Multimedia Tools and Applications 3, 179–202 (1996)

Content-Based Representation and Retrieval of Visual Media: A State-of-the-Art Review

PHILIPPE AIGRAIN
Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier, 118, route de Narbonne, F-31062 Toulouse Cedex, France

HONGJIANG ZHANG[1]
Broadband Information Systems Lab., Hewlett-Packard Labs., 1501 Page Mill Road, Palo Alto, CA94304, USA

DRAGUTIN PETKOVIC
IBM Almaden Research Center, San Jose, CA 95120-6099, USA

Abstract. This paper reviews a number of recently available techniques in content analysis of visual media and their application to the indexing, retrieval, abstracting, relevance assessment, interactive perception, annotation and re-use of visual documents.

Keywords: content-based retrieval, image, video, texture, colour, editing, shot, sequence, indexing, visual interfaces, soundtrack, video browsers

1. Background

A few years ago, the problems of representation and retrieval of visual media were confined to specialized image databases (geographical, medical, pilot experiments in computerized slide libraries), in the professional applications of the audiovisual industries (production, broadcasting and archives), and in computerized training or education. The present development of multimedia technology and information highways has put content processing of visual media at the core of key application domains: digital and interactive video, large distributed digital libraries, multimedia publishing. Though the most important investments have been targeted at the information infrastructure (networks, servers, coding and compression, delivery models, multimedia systems architecture), a growing number of researchers have realized that content processing will be a key asset in putting together successful applications. The need for content processing techniques has been made evident from a variety of angles, ranging from achieving better quality in compression, allowing user choice of programs in video-on-demand, achieving better productivity in video production, providing access to large still image databases or integrating still images and video in multimedia publishing and cooperative work.

Content-based retrieval of visual media and representation of visual documents in human-computer interfaces are based on the availability of content representation data (time-structure for time-based media, image signatures, object and motion data). When it is possible, the human production of this descriptive data is so time consuming—and thus

costly—that it is almost impossible to generate it for large document spaces. There is some hope that for video documents, some of this data will be created at production time and coded in the document itself. Nonetheless it will never be available for many existing documents, and when considering the history of media and carriers one is lead to a very cautious estimate of how often this type of information will be really available even in future documents. Thus, there is a clear need for automatic analysis tools which are able to extract representation data from the documents.

The researchers involved in content processing efforts come from various backgrounds, for instance:

- the publishing, entertainment, retail or document industry where researchers try to extend their activity to visual documents, or to integrate them in hypertext-based new document types,
- the AV hardware and software industry, primarily interested by digital editing tools and other programma production tools,
- academic and national research laboratories where research had been conducted for some time on computer analysis and access to existing visual media, such as the MIT Media Laboratory [48], the Institute of System Sciences in Singapore [71], or IRIT in France [3].
- large telecommunication company laboratories, where researchers are primarily interesting in cooperative work and remote access to visual media,
- the robotics vision, signal processing, image sequence processing for security, or data compression research communities who try to find new applications for their models of images or human perception.
- computer companies developing digital library, multimedia databases and other multimedia applications or home appliances, such as IBM Almaden Research Center [23], NTT [61], Hitachi [63], Siemens Research [14] and Virage.

These researchers originally used very different models and techniques and often conflicting vocabulary. After a few years of lively confusion and exciting achievements, it is now possible to draw a clearer panorama of the state of this emerging field, and to outline some of its possible directions of development.

In this paper, we review the methods available for different types of visual content analysis, representation and their application and survey some open research problems. Section 2 covers various visual features for representing and comparing image content. Sections 3 reviews video content parsing and representation algorithms and schemes, including temporal segmentation, video abstraction, shot comparison and soundtrack analysis. Section 4 presents applications of visual representation schemes in content-based image and video retrieval and browsing. Finally, Section 5 summaries our survey and current research directions.

2. The many facets of image similarity

Retrieval of still images by similarity, i.e., retrieving images which are similar to an already retrieved image (retrieval by example) or to a model or schema is a relatively old

idea. Some might date it to the mnemotechnical ideas of the antiquity, but more seriously it appeared in specialized geographical information systems databases around 1980, in particular in the Query by Pictorial Example system of IMAID [19]. From the start, it was clear that retrieval by similarity called for specific definitions of what it means to be similar. In the mapping system, a satellite image was matched to existing map images from the point of view of similarity of road and river networks, easily extracted from images by edge detection. Apart from paper models [2, 8], it was only in the beginning of the 90s that researchers started to look at retrieval by similarity in large set of heterogeneous images with no specific model of their semantic contents. The prototype systems of Kato [38], followed by the availability of the QBIC commercial system using several types of similarities [23] contributed to making this idea more and more popular.

A system for retrieval by similarity rests on 3 components:

- extraction of features or image signatures from the images, and an efficient representation and storage strategy for this precomputed data,
- a set of similarity measures, each of which captures some perceptively meaningful definition of similarity, and which should be efficiently computable when matching an example with the whole database,
- a user interface for the choice of which definition(s) of similarity should be applied for retrieval, and for the ordered and visually efficient presentation of retrieved images and for supporting relevance feedback.

Recent work has made evident that:

- A large number of meaningful types of similarity can and must be defined. Only part of these definitions are associated with efficient feature extraction mechanisms and (dis)similarity measures.
- Since there are many definitions of similarity and the discriminating power of each of the measures is likely to degrade significantly for large image databases, the user interface and the feature storage strategy components of the systems will play a more and more essential role. We will come back to this point in Section 4.1.
- Visual content based retrieval is best utilized when combined with the traditional search, both at user interface and the system level. The basic reason for this is that we do not see the possibility of content based retrieval replacing the ability of parametric (SQL) search, text and keywords to represent the rich semantic content of the visual material (names, places, action, prices, etc.). The key is to apply content based retrieval where appropriate, and this is where the use of text and keywords is suboptimal. Examples of such applications are where visual appearance (e.g., color, texture, shape, motion) are important search arguments like in stock photo/video, art, retail, on-line shopping etc. Not only content based retrieval reduces the high variability among human indexers, but it also enables more "fuzzy" browsing and search which in many application is an essential part of the process. It is obvious then that content based retrieval involves strong user interaction, thus necessitating the development of special fast browsers and UI techniques.

In this section we briefly survey the various types of similarity definitions and associated feature extraction and measures for systems which do not assume any specific image domain or a-priori semantic knowledge on the images.

Gudivada has listed possible types of similarity for retrieval in [26]: color similarity, texture similarity, shape similarity, spatial similarity, etc. Some of these types can be considered in all or only part of one image, can be considered independently of scale or angle or not, depending on whether one is interested in the scene represented by the image or in the image per se.

2.1. *Color similarity*

Color distribution similarity has been one of the first choices [23, 33] because if one chooses a proper representation and measure it can be partially reliable even in presence of changes in lighting, view angle, and scale. For the capture of properties of the global color distribution in images, the need for a perceptively meaningful color model leads to the choice of HLS (Hue-Luminosity-Saturation) models, and of measures based on the 3 first moments of color distributions [57] preferably to histogram distances. It has been proposed in [11] to use hue and saturation distributions only when one wants to capture lighting-independent color distribution properties which are good signatures of a scene when the scale does not change too much. In this case one can identify the hue-saturation perceptive space with the complex unit discuss and define measures using statistical moments in this space. This is useful to avoid biases of measures which do not take in account the circular nature of hue, and could be further refined to distinguish between true spectral hues and the purples. Stricker and Orengo have argued in [57] for the importance of including the 3rd moment (distribution skewness) in the definition of the similarity measure.

One important difficulty with color similarity is that when using it for retrieval, an user will often be looking for an image "with a red object such as this one". This problem of restricting color similarity to a spatial component, and more generally of combining spatial similarity and color similarity is also present for texture similarity. It explains why prototype and commercial systems have included complex ad-hoc mechanisms in their user interfaces to combine various similarity functions.

