PROGRESSIVE RESOLUTION MOTION INDEXING OF VIDEO OBJECT

Jeho Nam and Ahmed H. Tewfik

Department of Electrical and Computer Engineering
University of Minnesota, Minneapolis, MN 55455 USA
e-mail: jnam@ece.umn.edu, tewfik@ece.umn.edu

ABSTRACT

We present a novel motion-based video indexing scheme for fast content-based browsing and retrieval in a video database. The proposed technique constructs a dictionary of prototype objects to support *query by motion*. The first step in our approach extracts moving objects by analyzing layered images constructed from the coarse data in a *3-D wavelet decomposition* of the video sequence. These images capture motion information only. Moving objects are modeled as collections of interconnected rigid polygonal shapes in the motion sequences that we derive from the wavelet representation. The motion signatures of the object are computed from the rotational and translational motions associated to the elemental polygons that form the objects. These signatures are finally stored as potential query terms.

1. INTRODUCTION

The fundamental step in content-based multimedia database system is to index individual multimedia objects with searchable compactly represented descriptors [1]. Well organized indexing structures facilitate the fast and practical browsing and direct access to the desired data objects in response to object-oriented searching. Such indexing procedure can be constructed by creating *semantic descriptors* based on information that identify and represent each multimedia object model. Therefore, to define and extract a single or multiple key features from each object data is an essential task in indexing construction. The *motion feature* is an important characteristic contained in video data. This unique property differentiates the dynamic video data from the static imagery data.

In this paper, we propose a method to index the *motion features* of video objects for content-based searching and browsing. To attack this task, our proposed scheme performs the motion analysis of video objects in multiresolution/multiscale manner. We extract moving objects by analyzing layered images constructed from the coarse data in a *3-D wavelet decomposition* of the video sequence. These images capture motion information only. Next, we fit a polygon to each connected region in the motion sequence (the wavelet domain images which capture the motion information). The rotational and translational motions associated to the center of gravity of these polygons are extracted. Finally, these motions are stored and indexed using a progressive resolution approach similar to [10]. Specifically, we combine a VQ approach with a coarse wavelet representation of the motion data to perform indexing and retrieval.

Most previous motion indexing approaches are based on global motions at frame level (e.g., camera zooming, panning and tilting). Although useful for some purpose, it is hard to investigate the local motion properties of video objects. Recently, some object-based motion indexing techniques have been presented [2, 3, 4, 5]. A variety of motion features are derived by several different methods. The motion vectors have been most widely adopted as an indexing clue for representing the motion properties. Traditionally, to examine camera operations, the motion vectors are computed by the optical flow, block-matching method and the Hough transform [6, 7, 8, 9]. Unfortunately, these procedures have a drawback of computational complexity, especially on processing of a long video data.

Note that our approach, like that of [5], relies ultimately on a low resolution wavelet representation of the motion for indexing and retrieval. Unlike the approach of [5], we extract moving objects by analyzing layered images constructed from the coarse data in a 3-D wavelet decomposition of the video sequence. This leads to a simple object identification approach that avoids matching problems. Furthermore, unlike all previous motion indexing methods, we extract the motion vectors by modeling moving objects as an association of rigid polygons directly in the 3-D wavelet motion sequence that we derive from the video.

2. PROGRESSIVE SPATIO-TEMPORAL SEGMENTATION OF MOVING OBJECTS

2.1. Progressive Resolution Analysis of a Video

The first step in the object-oriented video indexing is to locate a particular *VO (video object)* of interest. The VOs are carried by *video shot* that is regarded as a basic temporal segment unit of a long video sequence. We decompose a continuous video sequence into isolated shots using the temporal segmentation method proposed in [11]. In this method, shot boundaries are determined by examining the coarse data in a *3-D wavelet transform* of the video track. To reduce the computational complexity (especially for long video sequences), we begin by computing coarse reduced frames from a 2-D wavelet transform of a temporally sub-sampled video sequence. Through this spatio-temporal progressive analysis scheme, a set of isolated video shots are efficiently decomposed from video sequences. We extend this 3-D wavelet transform of a video data to locate and segment the motion contour of video objects within

each video shot.

