathcal{D}_{\tau+\delta})} = \\ &\quad = \frac{P(\mathcal{D}_{\tau+\delta}|\mathcal{S}_\tau = \mathbf{0}, \mathcal{S}_{\tau,\tau+\delta} = 1) P(\mathcal{S}_{\tau,\tau+\delta} = 1|\mathcal{S}_\tau = \mathbf{0})}{P(\mathcal{D}_{\tau+\delta}|\mathcal{S}_{\tau+\delta} = \mathbf{0}) P(\mathcal{S}_{\tau,\tau+\delta} = 0|\mathcal{S}_\tau = \mathbf{0})} \\ &\quad = \frac{P(\mathcal{D}_{\tau,\tau+\delta}|\mathcal{S}_{\tau,\tau+\delta} = 1)}{P(\mathcal{D}_{\tau,\tau+\delta}|\mathcal{S}_{\tau,\tau+\delta} = 0)} \frac{P(\mathcal{S}_{\tau,\tau+\delta} = 1|\mathcal{S}_\tau = \mathbf{0})}{P(\mathcal{S}_{\tau,\tau+\delta} = 0|\mathcal{S}_\tau = \mathbf{0})} \\ &\quad = \frac{P(\mathcal{D}_{\tau,\tau+\delta}|\mathcal{S}_{\tau,\tau+\delta} = 1)}{P(\mathcal{D}_{\tau,\tau+\delta}|\mathcal{S}_{\tau,\tau+\delta} = 0)} \frac{P(\mathcal{S}_{\tau,\tau+\delta} = 1, \mathcal{S}_\tau = \mathbf{0})}{P(\mathcal{S}_{\tau+\delta} = \mathbf{0})}, \end{aligned} \tag{9}$$

where we have assumed that, given $\mathcal{S}_{\tau,\tau+\delta}$, $\mathcal{D}_{\tau,\tau+\delta}$ is independent of all other $\mathcal{D}$ and $\mathcal{S}$. In this expression, while the first term on the right hand side is the ratio of the conditional likelihoods of activity given the state sequence, the second term is simply the ratio of probabilities that there may (or not) be a shot transition τ units of time after the previous one. In the Bayesian terminology, the shot duration density becomes a *prior* for the segmentation process. This is intuitive since knowledge about the shot duration is a form of prior knowledge about the structure of the video that should be used to favor segmentations that are a priori more plausible.

Assuming further that $\mathcal{D}$ is stationary, defining $\Delta_\tau = [t+\tau, t+\tau+\delta]$, considering the probability density function $p(\tau)$ for the time elapsed until the first scene change after t, and taking logarithms, leads to a *log posterior odds ratio* $\mathcal{L}_{post}$ of the form

$$\mathcal{L}_{post} = \log \frac{P(\mathcal{D}_{\Delta_\tau}|\mathcal{S}_{\Delta_\tau} = 1)}{P(\mathcal{D}_{\Delta_\tau}|\mathcal{S}_{\Delta_\tau} = 0)} + \log \frac{\int_\tau^{\tau+\delta} p(\alpha) d\alpha}{\int_{\tau+\delta}^\infty p(\alpha) d\alpha}. \tag{10}$$

The optimal answer to the question if a shot change occurs or not in $[t+\tau, t+\tau+\delta]$ is thus to declare that a boundary exists if

$$\log \frac{P(\mathcal{D}_{\Delta_\tau}|\mathcal{S}_{\Delta_\tau} = 1)}{P(\mathcal{D}_{\Delta_\tau}|\mathcal{S}_{\Delta_\tau} = 0)} \geq \log \frac{\int_{\tau+\delta}^\infty p(\alpha) d\alpha}{\int_\tau^{\tau+\delta} p(\alpha) d\alpha} = \mathcal{T}(\tau), \tag{11}$$

and that there is no boundary otherwise. By comparing this equation with equation (8), it is clear that the inclusion of the prior model for shot duration transforms the fixed thresholding approach of that equation into an adaptive one, where the threshold depends on how much time has elapsed since the previous shot boundary. We next study the behavior of this threshold for each of the shot duration models introduced in section III.

1) The Erlang model: Using the results of Appendix , it can be seen that the threshold of equation (11) is particularly simple to compute under the Erlang assumption, where

$$\int_a^b \epsilon_{r,\lambda}(\tau) d\tau = \frac{1}{\lambda} \sum_{i=1}^r [\epsilon_{i,\lambda}(a) - \epsilon_{i,\lambda}(b)]. \tag{12}$$

The log posterior odds ratio test for detecting a shot change - equation (11) - then becomes

$$\log \frac{P(\mathcal{D}_{\Delta_\tau}|\mathcal{S}_{\Delta_\tau} = 1)}{P(\mathcal{D}_{\Delta_\tau}|\mathcal{S}_{\Delta_\tau} = 0)} \geq \mathcal{T}_\epsilon(\tau), \tag{13}$$

where

$$\mathcal{T}_\epsilon(\tau) = \log \frac{\sum_{i=1}^r \epsilon_{i,\lambda}(\tau+\delta)}{\sum_{i=1}^r [\epsilon_{i,\lambda}(\tau) - \epsilon_{i,\lambda}(\tau+\delta)]}. \tag{14}$$

The typical temporal behavior of this threshold is presented in Figure 5. While in the initial segment of the shot, the threshold is very high and shot changes are very unlikely to be accepted, the threshold decreases as the scene progresses increasing the likelihood that shot boundaries will be declared.

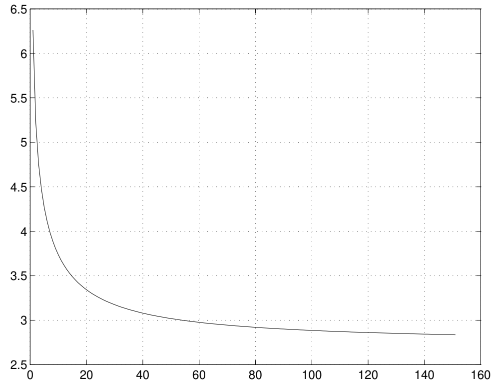

Fig. 5. Log posterior odds ratio threshold as a function of the time elapsed since the beginning of the current shot.

Even though the qualitative behavior of the threshold is what one would desire, a closer observation of the figure reveals the major limitation of the Erlang prior: its steady-state behavior. Ideally, in addition to decreasing monotonically over time, the threshold should not be lower bounded by a positive value as this may lead to situations in which its steady-state value is high enough to miss several consecutive shot boundaries. Instead, the threshold should at some point become negative, guaranteeing that, in steady-state, any shot boundary detectable without a prior is still detectable when the prior is introduced. Unfortunately, such a steady-state behavior is not achievable with an Erlang prior, for which

$$\lim_{\tau \to \infty} \mathcal{T}_\epsilon(\tau) = \frac{e^{-\lambda\delta}}{1 - e^{-\lambda\delta}} > 0.$$

This limitation is a consequence of the constant arrival rate assumption discussed in section III and can be avoided by relying instead on the Weibull model.

