Movie Maps

Heimo Müller, Ed Tan

Vrije Universiteit Amsterdam, Fac. of Arts, Word & Image studies - de Boelelaan 1105, 1081 HV Amsterdam
e-mail: mue@smr.nl

Abstract: This paper presents methods for moving image sequence visualization and browsing based on algorithms from computer vision, information visualization and on a hierarchical model for content semantics. We introduce a new method, OM-Images, for the visualization of temporal changes in a moving image sequence. Together with interactive browsing techniques the visualization methods can be used for the exploration of a movie at different levels of abstraction. The proposed levels of abstraction are the physical, image, object or discourse level. The visualization is used to generate 1) static descriptions, which printed on paper yield a "movie book" and 2) interactive documents, e.g. web pages or special movie browsers. Finally, we give examples of a movie book of a feature length film.

1 Introduction

Starting from the question "How can we describe and visualize a movie?" we encountered only text oriented and simple visual techniques, e.g. keyframes, for the description and visualization of the rich information space embedded in a movie. In most cases descriptions of a movie are only beneficial for a user, who already has seen the movie, and neither suitable for first time viewing nor for visualization of the structure and semantics of a movie in a more than superficial way.

When we investigate a movie, we usually play the movie and optional capture descriptions for permanent access. In most cases the interaction is restricted to simple play controls (stop, pause, play, forward, rewind, etc.) and jump functionality if a non-linear media is used. This process can not be sped up very much (the only way is to play fast through the movie without sound information) nor can the linear ordering of the movie elements (shot, scenes) be altered. In order to extend the interactivity of movie browsing we use pre-computed descriptions and allow for interactive exploration depending on the narrative structure, sound, location, etc. of movie elements. Such detailed descriptions can be generated by human annotations, as for example done in the archives of all television broadcasters, or by an automatic system [1-4]. However, if the descriptions become very detailed we have to apply visualization methods to arrange the large quantities of information and to explore the information space. In the following we describe a hierarchical model for the description of movies and methods for the visualization of movie descriptions in so called movie maps. Application areas for movie maps can be seen in

- film analysis – movie maps can visualize the internal structure of films and support the interactive exploration of a movie.
- post production – movie maps can be seen as a visual script, similar to a score or storyboard used in the final post production of music or animation movies.
- archiving and film information systems – movie maps can deal as a visual interface for browsing and search/retrieval applications.

2 Description Levels

We group descriptions into four levels: The physical level, the image level, the object level and the discourse level. We assume, that each level is self-contained and complete, that is, the set of all possible descriptions covers all movies hierarchically. That means that higher levels are based on the descriptions of lower levels.

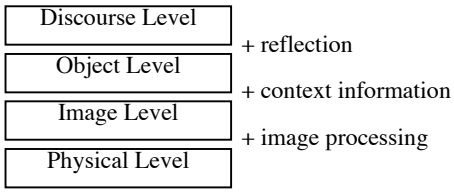


Figure 1 – Description Levels

2.1 Physical Level

At the physical level descriptions of a movie are given by attributes of the storage media. Examples for attributes are the storage location, owner, and depending on the storage media: film-format, aperture, film-stock, sound format, etc. for celluloid media and tape-format, color-depth, video-format, color-coding, etc. for electronic media. Attributes can be grouped into location, media type, history (generation), commercial and process specific parameters of a physical representation.

In most cases descriptions at the physical level are only used in commercial applications, e.g. in an archive or in postproduction environments. For the visualization of a movie these descriptions are not very useful, and we will therefore not go into detail on this subject. However for a metadata model as MPEG-7 the physical level builds the basis of all movie annotations.

2.2 Image Level

At the image level a movie can be seen as a sequence of frames, either in analogue or digital form. Descriptions at the image level can be based on audiovisual features generated by image/audio processing methods, e.g. sonograms [5], histograms [6,7,8], shapes [9,10], textures [11], camera parameters and motion representations [12-16].