2.2. *Texture similarity*

For texture as for color, it is essential to define a well-funded perceptive space. Picard and Liu [49] have shown that it is possible to do so using the Wold decomposition of the texture considered as a luminance field. One gets three components (periodic, evanescent and random) corresponding to the bi-dimensional periodicity, mono-dimensional orientation, and complexity of the analyzed texture. Experiments have shown that these independent components agree well with the perceptive evaluation of texture similarity [59]. The related similarity measures has lead to remarkably efficient results including for the retrieval of large-scale textures such as images of buildings and cars [47]. In QBIC system, Tomura texture features, contrast, compactness and direction, are used [23]. But of course one is again

confronted to the problem of combining texture information with the spatial organization of several textures (see below).

2.3. *Shape similarity*

A proper definition of shape similarity calls for the distinction between shape similarity in images (similarity between actual geometrical shapes appearing in the images) and shape similarity between the objects depicted by the images, i.e., similarity modulo a number of geometrical transformations corresponding to changes in view angle, optical parameters and scale. In some cases, one wants to include even deformation of non-rigid bodies. The first type of similarity has attracted research work only for calibrated image databases of special types of objects, such as ceramic plates. Even, in this case, the researchers have tried to define shape representations which are scale independent, resting on curvature, angle statistics and contour complexity. Systems such as QBIC [23] use circularity, eccentricity, major axis orientation (not angle-independent) and algebraic moment. It should be noted that in some cases the user of a retrieval system will want a definition of shape similarity which is dependent on view angle (for instance will want to retrieve trapezoids with an horizontal basis and not the other trapezoids).

In the general case, a promising approach has been proposed by Sclaroff and Pentland [53] in which shapes are represented as canonical deformations of prototype objects. In this approach, a "physical" model of the 2D-shape is built using a new form of Galerkin's interpolation method (finite-element discretization). The possible deformation modes are analyzed using Karhunen-Loeve transform. This yields an ordered list of deformation modes corresponding to rigid body modes (translation, rotation), low-frequency non-rigid modes associated to global deformations and higher-frequency modes associated to localized deformations.

As for color and texture, the present schemes for shape similarity modelling are faced with serious difficulties when images include several objects or background. A preliminary segmentation as well as modelling of spatial relationships between shapes is then necessary (are we interested in finding images where one region represents a shape similar to a given prototype or to some spatial organization of several shapes?).

2.4. *Spatial similarity*

Gudivada and Raghavan [27] have treated spatial similarity in the situation in which it is assumed that images have been (automatically or manually) segmented into meaningful objects, each object being associated with is centroid and a symbolic name. Such a representation is called a *symbolic image*, and it is relatively easy to define similarity functions for such image modulo transformations such as rotation, scaling and translation. Efforts have also been made to address spatial similarity directly (without segmentation and object indexing). This was the case, for instance, in the original work of Kato [38], in the limited case of direct spatial similarity (without geometrical transformation), using a number of ad-hoc statistical features computed on very low resolution images.

2.5. *Object presence analysis*

Finding images in which a particular object or type of object appears—all images with cars, all shots in a video in which a given character is present—is a particular case of similarity computation. Once again, the range of applicable methods is defined by the invariants of the object to be recognized. For color images, and for images whose color does not change, local color distribution is efficient, and can be reliable even when changes in scale or angle occur [45]. In the general case, the best results so far have been obtained with texture-based models [48]. A pyramidal analysis of texture (with the whole image considered as the texture and then spatial subblocks, etc.) has been shown to detect efficiently the local presence of objects at various scales. This is true not only for objects which we naturally think of as "texture-like" such as grass, water or clouds, but also for objects which are textured at much larger scale such as buildings or cars. When the problem is to locate images with a particular object (a particular face, a particular building) and not any object of a given type, principal component analysis methods of more general features of the images is the only efficient method. But, as we have seen in the previous section, their application to complex images of which the searched objects are only a part is still a largely open problem. It is worth to point out that object annotation by keywords is and will continue to be an efficient index for objects in images. But, more research is needed on algorithms that will assist human operators in generating these keywords.

2.6. *Summary*

In conclusion to this review of image similarity techniques, several main problems remain to be addressed for these techniques to be easily applicable to the full-range of access problems to large image databases:

- Study of the distribution of measures for various feature spaces on large real-world sets of images. In particular, how well is the perceived similarity order preserved by the measure when the number of images grows?
- Study of ranking visual items that correspond to human perception.
- Definition of methods for the segmentation of images in homogeneous regions for various feature spaces, and definition of models of this spatial organization which could be robustly combined with the similarity of the local features.
- Detection of salient features to a type of images or objects, so that to free user from specifying a particular set of features in query process.
- Combination of multiple visual features in image query and search.
- Developing efficient indexing schemes based on image similarity features for managing large databases. It has been shown that traditional database indexing techniques like R-trees, etc., fail in the context of content based image search, and currently there is no technique that allows retrieval of similar objects in multi-dimensional space. Ideas from statistical clustering, multi-dimensional indexing, and dimensionality reduction may be useful in this area.

Apart from these issues in extraction of low level visual features and establishment of related search/matching functions, extraction of higher (semantic) level image attributes (such as recognition of objects, human faces and actions) and related search/matching functions are definitely a more challenging task. Only when the features extracted at both these levels are combined, can content-based image indexes be built.

In addition, formalization of the whole paradigm of content based image retrieval to bring it to a sufficient level of consistency and integrity is essential to the success of the field. Without this formalism it will be hard to develop sufficiently reliable and mission critical applications that are easy to program and evaluate. Some early applications may be implemented without such a rigorous formalism, but the progress in the field will require full understanding of the basic requirements in content-based retrieval.

3. Video parsing and representation

To interact with video data is hard using conventional VCR-like video manipulation tools. The problem is that, from the point of view of content, the resources managed by the conventional systems are unstructured, apart from time code. Thus, the effort has been on introducing for structural and content analysis (parsing) of video, so that video can be indexed and accessed on the basis of structural properties and content.

Video parsing encompasses two tasks: temporal segmentation of a video program into elemental units, and content extraction from those units, based on both video and audio semantic primitives [71]. Many effective algorithms are now available for temporal segmentation. But, fully automated content extraction is a much more difficult task, requiring both signal analysis and knowledge representation techniques; so human assistance is still needed. On the other hand, the most fruitful research approach may be to concentrate on facilitating tools, using low-level visual features and content information from audio and close caption data. Such tools are clearly feasible and research in this direction should ultimately lead to an intelligent video retrieval and browsing systems [75].

In this section, we review a variety of video parsing and analysis approaches, including temporal segmentation of video, camera motion analysis, video soundtrack analysis, video abstraction approaches and shot similarity for shot comparison and clustering.

3.1. Temporal segmentation of video sequences

A video shot can be defined as consecutive images which appear to have been continuously filmed. A collection of one or more adjoining shots that focus on an object or objects of interest may comprise a scene. Shots are obviously a fundamental unit of manipulation (production, indexing, representation) of video. Therefore, detecting shot boundaries has been one of the first issues address by many researchers in video content analysis, and many algorithms and schemes have been developed and published.

There are a number of different types of transitions or boundaries between shots. The simplest transition is a cut, an abrupt shot change which occurs between two consecutive frames. More sophisticated transitions include fades, dissolves and wipes, etc. A fade is a slow change in brightness of images usually resulting in or starting with a solid black frame. A dissolve occurs when the images of the first shot get dimmer and the images of the second

shot get brighter, with frames within the transition showing one image superimposed on the other. A wipe occurs when pixels from the second shot replace those of the first shot in a regular pattern such as in a line from the right edge of the frames.