2) The Weibull model: Similarly to the Erlang prior, the threshold of equation (11) is easy to compute under the Weibull model. As shown in Appendix ,

$$\int_a^b w_{\alpha,\beta}(\tau)d\tau = \exp\left[-\left(\frac{a}{\beta}\right)^\alpha\right] - \exp\left[-\left(\frac{b}{\beta}\right)^\alpha\right], \quad (15)$$

from which

$$\begin{aligned}\mathcal{T}_w(\tau) &= \log \frac{\exp\left[-\left(\frac{\tau+\delta}{\beta}\right)^\alpha\right]}{\exp\left[-\left(\frac{\tau}{\beta}\right)^\alpha\right] - \exp\left[-\left(\frac{\tau+\delta}{\beta}\right)^\alpha\right]} \\ &= -\log\left\{\exp\left[\frac{(\tau+\delta)^\alpha - \tau^\alpha}{\beta^\alpha}\right] - 1\right\} \quad (16)\end{aligned}$$

and the threshold has the temporal behavior illustrated by Figure 6. Unlike the threshold associated with the Erlang prior, $\mathcal{T}_w(\tau)$ tends to $-\infty$ when τ grows without bound. This guarantees that a new shot boundary will always be found if one waits long enough.

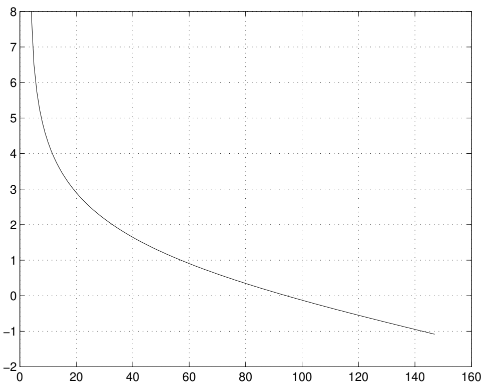

Fig. 6. Log posterior odds ratio threshold for Weibull distribution.

In summary, we see that both the Erlang and the Weibull prior lead to adaptive thresholds that are more intuitive than the fixed threshold commonly employed for shot segmentation. This suggests that the Bayesian approach may have higher accuracy. In the next section we confirm this intuition with detailed experimental results.

VI. Segmentation Results

The performance of the Bayesian shot detection module was evaluated on a database containing 23 promotional movie trailers for commercially released feature films. Each trailer consists of 2 to 5 minutes of video and the total number of shots in the database is 1959. The movie titles are presented in Table I. In all experiments, performance was evaluated by the *leave-one-out* method, i.e. one trailer was held for evaluation of the segmentation accuracy and all the remaining were included in a training set used to learn model parameters. Ground truth was obtained by manual segmentation of all the trailers.

We start by evaluating the relative performance of Bayesian models with different shot duration priors and compare it against the best possible performance achievable with a fixed threshold. For the latter approach, the value of the optimal threshold is obtained by brute-force, i.e. testing several values and selecting the one which performed the best. For completeness, besides the Erlang and Weibull priors we also tested the performance of the Poisson prior.

The first experiment was designed to evaluate the dependence of segmentation accuracy on the free variables of the model, namely the number of components in each of the mixture densities. Figure 7 presents graphs of the number of segmentation errors on the entire database for the three priors and two feature sets (histogram and tangent distance) considered. The curves in each graph depict the total error as a function of the number of Gaussians in the scene change activity model. Each curve corresponds to a different number of Erlang components in the activity model associated with regular frames. Also shown, as a constant line, in each graph is the best performance achieved with a fixed threshold.

Several observations can be made from the figure. First, while the Poisson prior leads to worse accuracy than that achievable with the static threshold, it is clear that both the Erlang and the Weibull priors lead to significant improvement over the performance achieved in the non-Bayesian setting. Second, the histogram distance performs better than the tangent distance. This was expected, given the discussion in section IV. The main point is however that the Bayesian framework is generic and improves the segmentation accuracy for both distance metrics.

Third, performance is insensitive to the number of Gaussians in the scene-change activity model. In fact, most of the curves are approximately constant. There is, however, some dependence on the number of Erlang components in the activity model for regular frames. In particular, a single component is insufficient for both feature sets, two components lead to the best results for the histogram distance, and three are required by the tangent distance. It is however relevant to notice that for a given set of distance features the model that performs best for all priors tends to be the same. This indicates that the Bayesian model is robust against possible mismatches in prior selection or poor parameter estimates.

Finally, the Weibull prior achieves the overall best performance. Not only it leads to the smallest number of errors for both distance metrics, but it is also more robust than the other priors. Notice how, once the minimum number of Erlang components required to obtain a good fit to the underlying density is reached, the curves are approximately flat and exhibit similar values. The better performance of the Weibull prior is also visible in Figure 8, which depicts the total number of errors, false positives, and missed boundaries for the best model associated with each prior. Notice that with this prior the Bayesian approach decreases the error rate of the standard static threshold by 15 to 20% and that the 20% gain is obtained with the best distance metric (histogram).

With respect to false positives and misses, the use of the Weibull prior leads to better performance than all the other approaches when the tangent distance features are used. In the