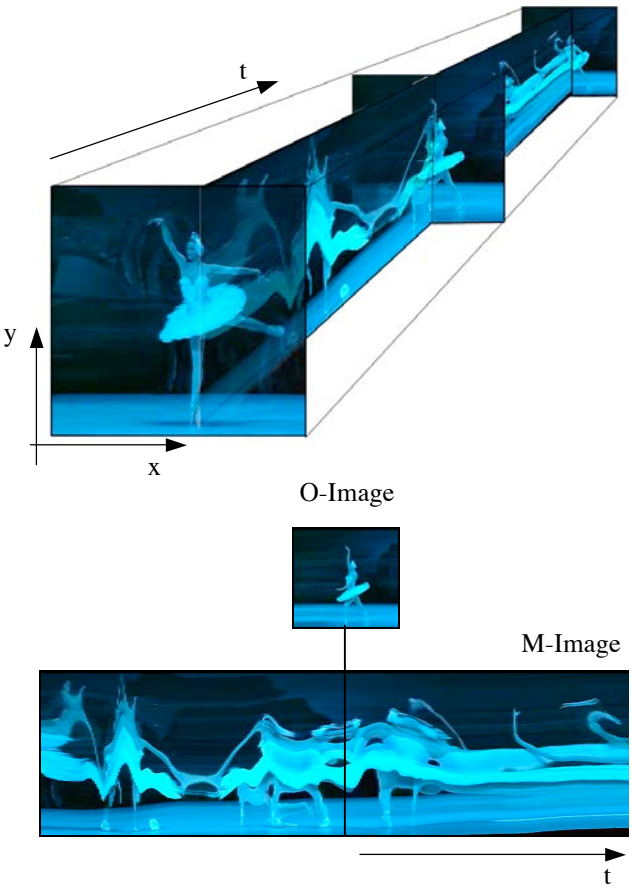


Figure 2 – OM Images

Image processing features have no "knowledge" about real world objects, e.g. there may be shape representations to recognize round objects and a segmentation by the color values "white" and "black", nevertheless there is no representation of the concept of a football at the image level.

In addition keyframes can be selected as representation of a sub-sequence of a movie. We use an extension of the classical keyframe concept, the so called OM, Object(s) + Movement(s) representation. An OM representation consists of a O-image (defined by the plane of maximum image flow in the 3D image volume) and a M-image orthogonal to the O-image. If an O-image is parallel to the x,y plane it is equivalent to a classical keyframe and the M-Image is a time section image [17] parallel either to the x,t or y,t plane. See Figure 2.

2.3 Object Level

At the object level a movie is described by the objects it contains. We can extract content objects automatically, e.g. by a face detection algorithm, or manually by traditional film and video annotation procedures. Content objects descriptions can be grouped into "atomic descriptions" and "structuring objects".

- *Atomic descriptions* – map the appearance of an object and its attribute on the time line of the image sequence. In film analysis atomic descriptions would correspond to denotative annotations and the time line would correspond to the screen duration of a movie [18,19]. Each content object can be visualized in a movie map, see Figure 3, which can easily be modeled by a relational database.

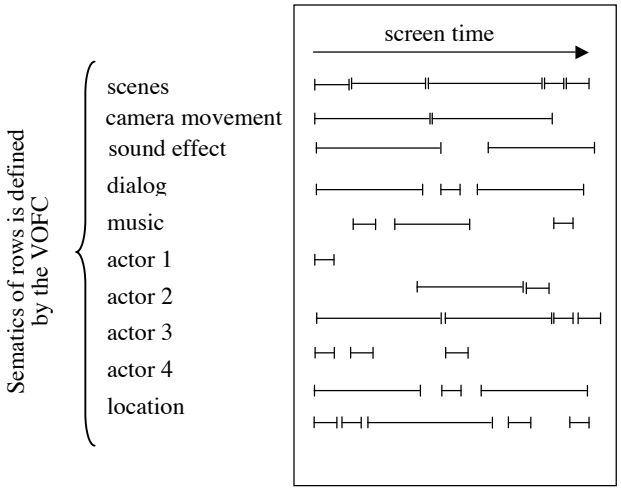


Figure 3 – A simple movie map (object layer)

We define the semantics of the rows by the so called Video Object Foundation Classes – VOFC. The overall class structure of the VOFC can be seen in Figure 4. VOFC describe the semantics of the content object by a

class hierarchy holding attributes and methods for the automatic generation of descriptions.

