MOTION AND SHAPE SIGNATURES FOR OBJECT-BASED INDEXING OF MPEG-4 COMPRESSED VIDEO

A. Müfit Ferman *Bilge Günsel* *A. Murat Tekalp*

Dept. of Electrical Engineering and Center for Electronic Imaging Systems
University of Rochester, Rochester, NY 14627
{ferman, gunsel, tekalp}@ee.rochester.edu

ABSTRACT

The emerging MPEG-4 standard enables direct access to individual objects in the video stream, along with boundary/shape, texture, and motion information about each object. This paper proposes an object-based video indexing method that is directly applicable to the MPEG-4 compressed video bitstreams. The method aims to provide object-based content-interactivity; thus, defines the audio-visual object as the indexing unit. The scheme involves object-based temporal segmentation of the video bitstream, selection of key-frames and key-video-object-planes, and characterization of the motion and/or shape of each video object including the background object. We also propose syntax and semantics for an indexing field to meet the content-based access requirement of MPEG-4. Experimental results are shown on two MPEG-4 test sequences.

1. INTRODUCTION

Ever increasing volume of multimedia databases available over networks or digital storage media renders any text-based search/access method incomplete. To remedy the shortcomings of text-based search engines, content-based techniques have been proposed in the recent years. Although the technology available today allows searches for visual data based on color, texture, and shape information, the difficulty associated with foreground/background separation prevents easy access to the "content" of a scene in terms of semantically meaningful objects in the traditional frame/field based video. With the emerging MPEG-4 standard, content-based interactivity with the input bitstream will be enabled at the object level [1]. This will facilitate direct access to individual audio-visual (A/V) video objects.

The object-based approach adopted by MPEG-4 creates a wide range of possibilities for improved video indexing. MPEG-4 introduces the concept of *Video Object* (VO), which may be of natural or synthetic origin, which allows blending of natural and synthetic content in a single sequence. Instances of VO at a given time are labeled *Video Object Planes* (VOPs). A scene is then described by its VOPs and the spatial/temporal composition information that represents the interactions between them. As each distinct A/V object is separately coded, detailed information about the shape, texture, and motion of individual VOs is available and directly accessible.

The capabilities that MPEG-4 will be designed to offer for content-based storage and retrieval, as described in [1], are shortly summarized below.

- MPEG-4 will support multiple layers of temporal segmentation, and thus a hierarchical representation, of the video stream. Associated with each temporal segment will be
  - *decision support representatives* (DSRs) that depict the asset of interest in condensed form; and
  - descriptive textual and/or numerical attributes.
- Frame-based user controls (rewind, ffw, etc.), as well as temporal-segment based ones ("skip to the next segment"), will be available to the user.
- Private data can be associated with any spatio-temporal object in the audio-visual stream. This can also be in the form of textual attributes defined by the user and geared towards describing scene content.
- Random access to any video object and any temporal segment will be supported.
- Composition information, which determines the spatial/temporal relationships between the VOs in the scene, will be readily available.
- Exact and approximate representations of an object's boundary will be available.
- Semantic information can be linked to each object in a scene; a scene representation that is based on semantic criteria will be thus possible.

In the following sections, we present an object-based video indexing framework that addresses these requirements, and propose an indexing field syntax for efficient content-based storage and retrieval. Specifically, Section 2 outlines a novel object-based temporal segmentation and key-frame/key-VOP selection algorithm. The indexing cues that will be part of the indexing field, along with the proposed syntax, are introduced in Section 3. Experimental results and discussions are presented in Section 4.

*This work is supported by a National Science Foundation SIUCRC grant and a New York State Science and Technology Foundation grant to the Center for Electronic Imaging Systems at the University of Rochester.

2. AN OBJECT-BASED VIDEO INDEXING SCHEME FOR MPEG-4

Temporal segmentation is in fact the first step in almost all video indexing methods in the literature [2]: the input stream is first broken down into smaller temporal units (in most cases, the *video shots*); several decision support representatives (DSR) are then selected from each of these units to represent the scene content in the temporal unit. Proper selection of these DSRs is essential for complete delineation of scene content, as subsequent analysis towards extraction of semantic structure of the sequence is done using these representative frames. The content of a scene is determined by the VOPs present and their interactions with each other; our object-based indexing approach exploits this fact and aims to pinpoint changes in content by detecting significant physical transformations each video object undergoes throughout its lifetime. Such a scheme would normally require segmentation, tracking, and identification of the individual objects in the scene. We do not address these issues, as our approach assumes that the MPEG-4 coded bitstream has already provided the object-based information.