Some form of shot change detection was incorporated in the hardware of some telecine devices as early as the 1970's. Software researchers tried to address this problem from the beginning of the 1990's. The first researchers [44, 63] worked only on cut detection, which was natural for them since they were working in the frame of editing systems, and shot change detection was applied to unedited rushes. Later work [7, 56, 69, 70] has addressed shot change detection including the case of progressive transitions. The scientific problem of shot change detection consists in separating various factors of image change:

- Motion, including both camera work-induced apparent motion and object motion;
- Luminosity changes and noise;
- Shot change, abrupt or progressive.

Two main types of methods have been developed for this separation:

- Methods using the difference in the statistical signatures of these various causes of image change;
- Methods resting on explicit modelling of motion.

The second type of methods has been used by researchers who wanted to perform motion analysis for other reasons (camera work and object motion analysis), who worked from motion vector-based video coding (MPEG for instance) or who preferred these methods because of their background [54, 69]. The first type of method has been favored by researchers for which low computation time was essential. The initial methods used color histogram distance computation between successive images [63], eventually made more robust to noise and object motion by dividing the image in blocks [44] and computing distances in each block. Robustness of the detection of shot change in presence of important object motion calls for either a motion detector [55, 70] or for some type of temporal filtering of the statistical signature of image change [5, 21]. Zhang et al., proposed a method for progressive transition detection using a combination of motion and statistical analysis [70]. Purely statistical methods were refined to detect progressive effects by using properties of the distribution of pixel-to-pixel change in successive low-resolution low-passed images (Such as the DC images extracted from JPEG-coded documents). A multi-pass approach has also been proposed to improve both detection accuracy and speed [70].

Recent work has also focussed on performing shot analysis in the compressed domain. Statistical methods are easy to apply to JPEG-coded videos [14]. Some researchers [52] have written specialized DC decoders to attain better performance in extracting DC images from JPEG-coded video which does not use progressive mode. Shot analysis of MPEG-coded video can proceed along two lines:

- Direct analysis using a combination of statistical analysis between **I**-frames and **P**-frames, and motion analysis in **B**-frames [42, 54, 71].
- Efficient extraction of DC images by approximated decoding of the MPEG-flow [66].

Experiments have shown that using only I frames in detecting shot boundaries usually results in higher rate of false positives. An optimal solution will be to use both I and B frames with a combination of statistical analysis of DC images from I frames and motion vectors associated with B frames, as proposed by Zhang et al. [73]. Processing time could be further reduced while retaining high detection accuracy by using the difference between consecutive I frames as the first filter for potential shot boundaries and applying motion analysis to confirm and refine the detection [75].

It may be argued that shot analysis should be done before MPEG-encoding, since it makes possible a much better encoding. The fact that current hardware encoders do not apply it explains why researchers have nonetheless to work on shot change detection from MPEG video. It is hoped that the MPEG-4 standard will provide features that indicate shot boundaries to facilitate video indexing.

In summary, shot change detection can be reliably[2] achieved by software only methods at frame rate using today's workstations, though detecting progressive shot changes is still less reliable than detecting abrupt cut. Comparison studies of various shot change detection algorithms and schemes have been performed by Dailianas et al. [22] and Boreczky and Rowe [16]. The latter study has shown that for methods which work at the level of measures of a parameter evaluated on successive images or on differences of successive images—without any temporal filtering—histogram based algorithms outperform other algorithms.

3.2. Camera work and object motion analysis

Camera work analysis in video is very useful for indexing and retrieval purposes, because it makes possible to segment long sequence shots into shorter homogeneous units defined by homogeneous camera work and can help in choosing good representative images or keyframes for a video shot. Also, some temporal filtering mechanism is necessary to eliminate camera motion noise when it is present in detecting shot changes [70]. The scientific problem of camera work analysis resides in the discrimination between camera work-induced apparent motion and object motion-induced apparent motion, followed by analysis of the camera work-induced motion in order to identify camera work. These are classical problems in computer vision, but they are made specific by the open environment (no calibrated data or knowledge about the scene contents) and by the need for very efficient computation.

Camera work analysis was first addressed through motion vector field analysis [12, 62], by matching motion vector fields to prototype models for various camera work. Very good results can be obtained by this type of analysis provided that some background/figure discrimination is used (for instance by estimating motion vectors through hierarchical block matching). Unfortunately these methods are computationally expensive, and the results obtained by direct global estimation from MPEG-type motion vectors are not very good. A simple, yet effective approach to camera work analysis has been proposed to distinguish the gradual transition sequences and classify camera pan and zoom operations [70]. The search for methods which are more computationally efficient has also lead to methods based on discrete tomography. Tonomura and Akutsu proposed to use X-ray images obtained by computing the average of each line and each column in successive images, first for

representation purposes [61] and then for camera work identification [13]. A simplified approach based on the same ideas has been used by Joly and Kim [37]. The distribution of the angles of edges in the X-ray images can be matched to camera work models, and camera noise filtering, camera motion classification and temporal segmentation can be obtained directly. X-ray images can be efficiently extracted from JPEG-compressed domains by accessing only DC, $AC_{0,1}$ and $AC_{1,0}$ coefficients [51].

Some difficult problems remain to be addressed in camera work analysis. All methods fail when very large object motion cannot be discriminated from background motion, even in cases in which the semantics processing of a human viewer succeeds in doing so. Discrimination between pan and lateral travelling and between zoom and booming can be achieved only through parallax analysis.

3.3. Framing and focus

The framing of a video shot (from wide shot to close shot) and its focus (narrow focus on object, total focus on the whole scene, etc.) are important parameters of video indexing as has been pointed out by Hampapur et al. [30]. Few researchers have tried to automatically recognize framing. Part of the difficulty comes from the fact that some framing types, such as a knee shot or a head-and-shoulder shot are defined in reference to the human characters present in the scene. Thus, framing analysis calls for some type of semantic analysis of images, or for finding statistical features which are strongly correlated with a given framing type. It seems that framing analysis is a reachable objective. Focus analysis can be useful not only for indexing but also by for the extraction of specially important objects in images with small depth of field. Existing defocus estimators such as those used in recovery of depth from focus [17, 22] should be applicable through some adaptation to the uncalibrated single image environment of video.

3.4. Video soundtrack analysis

Though there is general agreement on the fact that sound is an essential component of video, and that image/sound relationships are critical to the perception and understanding of video contents, video soundtrack analysis is still in its infancy. So far, it has been addressed from 3 different points of view:

- Speech, music and Foley sound[3] detection, segmentation, and representation;
- Locutor identification and retrieval in speech soundtracks;
- Word spotting and labelling for speech soundtracks when a textual transcription is available (close caption or speech recognition).

Video soundtracks are often very complex, resulting from the mixing of many sound sources. From the perceptive point of view, the discrimination between speech, music and natural noises (often synthesized as artifacts in video) defines the essential structure. It turns out that detection of speech, music and their combinations in presence of various types of noise is not such an easy problem, even discarding limited cases such as grumbling, purely

percussive music or recitative. Hawley [32] has designed an harmonicity detector which can be used for music detection, and after spectral substraction for speech detection when the music component has been deleted. A method for detection of isolated speech and music only, similar in its principle to Hawley's has been proposed by Wyse and Smoliar [65]. Some researchers investigate methods based on the micro-segmentation of sound signals [39] and evaluation of the degree of stationary in each micro-segment. The idea is to be able to build direct detectors of speech and music which function independently of the presence of the other component and are robust to noise.

Locutor labelling of speech segments seems to be possible through the adaptation of classical locutor identification techniques in order to train them on uncalibrated examples. It should be noted that recognition rates which would be unacceptable for identification purposes can still be very useful if the aim is to find "the next time this locutor is again present in the soundtrack".