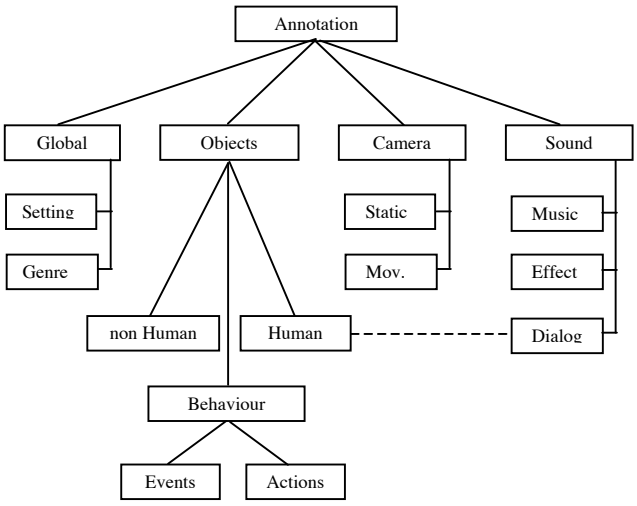


Figure 4 – Video Object Foundation Classes

- *Structuring Objects* – group and describe the hierarchical structure of content objects, e.g. grouping all shots together in which a certain person appears, or modeling the structure of a news program by clips and separators [20]. In film analysis the formal composition and the plot structure of a movie can be modeled with the help of structuring objects, see Figure 5.

2.4 Discourse Level

At the discourse level descriptions contain meta information about the movie, the nature of which depend on the particular form or genre. This means that the description of any particular sequence refers to the larger context of 1) the production (program, film, etc.) the sequence is a part of, 2) the form or genre to which the production belongs. For instance, the traditional description of a scene refers to the plot of an entire film ("last meeting of the Good and the Bad"), and to conventional events and objects in a genre ("final shoot out"). Discourse descriptions can be grouped into two classes (e.g. [21]):

- *Deep structure* - consists of propositions describing the complete contents including all logical relations (See the upright planes in Figure 5, each representing one discourse genre.) The nature of the predicates and arguments differs from one genre to another. Fictional, historical story, rhetorical argument and categorical exposition are four examples of major genres. For instance, in fictional stories, the predicates are causal and chronological relations, and the arguments are fictional events and characters. In categorical expositions (e.g. documentaries, demonstration films), predicates are part-whole and order relations, and arguments are real world categories of interest (airplane, butterfly, actions in cardiac surgery).
- *Surface structure* - describes the way the deep structure is represented in the object level, in terms of atomic objects as well as of structuring objects. (See dotted lines in Figure 5). As a description of a mapping, an application of transformations, it should be distinguished from the object level description itself.

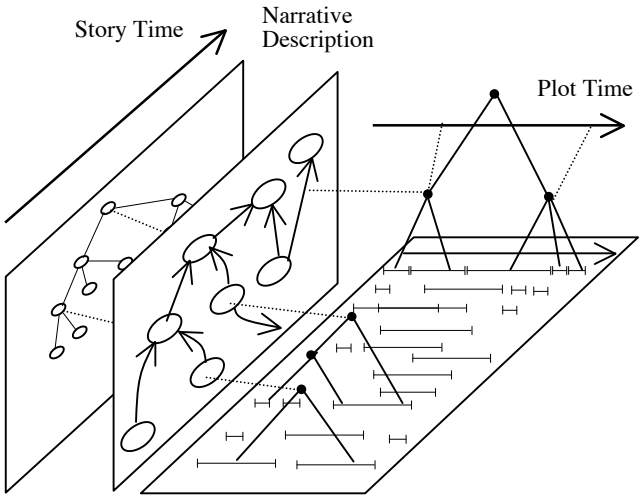


Figure 5 – A Movie Map with Structuring Objects and Narrative Descriptions

There is no one-to-one correspondence between the deep structure description (e.g. the complete story) and the object level description. The transformations inherent to surface structure description include (partial) deletion, summarization, embedding, restricting point of view, emphasizing, reordering and serialization. Profound knowledge exists as to the transformation of stories into film narratives, particularly as temporal and spatial aspects are concerned [20]. However, this knowledge is insufficient to render a complete mapping of story structures onto the object level. To complicate matters, phenomena such as irony and metaphor indicate that other, much more sophisticated transformations should be accounted for in the discourse structure description.

3 Movie Visualization

In order to support the navigation through a movie we visualize movie descriptions at various levels. Automatic video annotation algorithms produce, in the literature often called, "low-level" feature using computer vision algorithms. Low–level features are always descriptions at the image level. In addition "high-level" feature can be generated at the object and discourse level by AI methods (taking into account context and domain knowledge) and by human annotation.