In order to allow temporal as well as object-based access to the video stream, we propose the following framework for temporal video segmentation:

- Observe each object individually, and select as DSRs key VOPs that reflect the important changes in content for that particular object;
- Build an indexing field for every VO, which includes selected indexing cues that represent the VO;
- Construct a temporal segmentation of the video stream, using the temporal information about each VO;
- For each temporal segment, provide scene and composition information about the scene; include pointers to the indexing fields available for the VOs that are present in the temporal segment.

The first level of segmentation of the video stream is conducted on the VOs; the distinct temporal segments in the stream are then constructed by grouping together the information for all the VOs. We carry out the next level of temporal partitioning based on the observations made at the VO level. After the video stream is partitioned into exclusive temporal segments, further analysis is confined to individual segments themselves, as implied by MPEG-4 requirements. The number of indexing cues that may be extracted and computed for each temporal segment is vast; in the next section we present what we believe is a minimal yet sufficient set of attributes, that will successfully resolve commonly encountered queries and provide the desired content-based interactivity to the user. Once again, our main objective in the selection of indexing field parameters has been to provide immediate access to the individual VOs as well as the extracted temporal segments.

3. OBJECT-BASED INDEXING CUES

We propose that the indexing field contain the following parameters:

- The number of available temporal segments in the video stream
- For each temporal segment, the number of VOs, background VO ID, and a single DSR (in this case, a key-frame)
- For each VO in a temporal segment, the birth and death frames of the VO, and the number of DSRs (key-frames) available for the VO
- For each VO DSR, the frame number, shape and motion signatures

3.1. Background/Foreground VOs and VO ID

A background VO may provide valuable cues for correct scene interpretation and resolution of certain syntactic queries. Each VO, whether it be a foreground or background VO, is assigned a unique VO identifier (ID). The background VO provides an indirect link between multiple foregound VOs that are active simultaneously and enables analysis of interactions between multiple VOs within the scene.

3.2. Birth, Death, and Persistence of a VO

In MPEG-4, all VOs are distinct entities in the bitstream with known birth and death points. The search region for queries related to a specific VO can be strictly restricted to the frames within the lifespan of that VO by including this information in the VO indexing field. Note that a grace period is allowed for a VO to become invisible temporarily before it is declared dead: If a VO does not reappear within this grace period, its ID is expired.

3.3. Selection of Decision Support Representatives

Even though major scene content changes can be detected through birth and death frames of VOs, further temporal segmentation of the life span of a VO may be needed to capture and index prominent physical motions that a VO undergoes. Therefore, we propose to detect a number of key VOPs that successfully showcase the evolution of a VO throughout its life span, and use these as DSRs in the indexing field. This is achieved by observing the compression mode the encoder selects for each macroblock in the VOP [3]; this translates into determining whether the motion/texture characteristics of a VOP can be predicted from neighboring VOPs. We compute the ratio of the number of intra-coded macroblocks to the total number of (encoded) macroblocks on a given VOP; if this ratio exceeds a certain threshold the frame is labeled a key-frame/DSR. By changing the selected threshold the temporal segmentation process can be carried out in a hierarchical manner, and the temporal detail represented by the selected DSRs can be fine-tuned.

3.4. Shape and Motion Signatures for Each Temporal Segment

Once a VO is segmented into temporal subregions, shape and motion signatures for each individual subregion can be computed and stored with the frame number of each DSR. These features can be computed as a function of encoded shape and motion information associated with each VOP in the bitstream. Additional indexing attributes can also be

computed at the decoder by means of local analysis of the referenced DSRs.

After DSR selection, each VOP included in a DSR is indexed by their *motion* and *shape* signatures. The motion signature includes the global motion of the frame and the trace of local motion of each VOP. This requires the estimation of global motion observed in the *background*. The background is defined as a distinct VOP in MPEG-4; hence, its motion can be extracted from the bitstream. It is clear that the global motion between successive frames provides relevant information in determining the camera operation. We have used the Hough transform [4] to classify the global motion prevalent in the sequence. The Hough transform method has previously been adopted by Akutsu *et al.* [5] for indexing of uncompressed video; their approach relies on the transform for determining the divergence/convergence point of motion vectors, and ultimately represents camera movements. Instead, we consider the distribution of the global motion vectors in the Hough domain to discriminate between different types of camera operation. We extend the approach to the foreground VOs by first constructing the distribution of the object motion vectors in the Hough space using a predefined number of bins (i.e., directions). The bin that receives the most votes characterizes the object motion within the temporal segment of interest. Of particular note is the fact that motion of the background VO is compensated for when foreground object motion is determined; otherwise the obtained motion signature is biased and may lead to incorrect conclusions. It has been noted that quantizing the range of possible directions to 16 levels provides sufficient detail.