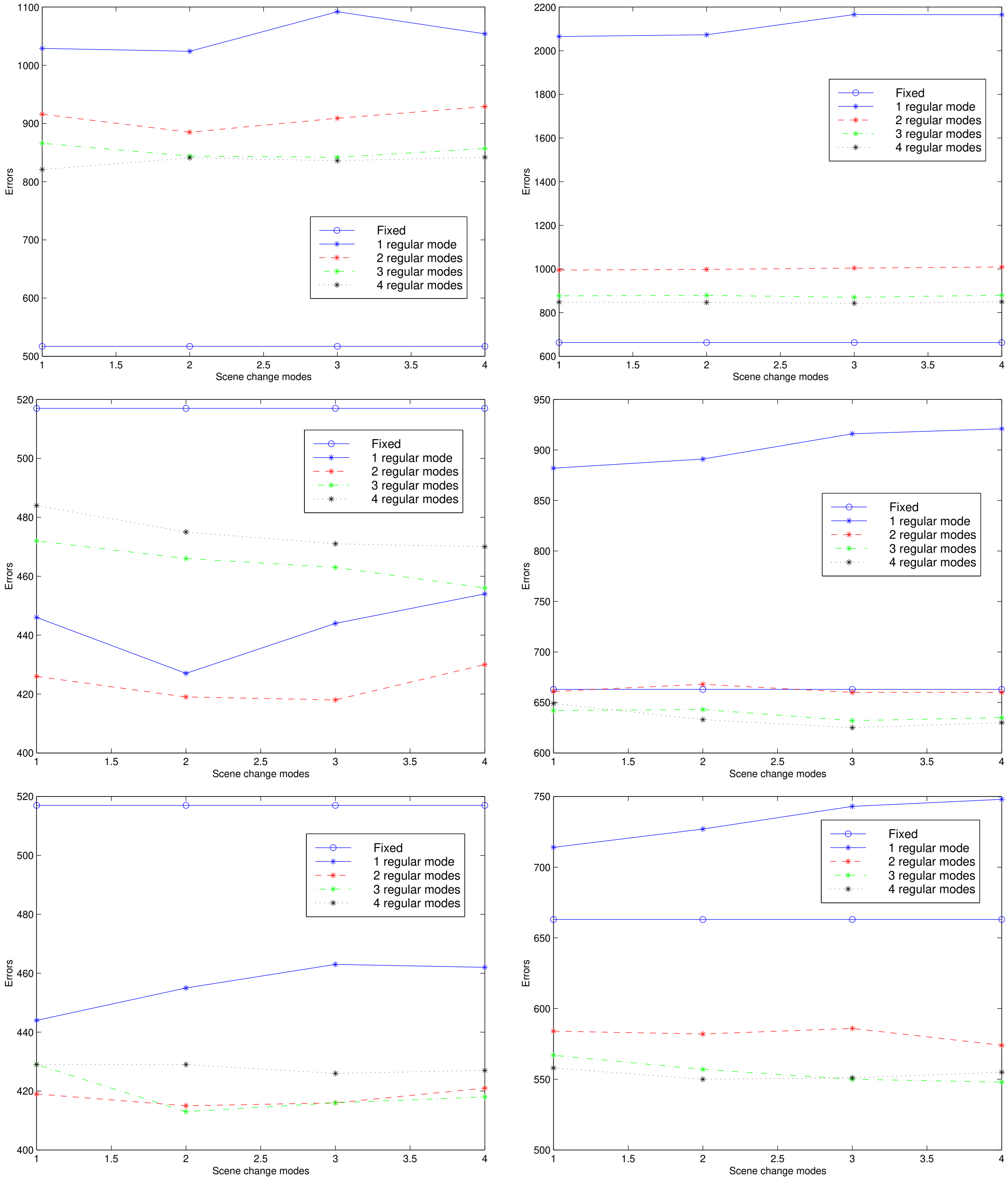

Fig. 7. Total error as a function of the number of model components. Left: histogram distance. Right: tangent distance. From top to bottom: Poisson, Erlang, and Weibull priors. Errors are shown as a function of the number of Gaussians in the scene change activity model given the number of Erlang components in the activity model for regular frames. “Fixed” is the best performance achieved with a static threshold.

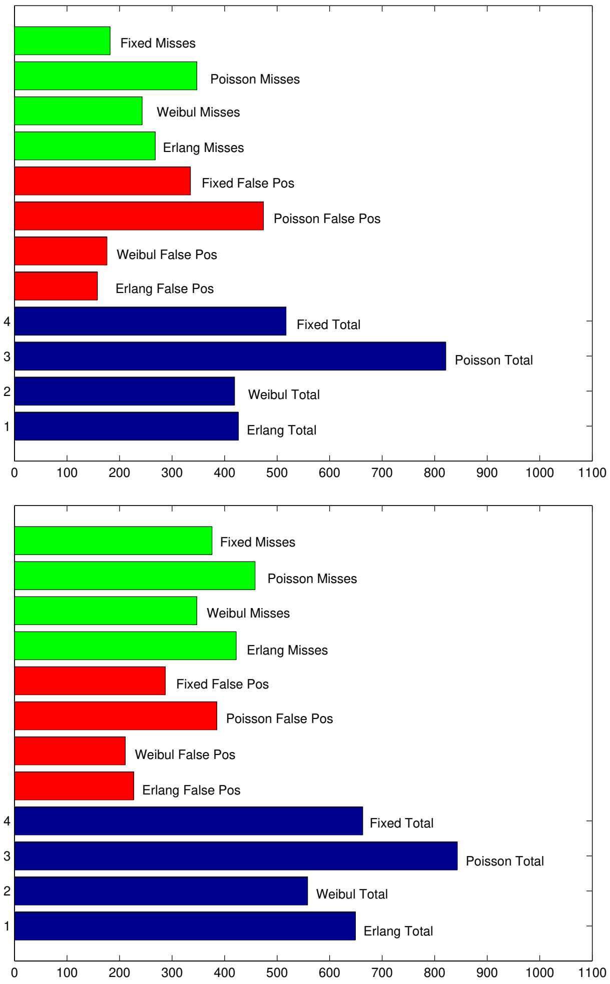

Fig. 8. Total number of errors, false positives, and missed boundaries for the various approaches. Left: histogram distance. Right: tangent distance.

case of histogram distances, there does not seem to be a clear cut winner. While the Erlang prior leads to the smallest number of false positives, the fixed threshold leads to the smallest number of misses. The Weibull prior ranks second in each class, providing a good compromise between the two types of errors. Overall, one can conclude that the Weibull prior has the best performance in terms of the trade-off between false positives and misses.

One final issue of practical concern is the robustness of the segmentation against inaccuracies in the estimates of the prior parameters. This is particularly relevant in the Weibull case since, as discussed in section III-B, it leads to a threshold that is unbounded and goes to $-\infty$ for large τ. It is thus possible that poor parameter estimates may lead to a threshold that decays too quickly originating too many false positives. Because we are relying on a robust estimate of the β parameter, the estimates of this parameter will be, by definition, insensitive to outliers. It remains to investigate the effect of poor estimates of the α parameter on the segmentation performance. For this, we have conducted a series of experiments where the performance was evaluated for pre-determined values of α ranging from 1 to 4 with intervals of 0.2, on the histogram distance feature set.

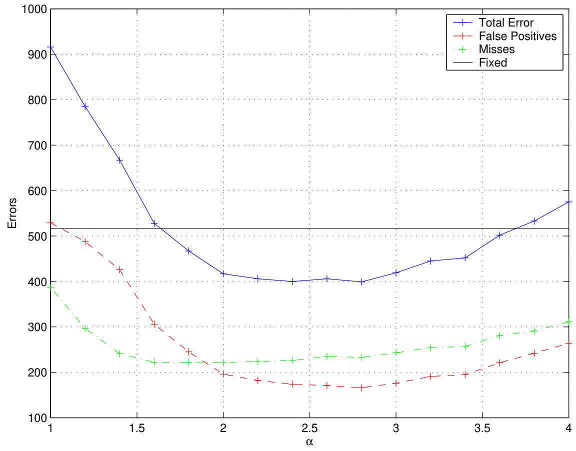

Fig. 9. Total number of errors, false positives, and missed boundaries as a function of α for the Weibull prior and the histogram distance features. The straight line is the best performance achieved with a fixed threshold.

Figure 9 presents the results of this experiment. It is clear that, for a large range ($\alpha \in [1.7, 3.7]$), the performance attained with the Weibull prior is superior to that of the fixed threshold and over a significant range ($\alpha \in [2, 3]$) it is very close to the optimal. Furthermore, both the number of false positives and misses are also approximately constant over a significant range of α.