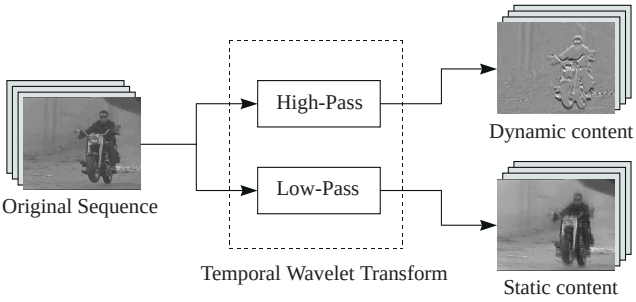

Figure 1: Temporal wavelet transform: decomposes a video sequence into dynamic content and static content

2.2. Temporal Wavelet Transform

The multiresolution nature of a wavelet analysis provides a compact representation of a signal localized in space, time or frequency domain. The wavelet transform can be implemented using a 2-band perfect reconstruction filter bank. This filter bank decomposes a signal into lower (low-pass) and upper (high-pass) subbands in the frequency domain. Thus, a video signal passed through a filter bank outputs both *static* (no motion) and *dynamic* (motion) content streams of the original signal. Such a *multiresolution temporal representation* of video sequence can effectively be exploited for spatio-temporal identification of a prominent object's motion contour within each isolated shot.

Due to the temporally high-passed filtering function of temporal wavelet transform, each frame of dynamic content stream has a zero-crossing value at pixels corresponding to motion boundary of moving objects. We compute the new intensity image sequence by thresholding the absolute value of each pixel in an individual frame of dynamic content stream. The resulting new image sequence has nonzero values on each pixel associated to motion boundary of moving objects. Here, we will refer to this sequence as the *motion sequence*.

Unlike the absolute difference image of two adjacent frames, this motion sequence contains a *time-localized history* of the temporal activity of moving objects. Note that we take the temporal wavelet transform of the spatially subsampled and filtered coarse frames by a 2-D wavelet transform of the original video data. The spatially smoothed and reduced frames have the effect of eliminating the minor spatial variations and time-varying illumination. In addition, the associated computational complexity is decreased due to the spatially reduced size of frames.

2.3. Layered Representation of Multiple Objects

We identify moving objects by investigating the amount of motion displayed by each object within a video shot. The multiresolution/multifrequency nature of the temporal wavelet transform can provide a coarse representation of multiple moving objects with layers. Intuitively, we expect the motion sequence corresponding to the most rapid moving object to appear in the higher successive subbands. Objects of lower motion are identified on several lower successive subbands.

We begin by separating prominent moving objects from the effect of global camera motions (e.g., zoom and pan) by combining the approach of [3] with our multiresolution motion decomposition. Next, we construct multi-layered image sequences [12] $Layer_k(m,n,t)$ of multiple moving objects using the direct temporal correlation of wavelet transform at several adjacent subbands image sequences $W_l(m,n,t)$ in time axis as follows:

$$Layer_k(m,n,t) = \prod_{l=i}^{j} W_l(m,n,t), \qquad l,i,j = 1,2,3,... \quad (1)$$

Here k is the layer number, l is the number of temporal subband and $i \leq j$. By examining the layered images, we can segment the video sequence into the individual moving objects. We tested this technique on the $TableTennis$ sequence. Figure 3 shows that two moving objects (ball and a player's arm) are effectively classified as different objects based on associated motion amounts.