Word spotting and speech soundtrack labelling are associated techniques (soundtrack labelling from a speech transcription reduces to a repeated word spotting problem). Speech recognition from video soundtrack is still difficult because of the presence of other sound components. The development of television programs with (unsynchronized) close captions makes soundtrack labelling particularly interesting. The Informedia project at Carnegie Mellon University is using both speech recognition techniques to convert speech into text and close caption text systematically for analyzing and abstracting news and documentary video programs [64]. Their work has shown that combining of speech, text and image analysis can provide much more information, thus, achieve higher performance in video content analysis and abstraction than any one media alone.

3.5. Video scene analysis

There can be from 500 to 1000 shots per hour in a typical moving image program. Thus, the production of a synoptical view of the video contents usable for browsing or for quick relevance assessment calls for the recognition of meaningful time segments of longer duration than a shot, or for abstracting a shot-by-shot segmentation by selecting specially relevant ones. In media production, the level immediately higher than the shot is the sequence or scene, a series of consecutive shots constituting a unit from the narrative point of view, either because they are shoot in the same location, or because they share some thematic visual contents. There are many different forms of sequences, from a field-counterfield sequence in motion picture (with many repetitions of 2 shot types interlaced with a few other shot types) to a sequence of 2 shots of outdoors reporting inserted in a TV news programs. The process of detecting video scene is analogous to paragraphing in text document analysis and requires higher level content analysis.

Two different kind of approaches have been proposed for the automatic recognition of sequences. Zhang et al. have used models of specific types of programs such as TV news [74]. They recognize (by simple image processing) specific shot types such as shots with an anchor person and an insert. They then use the model to analyze the succession of shot types and produce a segmentation in sequences. Such model or knowledge based approaches can also applied to, for instance, sport video parsing [24]. However,

when we extend the application domain, we are facing the same difficulties as in computer vision.

Aigrain et al. have used rules formalizing medium perception in order to detect local (in time) clues of macroscopical change [11]. These rules refer to transition effects, shot repetition, shot setting similarity, apparition of music in the soundtrack, editing rhythm and camera work. After detection of the local clues, an analysis of their temporal organization is done in order to produce the segmentation in sequences and to choose 1 or 2 representative shots for each sequence.

In summary, video scene analysis requires higher level content analysis and one cannot expect that it can be fully automatic based on visual content analysis using current image processing and computer vision techniques. Fusion of information extraction from video, audio and close caption or transcript text analysis may be the only solution and a successful example is the Informedia Project [64].

3.6. Video abstraction and representation

Considering the large amount of data necessary to code digital video, there is a high cost to video networking, both economically and in time. It is, thus, critical to offer means for quick relevance assessment of video documents. Also, while we tend to think of indexing supporting retrieval, browsing is equally significant for video source material. The task of browsing is actually very intimately related to retrieval, in formulating queries and examining retrieval results. A truly content-based approach to video browsing calls for an abstracted representation of video, even if a good index may be available (which is not the case now). How can we spend only 6 minutes to view an hour of video and still have a fairly correct perception of what its contents is like? Or how can we map an entire segment to some small number of representative images? This is the video abstracting problem. Obviously, we will not ask for the perception of the abstract to provide exhaustive information, but we would like salient features, typical style and all major subjects to be included.

3.6.1. Video icon construction. The construction of a statical icon representing a video shot is one of the basic brick of video representation in a abstracted manner and is very useful in video browsing. It has attracted much work, with two major approaches:

- Construction of a visual icon based on a frame extracted from the shot, eventually supplemented with pseudo-depth for the representation of the duration of the shot, and arrows and signs for the representation of object and camera motion;
- Synthesis of an image representing the global visual contents of the shot.

The first approach has been favored when the emphasis is on building a global structured view of a video document, fitted for quick visual browsing, and when it is not deemed possible for legal or cultural reasons to modify images from the document. This approach has been used for instance in the IMPACT system [63]. Some researchers have used icon spacing or image size instead of pseudo-depth for representation of the duration of the shot[4],

but this does not seem compatible with efficient screen space use. One can find examples of use of arrows and signs for representation of object and camera motion in [63].

Two interesting problems when one is using extracted frames for the representation of a video shot are image schematizing and smart shrinking. When media professionals or analysts draw (with paper and pencil) a representation of the visual contents of a shot, they outline a simplified view of the salient image features. This is a visually efficient representation and one might wonder if the same type of outlining could not be automatically produced. Unfortunately it is not an easy problem: intelligent outlining goes far beyond edge extraction. The second problem is more critical: there are limits on the resolution at which the visual contents of a video image are readable: most representations in visual interfaces use resolutions in the range 80×60 to 160×120, the upper figures being the most frequent. Occasionally, when one wants to display many images, resolutions down to 40×30 can be used, but many images become unreadable even from the global structure point of view. This creates limitations for the scope of video representations: only a limited number of shots can be simultaneously displayed, and we do not have the same type of zooming out capabilities that one finds for instance for sound visualization. It seems that it would be possible to use smarter "shrinking" strategies that the undersampling (with eventual filtering) presently used when decreasing resolution.

Teodosio and Bender [60], and Tonomura et al. [61] have proposed methods for the automatic construction of an image representing all the visible contents of a shot. Using camera work analysis and the geometrical transformations associated to each camera motion, the successive images are mapped into a common frame, and the synthetic image is progressively built. This image is not generally rectangular. Recently, Irani et al. working in the frame of image compression, have perfected this type of method on two points [35]:

- they use a more complete projective model, including parallax;
- they have shown that it is possible to compute what they call dynamic mosaic images with privilege being given to the moving parts of the image (action) instead of background oriented images.

The resulting images have been termed *salient stills* [60], *videospaceicons* [58] and *mosaic images* [35].

Some researchers have tried to combine key frame oriented shot representation with motion traces of the type produced by salient stills, but it still unclear whether a visually convincing result can be obtained by this combination.

3.6.2. Key-frame extraction. Key-frames are still images which best represent the content of the video sequence in an abstracted manner, and are extracted from original video data. Thus, key-frame based video representation views video abstraction as a problem of mapping an entire segment (both static and motion content) to some small number of representative images.

Key frames are frequently used to supplement the text of a video log [46], but there has been little work in identifying them automatically. The challenge is that the extraction of key-frames needs to be automatic and content based so that they maintain the important

content of the video while remove all redundancy. In theory semantic primitives of video, such as interesting objects, actions and events should be used; however, such general semantic analysis is not currently feasible. An approach to key-frame extraction based on low-level video features has been proposed by Zhang et al. [73, 75]. In this approach, key-frames are extracted at shot level using color and brightness features of frames and their variation, and dominant motion components resulting from camera operations and large moving objects. The algorithm has also been extended into MPEG compressed domain, in which DCT coefficients of I frames and motion vectors from B and P frames are used. In addition users can adjust several parameters to control the density of key frames in each shot. A user study [73] has shown that the algorithm performed satisfactory in terms of both high accuracy and low redundancy. It is also shown that the algorithm outperforms human operator in term of consistency.

Apart from browsing, key-frames can also be used in representing video in retrieval: video index may be constructed based on visual features of key-frames, and queries may be directed at key-frames using query by image content techniques [23, 75].

3.6.3. Video skimming. Video skimming is the scheme to answer the request of abstracting an hour of video, for instance, into 5 minutes highlights with a fair perception of the video contents. This is a relatively new research area and requires high level content analysis.

Joly and Kim [37] has used editing analysis to select series of shots which will be included in the abstract using rules. But, precise evaluation of the quality of the resulting abstracts has not been conducted yet. The more successful approach is to utilize information from text analysis of video soundtrack. Researchers [e.g., 58] working on documents with textual transcriptions have suggested to produce video abstracts by first abstracting the text by classical text skimming techniques and then looking for the corresponding parts in the video.