New ways for using this rich information in interactive exploration of movies has attracted many research efforts. In most cases they only integrate descriptions at the image level and are restricted to short (up to several minutes) video clips.

Techniques used are Rframes [22], scene transition graphs [23] or clustering methods [24]. In this paper a generalized approach, including all levels of descriptions is studied, and used to build static and interactive visual representations of a movie. The requirements for a interactive browsing tools are the following:

- It should be possible to handle *large amounts of data* (more than 1 hour, typical 20 hours, up to several 100 hours) in an efficient way. A real world application (e.g. a video archive) deals with such quantities in its daily work. Such quantities request hierarchical ordering and abstraction mechanisms.
- It should provide *good browsing and sorting capabilities*. If the user doesn't know exactly what he is looking for, a visual *movie mining* functionality is required.
- It should *be intuitive* to operate either through a WEB interface or using the common place GUI look and feel.
- The *response time* should be below a (psychological) limit and the degree of *interactivity* should be high. If time consuming operations are performed, e.g. automatic annotation or shot detection, the user should receive feedback about the operation's progress.

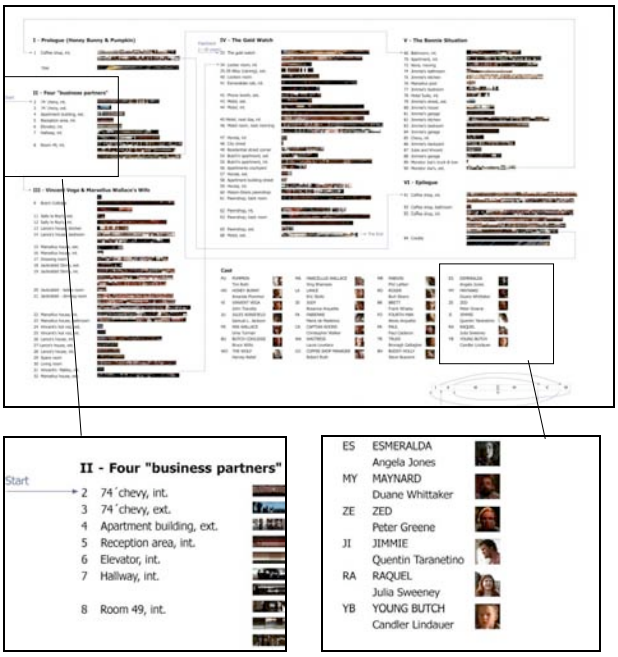
We use OM-Images and sonograms at the image level and graphical representations (arrows, different layers, font weights) at the object and discourse level to visualize structuring objects. All image level descriptions are directly generated from an MPEG-2 representation of a movie.

4 An Example

In Figure 6 the first 8 pages of the Pulp Fiction movie book can be seen. Due to very dense information in the OM-frames these pages are originally printed on A3 paper using a 600 dpi color laser printer.

Figure 6 – Navigation page and visualization of the first 14 minutes of Pulp Fiction

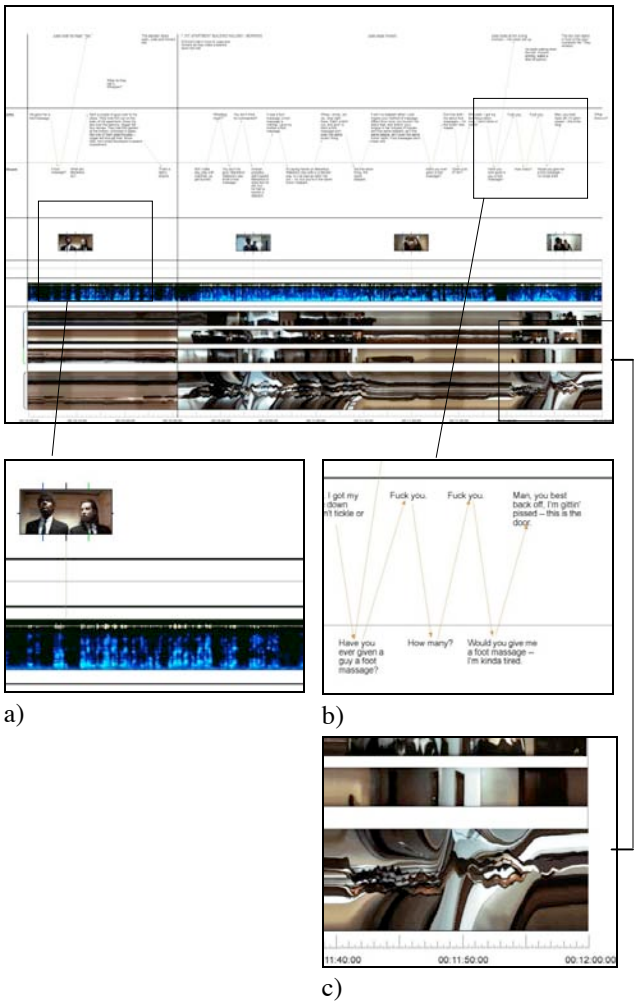
The overview page – the table of content of a movie – shows the top level structuring elements of a movie scenes and their respective M-Images (a), the cast (b) and the plot- versus story time visualized by a temporal ordering of the scenes.



