Foveated Shot Detection for Video Segmentation

Giuseppe Boccignone, *Member, IEEE*, Angelo Chianese, Vincenzo Moscato, and Antonio Picariello

***Abstract*—We view scenes in the real world by moving our eyes three to four times each second and integrating information across subsequent fixations (foveation points). By taking advantage of this fact, in this paper we propose an original approach to partitioning of a video into shots based on a foveated representation of the video. More precisely, the shot-change detection method is related to the computation, at each time instant, of a consistency measure of the fixation sequences generated by an ideal observer looking at the video. The proposed scheme aims at detecting both abrupt and gradual transitions between shots using a single technique, rather than a set of dedicated methods. Results on videos of various content types are reported and validate the proposed approach**

***Index Terms*—Attentive vision, dissolves, hard cuts, shot detection, video segmentation.**

I. Introduction

DETECTION of shot boundaries provides a base for nearly all video abstraction and high level video segmentation methods [27], [34]. In this paper, we propose a novel approach to partitioning of a video into shots based on a foveated representation of the video.

A shot is usually conceived in the literature as a series of interrelated consecutive frames taken contiguously by a single camera and representing a continuous action in time and space. In other terms, a shot is a subsequence generated by the camera from the time it "starts" recording images, to the time it "stops" recording [16]. However, shot segmentation is ill-defined. On the one hand, a video is generated by composing several shots by a process called *editing*, and due to edit activity different kinds of transitions from one shot to another, either *abrupt* or *gradual*, may take place. An abrupt transition, or hard cut, occurs between two consecutive frames and is the most common type. An example is provided in Fig. 1.

Gradual transitions such as fades, wipes and dissolves (see Fig. 2) are spread over several frames and are obtained using some spatial, chromatic or spatiochromatic effect; these are harder to detect from a purely data analysis point of view because the difference between consecutive frames is smaller.

It has been observed [2] from a study of video production techniques that the production process originates several constraints, which can be useful for video edit classification in the framework of a model based approach to segmentation. However, the use of such constraints implies high costs in designing shot models due to the high number of degrees of freedom available in shot production (for review and discussion, see [16] and [27]).

On the other hand, for the purposes of video retrieval, one would like to mark the case of any large visual change, whether camera stops or not (e.g., a large object entering the scene). Thus, from a general standpoint, shot detection should rely on the recognition of any significant discontinuity in the visual content flow of the video sequence [16]. Meanwhile, the detection process should be unaffected by less significant changes within the same shot, like object/camera motion and lighting changes, which may contribute to missed or false detections. In such a complex scenario, despite the number of proposals in the literature, robust algorithms for detecting different types of boundaries have not been found, where robustness is related to both detection performance and stability with minimum parameter tuning [21].

At the heart of our ability to detect changes from one view of a scene to the next is the mechanisms of visual attention. Film makers have long had the intuition that changes to the visual details across cuts are not detected by audiences, particularly when editing allows for smooth transitions [51]. In the movie *Ace Ventura: When Nature Calls*, the pieces on a chess board disappear completely from one shot to the next. In *Goodfellas*, a child is playing with blocks that appear and disappear across shots. In fact, almost every movie, and almost every cut, has some continuity mistake, yet, most of the time people are blind to these changes. It has been noted that change blindness is evident when mistakes occur far from the viewer's focus of attention [51].

The term *attention* captures the cognitive functions that are responsible for filtering out unwanted information and bringing to consciousness what is relevant for the observer [7], [23], [52]. Visual attention, in turn, is related to how we view scenes in the real world: moving our eyes (saccade) three to four times each second, and integrating information across subsequent fixations [60]. Saccades represent overt shifts of spatial attention that can be performed either voluntarily (top-down), or induced automatically (bottom-up) by salient targets suddenly appearing in the visual periphery and allow an observer to bring targets of interest onto the fovea, the retinal region of highest spatial resolution. Eye movements, though being characterized by some degree of randomness [6], [48], are likely to occur in a specific path (the *scanpath*, [37]) so as to focus areas that are deemed important. The scanpath can be conceived as a visuomotor pattern resulting from the perceptual coupling of observer and observed scene. An example generated on the third frame of Fig. 1 is illustrated in Fig. 3.

In the course of a scan, we have a rich visual experience from which we abstract the meaning or gist of a scene. During next

Manuscript received July 30, 2003; revised February 24, 2004 and March 9, 2004. This work was supported by the Italian Ministero per l' Universita' e la Ricerca Scientifica e Tecnologica and by the Instituto Nazionale Fisica della Materia (INFM). This paper was recommended by Associate Editor R. Lienhart.

G. Boccignone is with the Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica, University of Salerno, Salerno I-84084, Italy (e-mail: boccig@unisa.it).

A. Chianese, V. Moscato, and A. Picariello are with the Dipartimento di Informatica e Sistemistica, the University of Naples "Federico II," 21-I-80125 Naples, Italy (e-mail: angchian@unina.it; vmoscato@unina.it; picus@unina.it).

Digital Object Identifier 10.1109/TCSVT.2004.842603

Fig. 1. Example of hard cut effect from a TREC01 video. Abrupt transition occurs between the second and the third frame.

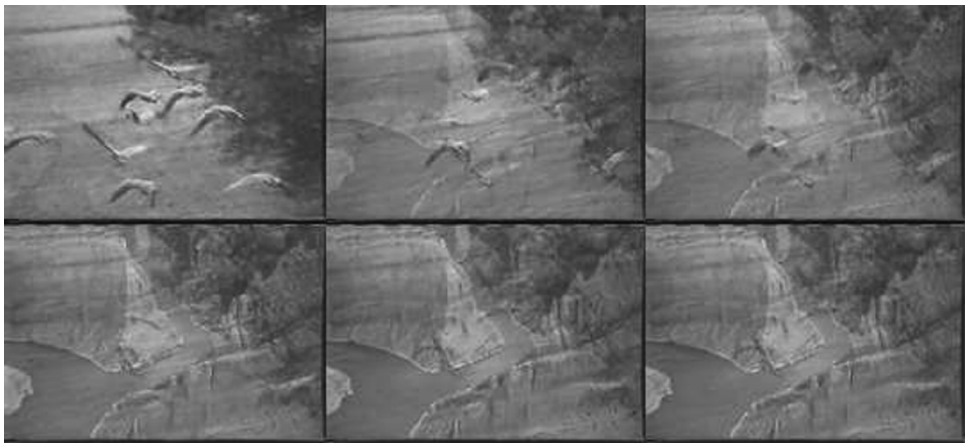

Fig. 2. Example of dissolve effect.

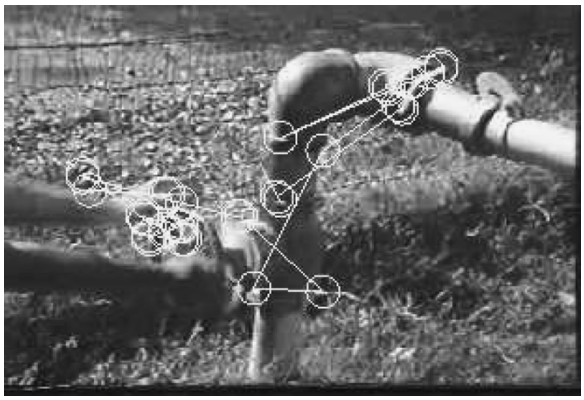

Fig. 3. Scanpath eye-tracked from a human observer while viewing the third frame presented in Fig. 1. The scanpath has been graphically overlapped on the original image: circles represent fixations, and lines trace displacements (saccades) between fixations.

scan, if the gist is the same our perceptual system assumes the details are the same. Clearly, this "sketch" representation not only serves the information reduction purpose of filtering unwanted information, but also, by integrating the gist from one view to the next, to achieve the impression of a stable world. However, the lack of a detailed representation of the outside world from one view to the next can rise failures of change detection [51].

The background question which motivates this work is whether these mechanisms that are useful to prevent audiences noticing the transitions, can conversely be exploited to detect such transitions, and thus help for video segmentation. Intuitively, one could argue that if the playback speed is reduced (or, equivalently, the saccade rate increased) change blindness effects would be reduced too. This corresponds to introducing an ideal observer or agent, capable of tuning his saccadic rate. In some sense, this is akin to Gargi's experimental study of human ground-truthing, where most consistent results in marking shot changes were obtained when subjects performed such task at half speed after viewing the sequence once at full speed [16].

The rationale behind our approach is that perceptual capacity of an observer can be defined at two levels [38]. At the first level there is the ability of the agent to explore the scene in ways mediated by knowledge of patterns of visuomotor behavior, that is the ability to exploit the interdependence between incoming sensory information and motor behavior (eye movements). At the second, higher level, there is the accessing by the observer of information related to the nature of observer's own exploration.

For example, while viewing a video sequence, it is reasonable that in the presence of similar visual configurations, and in the absence of an habituation mechanism, an observer should consistently deploy attention to visually similar regions of interest and by following a similar motor pattern; clearly, when the gist of the world observed undergoes a significant change, the visuomotor pattern cannot be exploited further, since inconsistent, and a new scanpath will be generated. Such an intuitive assumption can be theoretically motivated on the basis that after an abrupt transition the video signal is governed by a new statistical process [28]. Indeed, it has been shown [6] that gaze-shift is strongly constrained by structure and dynamics of the underlying random field modeling the image. Quantitatively, if a measure $\mathcal{M}$ of *attention consistency* is defined, $\mathcal{M}$ should decrease down to a minimum value. For instance, this is what is likely to occur when a view abruptly changes.