A more sophisticated approach has been proposed by the Informedia project team in which text and visual content information are fused to identify video sequences that highlight the important contents of video [31]. More specifically, low-level and mid-level visual features, including shot boundary, human face, camera and object motion and subtitles of video shots are integrated with keywords, spotted from text obtained from close caption and speech recognition, following the procedures as below:

- Keyword selection using the well-known TF-IDF[5] technique to skim audio;
- Sequence characterization by low-level and mid-level visual features;
- Selecting number of keywords according to required skimming factor;
- Prioritizing image sequences located closely to each selected keyword:
  - — Frames with faces or text;
  - — Static frames following camera motion;
  - — Frames with camera motion and human faces or text;
  - — Frame at the beginning of the scene;
- Compose a skimmed highlight sequence with selected frames.

Experiments of this skimming approach has shown impressive results on limited types of documentary video which have very explicit speech or text (close caption) contents, such

as education video, news or parliament debates. However, satisfying results may not be achievable using such a text (keyword) driven approach to other videos with a soundtrack containing more than just speech, or stock footage without soundtrack.

3.7. Shot similarity and content-based retrieval of clips

Defining video shot or sequence similarity is a key issue in building content-based indices, retrieving sequences of similar visual content and clustering similar shots to construct synoptical views or visual summary. Most researchers have computed shot similarity from similarity of the images chosen to represent each shot, which obviously leaves apart all the dynamical dimension of the shots if only a single representative image is used. This can be improved by using keyframes extracted based on both static and temporal features of video [73, 75]. Based on keyframes, Yeung and Liu have recently formalized a slightly extended definition of shot similarity for shot clustering [68].

Though a set of keyframes will represent temporal content of shots to some extent, more precise measure of shot similarity should incorporate motion features, apart from features of static images. In response to such requirement, a set of statistic measures of motion features of shots has been proposed and applied in news anchor detection and shot clustering for browsing and annotation [74, 76]. However, defining more quantitative measures of shots similarity that capture the motion nature of video still remains a challenging research topic.

Techniques for the content-based retrieval of video will need much progress before they can be applied to the full search of a complete video archive. With existing techniques for feature extraction, similarity computation and explicit or implicit user query formulation, content-based retrieval of video is restricted to two already very important situations:

- finding a shot in a document;
- retrieval of single-shot clips in an archive of such clips.

The techniques described in Sections 2 and 3.7 are all applicable in order to find, for instance, the next shot which is similar from some point of view to some "present" shot. Though these techniques have been applied in some prototypes [44, 48, 75], no precise comparison between this type of navigation by visual content retrieval and direct user browsing of shot representations has been made. In-document navigation by locutor or word retrieval in soundtrack would clearly be very useful: we have already mentioned that the Informedia project plans to make it possible. For the second type of application, in the case of video stock footage, Gordon and Domeshek [25] have proposed indexing schemes which could be combined with automatic feature extraction and similarity measures in order to build a retrieval system.

3.8. Visual presentation and annotation techniques for video

Visual presentation, interactive perception and annotation of video contents raise many difficult issues. A complete treatment of the related techniques is not possible in the limits of this review. One can refer to [3] for a discussion of the adaptations in the information

infrastructure which are necessary to make possible some key functionalities of interactive perception of audiovisual contents. Let us mention some interesting approaches to visual information display and user annotation techniques.

Video contents are multi-dimensional, with image, sound, and many semantic dimensions. Video representation is screen space consuming, and calls for representations at various time scales. The temporal dimension of video is particularly difficult to accommodate:

- because of the plurality of time domains (sequential time in the document, complex narrative time space),
- because the image component is better represented in a non-linear manner, i.e., using the shots or sequences as equally spaced units, whose duration is figured by pseudo-3D-depth, and the sound components are more easy to represent with a linear time-scale.

Solutions for combining non-linear time scales for image with pseudo-linear time scales for sound have been proposed by Aigrain and Joly in [6, 10]. Butler and Parkes have proposed what they call *time-space filmic diagrams* for the on-screen representation of narrative time [18].

User annotation is of course a key aspect of interactive access to video as was pointed out since the Experimental Video Annotator of Mackay and Davenport [41]. Due to the improperness of input physical devices, on-screen annotation is still very far from seriously competing with the naturalness and expressiveness of pencil annotation of a printed representation. An interesting approach to finding more adequate on-screen annotation mechanisms is icon palette annotation in which the user chooses pre-defined icons in icon palettes. Haase has proposed a model for icon palette annotation in [29].

3.9. Video indexing and visual cataloging

Traditional video indexing must be rethought in the new situation opened by the availability of content analysis techniques. In video archives and databases, 3 levels of descriptive coverage are traditionally considered:

- Minimal cataloguing with indications concerning the origin and carrier of the document, its title, duration, and more generally all information which is directly accessible on labels or package of the carrier, without any research or viewing of the document;
- More complete cataloguing with indications of various authors or contributors, eventually calling for some research or viewing of the credits in the document itself;
- Real contents indexing, for instance shot-by-shot indexing and description.

Human contents indexing is very time consuming and thus costly. It has been estimated [28] that the ratio of indexing time to document time is of the order of 10 : 1. It is, thus, used only for limited samples of particularly valuable documents, such as news archives of the broadcasting industry. Video editing analysis and representation techniques make possible a completely different organization of the video indexing process. It is now possible to build automatic storyboarding software which will produce shot-by-shot storyboard in real-time

during transfer or digitizing of a document (see [10, 11] for presentations of such a software). It is thus possible for human indexers to concentrate on the truly semantic information which will not be readable on the automatically produced storyboards, and hard to retrieve for content-based retrieval systems. These storyboards can be either printed or on-line accessed. In both cases, access to a full shot-by-shot storyboard maybe too long, and synoptic representations constructed from large-scale analysis (see Section 3.5) or browsers (see below) can be used to provide users with a quicker glimpse at the document visual contents.

3.10. Computer-assisted video annotation and transcription

Users who access video documents frequently need to produce speech transcriptions of the dialogues or shot-by-shot annotation, for their own use or for communication purposes. When no close caption is available, full automatic speech transcription in a multi-locutor noisy environment of continuous speech such as video is out of reach of speech recognition research. But video analysis can help in building softwares which make dialogue transcription and shot-by-shot annotation much easier. The simple fact of being able to playback a shot in a loop while typing, with easy navigation to the next of previous shots, can save time by a factor of 2 or 3 compared to transcription using a tape-recorder like software with no shot segmentation information. Pages of transcript or notes illustrated by images extracted from shots can be automatically printed or accessed on-screen [10]. When close caption is available, such pages can be automatically produced, by combining shot analysis with speech labelling. Shahraray and Gibbon [56] have developed a software which produces HTML-coded video transcripts for close-captioned news.

3.11. Summary

The majority of the techniques reviewed in this section address problems in recovering low level structure of video sequences, though these techniques are basic and very useful in facilitating intelligent video browsing and search. Successful efforts in extracting semantic content are still very limited and there is a long way to go before we can achieve our goal of content-based video retrieval. On the other hand, what levels of understanding of content are most important to recover from video? Should recovering high-level intentional descriptions (what the characters are thinking or trying to do) as is needed for discourse and semantic analysis be our goal? It was observed during NSF/ARPA Workshop on Visual Information Management Systems that there seemed to be good agreement that a focus on human action is the most important topic to address over the next decade, since over 95% of all video the primary camera subject is a human or group of humans (who may be discussing some more abstract subject) [36].

On the other hand, it should be pointed out that, one of the most important points is that video content analysis, retrieval and management should *not* be thought of a fully automatic process. We should focus in developing video analysis tools to facilitate human analysts, editors and end users to manage video more intelligently and efficiently. While solving the general problem of video analysis and retrieval is difficult, we believe applications to particular needs of specific users and specific domains may be very successful.