The shape signature consists of *eigenshape* representations extracted from the VOP shape information given in the MPEG-4 bitstream. These representations can be efficiently used in shape matching and similarity ranking for query-by-example [6]. Eigenshape-based methods involve decomposition of the shape information into an ordered basis of orthogonal principal components, supporting efficient data reduction, robust reconstruction, and a unique description of the input shape. The shape-related indexing information stored for each DSR is the eigenshape representation; during retrieval, a certain *mismatch* matrix for the query example and the objects in the DSR needs to be constructed. A shape similarity metric of choice can then be used to determine a match value between query example and a given VO.

The syntax proposed for the indexing field of each temporal video segment is given in Table I; it should be noted that this field is not a part of the standard MPEG-4 syntax and is intended to complement the existing MPEG-4 structure.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

We illustrate the proposed methods on the 300-frame MPEG-4 test sequences "*cyclamen*" (SIF format) and "*coastguard*" (CIF format). The software developed for automatic temporal segmentation and DSR selection is based on the MPEG-4 Video Verification Model 4.0 Software by Microsoft.

We first present the results of temporal segmentation on the "*cyclamen*" sequence using the object-based approach. Three example frames from the sequence are depicted in Fig. 1(a)-(c), while the corresponding segmentation masks are shown in Fig. 1(d)-(f). Note that the segmentation mask provided by MPEG-4 regards all the flowers as a single object, even though the individual flowers in the scene change constantly. For this sequence no motion is present at the background and foreground objects except the camera motion. Using the motion information, a single DSR for each VO-the birth frames-was selected by the proposed object-based indexing approach. This result is consistent with the object-based paradigm: The whole sequence involves flowers only; according to the MPEG-4 bitstream no new objects enter the scene, nor is any radical physical transformation registered for the objects present. One can argue that the color composition in the scene changes, since the individual flowers are different, and that the object-based approach fails to capture this evolution, which would be noted by a scene-based segmentation technique. The object-based approach can be refined to handle such subtleties by integrating the proposed scene change detection scheme with frame-based compressed scene analysis techniques [2], [7], using the compressed YUV information in the MPEG-4 bit streams. The number of foreground objects can also be increased by defining new VOs for individual flowers.

Flag/Parameter	# Bits
#_Temporal_Segments	8
for each Temporal_Segment {	
#_VOs	8
for each VO {	
Background_VO_flag	1
if(Background_VO_flag == 1) {	
Background_VO_ID	12
#_Foreground_VOs	8
for each(Foreground_VO) {	
Foreground_VO_ID	8
}	
}	
else {	
Foreground_VO_ID	12
Background_VO_ID	12
}	
VO_Birth_frame	16
VO_Death_frame	16
VO_textual_attribute	
num_VO_DSRs	8
for each(VO_DSR) {	
VO_DSR_frame	8
Shape_Signature	
Motion_Signature	
DSR_textual_attribute	
}	
}	
}	

Table I. The proposed indexing field for individual temporal segments selected in the MPEG-4 bitstream.

The results obtained by Hough-based motion analysis are shown in Fig. 2 for the "*coastguard*" sequence. The camera operation present in the sequence are "pan left," "tilt up," and "pan right." In Fig. 2(c)-(d) the distribution of the motion vectors of the background VO for the whole sequence are presented, obtained using motion vectors provided by the MPEG-4 VM and hierarchical block-matching with three-step search [8], respectively. Comparison of the plots reveals that while the motion information provided by MPEG-4 is good enough to capture the overall nature of the global motion (and, consequently, the camera operations in the sequence), it is not as reliable as that given by a more sophisticated, hierarchical block matching technique. This effect is more evident in the single-VOP distributions shown in Fig. 2 (e)-(f), computed for the small speedboat in the foreground. The distribution of MPEG-4 motion vectors, in Fig. 2(e), is confusing and disorienting, while hierarchical block-matching using three-step search generates a correct depiction of the "motion signature" for the object in question. For the assignment of "shape signature," a B-spline is first fitted to the closed object contours extracted from the shape information provided by MPEG-4 bit stream. Eigen decomposition is then performed to obtain eigenshapes for each B-spline control point. To reduce the computational complexity, the number of control points is limited to 100, which is adequate for a smooth shape representation. Eigenshape representations corresponding to low frequency components are specified and stored as shape indexing features.