The reasons for the improved performance of Bayesian segmentation are illustrated by Figure 10, which presents the evolution of the thresholding process for one of the trailers in the database ("blankman"). Two thresholding approaches are depicted: the one derived from the Bayesian formulation with the Weibull prior and standard fixed thresholding. The activity features are, in both cases, histogram distances.

Notice that the adaptive behavior of the Bayesian threshold significantly increases the robustness against spurious peaks of the activity metric originated by events such as very fast motion, explosions, camera flashes, etc. This is visible, for example, in the shot between frames 772 and 794 which depicts a scene composed of fast moving black and white graphics where figure and ground are frequently reversed. While the Bayesian procedure originates a single false-positive, fixed thresholding produces several. This is despite the fact that we have complemented the plain fixed threshold with commonly used heuristics, such as rejecting sequences of consecutive shot boundaries [7]. The vanilla fixed threshold would originate many more errors. Figure 11 presents a few frames from this shot, illustrating the type of motion and the intensity variations contained in it.

VII. Semantic characterization

We have argued in sections I and IV that a coarse semantic characterization can be achieved by mapping each video stream in a video library into a two-dimensional feature space capturing the average shot duration and activity. In order to evaluate the characterization ability of this transformation, we applied it to the trailer library introduced above.

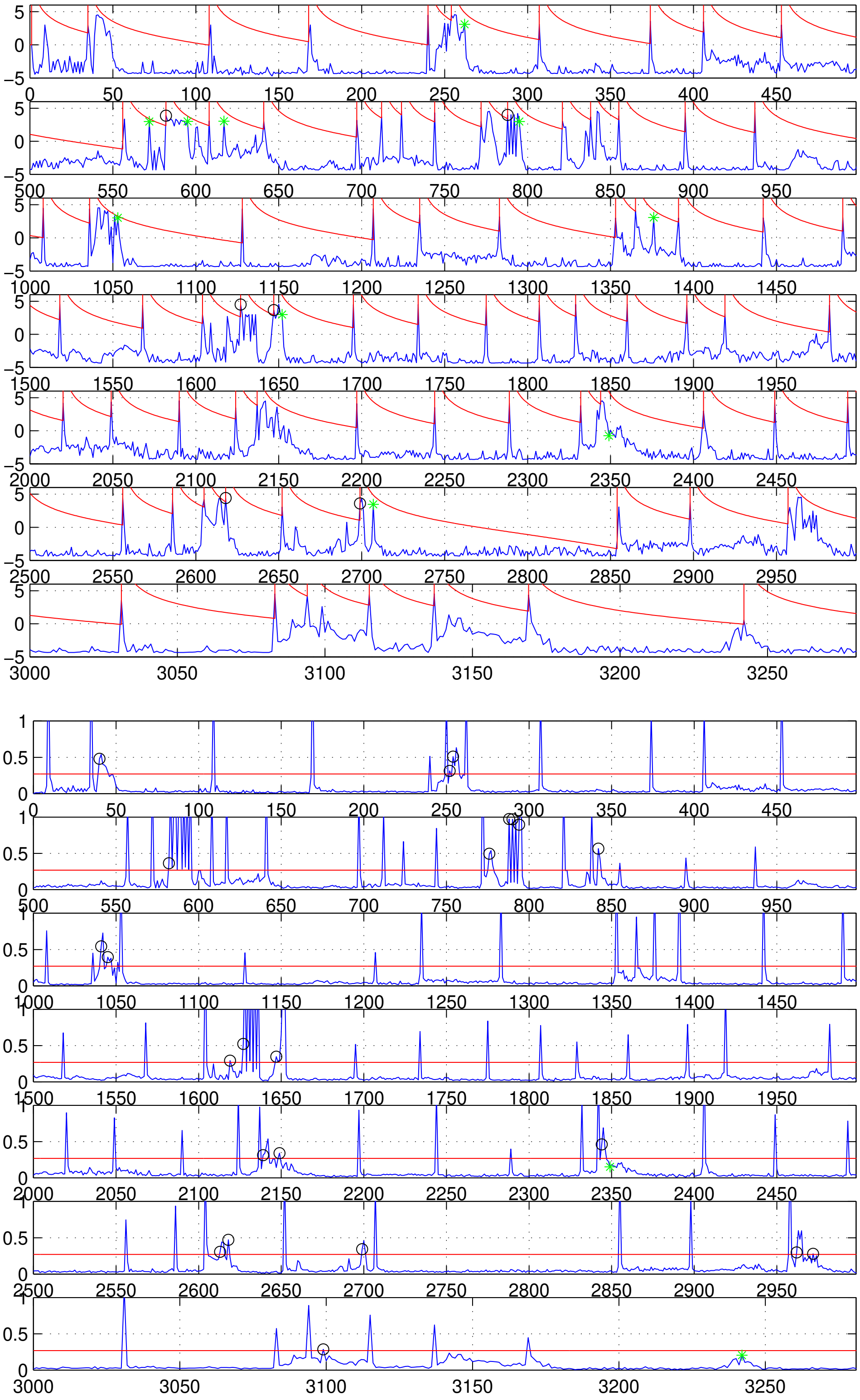

Fig. 10. Evolution of the thresholding process for a challenging trailer. Top: Bayesian segmentation. The likelihood ratio and the Weibull threshold are displayed. Bottom: Fixed thresholding. The histogram distances and the optimal threshold (determined by the *leave-one-out* method using the other trailers in the database) are presented. In both graphs, missed boundaries are signaled by circles, while false positives are indicated by stars.

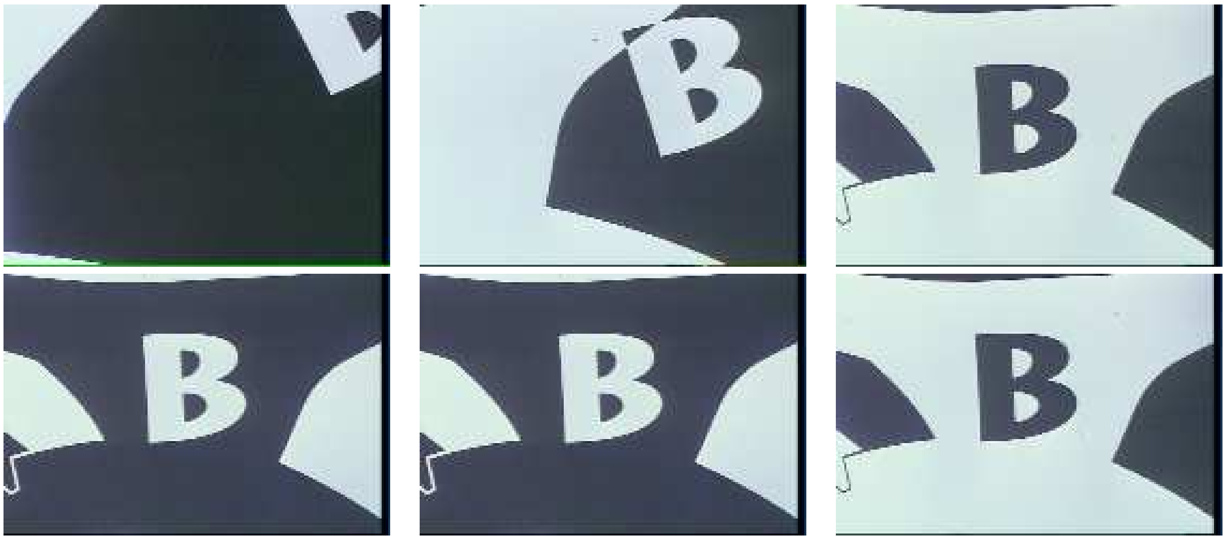

Fig. 11. Six Frames from a shot in the database displayed in raster-scan order. Notice the fast motion and the reversals of foreground/background color.