3. MOTION ANALYSIS OF OBJECTS

3.1. Rigid Moving Object

In this section, we describe a model for temporal variations of the spatial intensity pattern on the motion sequence. Generally, a 3-D moving object is analyzed by projecting its time-varying activities onto 2-D image plane. The motion characteristics, such as *translation* and *rotation*, of a moving object can be represented as follows,

$$\vec{\mathbf{v}}_{\mathbf{i+1}} = \mathbf{A}\vec{\mathbf{v}}_{\mathbf{i}} + \vec{\mathbf{d}}. \quad (2)$$

where $\mathbf{A}$ contains a scale and rotation factor, and $\vec{\mathbf{d}}$ is a displacement vector ($\vec{\mathbf{d}} = [\Delta x\ \Delta y]^T$). In order to obtain these motion parameters, we first compute *the motion gravity* from the *intensity values within the area of motion contour*. Such a motion gravity is located by a method similar to [13]. Then, a polygon (e.g., ellipse or rectangle) is fitted to the motion contour of each object of interest. An object encompasses the distribution of the motion intensity around the centroid of motion (motion gravity). The major axis of such a polygon is simply derived based on the relative position between the furthest intensity pixel and the centroid (Figure 2). The size of a polygon may be changed in each successive frames according to the amount and distribution of the motion intensities. However, because we focus on the positions of a motion centroid and the major axis, the time-varying size of a polygon has no effect on our motion feature analysis. Recall that all our methods of identifying the motion contour and deriving the fitting polygon are performed on the reduced coarse image sequence computed by a 3-D wavelet transform of a video data. Thus, we can achieve the highly reduced computational complexity in even a large video database.

3.2. Multiresolution Analysis of Motion Features

From the location of a centroid and axes of a fitting polygon, we compute the motion indexing features representing

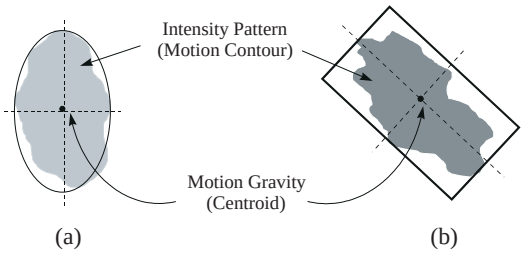

Figure 2: Fitting polygon: (a) ellipse (b) rectangle

the motion behavior of objects. First, an indexing feature is extracted from translational motion $\vec{\mathbf{d}}$. The position tracking of a centroid effectively reflects the trajectory of objects in the translational movement. In addition, from the position of the polygon's axes, it is possible to derive the rotational motion feature $\mathbf{A}$ from the angle value of a moving rigid object around the hinge. Due to the inherent spatially coarse representation in our proposed approach, these motion signatures may be slightly inaccurate in describing the temporal motion behavior of objects. Furthermore, very slowly moving objects in a long video sequence need to be efficiently represented in a compact form (i.e., in shorter time-scale). For this purpose, we take a wavelet transform of $\mathbf{A}$ and $\vec{\mathbf{d}}$ since it is capable of compactly represent their temporal variation at different scales and resolution. Note that we keep track of the lowest 3 scales. This provides a filtering of inaccurate motion vectors (Figure 4).

As explained earlier, the dictionaries are constructed in terms of these multiresolution subbands of motion features from video objects. Based on the dictionaries, progressive searching and matching to the query by motion is performed in an efficient manner.

4. EXPERIMENTAL RESULTS

We tested our proposed technique on two different categorized motion (translation and rotation) of video sequences. All image sequences are spatially filtered and subsampled by 2-D wavelet transform at 30 × 40 resolution from the original size of 320 × 240. Figure 3 shows the effectiveness of the first step of our procedure for detecting and classifying the changing parts of moving objects using the temporal 1-D wavelet transform. On the correlation between adjacent subbands, a rough classification between a ball and player's arm is shown. In Figure 4, a vertically moving (up-and-down) ball is successfully captured by an ellipse centered on the gravity of motion intensity. A simple trajectory of gravity's position is also illustrated. Such a global view of moving object provides a user with effective and efficient cue for motion-based indexing of a large video database. The rotational movement of a particular object around a hinge is shown in Figure 5. A salesman's ankle and a part of left upper-arm are captured by a rectangle centered on the motion intensity. From the rotational movement of the axes of this rectangle around the hinge (salesman's elbow), the rotational motion parameter (i.e., the angle value) is computed without any additional work.