Our present research focuses on motion estimation techniques to improve the accuracy of the motion data that MPEG-4 provides. This includes motion compensation through the estimation of camera motion parameters to allow the computation of individual object motions. We are also looking for new approaches to model video objects efficiently by 2-D/3-D meshes.

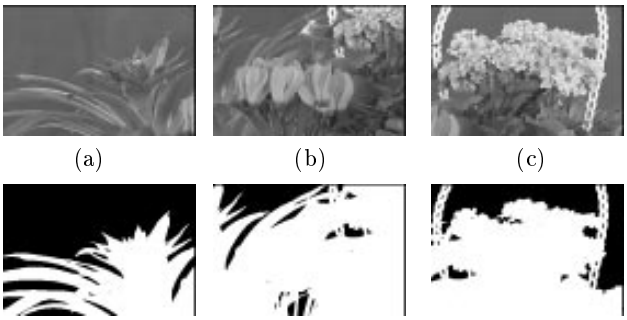

Figure 1. Example frames from the MPEG-4 test sequence "*cyclamen.*" The frames depicted are (a) 10; (b) 150; and (c) 260. The corresponding segmentation masks are given in (d) through (f), respectively.

Acknowledgements: The authors wish to thank Dr. P. J. L. van Beek for installing and updating the Microsoft MPEG-4 VM4.0 encoder/decoder.

REFERENCES

[1] MPEG-4 Profiles v.1.2-Req. ISO/WG11 N1494, Nov. 1996.

[2] B. Furht, S.W. Smoliar, and H. Zhang. *Video and Image Processing in Multimedia Systems.* Kluwer Academic Publishers, 1995.

[3] A. M. Ferman, B. Günsel, and A. M. Tekalp. Object-based indexing of MPEG-4 compressed video. In *to be presented at VCIP'97*, Feb. 1997.

[4] P. V. C. Hough. Methods and means for recognizing complex patterns. U.S. Patent 3069654, 1962.

[5] A. Akutsu, Y. Tonomura, H. Hashimoto, and Y. Ohba. Video indexing using motion vectors. In *Proc. SPIE: VCIP'92*, pages 1522–1529, Nov. 1992.

[6] B. Günsel and A. M. Tekalp. Similarity analysis for shape retrieval by example. In *Proc. ICPR'96*, pages 330–334, Austria, Aug. 1996.

[7] B.-L. Yeo and B. Liu. Rapid scene analysis on compressed video. *IEEE Trans. on Circuits and Systems for Video Technology*, 5:533–544, Dec. 1995.

[8] M. Bierling. Displacement estimation by hierarchical blockmatching. In *Proc. SPIE VCIP'88, Vol. 1001*, pages 942–951, 1988.

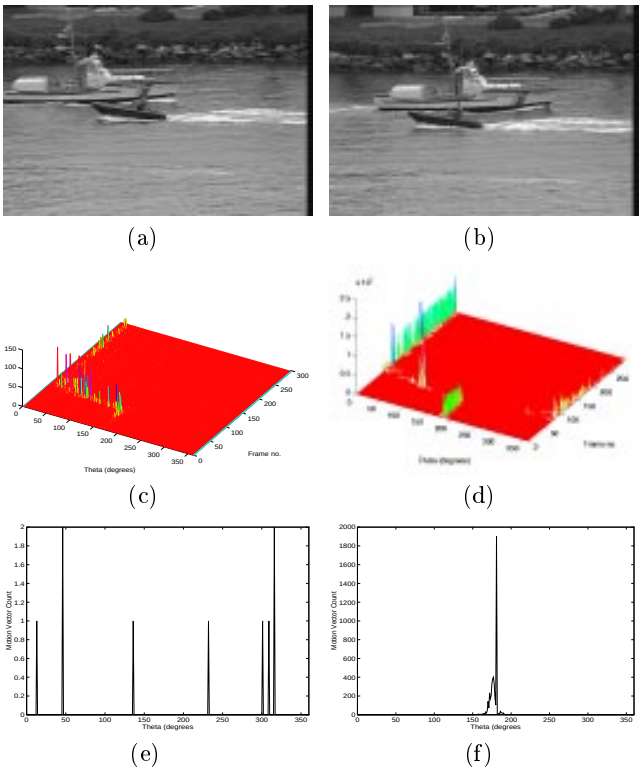

Figure 2. The Hough-based motion representation for the test sequence "*coastguard.*" Frame no.s 50 and 68 are shown in (a) and (b), respectively. The distribution of background VO motion for 300 frames: (c) MPEG-4, (d) hierarchical block matching. The distribution of foreground object motion for a single frame: (e) MPEG-4, (f) hierarchical block matching.