TABLE I
TITLES OF THE ENTRIES IN THE MOVIE DATABASE AND NAMES THAT APPEAR ON FIGURE 12.

Movie	Legend
"Circle of Friends"	circle
"French Kiss"	french
"Miami Rhapsody"	miami
"The Santa Clause"	santa
"Exit to Eden"	eden
"A Walk in the Clouds"	clouds
"While you Were Sleeping"	sleeping
"Bad Boys"	badboys
"Junior"	junior
"Crimson Tide"	tide
"The Scout"	scout
"The Walking Dead"	walking
"Ed Wood"	edwood
"The Jungle Book"	jungle
"Puppet Master"	puppet
"A Little Princess"	princess
"Judge Dredd"	dredd
"The River Wild"	riverwild
"Terminal Velocity"	terminal
"Blankman"	blankman
"In the Mouth of Madness"	madness
"Street Fighter"	fighter
"Die Hard: With a Vengeance"	vengeance

Figure 12 shows how the movies populate the feature space. The features were obtained by segmenting the video into shots and simply measuring the average duration of each shot and the average value of the activity feature for the regular frames in the shot. Results are presented for the two activity metrics discussed in the previous sections and were normalized to $[0, 1]$ by dividing by the maximum value along each axis. The segmentations that ranked best in the previous section were the ones employed. We also performed a search in the *Internet Movie Database* (IMDB) [1] for the *genre* assigned to each movie by the *Motion Picture Association of America*. Three major classes were identified: *romance/comedy*, *action*, and *other* (which includes *horror*, *drama*, and *adventure*). There were not enough points in the movie sample to further subdivide the other class in a meaningful way. The genre classes are indicated in the plots by the symbol used to represent each movie.

While there are not enough points in the sample to take definitive conclusions, several interesting observations can be made from the figure. First, while there are differences in the details, the overall behavior is the same for the two graphs. In particular, the points seem to obey a law of the type $length \times activity = constant$. This is particularly interesting because the existence of a related law, $character \times action = constant$ has been postulated in the film theory literature [28]. This seems to 1) confirm the fact that the metrics of activity on which we have relied are a good indicator for the *action* content of a movie, and 2) indicate that the shot length is a good metric for the amount of *character* in a movie. The later relationship is intuitive, since the construction of complex characters requires extensive use of dialogue and this is inherently lengthy.

Second, there seems to be a clear separation between the three semantic classes in the activity/length feature space. In particular, movies of the *romance* and *comedy* genres are mostly above the top dashed line, *action* movies below the bottom one, and the other genres in between. There are only two movies that consistently violate these rules: "jungle" and "madness". Further investigation reveals the reason for these outliers: while the romances above the top line either belong to the category *drama/romance* or *comedy/romance*, "jungle" is categorized as *adventure/romance* indicating a degree of action which is unusual for movies in the *romance* class. On the other hand, while "madness" is assigned to the *horror* genre, it is full of action-packed scenes. More samples from the *horror* class would be necessary for a deeper analysis of the interplay between these two genres.

In addition to these two consistent outliers, there are two additional ones in each graph. When the histogram distance is used as activity feature, both "riverwild" (*action*) and "scout" (*drama*) incorrectly appear in the *romance/comedy* region.

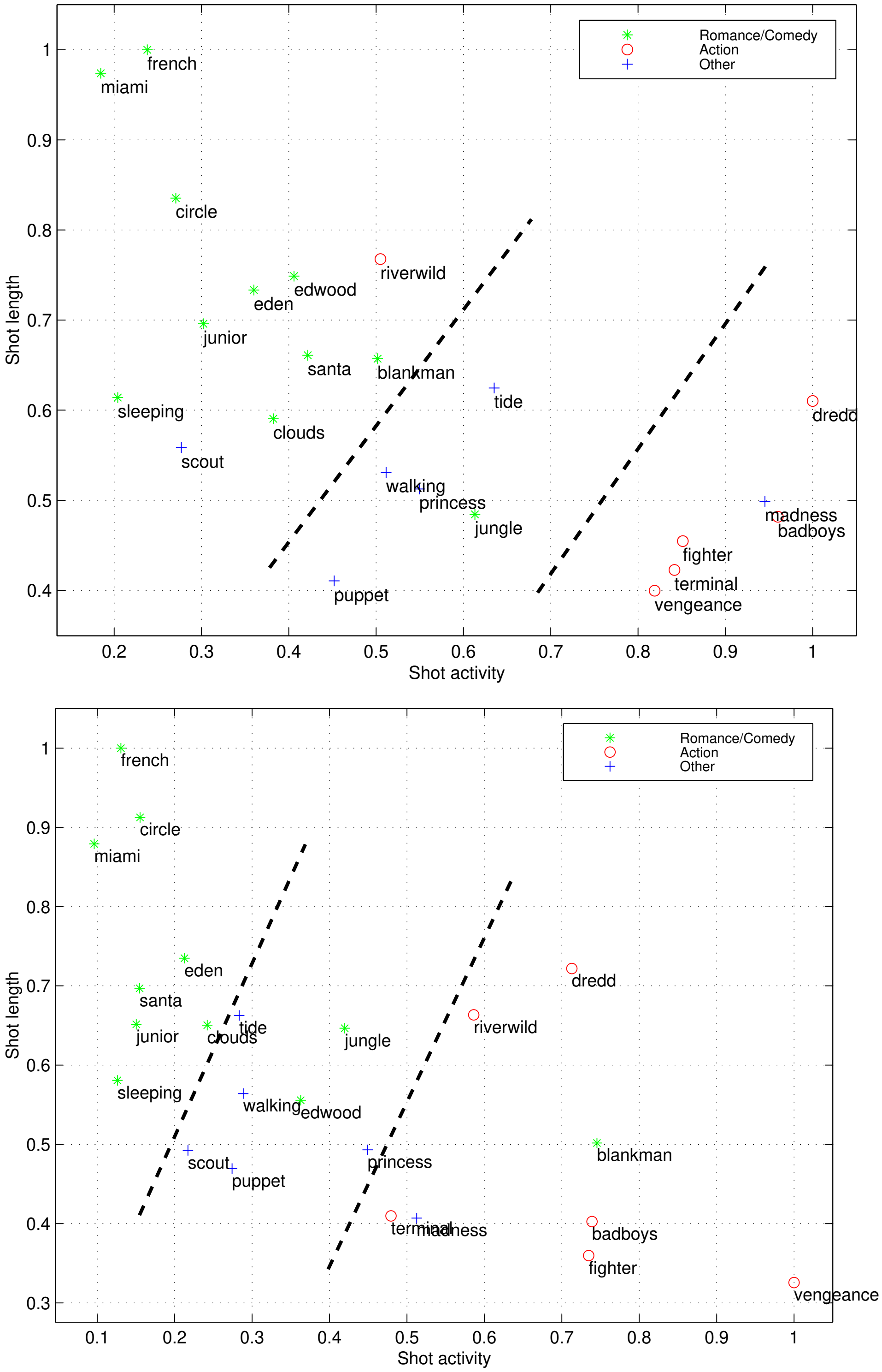

Fig. 12. Population of the feature space by the movies in our database. Top: histogram distance. Bottom: tangent distance. Movie names are listed in Table I.