a) b)

Figure 7 – The overview Page

Figure 8 shows the visualization of minutes 10 to 12 of the movie by O-Images and sonograms (a), the script of the movie in relation to the movie plot time (b) and 3 vertical and one horizontal M-Images (c).



a) b) c)

Figure 8 – Minutes 10 to 12

In Figure 9 the user interface of an interactive movie browser (including a re-ordering functionality for virtual shots) is shown. This user interface is a design study and will be implemented by the VICAR project in the near future. Within the browser a movie is divided into *shots* (a sequence of images typically equivalent to a camera shot), s*cenes* (a sequence of shots with a common theme, location, action, etc.) and *parts* (structuring elements inside movies). A movie is defined as a continuous stream of images. The ordering of shots, scenes and parts in the visualization of a movie can not be altered. In contrast to this a *virtual movie* (a movie without physical representation) is defined by a sequence of links static movie object. Within a *virtual movie* elements can be arbitrarily reordered and grouped. Virtual movies allow for different views to a subset of a movie and can be used for the organization of special collections or search results.

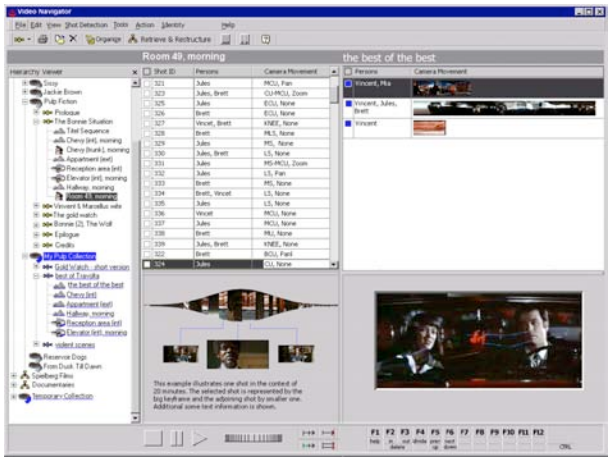


Figure 9 – An interactive Movie Browser

5 Summary and Future Work

As the use of movies in digital representations, e.g. MPEG-2 and the quality of automatic annotation tools are speedily increasing, browsing through huge content collection is a major problem. This development can be compared to the first steps in scientific visualization. In this paper we address the problem of movie description and visualization. We propose to structure descriptions in a hierarchical way:

- Physical Level – image sequences are described by a physical representation.
- Image Level – image sequences are described by the feature space of the used image processing algorithm.
- Object Level – image sequences are described by a denotative annotation (human or automatic generated) and structuring objects.
- Discourse Level, where the descriptions are at time only given in natural language describing the discourse content addressed by the sequence and its expression in the object level.

We describe the principle of movie maps at the object and discourse level, and present a new visualization method at the image level. Finally, we demonstrate the concepts behind a movie book of Pulp Fiction and show the user interface of a interactive image browser. Our future work will address hypermedia visualizations (hyper movies) and new human computer interaction techniques using high resolution displays (dynamic paper).

6 Acknowledgments

The work reported here was carried out under the Training and Mobility of Researcher (TMR) program of the EC. Our thanks are due to Marten den Uyl, and Herre Kuipers of Sentient Machine Research, Werner Haas and Peter Uray of Joanneum Research for their critical reviews and various discussions. Furthermore we like to thank to MIRAMAX for their support of the project.