4. User interfaces and browsing tools

4.1. User interface for still image retrieval

The variety of similarity types creates user interface problems for content-based retrieval of still images. It may become almost as difficult to control the retrieval process by choosing and combining similarity types as it is to formulate a textual query. Combining navigation techniques with retrieval by similarity have been suggested as a potential solution for this problem [2]. It is much easier to choose between examples of images which are similar from some point of view than to explicitly state the nature of this similarity. Images which are *dissimilar* from some point of view can also be presented: Romer [36] has pointed out that image search is also a matter of looking for an image which is "not like this one from this point of view". When similarity is computed from image descriptions and not from the image content itself, retrieval by navigation along similarity/dissimilarity links has proved to be faced with serious limits at least when the user is looking for a particular image in a large set of images [7, 8]. But looking back at similarity navigation now that feature extraction and similarity measures are more mature may be worthwhile.

Therefore, advanced user interfaces that allow relevance feedback by example, learning, and very fast browsers are essential to the success of content-based image retrieval and user and task learning or adaptation is, therefore, a key issue in image retrieval. A practical system should have the flexibility to adaptably select feature sets, similarity measures, and search methods to suit particular images. A very promising example of such methodology is the "society of models" approach proposed by Picard and Minka [50].

4.2. Video browsing tools

The volume of video data requires techniques to present information landscape or structure to give idea of what is out there. Interactive browsing of full video contents is probably the essential feature of new forms of interactive access to video. Three basic models have been proposed for video presentation and browsing:

- time-line and strata browsers;
- hierarchical browsers;
- graph browsers.

Time-line based browsers have been favored by researchers interested in video production and editing systems, for which time-line interfaces are classical. Some browsers rest on a single shot-based image component line [7, 63], but the multidimensional dimension character of video, calling for multi-line representation of the contents has been stressed by researchers working in the frame of the Muse toolkit [34, 41]. This has been systematized in the *strata* models proposed by Aguierre-Smith and Davenport [1]. Recent work on time-line browsers has focussed on combining various automatically produced lines representing a particular angle of view on the document (shots, transition effects, speech, music, Foley sounds), with user lines used for annotation. Examples of this type of browsers can be

found in [11, 63]. A limitation of time-line browsers is that, since it is difficult to zoom out while keeping a good image visibility, the time-scope of what is actually displayed at a given moment on screen is relatively limited. This has lead researchers to introduce multi-scale or hierarchical aspects in the time-line browsers.

A first attempt at building hierarchical browsers—called the Video Magnifier [43] simply used successive horizontal lines each of which offered greater time detail and narrower time scope by selecting images from the document. There was no underlying structure. Recent hierarchical browsers [72] are based on levels corresponding to the actual structure of the video contents: for instance shots, sequences, scenes, etc. Shots are automatically recognized, while higher level objects are manually defined by a human indexer. The video contents are accessed as a tree each node corresponding to a segment at a given level and being represented by an image (with timing data and eventual name). Such hierarchical browsers has been further enhanced to provide similarity based browsing. That is, visual similarity of shots based on keyframes, instead of temporal relations only, are used to cluster shots automatically into scenes to provide a more content based overview at scene levels [76].

An alternative approach to hierarchical browsers has been proposed by Yeung et al. [67]. Using the clustering of visually similar shots, they construct a directed graph whose nodes are clusters of shots. Cluster A is linked to cluster B if one of the shots in A is immediately followed by a shot in B. The resulting graph is displayed for browsing, each node being represented by a key-frame extracted from one of the shots in the node. The graph can be edited for simplification by a human indexer. The drawbacks of this approach lie in the difficulty of the graph layout problem, resulting in poor screen space use, and in the fact that the linear structure of the document is no longer perceptible.

4.3. Playback control and computer-assisted perception

There has been surprisingly little work on making possible new modes of playback control in interaction with video. Most systems use content analysis and representation mostly for the retrieval or visual browsing of segments, each of which is then played using simple "tape recorder-like" control panels. Researchers working on viewing systems for scholars and media specialists have proposed models and realized systems [11] in which:

- playback is synchronized with the display of cursors on the contents representation,
- temporal programming of playback of segments is possible using visual programming from examples and direct manipulation.

Temporal programming of playback is essential for the analytic and comparative perception of video. It can be combined with spatial layout, by playing back for instance two extracts of different or the same document side by side.

5. Conclusions

In this paper, we have reviewed to a great extent many research efforts and techniques addressing visual content representation and content based retrieval of visual data. It is

clear from this survey that the number of research issues and their scope are rather large and expanding rapidly with advances in computing and communication. As a result, visual data is becoming the center of multimedia computing, and more and more researchers from different fields are attracted and start to explore these issues. On the other hand, how to extract and manage semantics information of visual data remains a major bottleneck, which call for not only more research efforts, but most critically right research approaches.

An application-oriented approach is critical to the success of visual data representation and retrieval researches and will prevent it from being too theoretical. By working on strongly focused applications, the research issues reviewed in this paper can be addressed in the context of well defined applications and will facilitate the applications, while achieving general solutions remains long term research topics.

When we strive for visual content analysis and representation, it should be pointed out again that integration of different information sources, such as speech, sound and text is as important as visual data itself in understanding and indexing visual data. Keywords and conceptual retrieval techniques are and will always be an important part of visual information systems. What we should focus on should include techniques to help annotation of image and video, and other meta-data.

Notes

1. This work was performed while this author was with Institute of Systems Science, Singapore.
2. With an error rate comparable to human indexing.
3. Foley sounds, named after Jack Foley [1891–1967], sound engineer for Universal Studios, are artifically produced sounds imitating natural noises. There are databases of Foley sounds and similar natural or instrumental sounds and indexing and retrieval in these databases has lead to interesting research [15].
4. One can find examples of use of image spacing or size for representation of duration in drawn storyboards; see [47].
5. TF-IDF technique: Term Frequency Inverse Document Frequency.

References

1. T.G Aguierre-Smith and G. Davenport, "The stratification system: A design environment for random access video," Proc. 3rd Int. Workshop on Network and Operating System Support for Digital Audio and Video, La Jolla, CA, USA, Nov. 1992, pp. 250–261.
2. P. Aigrain, "Organizing image banks for visual access: Model and techniques," OPTICA'87 Conf. Proc., Amsterdam, Learned Information, April 1987, pp. 257–270.
3. P Aigrain, "Image and sound digital libraries need more than storage and networked access," Proc. International Symposium on Digital Libraries, ULIS, Tsukuba, Japan, Aug. 1995, pp. 112–118.
4. P. Aigrain, "Software research for video libraries and archives," IFLA Journal, special issue on the UNESCO Memory of the World Project, Vol. 21, No. 3, pp. 198–202, 1995.
5. P. Aigrain and P. Joly, "The automatic real-time analysis of film editing and transition effects and its applications," Computers & Graphics, Vol. 18, No. 1, pp. 93–103, Jan.–Feb 1994.
6. P. Aigrain and P. Joly, "Discrete visual manipulation user interfaces for video," Proc. RIAO'94 Conference, New-York, Oct. 1994, Vol. 2, pp. 12–17.
7. P. Aigrain and V. Longueville, "A connection graph for user navigation in a large image bank," Proc. RIAO'91, Barcelona, Spain, April 1991, Vol. 1, pp. 67–84.
8. P. Aigrain and V. Longueville, "Evaluation of navigational links between images," Information Processing and Management, Vol. 28, No. 4, pp. 517–528, 1992.