On the other hand, a view change may occur across long delay intervals, as in gradual transitions. In this case, $\mathcal{M}$ should account for a behavior similar to that experienced in change blindness experiments, where subjects fail to detect a slow, global spatiochromatic editing of a sequence presenting the same image [38], but suddenly succeed when the frame rate of presentation is increased, due to the reduction of the time lag between the first and the last frames of the transition. In this case, the $\mathcal{M}$ function should vary smoothly across the interval, while decreasing rapidly if measured on the first and the last frames of the same interval. It is worth remarking that shots involved in a dissolve transition may have similar color distribution, which a color histogram would hardly detect [28], while differing in structural information that can be detected by appropriate algorithms (e.g., edge based); as in the case of hard cuts, the sequence of attention shifts can be suitably exploited, since its dynamics [6] is strongly intermingled with the complexity of the statistical process modeling the signal (e.g., two-source model for a dissolve [28]).

As regards the second level, namely the evaluation of information about the nature of visual exploration itself, it can be stated as an inference drawn by the observer from its own sensorimotor behavior under prior knowledge available. On such assumption, the problem of detecting a shot change given the change of the observer's behavior $\mathcal{M}$, naturally leads to a Bayesian formulation, and can be conceived as a signal detection problem where the probability that a shot boundary B occurs, given a behavior $\mathcal{M}$, is compared against the probability that a shot boundary is not present.

The introduction of this approach has several advantages, both theoretical and practical. First, it allows to find a uniform method for treating both abrupt and gradual transitions. As discussed previously, this result stems from relations occurring between the dynamics of gaze-shifts and statistical processes modeling the observed image [6]; also, the method is well grounded in visual perception theories [37], [38]. As such, it is suitable to overcome usual shortcomings of other simpler techniques proposed so far (e.g, histogram manipulations). In this sense, higher robustness can be achieved, as regards performance and stability in detecting important visual changes while discarding negligible ones. Then, once the distinctive scanpath has been extracted from a frame, subsequent analysis needs only to process a sparse representation of the frame. Eventually, attentive analysis can, in perspective, provide a sound and unitary framework at higher levels of video content analysis. For instance, key frame selection/generation could be conceived in terms of average scanpath of shot frames; multimodal processing for deriving semantic properties of a scene, can be stated in terms of attentive audio/visual integration.

In Section II, we briefly discuss background and related work on shot segmentation. In Section III, we outline the model for foveated analysis of a video sequence. In Section IV, the computation of patterns of visuomotor behavior is discussed. In Section V, we derive the procedure to calculate the $\mathcal{M}$ function and the boundary detection algorithm. Sections VI presents the experimental protocol and results obtained. Some concluding remarks are given in Section VII.

II. Background and Related Work

Assume as input to a segmentation system a video sequence, that is a finite sequence of time parameterized images, $(f(t_0), f(t_1), \ldots, f(t_N))$, where each image $f(t_n)$ is called a frame. Each frame is a color image, namely a mapping from the discrete image support $\Omega \subseteq Z^2$ to an m-dimensional range, $f : \Omega \rightarrow Q \subseteq Z^m$; in other terms, it is a set of single-valued images, or channels, sharing the same domain, i.e., $f(x,y) = (f_i(x,y))^T$, where the index $i = 1, \ldots, m$, defines the ith color channel and (x,y) denotes a point in the Ω lattice. $Q = \{q_1, \ldots, q_N\}$ is the set of colors used in the image. Each frame displays a view, a snapshot, of a certain visual configuration representing an original world scene.

A time segmentation of a video f defined on the time interval $[t_0, t_N]$ is a partition of the video sequence into N_b subsequences or blocks. One such partition can be obtained in two steps. First, a mapping $\mathcal{T} : Z^m \rightarrow F$ of the frame $f(t_n) \in Z^m$ to a representation $\mathcal{T}(f(t_n)) \in F$, F being a suitable feature space, is performed. Then, given two consecutive frames $f(t_n)$ and $f(t_{n+l})$, where $l \geq 1$ is the skip or interframe distance, a discriminant function $\mathcal{D} : F \times F \rightarrow R^+$ is defined to quantify the visual content variation between $\mathcal{T}(f(t_n))$ and $\mathcal{T}(f(t_{n+l}))$, such that a boundary occurs at frame $f(t_n)$ if $\mathcal{D}(\mathcal{T}(f(t_n)), \mathcal{T}(f(t_{n+l}))) > T$, where T is a suitable threshold.

Thus, in principle, to solve the shot detection problem three steps must be undertaken: choose an appropriate mapping $\mathcal{T}$; define a robust discriminant function $\mathcal{D}$; devise a (universal) threshold T.

As regards the first two points, different techniques have been used: pixel based methods, such as the mean absolute value of intensity between frames [24], [39], or block matching [21], [50], histograms difference [15], [16], [34], [61], [64] motion difference [63] and perceived motion energy [31], and differential geometry [58].

For what concerns the third point, heuristically chosen thresholds have been proposed [34], [39]. However a fixed thresholding approach is not feasible especially when considering gradual transitions. In particular, dissolve effects are reputed the most common ones, but also the most difficult to detect [13], [57]. A dissolve can be obtained as a combination of fade-out and fade-in, superimposed on the same film strip; fade-out occurs when the visual information gradually disappears, leaving a solid color frame, while fade-in takes place when the visual information gradually appears from a solid color frame (refer again to Fig. 2).

Dissolve detection is still a challenging problem. Few techniques have been published [29]. Variable thresholding has been proposed in [61] and [64], the latter relying on gaussian distribution of discontinuity values. For instance, in [64], the twin-comparison approach using a pair of thresholds, for detecting hard cuts and gradual transitions, respectively, has been introduced. More significant improvements have been achieved by recasting the detection problem in a statistical framework. A novel and robust approach has been presented by Lienhart [30], which relies on multiresolution analysis of time series of dissolve probabilities at various time scales; experiments achieved a detection rate

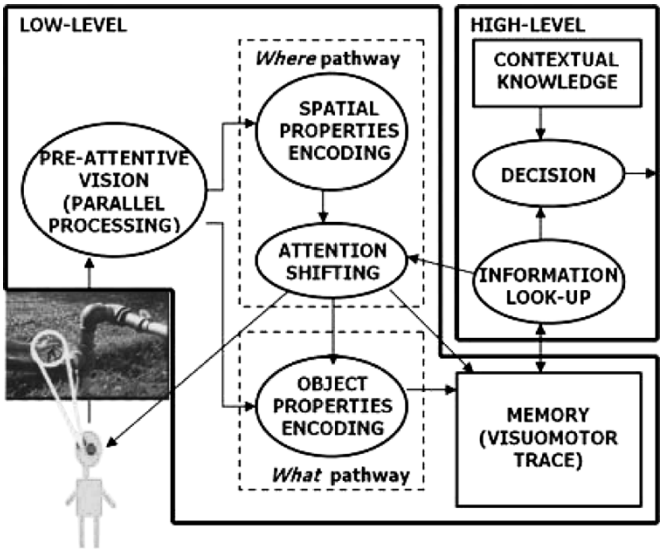

Fig. 4. General model of attentive/foveated video analysis. At a lower level, the observer generates visuomotor patterns, related to the content of the sequence. At a higher level, the observer detects scene changes by judging his own visuomotor behavior in the context of prior knowledge available.

of 75% and a false alarm rate of 16% on a standard test video set. Further, it has been recently argued that a statistical framework incorporating prior knowledge in model based [20], statistical approaches leads to higher accuracy for choosing shot boundaries [21], [58].

III. Outline of the Model for Foveated Video Analysis

The evaluation of attention consistency relies on the model of foveated analysis outlined in Fig. 4.

In the *preattentive stage*, salient features are extracted by specific detectors operating in parallel for all points of the image, at different spatial scales, and organized in the form of contrast maps. In order to obtain such a representation different methods can be adopted, e.g., [14], [22], and [42]. It is worth remarking that the model presented in this paper is unaffected by the method chosen to implement such preattentive stage. We experimented with schemes proposed in [14] and [22], and opted for the latter due to simplicity and limited computational complexity. Precisely, low-level vision features are derived from the original color image decomposed at several spatial scales using Gaussian [8] and oriented pyramids (via convolution with Gabor filters [19]). Note that pyramid computation is an $O(|\Omega|)$ method, where $|\Omega|$ represents the number of samples in the image support Ω.

The features considered are: brightness (I); color channels tuned to red (R), green (G), blue (B), and yellow (Y) hues; and orientation (O). From color pyramids, red/green (RG) and blue/yellow (BY) pyramids are derived by subtraction. Then, from each pyramid a contrast pyramid is computed encoding differences between a fine and a coarse scale for a given feature. As a result, one contrast pyramid encodes for image intensity contrast, four encode for local orientation contrast, and two encode for RG and BY contrast (see [22], for details).

Successively, this preattentive representation, undergoes specialized processing through a "*Where*" system devoted to localizing objects, and a "*What*" system tailored for identifying them. Clearly, tight integration of these two information pathways is essential, and indeed attentive mechanisms play a fundamental role. A plausible assumption is that, in the "What" pathway, early layers provide feature extraction modules, whose activity is subjected to temporal modulation by the "Where" pathway and the related attention shifting mechanism, so that unmodulated responses are suppressed.

In the "Where" pathway, the preattentive contrast maps are combined into a master or saliency map [1], [22], [33], which is used to direct attention to the spatial location with the highest saliency (attention shifting stage). The region surrounding such location represents the current focus of attention (FOA), C_s. By traversing spatial locations of decreasing saliency, a scanpath $(C_s)_{s=1,2,...}$ is obtained by connecting a sequence of FOAs, and stored.

It is important to note that, in general and specifically in this work, such "working memory" retains either a representation of a set of visual features (measured at FOAs) and a motor map of how such features have been explored; indeed, the memory of an attentive system is a *visuomotor trace* of a world view [18], [38], rather than a classical feature representation of the original scene, and any subsequent information-lookup task entails a prediction/confirmation upon such visuomotor scheme. Denote $\mathcal{T}(f(t))$ the visuomotor trace (simply, the trace) of frame $f(t)$.