"riverwild" is a good example of why the histogram distance might fail as a metric of activity. It is an action movie whose plot revolves around white-water rafting and contains numerous shots depicting this sport. While these shots exhibit a significant amount of motion from frame-to-frame, the color histograms tend not to change too much, because there is always plenty of water in the background. The action content cannot, therefore, be captured well by the histogram distance. Since, as discussed in section IV, the tangent distance performs significantly better under these situations it is not surprising that the movie is correctly classified when the latter is used as activity metric.

In fact, the outliers created by the tangent distance are much more intuitive than those originated by the histogram distance. In addition to "jungle" and "madness", it misplaces the comedies "blankman" and "edwood". However, while the comedies above the top dashed line are typically categorized as *comedy/romance* or simply *comedy*, "edwood" receives the awkward categorization of *comedy/drama* (indicating that characterizing its content is probably a difficult task), and "blankman" that of *comedy/screwball/super hero* confirming the fact that it is an action-packed comedy, which could easily fall in the *action* category. Thus while, strictly speaking, the placement of these movies on the *action* and *other* classes is incorrect, it is semantically plausible.

In summary, it appears that the lines superimposed on the shot length/activity space are likely to have a semantic meaning and that this space is one where the movies are nicely laid out according to the degree of action in their scripts, providing discrimination between several genres of content. This property could prove useful for applications where a coarse characterization of the content into semantic classes can be used as a quick filtering mechanism, allowing users to rapidly eliminate items in which they clearly have no interest. For example, a graph such as the ones above could be used as a graphical front-end to a movie library providing an intuitive way for the identification of movies from different genres. On the other hand, the semantic characterization could also be used to provide constraints for a query-by-example content-based retrieval system. This has in fact been recently reported in [22], where a two-way classification of shots into *action* and *character* classes, based on the features introduced above, is combined with a more traditional metric of visual similarity for the purposes of retrieval. It is shown that the inclusion of semantic constraints improves the retrieval accuracy.

VIII. Discussion

In this paper, we have argued that the analysis of video structure plays an important role for its semantic understanding. In particular, knowledge about structure can be used both as a means to improve low-level video processing tasks and to extract features that convey semantic information. We have presented statistical models for two important aspects of video structure and used them as the basis for a Bayesian formulation of shot segmentation. This formulation was shown to extend the standard thresholding model in an adaptive and intuitive way, leading to improved shot segmentation. By applying the feature transformation required for Bayesian segmentation to a database of movie clips we have illustrated how the Bayesian model captures important semantic properties of the content that can be useful for content-based access to movie libraries.

There are several issues in semantic characterization that we did not address here. In particular, it would be interesting to build a movie classifier based on the feature space of Figure 12. Unfortunately, to have high confidence in the resulting classification rates, we would have to process hundreds or even thousands of movies, a requirement that is beyond the reach of our current computational resources. There are however questions that will be feasible to study in the near term. First, it would be interesting to extend the segmentation framework presented here to the more difficult topic of scene segmentation. While, in principle, the framework is still valid, it remains to be seen if the models now proposed can also account for scene duration, if there are better alternatives, what features should be used to detect scene boundaries, and how would the framework compare against other methods presented in the literature [24], [54]. Second, it would be interesting to develop models for elements of mise-en-scene other than activity. While there has been a significant amount of work in the machine vision, image processing, and content-based retrieval communities towards the development of modules capable of extracting semantic information from images [15], [43], [45], a much smaller amount of work has been devoted to the development of a framework for the integration of those modules into a complete retrieval system. We have recently argued that Bayesian inference is the appropriate tool for this integration [25], [50], and are currently investigating the use of Bayesian principles to achieve more extensive semantic characterization.

Appendix

In this appendix, we derive expressions for the ML estimation of the parameters of the Erlang distribution from a set of training data, and the integral of its density function over an interval. We start by deriving, from equation (1), the log-likelihood of the Erlang density

$$\log \epsilon_{r,\lambda}(\tau) = r \log \lambda + (r-1) \log \tau - \lambda\tau - \log((r-1)!). \quad (17)$$

Given a sample $\tau = \{\tau_1, \ldots, \tau_N\}$ of N independent Erlang random variables, the ML parameter estimates are obtained from

$$\{r^*, \lambda^*\} = \arg\max_{r,\lambda} \sum_i \log \epsilon_{r,\lambda}(\tau_i). \quad (18)$$

This is a maximization problem over both discrete (r) and continuous (λ) variables, and such problems are typically hard to solve. However, because it is a sum of r independent arrival times of a Poisson process, by the central limit theorem [14], the Erlang density can be approximated by a Gaussian for large r. Therefore, one can either rely on the Gaussian approximation or assume that the range of interest of r is small. Since in all our experiments we have found that the histograms for shot duration are never close to those of a Gaussian random variable, we have indeed relied on this assumption.

For small r, the maximization becomes significantly simpler since it is possible to do an exhaustive search over all r. For this:

1) Select a range for r. In our experiments we have used r between 1 and 10.
2) For each r, find the λ that maximizes equation (18).
3) Select the value of r that leads to the largest overall log-likelihood.