9. P. Aigrain, P. Joly, and V. Longueville, "Medium-knowledge-based macro-segmentation of video into sequences," Working Notes of IJCAI Workshop on Intelligent Multimedia Information Retrieval, Montreal, Aug. 1995, pp. 5–14.
10. P. Aigrain, P. Joly, H.-K. Kim, and P. Lepain, Software Tools for Moving Image Archives: Access, Indexing and User Interfaces, G. Boston (Ed.), Proc. Joint Technical Sympoisum on Technology and Our Audiovisual Heritage, FIAF/FIAT/IASA/IFLA/ICA, London, Jan. 1995.
11. P. Aigrain, P. Joly, P. Lepain, and V. Longueville, "Representation-based user interfaces for the audiovisual library of year 2000," Proc. IS&T/SPIE'95 Multimedia Computing and Networking, San Jose, Feb. 1995, pp. 35–45.
12. A. Akutsu, Y. Tonomura, H. Hashimoto, and Y. Ohba, "Video indexing using motion vectors," Proc. Visual Communication and Image Processing, SPIE, Amsterdam, 1992, Vol. 1818, pp. 1522–1530.
13. A. Akutsu and Y. Tonomura, "Video tomography: An efficient method for camerawork extraction and motion analysis," Proc. A.C.M. Multimedia Conference, San Francisco, Oct. 1994.
14. F. Arman, A. Hsu, and M.Y. Chiu, "Feature management for large video databases," Proc. Storage and Retrieval for Image and Video Databases I, SPIE, Feb. 1993, Vol. 1908, pp. 2–12.
15. T. Blum, D. Keislar, J. Wheaton, and E. Wold, "Audio databases with content-based retrieval," Working Notes of IJCAI Workshop on Intelligent Multimedia Information Retrieval, Montreal, Aug. 1995, pp. 71–92.
16. J.S. Boreczky and L.A. Rowe, "Comparison of video shot boundary detection techniques," Proc. SPIE Conf. Storage and Retrieval for Video Databases IV, San Jose, CA, USA, Feb. 1995.
17. V.M. Bove, Jr., "Entropy-based depth from focus," Journal of the Optical Society of America A, Vol. 10, pp. 561–566, April 1993.
18. S. Butler and A.P. Parkes, "Filmic spacetime diagrams for video structure representation," to appear in Image Communication special issue on Image and Video Semantics: Processing, Analysis and Application, 1996.
19. N.-S. Chang and K.-S. Fu, "Query by pictorial example," IEEE Transactions on Software Engineering, Vol. 6, No. 6, pp. 519–524, Nov. 1980.
20. M. Cherfaoui and C. Bertin, "Two-stage strategy for indexing and presenting video," Proc. SPIE Conf. Storage and Retrieval for Video Databases III, San Jose, CA, USA, Feb. 1994, Vol. 2185.
21. A. Dailianas, R. Allen, and P. England, "Comparison of automatic video segmentation algorithms," Proceedings of SPIE Photonics West, Philadelphia, Oct. 1995.
22. J. Ens and P. Lawrence, "An investigation of methods determining depth from focus," IEEE Transactions on Pattern Matching and Machine Intelligence, Vol. 15, pp. 97–108, Feb. 1993.
23. M. Flickner et al., "Query by image and video content," IEEE Computer, pp 23–32, Sept. 1995.
24. Y. Gong, L.T. Sin, H.C. Chuan, H.J. Zhang, and M. Sakauchi, "Automatic parsing of TV soccer programs," Proc. Second IEEE International Conference on Multimedia Computing and Systems, Washington DC, May 15–18, 1995, pp. 167–174.
25. A.S. Gordon and E.A. Domeshek, "Conceptual indexing for video retrieval," Working Notes of IJCAI Workshop on Intelligent Multimedia Information Retrieval, Montreal, Aug. 1995, pp. 23–38.
26. V.N. Gudivada, "On spatial similarity measures for multimedia applications," Proc. SPIE Conf. Storage and Retrieval for Image and Video Databases III, San Jose, CA, USA, Feb. 1994, Vol. 2420, pp. 363–380.
27. V.N. Gudivada and V.V. Raghavan, "Design and evaluation of algorithms for image retrieval by spatial similarity," ACM Transactions on Information Systems, Vol 13, No. 2, pp. 115–144, April 1995
28. V. Guigueno, "L'identité de l'image: Expression et systémes documentaires," rapport d'option, Ecole Polytechnique, Palaiseau, France, Juillet, 1991.
29 K. Haase, "Framer: A persistent portable representation library," Proc. of ECAI'94, 1994.
30. A. Hampapur, R. Jain, and T.E. Weymouth, "Production model based digital video segmentation," Multimedia Tools and Applications, Vol. 1, No. 1, pp. 9–46, 1995.
31. A.G. Hauptmann and M. Smith, "Text, speech and vision for video segmentation: The Informedia project," Working Notes of IJCAI Workshop on Intelligent Multimedia Information Retrieval, Montreal, Aug. 1995, pp. 17–22.
32. M. Hawley, Structure Out of Sound, Ph.D. Dissertation, MIT Media Laboratory, Cambridge, Mass., USA, 1993.
33. K. Hirata and T. Kato, "Query by Visual Example: Content-Based Image Retrieval," Proc. E.D.B.T.'92 Conf. on Advances in Database Technology, in Pirotte, Delobel, and Gottlob (Eds.), Springer-Verlag, Lecture Notes in Computer Science, Vol. 580, pp. 56–71, 1994.

34. M.E. Hodges, R.M. Sassnett, and M.S. Ackerman, "A construction set for multimedia applications," IEEE Software, pp. 37–43, Jan. 1989.
35. M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu, "Mosaic based representations of video sequences and their applications," Image Communication special issue on Image and Video Semantics: Processing, Analysis and Application, 1996.
36. R. Jain, A. Pentland, and D. Petkovic (Eds.), Workshop Report: NSF-ARPA Workshop on Visual Information Management Systems, Cambridge, Mass., USA, June 1995.
37. P. Joly and H.-K. Kim, "Efficient automatic analysis of camera work and micro-segmentation of video using spatio-temporal images," Image Communication special issue on Image and Video Semantics: Processing, Analysis and Application, 1996.
38. T. Kato, "Database architecture for content-based image retrieval," Proc. of SPIE Conf. on Image Storage and Retrieval Systems, San Jose, Feb. 1992, Vol. 1662, pp. 112–123.
39. P. Lepain and R. André-Obrecht, "Micro-segmentation d'enregistrements musicaux," Actes des Journées d'Informatique Musicale, Vol. 95-13, pp. 81–90, 1995.
40. W.E. Mackay and G. Davenport, "Virtual video editing in interactive multimedia applications," Communications of the A.C.M., Vol. 32, No. 9, July 1989.
41. J. Meng, Y. Juan, and S.-F. Chang, "Scene change detection in an MPEG compressed video sequence," IS&T/SPIE'95 Digital Video Compression: Algorithm and Technologies, San Jose, Feb. 1995, Vol. 2419, pp. 14–25.
42. M. Mills, J. Cohen, and Y.Y. Wong, "A magnifier tool for video data," Proc. INTERCHI'92, ACM, May 1992, pp. 93–98.
43. A. Nagasaka and Y. Tanaka, "Automatic scene-change detection method for video works," E. Knuth and I.M. Wegener (Eds.), Proc. 40th National Con. Information Processing Society of Japan, 1990.
44. A. Nagasaka and Y. Tanaka, Automatic Video Indexing and Full-Search for Video Appearances, E. Knuth and I.M. Wegener (Eds.), Visual Database Systems, Elsevier Science Publishers: Amsterdam, Vol. II, pp. 113–127, 1992.
45. B.C. O'Connor, "Selecting key frames of moving image documents: A digital environment for analysis and navigation," Microcomputers for Information Management, Vol. 8, No. 2, pp. 119–133, 1991.
46. B. Peeters, J. Faton, and P. de Pierpont, Storyboard-Le Cinéma Dessiné, Editions Yellow Now, 1992.
47. A. Pentland, R.W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," Proc. Storage and Retrieval for Image and Video Databases II, San Jose, CA, USA, Feb. 1994, Vol. 2185.
48. R. Picard and Fang Liu, "A new World ordering for image similarity," Proc. Int. Conf. on Acoustic Signals and Signal Processing, Adelaide, Australia, March 1994, Vol. 5, p. 129.
49. R.W. Picard and T.O. Minka T., "Vision texture for annotation," Multimedia Systems, ACM-Springer, Vol. 3, No. 3, pp. 3–14, Feb. 1995.
50. F. Salazar, "Analyse automatique des mouvements de caméra dans un document vid'eo," IRIT, rapport de recherche, 95-33-R, Universit'e Paul Sabatier, Toulouse, France, Sept. 1995.
51. F. Salazar and F. Valéro, "Analyse automatique de documents vidéo," IRIT, rapport de recherche, 95-28-R, Université Paul Sabatier, Toulouse, France, Juin 1995.
52. S. Sclaroff and A. Pentland, "Modal matching for correspondence and recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17, No. 6, pp. 545–561, June 1995.
53. I.K. Sethi and N. Patel, "A statistic approach to scene change detection," Proc. SPIE Storage and Retrieval for Image and Video Databases III, San Jose, CA, USA, Feb. 1995, Vol. 2420, pp. 329–338.
54. B. Shahraray, "Scene change detection and content-based sampling of video sequences," IS&T/SPIE'95 Digital Video Compression: Algorithm and Technologies, San Jose, Feb. 1995, Vol. 2419, pp. 2–13, SPIE Proceedings.
55. B. Shahraray and D.C. Gibbon, "Automatic generation of pictorial transcripts of video programs," IS&T/SPIE'95 Digital Video Compression: Algorithm and Technologies, San Jose, Feb. 1995, Vol. 2417, pp. 512–519, SPIE Proceedings.
56. M. Stricker and M. Orengo, "Similarity of color images," Proc. Storage and Retrieval for Image and Video Databases III, San Jose, CA, USA, Feb. 1995, Vol. 2420, pp. 381–392, SPIE Conference Proceedings.
57. A. Takeshita, T. Inoue, and K. Tanaka, "Extracting text skim structures for multimedia browsing," in M. Maybury (Ed.), Working Notes of IJCAI Workshop on Intelligent Multimedia Information Retrieval, Montreal, Aug. 1995, pp. 46–58.

