EURASIP Journal on Applied Signal Processing 2003:1, 41–47

Retrieval by Local Motion

Berna Erol

Ricoh California Research Center, 2882 Sand Hill Road, Suite 115, Menlo Park, CA 94025-7022, USA
Email: berna_erol@rii.ricoh.com

Faouzi Kossentini

Department of Electrical and Computer Engineering, University of British Columbia,
2356 Main Mall, Vancouver, British Columbia, Canada V6T 1Z4
Email: faouzi@ece.ubc.ca

Received 15 May 2002 and in revised form 30 September 2002

Motion feature plays an important role in video retrieval. The current literature mostly addresses motion retrieval only by camera motion and global motion of individual video objects in a video scene. In this paper, we propose two new motion descriptors that capture the local motion of the video object within its bounding box. The proposed descriptors are rotation and scale invariant and based on the angular and circular area variances of the video object and the variances of the angular radial transform coefficients. Experiments show that ranking obtained by querying with our proposed descriptors closely match with the human ranking.

Keywords and phrases: video databases, video indexing and retrieval, object-based video, motion descriptor, MPEG-4, MPEG-7.

1. INTRODUCTION

As the advancements in digital video compression resulted in the availability of large video databases, indexing and retrieval of video became a very active research area. Unlike still images, video has a temporal dimension that we can associate with motion features. We use this information as one of the key components to describe video sequences; for example, “this is the part where we were salsa dancing” or “this video shows my daughter skating for the first time.” Consequently, motion features play an important role in content-based video retrieval.

It is possible to classify the types of video motion features into three groups.

(i) Global motion of the video or camera motion (e.g., camera zoom, pan, tilt, roll).
(ii) Global motion of the video objects within a frame (e.g., an object is moving from the left to the right of the scene).
(iii) Local motion of the video object (e.g., a person is raising his/her arms).

Camera operation analysis is generally performed by analyzing the directions of motion vectors that are present in compressed video bit stream [1, 2, 3] or optical flow analysis in the spatial domain [4]. For example, panning and tilting motions are likely to be present if most of the motion vectors inside a frame are in the same direction. Similarly, zooming motion can be identified by determining whether or not the motion vectors at the top/left of the frame have opposite directions than the motion vectors at the bottom/right of the frame [5, 6].

Global motion of video objects is represented with their motion trajectories, which are formed by tracking the location of video objects (object’s mass center or some selected points on the object) over a sequence of frames. Forming motion trajectories generally requires segmentation of video objects in a video scene. In MPEG-4, the location information of the video object bounding box (the upper-left corner) is already available in the bit stream making the formation of the trajectory a simple task [7]. The classification and matching of object motion trajectories is a challenging issue as the trajectories contain both the path and the velocity information of the objects. In [8], Little and Gu proposed to extract separate curves for the object path and speed and match these two components separately. Rangarajan et al. [9] demonstrated a two-dimensional motion trajectory matching through scale-space and Chen and Chang [10] proposed to match the motion trajectories via a wavelet decomposition.

Most available content-based video retrieval systems in the literature employ camera motion features and/or global object motion for retrieval by motion. For example, the Jacob system [11] supports queries using common camera motion changes such as pan, zoom, and tilt. Another retrieval system, VideoQ, employs a spatio-temporal segmentation algorithm in order to retrieve individual objects with their global motion inside a scene [12]. It allows the user to specify an arbitrary polygonal trajectory for the query object and retrieves the video sequences that contain video objects with similar trajectories. Similar to VideoQ, NeTra-V supports spatio-temporal queries and utilizes motion histograms for global camera and video object motion retrieval [13]. Moreover, the content-based description standard MPEG-7 [14, 15] supports motion descriptors, in particular, camera motion which characterizes the 3D camera operations, motion trajectory which captures 2D transitional motion of objects, parametric motion which describes the global deformations, and motion activity which specifies the intensity of action.