Given r, the maximization of equation (18) is straightforward. Substituting equation (17) in equation (18), taking the derivative with respect to λ and setting it to zero, we obtain

$$\lambda^* = \frac{Nr}{\sum_i \tau_i}. \tag{19}$$

We next consider the evaluation of

$$\int \epsilon_{r,\lambda}(\tau)d\tau = \int \frac{\lambda^r \tau^{r-1} e^{-\lambda\tau}}{(r-1)!} d\tau \tag{20}$$

and use integration by parts to obtain

$$\int \tau^{r-1} e^{-\lambda\tau} d\tau = -\frac{1}{\lambda}\tau^{r-1}e^{-\lambda\tau} + \frac{r-1}{\lambda}\int \tau^{r-2}e^{-\lambda\tau}d\tau,$$

from which

$$\begin{aligned}\int \epsilon_{r,\lambda}(\tau)d\tau &= -\frac{\lambda^{r-1}\tau^{r-1}}{(r-1)!}e^{-\lambda\tau} + \int \frac{\lambda^{r-1}\tau^{r-2}e^{-\lambda\tau}}{(r-2)!}d\tau \\ &= -\frac{1}{\lambda}\epsilon_{r,\lambda}(\tau) + \int \epsilon_{r-1,\lambda}(\tau)d\tau.\end{aligned}$$

This equation can be solved recursively, leading to

$$\int \epsilon_{r,\lambda}(\tau)d\tau = -\frac{1}{\lambda}\sum_{i=1}^{r}\epsilon_{i,\lambda}(\tau). \tag{21}$$

In this appendix, we derive expressions for the ML estimation of the parameters of the Weibull distribution from a set of training data, and the integral of its density function over an interval. We start by deriving, from equation (2), the log-likelihood of the Weibull density

$$\log w_{\alpha,\beta}(\tau) = \log\frac{\alpha}{\tau} + \alpha\log\frac{\tau}{\beta} - \left(\frac{\tau}{\beta}\right)^{\alpha}. \tag{22}$$

Given a sample $\tau = \{\tau_1, \ldots, \tau_N\}$ of N independent Weibull random variables, the ML parameter estimates are obtained from

$$\{\alpha^*, \beta^*\} = \arg\max_{\alpha,\beta}\sum_i \log w_{\alpha,\beta}(\tau_i). \tag{23}$$

Taking derivatives with respect to α and β, we obtain

$$\begin{aligned}\frac{\partial}{\partial\alpha}\sum_{i=1}^{N}\log w_{\alpha,\beta}(\tau_i) &= \frac{N}{\alpha} + \sum_{i=1}^{N}\log\frac{\tau_i}{\beta} - \sum_{i=1}^{N}\left(\frac{\tau_i}{\beta}\right)^{\alpha}\log\frac{\tau_i}{\beta} \\ \frac{\partial}{\partial\beta}\sum_{i=1}^{N}\log w_{\alpha,\beta}(\tau_i) &= -\frac{\alpha N}{\beta} + \sum_{i=1}^{N}\alpha\left(\frac{\tau_i}{\beta}\right)^{\alpha-1}\frac{\tau_i}{\beta^2} \\ &= \frac{\alpha}{\beta}\sum_{i=1}^{N}\left[\left(\frac{\tau_i}{\beta}\right)^{\alpha} - 1\right]. \end{aligned} \tag{24}$$

Setting the derivatives to zero leads to a system of equations that cannot be solved in closed form. The maximization can, however, still be performed numerically by relying on gradient descent techniques such as Newton's method [5]. In our experiments, we adopted a simpler variation of this process which consisted of restricting the parameter α to be a multiple of 0.5. This variation was inspired by the method introduced above to deal with the Erlang case and can be solved in similar fashion. Namely, we find the optimal value for β assuming that α is known and then perform an exhaustive search over α. Given α the optimal β is the one that sets equation (24) to zero, i.e.

$$\beta^* = \left[\frac{1}{N}\sum_{i=1}^{N}\tau_i^{\alpha}\right]^{\frac{1}{\alpha}}. \tag{25}$$

We can thus see that the optimal β is a function of the sample mean of the α power of the sample values τ_i. It is well known in the statistics literature [37] that the sample mean is very sensitive to the presence of outliers in the data. More robust estimates can be achieved by replacing the sample mean by the sample median, leading to

$$\beta^* = [median(\tau_i^{\alpha})]^{\frac{1}{\alpha}} = median(\tau_i). \tag{26}$$

We next consider the evaluation of

$$\int w_{\alpha,\beta}(\tau)d\tau = \int \frac{\alpha\tau^{\alpha-1}}{\beta^{\alpha}}\exp\left[-\left(\frac{\tau}{\beta}\right)^{\alpha}\right]d\tau.$$

Using the change of variable

$$\lambda = \left(\frac{\tau}{\beta}\right)^{\alpha}, \qquad d\lambda = \frac{\alpha}{\beta^{\alpha}}\tau^{\alpha-1}d\tau$$

it is straightforward to show that

$$\int w_{\alpha,\beta}(\tau)d\tau = -e^{-\left(\frac{\tau}{\beta}\right)^{\alpha}}. \tag{27}$$

Acknowledgments

We thank Giri Iyengar for several stimulating discussions about this work. We also thank comments from two anonymous reviewers that lead to improvements in the paper.

References

[1] *Internet Movie Database*. http://us.imdb.com/.
[2] P. Anandan, J. Bergen, K. Hanna, and R. Hingorani. Hierarchical Model-Based Motion Estimation. In M. Sezan and R. Lagendijk, editors, *Motion Analysis and Image Sequence Processing*, chapter 1. Kluwer Academic Press, 1993.
[3] Y. Ariki and Y. Saito. Extraction of TV News Articles based on Scene Cut Detection using DCT clustering. In *IEEE Int. Conf. on Image Processing*, pages 847–850, Lausanne, 1996.
[4] S. Belongie, C. Carson, H. Greenspan, and J. Malik. Color-and texture-based image segmentation using EM and its application to content-based image retrieval. In *International Conference on Computer Vision, Bombay, India*, pages 675–682, 1998.
[5] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1995.
[6] D. Bordwell and K. Thompson. *Film Art: an Introduction*. McGraw-Hill, 1986.
[7] J. Boreczky and L. Rowe. Comparison of Video Shot Boundary Detection Techniques. In *Proc. SPIE Conf. on Visual Communication and Image Processing*, 1996.
[8] V. Bove, J. Dakss, S. Agamanolis, and E. Chalom. Adding Hyperlinks to Digital Television. In *Proc. SMPTE 140th Technical Conference*, 1998.
[9] R. Castagno, T. Ebrahimi, and M. Kunt. Video Segmentation Based on Multiple Features for Interactive Multimedia Applications. *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 8(5):562–571, September 1998.

[10] E. Chalom. *Statistical Image Sequence Segmentation using Multidimensional Attributes*. PhD thesis, Massachussetts Institute of Technology, 1998.

[11] S. Chang, W. Chen, H. Meng, H. Sundaram, and D. Zhong. A Fully Automated Content-based Video Search Engine Supporting Spatiotemporal Queries. *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 8(5):602–615, September 1998.

[12] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from Incomplete Data via the EM Algorithm. *J. of the Royal Statistical Society*, B-39, 1977.

[13] Y. Deng and B. Majunath. NeTra-V: Toward an Object-based Video Representation. *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 8(5):616–627, September 1998.

[14] A. Drake. *Fundamentals of Applied Probability Theory*. McGraw-Hill, 1987.

[15] D. Forsyth and M. Fleck. Body Plans. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico*, pages 678–683, 1997.

[16] B. Funt and G. Finlayson. Color Constant Color Indexing. *IEEE Trans. on Pattern. Analysis and Machine Intelligence*, 17(5):522–529, May 1995.

[17] U. Gargi, R. Kasturi, and S. Antani. Performance Characterization and Comparison of Video Indexing Algorithms. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Santa Barbara, California*, pages 559–565, 1998.