At the higher perceptual level, the observer infers scene changes by judging his own visuomotor behavior. To this end, given two frames $f(t)$ and $f(t+l)$ (for notational simplicity, $t = t_n$), an effective procedure is needed to compute the function $\mathcal{M}(t))$ which gauges the consistency between the two traces $\mathcal{T}(f(t))$ and $\mathcal{T}(f(t+l))$. A way to solve this problem, is suggested by experiments performed by Walker and Smith [59], who provided evidence that when observers are asked to make a direct comparison between two simultaneously presented pictures, a repeated scanning, in the shape of an FOA by FOA comparison, occurs. Using this procedure, which we name *information look-up loop*, the consistency $\mathcal{M}$ will eventually be calculated as the average of the local consistencies measured on pairs of FOA's, iteratively selected from the two traces $\mathcal{T}(f(t))$ and $\mathcal{T}(f(t+l))$ according to a "best fit" prediction strategy. The behavior of the $\mathcal{M}$ function, is then used by a detection module, based on Bayesian decision theory, which, under prior contextual knowledge available, infers from $\mathcal{M}$ the presence of a scene transition, either abrupt or gradual.

In Sections IV and V, we detail how the different levels of our attentive system have been designed.

IV. Low Perceptual Level: Generation of Visuomotor Patterns

At this lower perceptual level, the agent observes the sequence and generates patterns of visuomotor behavior (traces).

A. Where System: From Preattentive Features to Attention Shifting

The goal of the "Where" system is to build a saliency map of the frame and define over this map the motor trace, that is

the sequence of FOAs $(C_s)_{s=1,2,...,K}$. To this end, the contrast features for intensity, color, and orientation, obtained from the preattentive stage, are summed across scales (pyramid levels) into three separate contrast maps, one for intensity, one for color and one for orientation. Eventually, the three maps, normalized between 0 and 100, are linearly summed into a unique master map (for simplicity, we compute the latter as the average of the three maps), or saliency map (SM).

By using the SM map, the attention shifting mechanism could be implemented through a variety of ways (e.g., [6], [17], [22], [42], and [54]). One intuitive method for traversing spatial locations of decreasing saliency, is to use a winner-take-all (WTA) strategy [22], [54], in which the most salient location "wins" and determines the setting of the FOA; the winner is subsequently inhibited in order to allow competition among less salient locations, for predicting the next FOA. A simple and efficient way of implementing such strategy is through a WTA neural network, e.g., an array of integrate-and-fire neurons with global inhibition [22]. It is worth noting that a WTA algorithm, due to fast convergence properties, has $O(n)$ time complexity, n being the number of processing elements (neurons) of the network. In our scheme, the number of neurons is constrained by the number of samples in the saliency map (each point of the map, represents the input to one neuron). Since the map resides at an intermediate scale between the highest and the lowest resolution scales, namely at scale 4, a reduction factor 1:16 is achieved with respect to the original image, thus the time complexity of the WTA stage is given by $|\Omega|/16$ time units.

This solution has the advantage of providing information on the fixation time spent on the FOA (the firing time of WTA neurons) and our model, differently from others proposed in the literature, explicitly exploits such information. After the "Where" processing, the frame $f(t)$ is represented by a spatiotemporal, or motor trace representing the stream of foveation points $(C_s^t(p_s;\tau_s))_{s=1,2,...,K}$, where $p_s = (x_s, y_s)$ is the center of FOA s, and the delay parameter τ_s is the observation time spent on the FOA before a saccade shifts to C_{s+1}.

As outlined in Fig. 5 the generation of spatiotemporal information is basically an information reduction step in which we assume that the "Where" system "projects" toward the "What" system and signals the *FOA* to be analyzed.

B. What Pathway: Properties Encoding

In the "What" pathway, features are extracted from each highlighted *FOA*, relative to color, shape and texture. An *FOA* is represented in the intensity and color opponent pyramids, at the highest resolution scale. Note that in biological vision, the spatial support of the FOA is usually assumed as circular; here, for computational purposes, each FOA is defined on a square support $D_{p_s} \subseteq \Omega$, centered on p_s, of dimension $|D_{p_s}| = (1/36)|\Omega|$. In the following, we drop the τ parameter for sake of simplicity.

Color features. Given a set of representative colors $Q = \{q_1, \ldots, q_B\}$, a color histogram $h(C(p)) = \{h_b\}$ of the FOA $C(p)$ is defined on bins b ranging in [1,B], such that h_b given for any pixel in D_p, is the probability that the color of the pixel is $q_b \in Q$. Here, $B = 16 \times 16 \times 16$ is used. For a three-channel frame, the FOA histogram calculation time is $|D_p| \times 3$.

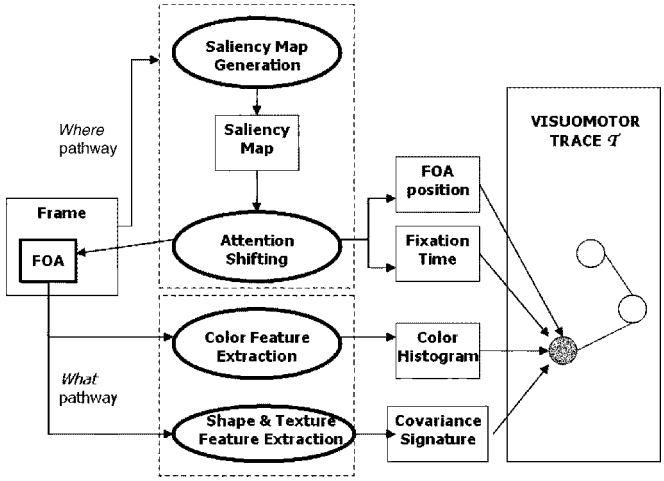

Fig. 5. Generation of the visuomotor trace of a single frame. The scheme shows the selection of a FOA by the "Where" pathway, and the extraction of FOA information by the "What" pathway. For visualization purposes, the trace is represented as a graph-like structure: each node corresponds to a single FOA, and the arc joining two FOAs denotes a saccade.

Shape and texture features. A wavelet transform (WT) of the FOA has been adopted [32]. Denote the wavelet coefficients as $w_l^k(x,y)$, where $(x,y) \in D_p$, l indicates the decomposition level and k indexes the sub-bands. In our case, due to the limited dimension of the FOA, only a first level decomposition $(l = 1)$ is considered, and in the sequel, for notational simplicity the index l is dropped. Decomposition gives rise to four subregions of dimension $|D_p|/4$. Then, only the details components of the WT are taken into account, in order to characterize shape (edges) and texture. Namely, for $k = 1, 2, 3$, the detail sub-bands contain horizontal, vertical and diagonal directional information, respectively, and are represented by coefficient planes $[\{w^k(x,y)\}]_{k=1,2,3}$. Next, the Wavelet Covariance Signature is computed, i.e., the feature vector of coefficient covariances $\Sigma^2_{C_s^m} = \{\sigma^2_{X,Y}\}$, where

$$\sigma^2_{X,Y} = \sum_{x,y} \left\{ \frac{1}{\frac{|D_p|}{4}} \sum_{k=1}^{3} X_k(x,y) Y_k(x,y) \right\}. \quad (1)$$

The pair (X_k, Y_k) is in the set of coefficient plane pairs $\{(w_i^k, w_j^k)\}$, i and j being used to index the three channels, and (x,y) span over the sub-band lattice of dimension $|D_p|/4$. Note that, the FOA wavelet representation at level 1 can be obtained through $2h|D_p|$ operations, where h is the size of convolution filters (here $h = 3$) [32], while calculation of covariances can be accomplished in $|D_p|^2$ operations. Clearly, $|\Sigma^2| = 18$.

As summarized in Fig. 5, the saccadic movements together with their resultant fixations, and feature analysis of foveated regions, allow the formation of the trace $\mathcal{T}(f(t))$, briefly $\mathcal{T}(t)$, of the view observed in frame $f(t)$

$$\mathcal{T}(t) = \left(\mathcal{T}_s^t\right)_{s=1,...,K} \quad (2)$$

where $\mathcal{T}_s^t = (C_s^t, h_b(C_s^t), \Sigma^2_{C_s^t})$.

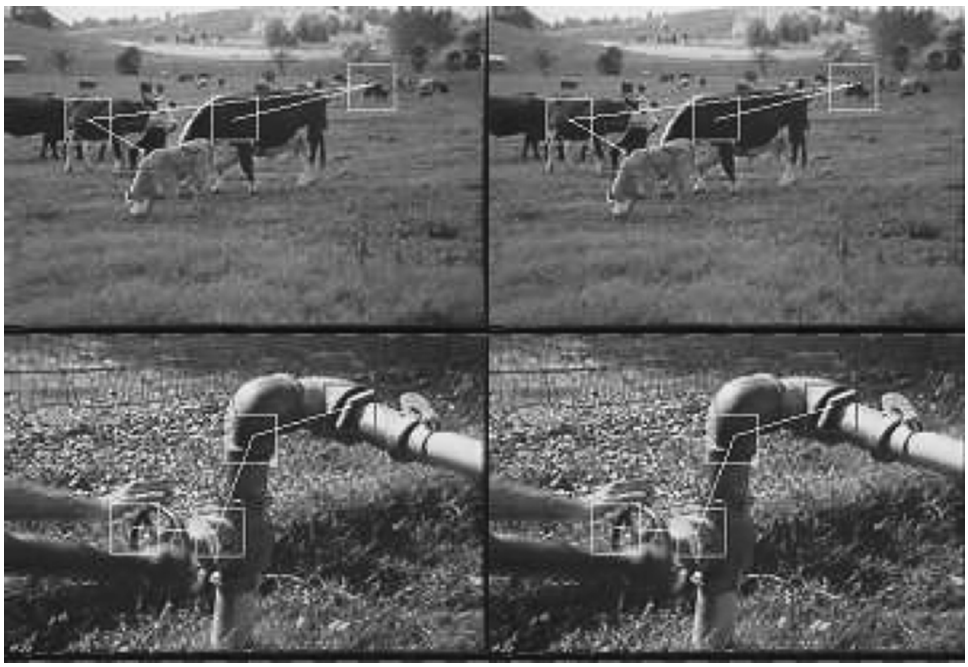

Fig. 6. Traces generated on four frames embedding an hard cut. The first four FOAs are shown for each frame. The red rectangle represents the first FOA of the trace. The trace sequence abruptly changes between frames 2 and 3.