5. CONCLUSIONS

We presented a new object-based motion indexing scheme of video objects. Through its progressive multiresolution and multiscale properties, the proposed approach speeds up the process of constructing the compact description of moving objects and is capable of prompt responding the query by motion in an effective manner. Experimental results show the effectiveness of our proposed approach.

6. REFERENCES

[1] S. W. Smoliar and H. Zhang, "Content-Based Video Indexing and Retrieval," *IEEE Multimedia*, vol. 1, no. 2, pp. 62-72, 1994.

[2] M. Ioka and M. Kurokawa, "A Method for Retrieving Sequences of Images on the basis of Motion Analysis," *Proc. of SPIE: Images Storage and Retrieval Systems*, pp. 35-46, 1992.

[3] A. M. Ferman, B. Günsel and A. M. Tekalp, "Motion and Shape Signatures for Object-based Indexing of MPEG-4 Compressed Video," *Proc. of ICASSP '97*, vol. 4, pp. 2601-2604, 1997.

[4] D. Zhong and S-F. Chang, "Spatio-Temporal Video Search using the Object Based Video Representation," *Proc. of ICIP '97*, vol. 1, pp. 21-24, 1997.

[5] E. Sahouria and A. Zakhor, "Motion Indexing of Video," *Proc. of ICIP '97*, vol. 2, pp. 526-529, 1997.

[6] A. Akutsu, Y. Tonomura, H. Hashimoto and Y. Ohba, "Video Indexing using Motion Vectors," *Proc. of SPIE: VCIP '92*, pp. 1522-1529, 1992.

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

[8] G. Iyengar and A. B. Lippman, "Videobook: An Experiment in Characterization of Video," *Proc. of ICIP '96*, vol. 3, pp. 855-858, 1996.

[9] F. M. Idris and S. Panchanathan, "Spatio-Temporal Indexing of Vector Quantized Video Sequences," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 7, no. 5, Oct. 1997.

[10] M. D. Swanson and A. H. Tewfik, "Embedded Object Dictionaries for Image Database Browsing and Searching," *Proc. of ICIP '96*, vol. 3, pp. 875-878, 1996.

[11] J. Nam and A. H. Tewfik, "Combined Audio and Visual Streams Analysis for Video Sequence Segmentation," *Proc. of ICASSP '97*, vol. 4, pp. 2665-2668, 1997.

[12] J. Y. A. Wang and E. H. Adelson, "Representing Moving Images with Layers," *IEEE Trans. on Image Processing*, vol. 3, No. 5, pp. 625-638, Sept. 1994.

[13] G. Rigoll and A. Kosmala, "New Improved Feature Extraction Methods for Real-Time High Performance Image Sequence Recognition," *Proc. of ICASSP '97*, vol. 4, pp. 2901-2904, 1997.