58. H. Tamura, S. Mori, and T. Yamawaki, "Texture features corresponding to visual perception," IEEE Trans. on Syst., Man, and Cybern.,Vol. 6, No. 4, pp. 460–473, 1979.
59. L. Teodosio and W. Bender, "Salient video stills: Content and context preserved," Proc. ACM Multimedia'93, Anaheim, CA, USA, Aug 1993.
60 Y. Tonomura, A. Akutsu, K. Otsuji, and T. Sadakata, "VideoMAP and VideoSpaceIcon: Tools for anatomizing video content," Proc. InterChi'93, ACM, 1993, pp. 131–136.
61. Y.T. Tse and R.L. Baker, "Global zoom/pan estimation and compensation for video compression," Proc. ICASSP'91, May 1991, Vol. 4.
62. H. Ueda, T. Miyatake, and S. Yoshisawa, "IMPACT: An interactive natural-motion-picture dedicated multimedia authoring system," Proc. CHI'91, ACM, 1991, pp. 343–350.
63. H.D. Wactlar, D. Christel, A. Hauptmann, T. Kanade, M. Mauldin, R. Reddy, M. Smith, and S. Stevens, "Technical challenges for the informedia digital video library," Proc. International Symposium on Digital Libraries, Tsukuba, Japan, Aug. 1995, pp. 10–16
64. L. Wyse and S.W. Smoliar, "Towards content-based audio indexing and retrieval," Proc. IJCAI Workshop on Computational Auditory Scene Analysis, D. Rosenthal and H.G. Okuno (Eds.), Montréal, Aug. 1995, pp. 149–152.
65. B.-L. Yeo and B. Liu, "On the extraction of DC sequence from MPEG compressed video," International Conference on Image Processing (ICIP'95), Washington, DC, USA, Oct. 1995, IEEE.
66. M.M. Yeung, B.-L. Yeo, W. Wolf, and B. Liu, "Video browsing using clustering and scene transitions on compressed sequences," IS&T/SPIE'95 Multimedia Computing and Networking, San Jose, Feb. 1995, Vol. 2417, pp. 399–413.
67. M.M. Yeung and B. Liu, "Efficient matching and clustering of video shots," International Conference on Image Processing (ICIP'95), Washington, DC, USA, Oct. 1995, IEEE
68. R. Zabih, K. Mai, and J. Miller, "A robust method for detecting cuts and dissolves in video sequences," Proc. ACM Multimedia'95, San Francisco, Nov. 1995.
69. H J. Zhang, A. Kankanhalli, and S.W. Smoliar, "Automatic partitioning of full-motion video," Multimedia Systems, ACM-Springer, Vol. 1, No. 1, pp. 10–28, 1993.
70. H.J. Zhang and S.W. Smoliar, "Developing power tools for video indexing and retrieval," Proc. SPIE'94 Storage and Retrieval for Video Databases, San Jose, CA, USA, Feb. 1994.
71. H.J. Zhang, S.W. Smoliar, and J.H. Wu, "Content-based video browsing tools," Proceedings of IS&T/SPIE'95 Multimedia Computing and Networking, San Jose, Feb. 1994, Vol 2417.
72. H.J Zhang, C.Y. Low, Y. Gong, and S W. Smoliar, "Video parsing using compressed data," Proc. SPIE'94 Image and Video Processing II, San Jose, CA, USA, Feb. 1994, pp. 142–149.
73. H.J. Zhang, S Y Tan, S.W. Smoliar, and Y. Gong, "Automatic parsing and indexing of news video," Multimedia Systems, Vol. 2, No. 6, pp. 256–265, 1995.
74. H.J. Zhang, C.Y. Low, S.W. Smoliar, and J.H. Wu, "Video parsing, retrieval and browsing: An integrated and content-based solution," Proc. ACM Multimedia'95, San Francisco, Nov. 5–9, 1995, pp. 15–24.
75. D. Zhong, H.J. Zhang, and S.-F. Chang, "Clustering methods for video browsing and annotation," Proc. Storage and Retrieval for Image and Video Databases IV, San Jose, CA, USA, Feb. 1995.

Philippe Aigrain graduated in computer science from University of Toulouse, and holds a "Doctorat" in theoretical computer science from University Paris 7 (1980). In 1982 he was a research fellow at University of

California, Berkeley. Since 1983, he has been conducting research on image and sound archives and interaction with audiovisual media He is presently head of the Media Analysis and Interaction research group in the Institut de Recherche en Informatique de Toulouse, France. His research interests are in content analysis of video and musical recordings and discrete manipulation interfaces for time-based media.

HongJiang Zhang obtained his Ph.D. from the Technical University of Denmark in 1991, B.Sc. in 1982 from Zhengzhou University, Zhengzhou, China, both in Electrical Engineering. From December, 1991, he has been with the Institute of Systems Science, National University of Singapore, led the work on video/image content analysis, representation indexing and retrieval. He joined the Broadband Information System Lab. of Hewlett-Packard Labs., Palo Alto, in October, 1995. His current research interests are in video/image content analysis and retrieval, interactive video, and image processing. He has published over 40 papers and book chapters in these areas and is a co-author of "Image and Video Processing in Multimedia Systems", a book published by Kluwer Academic Publishers. He serves on program committees of several international conferences on multimedia. He is also a member of the Editorial board of Kluwer's international journal "Multimedia Tools and Applications".

Dragutin Petkovic is the manager of the Advanced Algorithms, Architectures and Applications Department at IBM Almaden Research Center. Despite all the managerial responsibilities, he is still involved in marketing, applications, and testing much of the software his group creates. His research interests include image analysis applied to industrial, commercial, and biomedical problems, content-based search, large-image and multimedia databases, and advanced user interfaces.