V. High Perceptual Level: Inference of Scene Changes From Visuomotor Behavior

At this level, the observer evaluates the information regarding the nature of visual exploration itself and infers the presence of a shot boundary from its own sensorimotor behavior under prior knowledge available on the kinds of transitions with which he is dealing.

A. Attention Consistency via Information Look-Up

An agent observing views that present similar configurations of objects will generate consistent traces until a transition occurs. An example of such behavior is provided in Fig. 6 where a trace of three FOAs is tracked in a subsequence embedding a hard cut.

Formally, we need to define a measure of consistency $\mathcal{M} : F \times F \rightarrow R^{+}$, such that $\mathcal{M}(\mathcal{T}(m), \mathcal{T}(n))$, where the traces $\mathcal{T}(m)$ and $\mathcal{T}(n)$ have been generated by observing frames $f(t_m) = f(m)$ and $f(t_m + l) = f(n)$. A strategy to solve this problem is to make an FOA by FOA comparison [59]. This information look-up loop is summarized in the scheme of Fig. 7.

The procedure, which we denote attention consistency (AC), given a fixation point $C_r^m(p_r; \tau_r)$ in a first frame, selects the homologous point $C_s^n(p_s; \tau_s)$ in a second frame among those belonging to a local temporal window defined in the interval $[s - H, s + H]$, i.e., $[C_s^n(p_s; \tau_r), C_{s\pm1}^n(p_{s\pm1}; \tau_{s\pm1}), \ldots, C_{s\pm H}^n(p_{s\pm H}; \tau_{s\pm H})]$. The choice is performed by computing, for the pair C_r^m and C_s^n, the FOA consistency

$$\mathcal{M}^{r,s} = \alpha \mathcal{M}_{\text{spatial}}^{r,s} + \beta \mathcal{M}_{\text{temporal}}^{r,s} + \gamma \mathcal{M}_{\text{visual}}^{r,s} \qquad (3)$$

where α, β, $\gamma \in [0, 1]$, and by choosing the FOA s as $s = \arg\max\{\mathcal{M}^{r,s}\}$. Such "best fit" is retained and eventually used

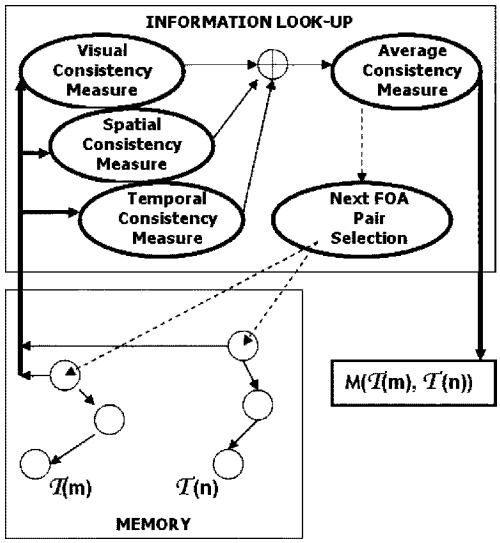

Fig. 7. Information look-up loop for determining the attention consistency $\mathcal{M}$ related to frames m and n, by exploiting the visuomotor traces $\mathcal{T}(m)$, $\mathcal{T}(n)$.

to compute $\mathcal{M}(\mathcal{T}(m), \mathcal{T}(n))$ as the average of the first K FOA consistencies

$$\mathcal{M} = \frac{1}{K} \sum_{f=1}^{K} \mathcal{M}_f^{r,s}. \qquad (4)$$

This "best fit" strategy has been chosen in order to reduce the sensitivity of the algorithm both to the starting FOA point and to the fact that, in similar images, some FOAs could be missing due to lighting changes and noise, even if this is unlikely to occur for small interframe distances.

The right-hand terms of (3), namely $\mathcal{M}_{\text{spatial}}^{r,s}$, $\mathcal{M}_{\text{temporal}}^{r,s}$, and $\mathcal{M}_{\text{visual}}^{r,s}$, account for local measurements of spatial tem-

poral and visual consistency, respectively. These are calculated as follows.

Local spatial consistency. $\mathcal{M}^{r,s}_{\text{spatial}}$ is gauged through the ℓ^1 distance between homologous *FOAs* centers

$$d(p_r, p_s) = |x_r - x_s| + |y_r - y_s|. \tag{5}$$

The distance is "penalized" if, for the two frames, the displacement between the current *FOA* and the next one is not in the same direction

$$\hat{d}(p_r, p_s) = d(p_r, p_s) \cdot e^{-\Delta(p_r, p_s)} \tag{6}$$

Δ being the difference of direction between two *FOAs*

$$\Delta = \zeta \cdot \text{sgn}\left[(x_r - x_{r-1}) \cdot (x_s - x_{s-1})\right] \cdot \text{sgn}\left[(y_r - y_{r-1}) \cdot (y_s - y_{s-1})\right] \tag{7}$$

where ζ is a penalization constant. Thus, after $\hat{d}(p_r, p_s)$ normalization

$$\mathcal{M}^{r,s}_{\text{spatial}} = 1 - \hat{d}(p_r, p_s). \tag{8}$$

Local temporal consistency. $\mathcal{M}^{r,s}_{\text{temporal}}$ takes into account the difference of time that the observer gazes at two different fixation points. To this end, the ℓ^1 distance is introduced

$$d(\tau_r, \tau_s) = |\tau_r - \tau_s|. \tag{9}$$

The distance measured in (9) is normalized with respect to the maximum fixation time of the scanpath. Then temporal consistency is calculated as

$$\mathcal{M}^{r,s}_{\text{temporal}} = 1 - d(\tau_r, \tau_s). \tag{10}$$

Local visual consistency. $\mathcal{M}^{r,s}_{\text{visual}}$ is defined using either color and texture/shape properties. Evaluation of consistency in terms of color is performed by exploiting well-known histogram intersection, which again is an ℓ^1 distance on the color space [53]. Given the two color histograms $h(C^m_r)$ and $h(C^n_s)$, defined on the same number of bins $b = [1, \ldots, B]$

$$d^{r,s}_{\text{col}} = \frac{\sum_{b=1}^{B} \left(\min\left(h_b\left(C^m_r\right), h_b\left(C^n_s\right)\right)\right)}{\sum_{b=1}^{B} h_b\left(C^m_r\right)} \tag{11}$$

where $\sum_{b=1}^{B} h_b(C^m_r)$ is a normalization factor. Then

$$\mathcal{M}^{r,s}_{\text{col}} = 1 - d^{r,s}_{\text{col}}. \tag{12}$$

Computational complexity for the histogram analysis part is proportional to the number of bins in the histogram, thus taking B time units.

Shape and texture consistency is measured as

$$\mathcal{M}^{r,s}_{\text{tex}} = 1 - \frac{1}{R} \sum_{i=1}^{|\Sigma^2|} \frac{\left|\Sigma^2_{C^m_r}[i] - \Sigma^2_{C^n_s}[i]\right|}{\min\left(\left|\Sigma^2_{C^m_r}[i]\right|, \left|\Sigma^2_{C^n_s}[i]\right|\right)} \tag{13}$$

where R is a normalization factor to bound the sum in [0,1], and $|\Sigma^2|$ the number of features in the feature vector Σ^2 computed through (1). Eventually, *FOAs* visual content consistency is given from the weighted mean of terms calculated via (12) and (13)

$$\mathcal{M}^{r,s}_{\text{visual}} = \mu_1 \mathcal{M}^{r,s}_{\text{col}} + \mu_2 \mathcal{M}^{r,s}_{\text{tex}}. \tag{14}$$

The computation cost of (3) is approximately linear in the number of histogram bins B, since $|\Sigma^2| = 18$, and (8) and (10), are performed in constant time units. Thus, the algorithm [see (4)] requires $(2H + 1)BK$ operations, which means that, once H and B have been fixed as in our case, the AC algorithm is linear in the number of FOA's K; in particular, a value of $H = 2$ for the best fit window provides suitable results. The value of $K = 10$ was chosen either because, in this way, each *FOA* is only visited once, and for the bottom-up importance of earliest FOAs [40]. For what concerns the setting of equation parameters, considering again (3), we simply use $\alpha = \beta = \gamma = 1/3$, granting equal informational value to the three kinds of consistencies; similarly, we set $\mu_1 = \mu_2 = 1/2$ in (14).

B. Using Attention Consistency and Prior Knowledge for Detecting Shot Transitions

The observer's behavior can be formalized as the attention consistency gauged over subsequences of the video sequence f. To this end, let us generalize the local attention consistency measure $\mathcal{M}$ to a parametrized family $\mathcal{M} : F \times F \times N^+ \to R^+$, which accounts for the attentive behavior over the full sequence f , namely $(\mathcal{M}(\mathcal{T}(i), \mathcal{T}(i + l)))_{i=0,l,\ldots,N/l}$.