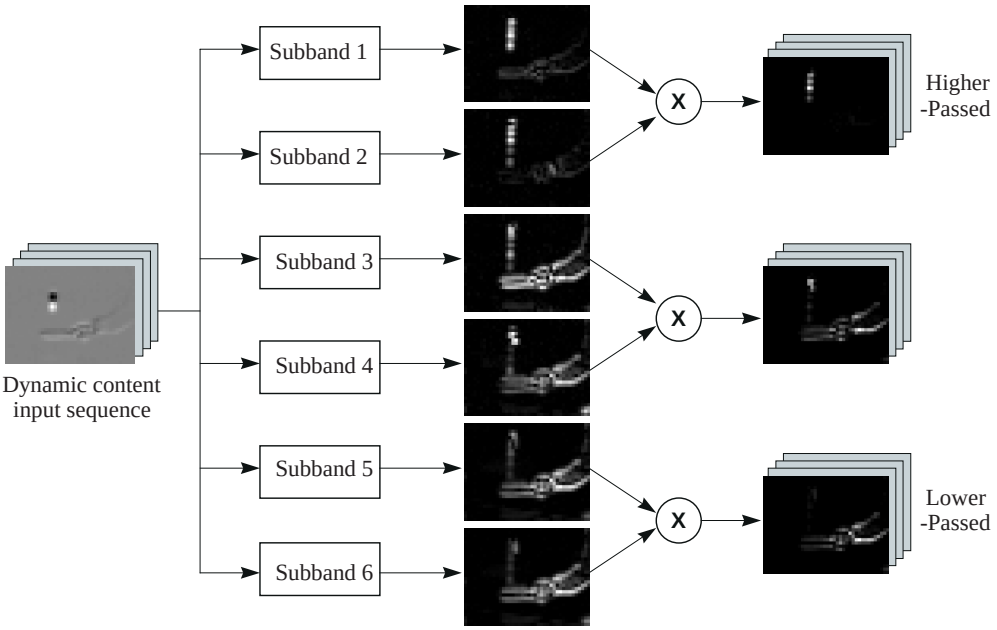

Figure 3: Layered segmentation of multiple moving objects

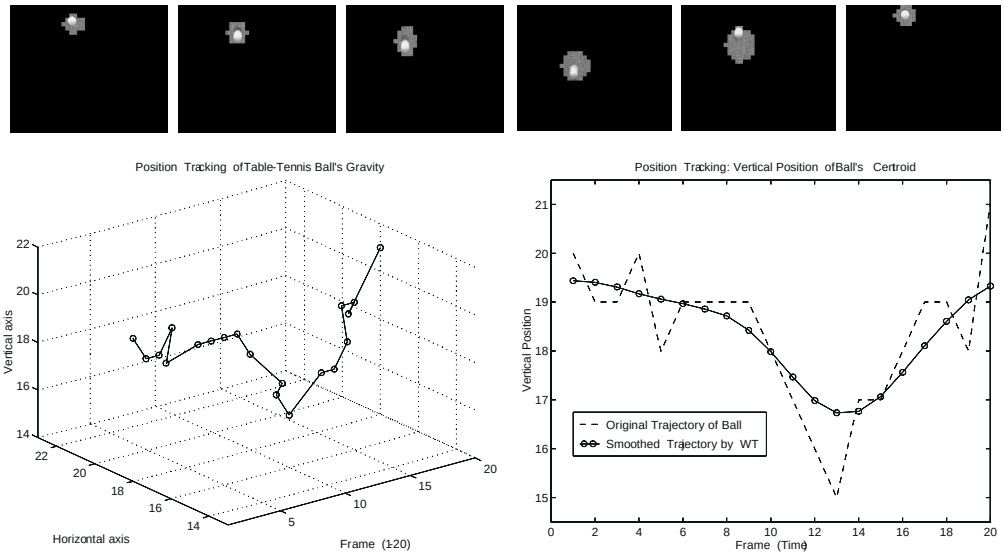

Figure 4: Translational motion: tracking of a ball (left). Its smoothed version by wavelet analysis (right)

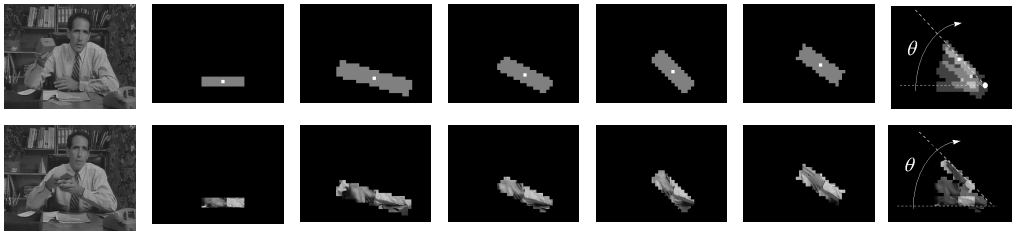

Figure 5: Rotational motion: rotational trajectory of left upperarm and its angle value