On the other hand, local motion, the motion video objects within their bounding box, could give valuable information about its articulated parts, elasticity, occlusion, and so forth. Classifying and identifying video objects using their local motion is potentially useful in many applications. For example, it could be useful to identify some suspicious human actions in surveillance video sequences. It could also be useful for efficient video compression, where the encoder can allocate more coding bits or a better communication channel for the video objects that demonstrates important actions, for example, a person running out of a store (there is a chance that the person might be a criminal) or a player scoring. Moreover, processing database queries such as “find a video sequence where people are dancing” would be possible only by enabling the retrieval of video objects with their local motion. The current research in detecting the local motion of video objects has been restricted mostly to specific domains. Stalidis et al. employed a wavelet-based model using boundary points of magnetic resonance images (MRI) to describe the cardiac motion in [16]. Miyamori and Iisaku [17] proposed to classify the actions of tennis players using 2D appearance-based matching. Hoey and Little suggested a method for the classification of motion, which is based on the representation of flow fields with Zernike polynomials in [18]. Their method is applied to the classification of facial expressions. In [19], Fujiyoshi and Lipton presented a process to analyze human motion by first obtaining the skeleton of the objects and then determining the body posture and motion of skeleton segments to determine human activities. Human motion classification was also studied by other researchers including Little and Boyd in [20], where they proposed to recognize individuals by periodic variation in the shape of their motion, and Heisele and Woehler in [21], where they suggested discriminating pedestrians by characterizing the motion of the legs. Moreover, Cutler and Davis [22] proposed to characterize the local motion by detecting periodicity of the motion by Fourier analysis on the gray scale video. Most of the work in this area focuses on “recognizing” the motion of specific objects and they assume prior knowledge about the video content.

As the video object content becomes more widely available, mostly due to the emergence of 3D video capture devices [23, 24], object-based MPEG-4 [25] video encoding standard, and the availability of the state of the art segmentation algorithms [26, 27], there is a need for more generic motion features that describe the local motion of video objects. In this paper, we propose two content-independent local motion descriptors. Motivated by the fact that any significant motion of video objects within their bounding box would very likely result in changes in their shape, our motion descriptors are based on the shape deformations of video objects. The first descriptor, angular circular local motion (ACLM), is computed by dividing the video object area into a number of angular and circular segments and computing the variance of each segment over a period of time. The other proposed descriptor is based on the variances of the angular radial transform (ART) coefficients. We assume that the segmented objects are obtained prior. The proposed descriptors are extracted using video objects’ binary shape masks. The rest of the paper is organized as follows. Sections 2 and 3 describe the proposed local motion descriptors as well as their extraction and matching. Experimental results that illustrate the retrieval performance of our methods and the associated trade-offs are presented in Section 4. Conclusions are given in Section 5.

2. ANGULAR CIRCULAR LOCAL MOTION (ACLM) DESCRIPTOR

Unlike the shape of visual objects in still images, the shape of a video object is not fixed and is very likely to change with time. Given that the camera effects, such as zooming, are compensated for, the shape deformations in an object’s lifespan could offer some valuable information about the object’s local motion, occlusion, articulated parts, and elasticity. The variance of the object area is a good measure for such shape deformations. Nevertheless, it may not be sufficient to capture the motion of the video objects in some cases, especially if the object motion does not have an effect on the area of the object. For example, if an object has an articulated part that is rigid in shape, then the object’s area may not change even if there is local motion. Here, we propose to divide the binary shape mask of a video object into M angular and N circular segments and use the variance of the pixels that fall into each segment to describe the local motion. Variances are computed for each angular circular segment in the temporal direction using the temporal instances of the video objects. Then, the local motion feature matrix is formed for each video object as follows:

$$\mathbf{R} = \begin{bmatrix} \sigma_{0,0}^2 & \cdots & \sigma_{0,m}^2 & \cdots & \sigma_{0,M-1}^2 \\ \vdots & & \vdots & & \vdots \\ \sigma_{n,0}^2 & \cdots & \sigma_{n,m}^2 & \cdots & \sigma_{n,M-1}^2 \\ \vdots & & \vdots & & \vdots \\ \sigma_{N-1,0}^2 & \cdots & \sigma_{N-1,m}^2 & \cdots & \sigma_{N-1,M-1}^2 \end{bmatrix}, \tag{1}$$

where M and N are the number of angular and circular sections, respectively, and $\sigma_{n,m}^2$ is the variance of the pixels that fall into the segment (n, m) and computed as follows:

$$\sigma_{n,m}^2 = \frac{1}{A(n,m)K}\sum_{k=0}^{K-1}\sum_{\theta=\theta_m}^{\theta_{m+1}}\sum_{\rho=\rho_n}^{\rho_{n+1}}\left(\text{VOP}_k(\rho,\theta)-\mu_{n,m}\right)^2,$$
$$\mu_{n,m} = \frac{1}{A(n,m)K}\sum_{k=0}^{K-1}\sum_{\theta=\theta_m}^{\theta_{m+1}}\sum_{\rho=\rho_n}^{\rho_{n+1}}\text{VOP}_k(\rho,\theta), \tag{2}$$

where K is the number of the temporal instances of the video object, VOP_k is the binary shape mask of the video object plane (VOP) at kth instant, $\text{VOP}_k(\rho, \theta)$ is the value of the binary shape mask in VOP_k at the (θ, ρ) position in the polar coordinate system centered at the mass center of VOP_k, $A(n, m)$ is the area, θ_m is the start angle, and ρ_m is the start radius of the angular circular segment (n, m), and they are defined as

$$A(n,m) = \frac{\pi\left(\rho_{n+1}^2-\rho_n^2\right)}{M},$$
$$\theta_m = m\times\frac{2\pi}{M}, \qquad \rho_n = n\times\frac{\rho_{\max}}{N}, \tag{3}$$

where M and N are the number of angular and circular sections, respectively and $\rho_{\max}$ is found by

$$\rho_{\max} = \max_{\text{VOP}_k\in\text{VO}}\left\{\rho_{\text{VOP}_k}\right\}, \tag{4}$$

where VOP_k is the kth instant of the video object and ρ_{VOP_k} is the radius of the tightest circle around the VOP_k that is centered at the mass center of VOP_k.

The proposed descriptor is scale invariant since the number of angular and circular segments is the same for all video objects, and the size of each segment is scaled with $\rho_{\max}$. We attain an approximate rotation invariance of the descriptor by employing an appropriate query matching method similar to the one used for matching the contour-based shape descriptor in MPEG-7 [14]. That is, we provide the rotation invariance by reordering the feature matrix **R** so that the angular segment with the largest variance is in the first column of **R**. This is achieved by first summing the columns of the feature matrix **R** to obtain the $1 \times M$ projection vector $\vec{A}$ and then finding the maximum element of $\vec{A}$, which corresponds to the angular segment m_L that has the largest variance. Finally, we circularly shift to the left the columns of **R** by m_L to obtain a rotation invariant feature vector.

The trade-offs associated with using different numbers of angular and circular segments for this descriptor are presented in Section 4.

3. ART-BASED LOCAL MOTION DESCRIPTOR

Employing angular radial transform (ART)-based shape descriptors is an efficient way to retrieve shape information as they are easy to extract and match. Consequently, an ART-based descriptor was recently adopted by MPEG-7 [14]. Here, we propose to use the variance of the ART coefficients, computed for each object plane of a video object, as a local motion descriptor. As the ART descriptors describe the region of a shape, different than their contour-based counterparts such as curvature scale-space and Fourier descriptors, they are capable of representing holes and unconnected regions in the shape. Therefore, our proposed ART-based descriptor captures a large variety of shape region deformations caused by the local motion. The ART transform is defined as [14]