In such framework, the problem of inferring a shot change given the change of observation behavior $\mathcal{M}(t)$ can be conceived as a signal detection problem where the probability that a shot boundary B occurs, given a behavior $\mathcal{M}(t)$, $P(B|\mathcal{M}(t))$, is compared against the probability that a shot boundary is not present, $P(\bar{B}|\mathcal{M}(t))$. More precisely, the observer's judgement of his own behavior can be shaped in a Bayesian approach where detection becomes the inference between the following two hypotheses:

- $\mathcal{H}_0$: no shot boundary occurs between the two frames under analysis ($\bar{B}$);
- $\mathcal{H}_1$: a shot boundary occurs between the two frames (B).

In this setting, the optimal decision is provided by a test where $\mathcal{H}_1$ is chosen if $p(\mathcal{M}(t)|B)P(B) > p(\mathcal{M}(t)|\bar{B})P(\bar{B})$ and $\mathcal{H}_0$ is chosen, otherwise. Namely, a cut occurs if

$$\mathcal{L}(t) > \frac{P(\bar{B})}{P(B)} = \frac{1 - P(B)}{P(B)} \tag{15}$$

where $\mathcal{L}(t) = p(\mathcal{M}(t)|B)/p(\mathcal{M}(t)|\bar{B})$ represents a likelihood ratio.

In general, the prior shot probability $P(B)$ models shot boundaries as arrivals over discrete, nonoverlapping temporal

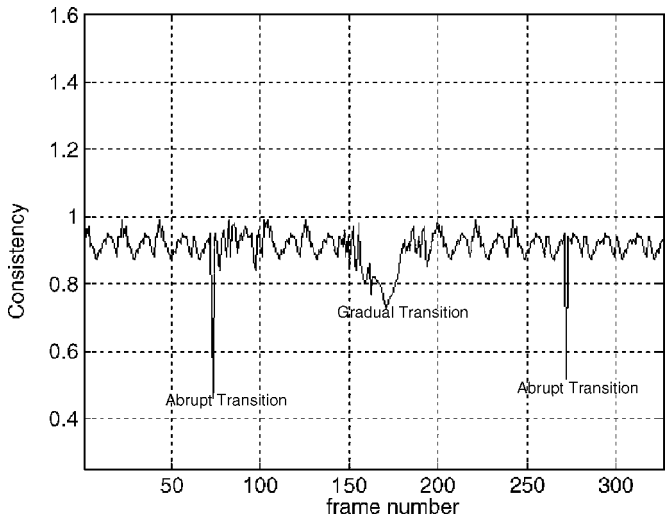

Fig. 8. Plot of $\mathcal{M}(t)$ function for a sequence characterized by one a dissolve region embedded between two abrupt transitions.

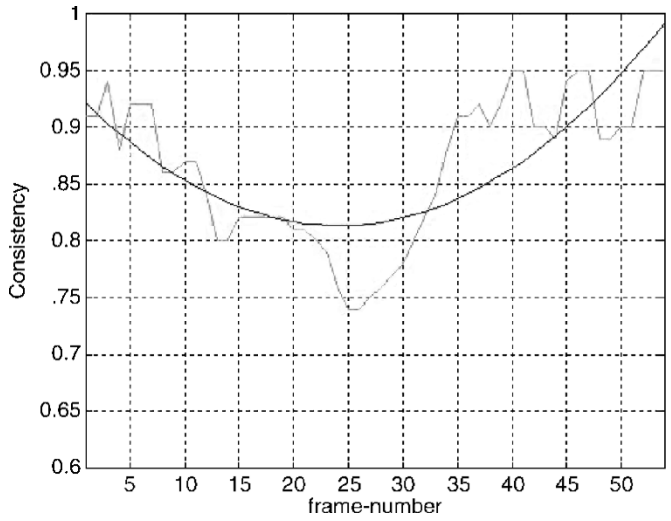

Fig. 9. Attention consistency $\mathcal{M}$ in a dissolve region and its parabolic fitting.

intervals, and a Poisson process seems an appropriate prior [21], [58], which is based on the number of frames elapsed since the last shot boundary. Hanjalic has suggested [21] that the prior $P(B)$ should be more conveniently corrected by a factor depending upon the structural context of the specific shot boundary, gauged through a suitable function.

It is possible to generalize this suggestion resorting to contextual Bayesian analysis [47] in which an occurrence of the property B is detected by taking into account the behavior $\mathcal{M}(t)$ given a context E, that is a set of events $\{e_1, e_2, \ldots, e_n\}$ characterizing B. Namely, $\mathcal{H}_1$ is chosen if $p(\mathcal{M}(t)|B,E)P(S|E) > p(\mathcal{M}(t)|\bar{B},E)P(\bar{B}|E)$. Thus, a cut is detected according to the likelihood ratio

$$\mathcal{L}(t) > \frac{1-P(B|E)}{P(B|E)} \tag{16}$$

where now the right-hand side of the equation defines the adaptive threshold

$$T(t) = \frac{1-P(E|B)P(B)}{P(E|B)P(B)}. \tag{17}$$

The prior probability $P(B)$ models the Poisson process of boundary arrival according to the cumulative probability $P(B) = 1/2 \cdot \sum_{w=0}^{\lambda(t)} (\mu^w/w!)\exp(-\mu)$ [21].

As regards $P(E|B)$, under weak coupling assumption [62] of structural events $e_1, e_2, \ldots, e_n$, we can set $P(e_1, e_2, \ldots, e_n|B) = \prod_i P(e_i|B)$. The events that constitute the structural context can be described as follows.

Consider the behavior of function $\mathcal{M}$ for both abrupt and gradual transitions. An example related to a video sequence characterized by the presence of two hard cuts embedding a dissolve is depicted in Fig. 8.

The first event we deal with is a *shape* event: when the gist of the world observed abruptly changes (hard cut), $\mathcal{M}$ decreases down to a minimum value.

Thus, as regards hard cuts, to calculate the probability $P(E|B)$, we use a sliding window of dimension $W = 10$, centered on the frame $f(t)$, thus including all frames in the temporal interval $[t-(W/2), t+(W/2)]$, chosen with the interframe distance $l = 5$. For each frame, we consider the probability that the difference between the first minimum of $\mathcal{M}$, $\mathcal{M}_{\min 1}$, and the second minimum $\mathcal{M}_{\min 2}$ detected within the temporal window, be significant

$$P(E|B) = P(\text{shape}|B_{\text{cut}}) = \frac{1}{1+\exp(\beta'\delta)} \tag{18}$$

where δ represents the normalized difference $(\mathcal{M}_{\min 1} - \mathcal{M}_{\min 2})/\mathcal{M}_{\min 1}$.

On the contrary, during a dissolve, the difference between consecutive frames is reduced, and a frame is likely to be similar to the next one. Thus, the consistency function will vary smoothly across the transition interval. Indeed, the behavior of $\mathcal{M}$ along a dissolve region is of parabolic type, and can be more precisely appreciated in Fig. 9, where $\mathcal{M}(t)$ decreases very slowly till a local minimum point (fade-out effect), then slowly increases (fade-in effect).

A second event, which we denote $d\mathcal{M}$, stems from the fact that the first derivative function of $\mathcal{M}$ is approximately constant and about zero in those frames characterized by dissolve effects (see Fig. 10). Clearly, previous events are not sufficient to completely characterize the context of a dissolve region: in fact, $\mathcal{M}$ could exhibit a similar trend, e.g., in shots featuring a slow zoom. Thus, the inconsistency between the edge frames, that is the first and last frames of an hypothetical dissolve region, must be taken into account. We denote this event a *change* event.

Summing up, in the case of dissolves we can assume

$$P(E|B) = P(\text{shape}|B_{\text{dis}})P(d\mathcal{M}|B_{\text{dis}})P(change|B_{\text{dis}}). \tag{19}$$

To calculate the probability $P(E|B)$, we use a sliding window of dimension $W = 20$, centered on the frame $f(t)$, which includes all frames in the temporal interval $[t-(W/2), t+(W/2)]$, chosen with the interframe distance $l = 5$. The first term on the right-hand side of (19) is defined as

$$P(\text{shape}|B_{\text{dis}}) = \frac{1}{1+\exp\left(\beta'(d_{\min \text{P}})\right)} \tag{20}$$

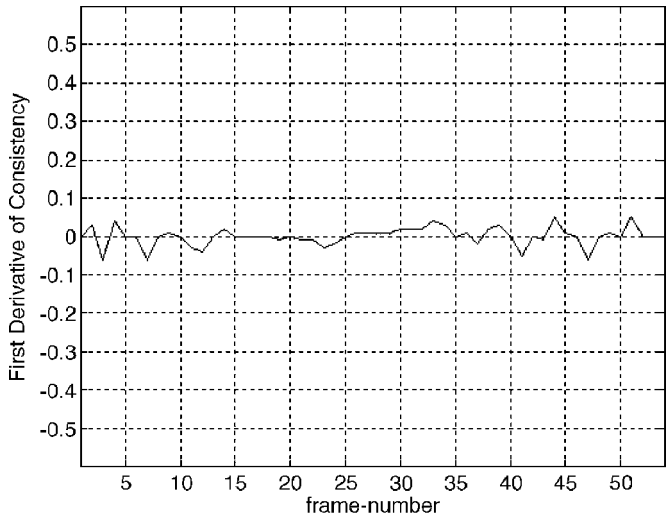

Fig. 10. First derivative of $\mathcal{M}(t)$ in the same region shown in Fig. 9.