7 References

[1] Flickner J.P., Sawnhey H., Niblack W, et al, Query by image and video content: the QBIC system, IEEE Computer, 28, Spetember 1995 pp. 23-30.

[2] Aigrain P., Zhang H.J., Petkovic D., Content-Based Representation and Retrieval of Visual Media – A State-of-the-Art Review, „Multimedia Tools and Applications 3(3), 1996, pp. 179-202.

[3] Bolle R.M., Yeo B.-L., Yeung M.M. Video Query: Research directions, IBM Journal of Research and Development, Vol 42. No. 2 – Multimedia Systems.

[4] Müller H., Haas W., Uray P., VICAR – An Automatic Video Annotation System, Proc. of EMMSEC 98, Bordeaux, Oct. 1998.

[5] Minami K, Akutsu A, Hamada H., Video Handling with Music and Speech Detection, IEEE Multimedia July-September 1998, pp. 17-25.

[6] Beigi M,, Benitez A., and Chang S.-F., MetaSEEk: A Content-Based Meta Search Engine for Images, SPIE Conference on Storage and Retrieval for Image and Video Database, Feb. 1998.

[7] Lienhart R., Kuhmünch C., Effelsberg. W. On the Detection and Recognition of Television Commercials, Proc. IEEE Conf. on Multimedia Computing and Systems, Ottawa, Canada, pp. 509-516, June 1997.

[8] Zhang H.J., Tan C.Y., Smoliar S.W., Gong Y., Automatic Parsing and Indexing of News Video, Multimedia Systems, Vol. 2, No. 6, 1995, pp. 256-265.

[9] Eakins J.P., Retrieval of trade mark images by shape feature, Proc of ELVIRA, 1994.

[10] Herodotou N., Plataniotis K.N., Venetsanopoulus A.N., A Content-Bases Storage and Retrieval Scheme for Image and Video Databases, SPIE Vol. 3309, pp. 697-708.

[11] Tamura H., Mori S., Textural Features Corresponding to Visual Perception, IEEE Trans. on Systems, Man and Cybernetics Vol. 8, No. 6, 1978.

[12] Akutsu A., Tonomura Y., Hashimoto H., Ohba Y., Video Indexing using Motion Vectors, Proc, Visual Communications and Image Processing, SPIE. Vol. 1818, 1992, pp. 1522-1530.

[13] Ardizzone E., Cascia M.La, Video Indexing using Optical Flow Field, Proc. of ICIP 96, vol 3, 1996, pp. 831-834.

[14] Kato T., Database Architecture for Content Based Image Retrieval, Proc. of SPIE Conference on Image Storage and Retrieval Systems, Vol. 1662, 1992, pp. 112-123.

[15] Nam J., Tewfik A.H., Motion-based Video Object Indexing using Multiresolution Analysis, SPIE Vol. 3309, pp. 688-696.

[16] Sahouria E., Zakhor A., Motion Indexing of Video, Proc of ICIP 97, vol. 2, 1997, pp. 526-529.

[17] Uray P., Müller H., Plaschzug W., Haas W.: Visualising Artefacts, Meta Information and Quality Parameters of Image Sequences, Proc. of the IS&T Conference on Visual Data Exploration and Analysis V, San Jose, pp. 145-152, 1998.

[18] Bordwell D., Thompson K., Film art – an introduction, 5th ed. McGraw-Hill, 1997.

[19] Faulstich, W., Einführung in die Filmanalyse. 2te Auflage. Tübingen: Narr, 1980.

[20] Zhong D., Zhang H., and Chang S.-F., Clustering Methods for Video Browsing and Annotation, SPIE Conference on Storage and Retrieval for Image and Video Database, Feb. 1996.

[21] van Dijk T.A., Macrostructures: An interdisciplinary study of global structures in discourse, interaction and cognition. Hillsdale NJ, L. Erlbaum, 1980.

[22] Arman et al., Content-Based Browsing of Video Sequences, *Proceedings of ACM Multimedia 94*.

[23] Yeung M. et al., Video Browsing using Clustering and Scene Transitions on Compressed Sequences, Multimedia Computing and Networking, San Jose, Feb. 1995.