[18] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman Hall, 1995.

[19] B. Gunsel and M. Tekalp. Content-based Video Abstraction. In *IEEE Int. Conf. on Image Processing*, pages 128–132, Chicago, Illinois, 1998.

[20] A. Hanjalic, R. Lagendijk, and J. Biemond. Template-based Detection of Anchorperson Shots in News Programs. In *IEEE Int. Conf. on Image Processing*, pages 148–152, Chicago, 1998.

[21] R. Hogg and E. Tanis. *Probability and Statistical Inference*. Macmillan, 1993.

[22] G. Iyengar and A. Lippman. Semantically Controlled Content-Based Retrieval of Video Sequences. In *SPIE Multimedia Storage and Archiving Systems III, Boston*, 1998.

[23] A. Jain and A. Vailaya. Image Retrieval using Color and Shape. *Pattern Recognition Journal*, 29:1233–1244, August 1996.

[24] J. Kender and B. Yeo. Video Scene Segmentation Via Continuous Video Coherence. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Santa Barbara, California*, pages 367–373, 1998.

[25] A. Lippman, N. Vasconcelos, and G. Iyengar. Humane Interfaces to Video. In 32^{nd} *Asilomar Conference on Signals, Systems, and Computers*, Asilomar, California, 1998.

[26] C. Low, Q. Tian, and H. Zhang. An Automatic News Video Parsing, Indexing, and Browsing System. In *ACM Multimedia Conference*, 1996, Boston.

[27] W. Ma and H. Zhang. Benchmarking of Image Features for Content-based Retrieval. In 32^{nd} *Asilomar Conference on Signals, Systems, and Computers, Asilomar, California*, 1998.

[28] Wallace Martin. *Recent Theories of Narrative*, chapter 5. Cornell University Press, Ithaca, NY, USA, 1986.

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

[30] A. Nagasaka and Y. Tanaka. Automatic Video Indexing and Full-Video Search for Object Appearences. In E. Knuth and L. Wegner, editors, *Visual Database Systems II*, pages 173–127. Elsevier, 1992.

[31] W. Niblack and 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, California.

[32] G. Pass, R. Zabih, and J. Miller. Comparing images using color coherence vectors. In *ACM Intl. Multimedia Conference*, pages 65–73. ACM, Nov. 1996.

[33] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Content-based Manipulation of Image Databases. *Int. Journal of Computer Vision*, Vol. 18(3):233–254, June 1996.

[34] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[35] R. Redner and H. Walker. Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review*, 26(2):195–239, April 1984.

[36] K. Reisz and G. Millar. *The Technique of Film Editing*. Focal Press, 1968.

[37] P. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. John Wiley, 1987.

[38] P. Salembier and M. Pardas. Hierachical Morphological Segmentation for Image Sequence Coding. *IEEE Trans. on Image Processing*, Vol. 3, September 1994.

[39] P. Simard, Y. Le Cun, and J. Denker. Efficient Pattern Recognition Using a New Transformation Distance. In *Proc. Neural Information Proc. Systems*, Denver, USA, 1994.

[40] J. Smith and S. Chang. VisualSEEk: a fully automated content-based image query system. In *ACM Multimedia, Boston, Massachussetts*, pages 87–98, 1996.

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

[42] M. Swain and D. Ballard. Color Indexing. *International Journal of Computer Vision*, Vol. 7(1):11–32, 1991.

[43] Martin Szummer and Rosalind Picard. Indoor-Outdoor Image Classification. In *Workshop in Content-based Access to Image and Video Databases*, 1998, Bombay, India.

[44] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley, 1985.

[45] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3, 1991.

[46] 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.

[47] N. Vasconcelos and A. Lippman. Multiresolution Tangent Distance for Affine Invariant Classification. In *Neural Information Processing Systems*, Denver, Colorado, 1997.

[48] N. Vasconcelos and A. Lippman. Empirical Bayesian EM-based Motion Segmentation. In *Proc. IEEE Computer Vision and Pattern Recognition Conf.*, San Juan, Puerto Rico, 1997.

[49] 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.

[50] N. Vasconcelos and A. Lippman. A Bayesian Framework for Semantic Content Characterization. In *Proc. IEEE Computer Vision and Pattern Recognition Conf.*, Santa Barbara, California, 1998.

[51] J. Wang and E. Adelson. Representing Moving Images with Layers. *IEEE Trans. on Image Processing*, Vol. 3, September 1994.

[52] Y. Weiss. Smoothness in Layers: Motion Segmentation Using Nonparametric Mixture Estimation. In *Computer Vision and Pattern Recognition Conf.*, San Juan, Puerto Rico, 1997.

[53] 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.

[54] M. Yeung and B. Yeo. Time-constrained Clustering for Segmentation of Video into Story Units. In *ICPR'96*, volume C, pages 375–380, 1996.

[55] H. Zhang, Y. Gong, S. Smoliar, and S. Tan. Automatic Parsing of News Video. In *Proc. Int. Conf. on Multimedia Computing and Systems*, May 1994, Boston, USA.

[56] H. Zhang, S. Smoliar, and J. Wu. Content-Based Video Browsing Tools. In A. Rodriguez and J. Maitan, editors, *Symposium on Electronic Imaging Science and Technology: Multimedia Computing and Networking*, pages 389–398, SPIE Vol. 2417, Feb. 1995, San Jose, California.

[57] H. Zhang, J. Wang, and Y. Altunbasak. Content-based Video Retrieval and Compression: A Unified Solution. In *IEEE Int. Conf. on Image Processing*, pages 13–16, Santa Barbara, California, 1997.

[58] J. Zhang, J. Modestino, and D. Langan. Maximum-Likelihood Parameter Estimation for Unsupervised Stochastic Model-Based Image Segmentation. *IEEE Trans. on Image Processing*, Vol. 3, July 1994.

PLACE
PHOTO
HERE

Nuno Vasconcelos received a *licenciatura* in electrical engineering and computer science from the Universidade do Porto, Portugal in 1988 and a Master of Science degree from the Massachusetts Institute of Technology in 1993. He is now concluding his PhD in the MIT Media Laboratory, where he has been a research assistant since 1991. During this period he has worked in several topics including video analysis and compression, motion estimation and segmentation, image representations, content-based image retrieval, semantic characterization of video content, and statistical modeling of visual data. His current interests are in the areas of image processing, machine vision, statistical learning, and multimedia.

PLACE
PHOTO
HERE

Andrew Lippman received the B.S. and M.S. degrees in electrical engineering from the Massachusetts Institute of Technology in 1971 and 1978, respectively, and a PhD from the EPFL, Lausanne, Switzerland in 1995. He is currently the Associate director of the MIT Media Laboratory and a Lecturer at MIT, where he has directed research programs on image processing, personal computers, entertainment, and graphics. Recently, he has established and directs a research consortium entitled “Digital Life” that addresses bits, people and community in a wired world. He holds seven patents in television and digital image processing, and has published widely in both the technical and lay literature. His current interests are in the design of flexible, interactive digital television infrastructures.