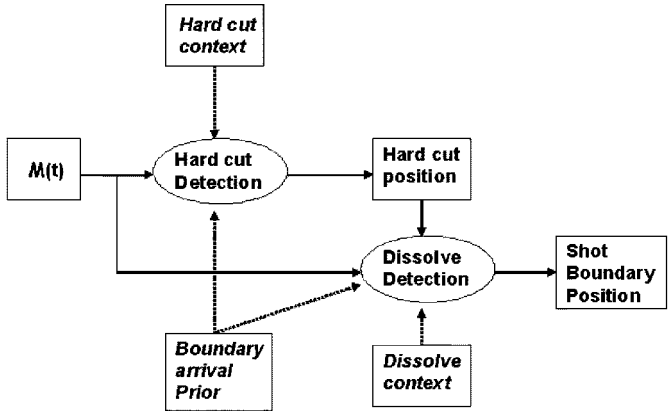

Fig. 11. Decision module for inferring boundary presence from $\mathcal{M}(t)$ behavior and prior/contextual knowledge.

where d_{minP} represents the distance between the absolute minimum of $\mathcal{M}$ within the temporal window and the minimum of the parabolic fitting performed on $\mathcal{M}$ values occurring in the same window.

The second term $P(d\mathcal{M}|B_{\text{dis}})$ accounting for the probability that derivative $d\mathcal{M}/dt$ be close to zero, is modeled as

$$P(d\mathcal{M}|B_{\text{dis}}) = \exp\left(-k\left|\frac{d\mathcal{M}}{dt} - \mu\right|\right) \tag{21}$$

where μ is the mean value of $d\mathcal{M}/dt$ within the time window. To compute derivatives, the $\mathcal{M}$ curve is preprocessed via median filtering [41] in order to avoid noise boost-up.

The third term $P(change|B_{\text{dis}})$, representing the probability that the first and the last frame of the dissolve be different, is given by

$$P(\text{change}|B_{\text{dis}}) = 1 - \frac{1}{1+\exp\left(-\beta\left(\mathcal{M}\left(\mathcal{T}(f_{\text{start}}), \mathcal{T}(f_{\text{end}})\right) - \delta'\right)\right)} \tag{22}$$

where f_{start} and f_{end} are the first and last frame of the sliding window, $f_{\text{start}} = f_{t-(W/2)}$ and $f_{\text{end}} = f_{t+(W/2)}$ respectively. The variation δ' is defined as

$$\delta' = \mathcal{M}_{\min} + \frac{(\mathcal{M}_{max} - \mathcal{M}_{\min})}{2} \tag{23}$$

where $\mathcal{M}_{\min}$ and $\mathcal{M}_{max}$ represent the absolute minimum and maximum values of the $\mathcal{M}$ function within the window, respectively.

The likelihood in (16) is estimated, on training sequences, by computing the histograms of the $\mathcal{M}(t)$ values within a shot and at its boundaries, respectively; then, ideal distributions are derived in non parametric form through Parzen windows [12] using kernels $\xi(1-\mathcal{M})\exp(-(1-\mathcal{M}))$ (boundaries) and $(1/\sigma\sqrt{2\pi})\exp(-((1-\mathcal{M})-\mu)^2/2\sigma^2)$ (within shot), where $\xi = 2.5$, $\mu = 1.1$, $\sigma = 0.4$, are the estimated parameters. Eventually, the decision module can be outlined as in Fig. 11.

The input is represented by the $\mathcal{M}(t)$ sequence computed by applying the AC algorithm on the video sequence, together with contextual knowledge. Boundary detection is accomplished according to a two-step procedure, which we denote inconsistency detection (ID).

In a first step abrupt transitions are detected by means of (16)–(18). At the end of this phase, we obtain the positions of hard cuts, which partition the original video in a sequence of blocks representing candidate shots.

In a second step, the frames interested in dissolve effects are detected. For each block, dissolve regions are individuated by means of (16), (17), and (20)–(22), computed through a sliding window centered on each frame of the block, chosen according to an interframe distance $l = 5$. Eventually, the output of the system is represented by the list of shot boundary positions, defining the shot segmentation of the original video.

The first step of the ID algorithm has complexity $O(N/l)$, N being the number of frames of the video sequence. The second step is $O(WN_bL_b/l)$, where W, N_b, L_b are the dimension of the sliding window, the number of blocks partitioned along the first step, and the maximum block length, respectively.

The dimensions of the sliding windows have been chosen by means of an analysis of *ROC* curves obtained for the training set in order to maximize true detections with respect to false alarms.

VI. Experiments and Results

To evaluate the performance of the proposed shot detection algorithm, a database of video sequences has been obtained from documentaries and news belonging to TREC01 video repository and from famous movies. The database represents a total of 1694 cuts and 440 dissolves in approximately 166 min of video. The selected sequences are complex with extensive graphical effects. Videos were captured at a rate of 30 frames/s, 640 × 480 pixel resolution, and stored in *AVI* format. These video sequences are also characterized for presenting significant dissolve effects. For each sequence a ground-truth was obtained by three experienced humans using visual inspection [16].

To obtain an estimate of parameters for detection, the training set, shown in Table I, has been used.

Experiments for performance evaluation were carried out on a test set including a total of 1304 cuts and 336 dissolves in 130 min of video, which is summarized in Table II.

TABLE I
DESCRIPTION OF THE VIDEO TRAINING SET

Sequence	Dur.(sec.)	Transitions (Abrupt-Gradual)
The School of Athens (Docum.)	60	0-9
BOR03 (Doc. TREC01)	330	34-21
ANNI006 (Doc. TREC01)	366	41-28
The Time Machine (Movie)	118	15-6
The Life is Beautiful (Movie)	600	100-30
Moulin Rouge (Movie)	700	200-10
Total	36 min	390-104

TABLE II
DESCRIPTION OF THE VIDEO SEQUENCES IN THE TEST SET

Sequence	Dur.(sec.)	Transitions (Abrupt-Gradual)
ANNI005 (Doc.TREC01)	245	38-8
BOR02 (Doc.TREC01)	328	20-9
BOR07 (Doc.TREC01)	420	45-22
BOR08 (Doc.TREC01)	350	42-18
NAD31 (Doc.TREC01)	516	51-19
NAD33 (Doc.TREC01)	310	41-8
NAD53 (Doc.TREC01)	692	62-36
NAD55 (Doc.TREC01)	485	49-24
NAD57 (Doc.TREC01)	420	43-23
SENSES111 (Doc.TREC01)	388	31-18
Desert Storm (News)	30	4-4
Mandela (News)	22	0-3
Dinosaurs (Movie)	600	205-50
Harry Potter (Movie)	661	176-21
Matrix (Movie)	617	162-0
The Fifth Element (Movie)	600	191-0
The Lord of the Rings II (Movie)	627	63-33
The Patriot (Movie)	500	81-47
Total	130 min	1304-336

The comparison between the proposed algorithm's output and the ground truth relies on the well-known recall and precision figures of merit [16]

$$\text{recall} = \frac{\text{detects}}{(\text{detects} + \text{MD})} \tag{24}$$

$$\text{precision} = \frac{\text{detects}}{(\text{detects} + \text{FA})} \tag{25}$$

where detects denotes the correctly detected boundaries, while MD and FA denote missed detections and false alarms, respectively. In other terms, at fixed parameters, recall measures the ratio between right detected shot changes and total shot changes in a video, while precision measures the ratio between right detected shot changes and the total shot changes detected by algorithm.

Results obtained are provided in Tables III and IV and summarized in Table V.

The proposed method achieves a 97% recall rate with a 95% precision rate on abrupt transitions, and a 92% recall rate with a 89% precision rate on gradual transitions (Table V). In order to provide an idea about the quality of this results, we refer to the discussion published by Hanjalic [21]. In particular, on dissolve detection, it is worth comparing with Lienahrt [30] and the works therein reported [29], [63].

Also, Table V provides results in terms of the $F1$ metric, $F1 = 2 \times \text{precision} \times \text{recall}/(\text{precision} + \text{recall})$, which is commonly used to combine precision and recall scores [43], $F1$ being high only when both scores are high. Summing up, the method proposed here achieves an overall average $F1$ performance of 0.91 when considering both kinds of transitions. This result can indicatively be compared to the performance of

TABLE III
ABRUPT TRANSITION PERFORMANCE OF THE FOVEATED DETECTION METHOD

Video	Cuts	Detections	MD	FA
ANNI005	38	40	0	2
BOR02	20	19	1	0
BOR07	45	50	1	6
BOR08	42	41	2	3
NAD31	51	52	0	1
NAD33	41	41	0	0
NAD53	62	65	1	4
NAD55	49	49	0	0
NAD57	43	42	2	1
SENSES111	31	31	2	2
Desert Storm	4	4	0	0
Dinosaurs	205	210	3	8
Harry Potter	176	176	0	0
Matrix	162	170	2	10
The Fifth Element	191	189	3	1
The Lord of the Rings II	63	67	0	4
The Patriot	81	83	1	3
Total	1304	1322	18	45

TABLE IV
GRADUAL TRANSITION PERFORMANCE OF THE FOVEATED DETECTION METHOD

Video	Dissolves	Detections	MD	FA
ANNI005	8	8	2	2
BOR02	9	10	3	4
BOR07	22	24	3	5
BOR08	18	16	2	0
NAD31	19	22	1	4
NAD33	8	8	0	0
NAD53	36	38	0	2
NAD55	24	27	2	5
NAD57	23	21	3	1
SENSES111	18	21	1	4
Desert Storm	4	4	0	0
Mandela	3	3	0	0
Dinosaurs	50	52	5	7
Harry Potter	21	22	1	2
The Lord of the Rings II	33	35	5	7
The Patriot	47	49	2	4
Total	336	360	30	43

TABLE V
PERFORMANCE OF THE METHOD

Type of Transition	Average Recall	Average Precision	$F1$
Abrupt	0.97	0.95	0.93
Gradual	0.92	0.89	0.90

a recently proposed method [43] that uses global and block wise histogram differences, camera motion likelihood, followed by k-nearest neighbor classification. Such method achieves an $F1$ performance of 0.94 and 0.69, for hard cuts and gradual transitions, respectively, resulting in an average performance of 0.82; interestingly enough, this result is higher than average scores (0.82 and 0.79) obtained by the two best performing systems at 2001 TREC evaluation [43]. It is worth noting that, in our case, the overall score of 0.91 also accounts for results obtained by processing movies included in our test set, which eventually resulted to be the most critical. For completeness sake, by taking into account only TREC01 video sequences, the overall performance of our method is 0.925.

As regards the efficiency of the method, recall that to obtain the visuomotor trace of the frame, main effort is spent on pyramid and WTA computation, which can be estimated as an $O(|\Omega|)$ step, where $|\Omega|$ represents the number of samples in the

TABLE VI
AVERAGE FRAME PROCESSING TIME FOR EACH STEP

Steps	Low-level	AC algorithm	ID algorithm
Elapsed time (msec)	26	6.2	2.8

image support Ω, while FOA analysis involves lower time complexity, since each of the K FOAs is defined on a limited support with respect to the original image ($1/36|\Omega|$) and only ten FOAs are taken into account to form a trace. The AC algorithm is $O(K)$, that is linear in the number of FOAs. The first step of ID algorithm has complexity $O(N/l)$, N and l being the number of frames of the video sequence and the interframe distance, respectively. The second step is $O(WN_bL_b/l)$, where W, N_b, L_b are the dimension of the sliding window, the number of blocks partitioned along the first step, and the maximum block length, respectively. From this analysis, by considering operations performed on a single frame, we can expect that most of the time will be spent in the low-level perception stage, while the AC and ID algorithms will have higher efficiency, the former only performing on a sparse representation of the frame ($K = 10$) and the latter working on $\mathcal{M}(t)$ values of the sliding window of dimension W. This is experimentally confirmed from the results obtained and reported in Table VI.

The system achieves a processing speed per frame of about 35 ms on a Pentium IV 2.4-GHz PC (1-GB RAM). It is worth noting that the current prototype has been implemented using the Java programming language, running in Windows XP operating system, without any specific optimization.

Clearly, for time critical applications, the bottleneck of the proposed method, that is the computing of visuomotor traces, could be easily reduced by resorting to existing hardware implementation of pyramidal representations ([9]) and more efficient realizations of the WTA scheme (e.g., in [4] a network is presented, which has $O(\lg n)$ time complexity).

VII. CONCLUSION

In this paper, we have defined a novel approach to partitioning of a video into shots based on a foveated representation of the video. To the best of our knowledge, foveation mechanisms have never been taken into account for video segmentation, while there are some recent applications to video compression (refer to [26]). The motivation for the introduction of this approach stems from the fact that success or failure in the perception of changes to the visual details of a scene across cuts are related to the attentive performance of the observer [51]. By exploiting the mechanism of attention shifting through saccades and foveations, the proposed shot-change detection method computes, at each time instant, a consistency measure $\mathcal{M}(t)$ of the foveation sequences generated by an ideal observer looking at the video. The problem of detecting a shot change given the change of consistency $\mathcal{M}(t)$ has been conceived as a Bayesian inference of the observer from his own visual behavior.

The main results achieved can be summarized as follows. The proposed scheme allows the detection of both cuts and dissolves between shots using a single technique, rather than a set of dedicated methods. Also, it is well grounded in visual perception theories and allows to overcome usual shortcomings of many other techniques proposed so far. In particular, features extracted are strictly related to the visual content of the frame; this, for instance is not true for simpler methods, such as histogram based methods, where, in general, totally different frames may have similar histograms (e.g., a frame generated by randomly flipping the pixels of another frame has the same histogram of the original one). Further, the FOA representation is robust with respect to smooth view changes: for instance, an object translating with respect a a background, gives rise to a sequence of similar visuomotor traces. Meanwhile, a large object entering the scene would be recognized as a significant discontinuity in the visual content flow of the video sequence; in this sense, the approach accounts for the more general definition of shot as a sequence of frames that was, or appears to be, continuously captured from the same camera [16]. Once the distinctive scanpath has been extracted from a frame, subsequent feature analysis need only to process a sparse representation of the frame; note that for each frame, we consider ten FOAs, each FOA being defined on a square support region whose dimension is 1/36 of the original image; further reduction is achieved at the detection stage, where only the $\mathcal{M}$ function is processed (cfr. Table VI). Last, the perceptual capacity of an observer to account for his own visual behavior, naturally leads, in this framework, to a Bayesian decision formulation for solving the detection problem, in a vein similar to [21] and [58]. In particular, by resorting to recently proposed contextual Bayesian analysis [47], we have generalized some suggestions introduced in [21] for exploiting structural information related to different types of transitions.

It is worth remarking that, with respect to the specific problem of gradual transitions, the present work focuses on dissolve detection. However, the detection scheme can be easily extended to other kinds of transitions; for instance, preliminary experiments performed on wipes (not reported here, because out of the scope of this paper) show a behavior of the $\mathcal{M}$ function characterized by a damped oscillatory pattern. Also, beyond the context of video segmentation, the proposed technique introduces some novelties per se with respect to the "Where" and "What" integration problem, the explicit use of the fixation time in building a visuomotor trace, and as regards the way to exploit the extracted information for comparing different views (information look-up problem).

Results on a test set representing a total of 1304 cuts and 336 dissolves in 130 min of video, including videos of different kinds are reported and validate the proposed approach. The performance of the currently implemented system is characterized by a 97% recall rate with a 95% precision rate on abrupt transitions, and a 92% recall rate with a 89% precision rate on gradual transitions. Meanwhile it exhibits a constant quality of detection for arbitrary complex movie sequences with no need for tuning parameters. Interestingly enough, the system has been trained on a small data set with respect to the test set used.

However, the introduction of an attention based approach not only is motivated by performance in shot-detection, but in perspective it could constitute an alternative to traditional approaches, and overcome their limitations for high-level video segmentation. Consider, for instance, the issue of scene change detection by jointly exploiting video and audio information. Audio and pictorial information play different roles and, to

some extent, complementary. When trying to detect a scene decomposition of the video sequence, the analysis of visual data may provides candidate cuts, which are successively validated through fusion with information extracted from audio data. How to perform such fusion, in a principled way, is unclear. However, behavioral studies and cognitive neuroscience have remarked the fundamental role of attention in integrating multimodal information [5]; and the approach proposed here could serve as a sound basis for such integration. In this way, the low-level and high-level video analysis could share the processing steps, making the entire content analysis process more effective and efficient.

ACKNOWLEDGMENT

The authors would like to thank the referees and Associate Editor for their enlightening and valuable comments which greatly improved the quality and clarity of an earlier version of this paper.

REFERENCES

[1] B. Adams, C. Breazeal, R. A. Brooks, and B. Scassellati, "Humanoid robots: A new kind of tool," *IEEE Intell. Syst. Mag.*, vol. 15, no. 4, pp. 25–31, Jul.-Aug. 2000.

[2] S. Antani, R. Kasturi, and R. Jain, "A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video," *Pattern Recogn.*, vol. 35, pp. 945–965, 2002.

[3] G. Backer, B. Mertshing, and M. Bollmann, "Data and model-driven gaze control for an active-vision system," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 12, pp. 1415–1429, Dec. 2001.

[4] M. Barnden and J. A. Srinivas, "Temporal winner-take-all networks: A time-based mechanism for fast selection in neural networks," *IEEE Trans. Neural Netw.*, vol. 4, no. 5, pp. 844–853, Sep. 1993.

[5] A. Berthoz, Le sens du mouvement, O. Jacob, Ed., 1997.

[6] G. Boccignone and M. Ferraro, "Gaze shift as a constrained random walk," *Physica A*, vol. 331, pp. 207–218, 2004.

[7] D. Broadbent, *Perception and Communication*. New York: Pergamon, 1958.

[8] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. COM-9, no. 4, pp. 532–540, Apr. 1983.

[9] P. J. Burt, "A pyramid-based front-end processor for dynamic vision applications," *Proc. IEEE*, vol. 90, no. 7, pp. 1188–1200, Jul. 2002.

[10] D. A. Chernyak and L. W. Stark, "TopDown guided eye movements," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 31, no. 4, pp. 514–522, Aug. 2001.

[11] J. Denzler and C. M. Brown, "Information theoretic sensor data selection for active object recognition and state estimation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 2, pp. 145–157, Feb. 2002.

[12] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: Wiley, 2001.

[13] W. A. C. Fernando, C. N. Canagarjah, and D. R. Bull, "Fade and dissolve detection in uncompressed and compressed video sequences," in *Proc. IEEE Int. Conf. Image Processing*, vol. 3, 1999, pp. 299–303.

[14] M. Ferraro, G. Boccignone, and T. Caelli, "Entropy-based representation of image information," *Pattern Recogn. Lett.*, vol. 23, pp. 1391–1398, 2002.

[15] B. Furht, S. W. Smoliar, and H. Zhang, *Video and Image Processing in Multimedia Systems*. Norwell, MA: Kluwer, 1995.

[16] U. Gargi, R. Kasturi, and S. H. Strayer, "Performance characterization of video-shot change detection methods," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 1, pp. 1–13, Jan. 2000.

[17] G. J. Giefing, H. Janssen, and H. Mallot, "Saccadic object recognition with an active vision system," in *Proc. 10th Eur. Conf. Artificial Intelligence*, 1992, pp. 803–805.

[18] M. A. Goodale and G. K. Humphrey, "The objects of action and perception," *Cognition*, vol. 67, pp. 181–207, 1998.

[19] H. Greenspan, S. Belongie, P. Perona, S. Rakshit, and C. H. Anderson, "Overcomplete steerable pyramid filters and rotation invariance," in *Proc. IEEE Computer Vision and Pattern Recognition*, Seattle, WA, 1994, pp. 222–228.

[20] A. Hampapur, R. Jain, and T. Weymouth, "Digital video segmentation," in *Proc. ACM Multimedia'94*, 1994, pp. 357–364.

[21] A. Hanjalic, "Shot-boundary detection: unraveled and resolved?," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 1, pp. 90–105, Jan. 2002.

[22] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, pp. 1254–1259, 1998.

[23] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Rev.*, vol. 2, pp. 1–11, 2001.

[24] T. Kikukawa and S. Kawafuchi, "Development of an automatic summary editing system for the audio visual resources," *Trans. Inst. Electron., Inform., Commun. Eng.*, vol. J75-A, no. 2, pp. 204–212, 1992.

[25] G. Krieger, I. Rentschler, G. Hauske, K. Schill, and C. Zetsche, "Object and scene analysis by saccadic eye-movements: An investigation with higher order statistics," *Spat. Vis.*, vol. 13, pp. 201–214, 2000.

[26] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video compression with optimal rate control," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 977–992, Jul. 2001.

[27] D. Li and H. Lu, "Model based video segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 5, pp. 533–544, May 1995.

[28] R. Lienhart, "Reliable transition detection in videos: a survey and practicioner's guide," *Int. J. Image Graph.*, vol. 1, no. 3, pp. 469–486, 2001.

[29] ——, "Comparison of automatic shot boundary detection algorithms," in *Proc. of SPIE 3656-29 Image and Video Processing*, vol. VII, 1999, pp. 1–12.

[30] ——, "Reliable dissolve detection," in *Proc. SPIE 4315*, 2001, pp. 219–230.

[31] T. M. Liu, H. J. Zhang, and F. H. Qi, "A novel video key frame extraction algorithm," in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 4, 2002, pp. 149–152.

[32] S. Mallat, *A Wavelet Tour of Signal Processing*. New York: Academic, 1998.

[33] R. Milanese, S. Gil, and T. Pun, "Attentive mechanisms for dynamic and static scene analysis," *Opt. Eng.*, vol. 34, pp. 2428–2434, 1995.

[34] A. Nagasaka and Y. Tanaka, "Automatic video indexing and full-videosearch for object appearances," *Visual Database Systems II*, pp. 113–127, 1995.

[35] W. W. Nelson and G. R. Loftus, "The functional visual field during picture viewing," *J. Exp. Psych., Human Learn. Mem.*, vol. 6, pp. 391–399, 1980.

[36] E. Niebur and C. Koch, *Computational Architectures for Attention*. Cambridge, MA: MIT Press, 1998.

[37] D. Noton and L. Stark, "Scanpaths in saccadic eye movements while viewing and recognising patterns," *Vis. Res.*, vol. 11, pp. 929–942, 1971.

[38] K. O'Regan, "Solving the 'real' mysteries of visual perception: the world as an outside memory," *Canad. J. Psychol.*, vol. 46, no. 3, pp. 461–488, 1992.

[39] K. Otsuji, Y. Tonomura, and Y. Ohba, "Video browsing using brightness data," in *Proc. SPIE-IST VCIP'91*, vol. 1606, 1991, pp. 980–989.

[40] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vis. Res.*, vol. 42, pp. 107–123, 2002.

[41] I. Pitas and A. N. Venetsanopoulos, *Non-Linear Filters*. Norwell, MA: Kluwer, 1989.

[42] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: comparison with eye fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 970–982, Sep. 2000.

[43] Y. Qi, A. Hauptmann, and T. Liu, "Supervised classification for video shot segmentation," in *Proc. IEEE Conf. Multimedia Expo (ICME03)*, vol. 2, 2003, pp. 689–692.

[44] R. P. N. Rao and D. H. Ballard, "Dynamic model of visual recognition predicts neural response properties in the visual cortex," *Neur. Comp.*, vol. 9, pp. 721–763, 1997.

[45] R. P. N. Rao, "An optimal estimation approach to visual perception and learning," *Vis. Res.*, vol. 39, pp. 1963–1989, 1999.

[46] R. A. Rensink, J. K. O'Regan, and J. J. Clark, "To see or not to see: The need for attention to perceive changes in scenes," *Psych. Sci.*, vol. 8, pp. 368–373, 1997.

[47] W. Richards, A. Jepson, and J. Feldman, "Priors, preferences and categorical percepts," in *Perception as Bayesian Inference*, D. C. Knill and W. Richards, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1996, pp. 93–122.

[48] R. D. Rimey and C. M. Brown, "Controlling eye movements with hidden Markov models," *Int. J. Comput. Vis.*, vol. 7, pp. 47–65, 1991.

[49] K. Schill, E. Umkehrer, S. Beinlich, G. Krieger, and C. Zetzsche, "Scene analysis with saccadic eye movements: top-down and bottom-up modeling," *J. Electron. Imag.—Spec. Issue. Human Vision Elect. Imag.*, vol. 10, no. 1, pp. 152–160, Jan. 2001.

[50] B. Shahraray, "Scene change detection and content-based sampling of video sequences," in *Proc. 1st SPIE*, vol. 2419, Feb. 1995, pp. 2–13.

[51] D. J. Simons and D. T. Levin, "Change blindess," *Trends Cogn. Sci.*, no. 7, pp. 261–267, 1997.
[52] E. A. Styles, *The Psychology of Attention*. Hove, U.K.: Psychology Press, 1997.
[53] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vision*, vol. 7, no. 1, pp. 11–32, 1991.
[54] J. K. Tsotsos *et al.*, "Modeling visual-attention via selective tuning," *Art. Intell.*, vol. 78, pp. 507–545, 1995.
[55] A. M. Tekalp, *Digital Video Processing*. Upper Saddle River, NJ: Prentice-Hall, 1995.
[56] S. Thorpe, A. Delorme, and R. Van Rullen, "Spike-based strategies for rapid processing," *Neural Netw.*, vol. 14, pp. 715–725, 2001.
[57] B. T. Truong, C. Dorai, and S. Venkatesk, "New enhancements to cut, fade, and dissolve detection processes in video segmentation," *ACM Multimedia*, pp. 219–227, 2000.
[58] N. Vasconcelos and A. Lippman, "Statistical models of video structure for content analysis and characterization," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 3–19, Jan 200.
[59] G. J. Walker-Smith, A. G. Gale, and J. M. Findlay, "Eye movement strategies involved in face perception," *Perception*, vol. 6, pp. 313–326, 1977.
[60] A. L. Yarbus, *Eye Movements and Vision*. New York: Plenum, 1967.
[61] B.-L. Yeo and B. Liu, "Rapid scene analysis on compressed video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 12, pp. 533–544, Dec. 1995.
[62] A. L. Yuille and H. H. Bulthoff, "Bayesian decision theory and psychophysics," in *Perception as Bayesian Inference*, D. C. Knill and W. Richards, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1996, pp. 123–162.
[63] R. Zabih, J. Miller, and K. Mai, "A feature-based algorith, for detecting and classifying scene breaks," in *Proc. ACM Int. Conf. Multimedia*, 1995, pp. 189–200.
[64] H. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Syst.*, vol. 1, pp. 10–28, 1993.

Giuseppe Boccignone (M'95) received the Laurea degree in theoretical physics from the University of Torino, Torino, Italy, in 1985.

In 1986, he joined Olivetti Corporate Research, Ivrea, Italy. From 1990 to 1992, he served as a Chief Researcher of the Computer Vision Lab at CRIAI, Naples, Italy. From 1992 to 1994, he held a Research Consultant position at Research Labs of Bull HN, Milan, Italy, leading projects on biomedical imaging. In 1994, he joined the Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica, University of Salerno, Salerno, Italy, where he is currently an Associate Professor of Computer Science. In 1997, he was appointed Research Associate at INFM (the National Institute for the Physics of Matter). He has been active in the field of computer vision, image processing, and pattern recognition, working on general techniques for image analysis and description, medical imaging, and object-oriented models for image processing. His current research interests lie in active vision, theoretical models for computational vision, and image and video databases.

Dr. Boccignone is a Member of the IEEE Computer Society.

Angelo Chianese received the Laurea degree in electronics engineering from the University of Naples, Naples, Italy, in 1980.

In 1984, he joined the Dipartimento di Informatica e Sistemistica, University of Naples "Federico II" as an Assistant Professor. He has been an Associate Professor of Computer Science and Engineering at the same university since 1992. Since 2003, he has been leading the e-Learning Laboratory at the University of Napoli. He has been active in the field of pattern recognition, optical character recognition, medical image processing, and object-oriented models for image processing. His current research interests lie in multimedia data base and multimedia content management for e-learning.

Mr. Chianese is a Member of the International Association for Pattern Recognition (IAPR).

Vincenzo Moscato received the Laurea degree in computer science and engineering from the University of Naples "Federico II," Naples, Italy, in 2002. He is currently working toward the Ph.D. degree in computer science and engineering at the same university.

His research interests are in the area of image processing (active vision) and multimedia database systems (image databases, video databases, and architectures for multimedia data sources integration).

Antonio Picariello received the Laurea degree in electronics engineering and the Ph.D. degree in computer science and engineering, both from the University of Naples, Naples, Italy, in 1991 and 1998, respectively.

In 1993, he joined the Istituto Ricerca Sui Sistemi Informatici Paralleli, The National Research Council, Naples, Italy. In 1999, he joined the Dipartimento di Informatica e Sistemistica, University of Naples "Federico II," and is currently an Assistant Professor of Data Base and Senior Researcher. He has been active in the field of computer vision, medical image processing and pattern recognition, object-oriented models for image processing, and multimedia database and information retrieval. His current research interests lie in knowledge extraction and management, multimedia integration and image and video databases.

Dr. Picariello is a Member of the International Association for Pattern Recognition (IAPR).