$$F_{nm} = \int_0^{2\pi}\int_0^1 V_{nm}(\rho,\theta)f(\rho,\theta)\rho\, d\rho\, d\theta, \tag{5}$$

where F_{nm} is an ART coefficient of order n and m, $f(\rho, \theta)$ is the binary shape map in polar coordinates, and $V_{nm}(\rho, \theta)$ is the ART basis function, which is separable along the angular and radial directions as follows:

$$V_{nm}(\rho,\theta) = A_m(\theta)R_n(\rho). \tag{6}$$

The angular and radial basis functions are given by

$$A_m(\theta) = \frac{1}{2\pi}e^{jm\theta}, \qquad R_n(\rho) = \begin{cases}1, & n = 0,\\ 2\cos(\pi n\rho), & n \neq 0.\end{cases} \tag{7}$$

The discrete ART coefficients of a binary shape map are found as follows. First, the size of the binary shape data is normalized by a linear interpolation to a predefined width W and height H, to obtain the size invariant shape map $I(x, y)$. The mass center of the binary shape map is aligned with the center of $I(x, y)$, that is, $I(W/2, H/2)$. Then, the discrete ART coefficients of the shape map of the object plane k (VOP_k) are computed by

$$F_{nm}(\text{VOP}_k) = \sum_{x=-W/2}^{W/2}\sum_{y=-H/2}^{H/2} V_{nm}\left(\sqrt{x^2+y^2}, \arctan\frac{y}{x}\right)$$
$$\times I_{\text{VOP}_k}\left(x+\frac{W}{2}, y+\frac{H}{2}\right). \tag{8}$$

The ART coefficients of the individual object planes are rotation variant. When ART coefficients are employed for still shape retrieval, the magnitude of the ART coefficients are employed for rotation invariance. Since we would like to capture any rotational changes that may be present in the shape of the video object when computing the variances in the ART coefficients, we employ the complex ART coefficients. The final ART-based local motion descriptor is defined as the magnitude of the complex variance computed over time, which is rotation invariant.

Because the area of the object shape is normalized for size prior to computing the ART coefficients, the local motion descriptor captures the real deformations of the shape, and it is robust to changes in the area of the video objects due to the events such as camera zooming, partial occlusion, and so on. If it is desired by the application that the motion descriptor capture such events, the size normalization of the descriptor should be done with respect to the largest object plane of the video object. The retrieval performance results of this descriptor, obtained by using a various number of angular and radial functions, are presented in Section 4.

4. EXPERIMENTAL RESULTS

4.1. Performance evaluation

We present our retrieval results by utilizing the normalized modified retrieval rank (NMRR) measure used in the MPEG-7 standardization activity [28]. NMRR not only indicates how much of the correct items are retrieved, but also how highly they are ranked among the retrieved items. NMRR is given by

$$\text{NMRR}(n) = \frac{\left(\sum_{k=1}^{\text{NG}(n)} \text{Rank}(k)/\,\text{NG}(n)\right) - 0.5 - \text{NG}(n)/2}{K + 0.5 - 0.5^{*}\,\text{NG}(n)}, \tag{9}$$

where NG is the number of ground truth items marked as similar to the query item and Rank(k) is the ranking of the ground truth items by the retrieval algorithm, where K is equal to min(4* NG(q), 2*GTM) where GTM is the maximum of NG(q) for all the queries. The NMRR is in the range of [0 1] and the smaller values represent a better retrieval performance. ANMRR is defined as the average NMRR over a range of queries.

4.2. Retrieval performance

Here, we demonstrate the performance of each of our proposed local motion descriptors. Our database contains over 20 arbitrarily shaped video objects, coded in 2 to 3 different spatial resolutions, each resulting in an MPEG-4 object database of over 50 bit streams. The ANMRR values presented in this section are obtained by averaging the retrieval results of 12 query video objects that have a large variety of local motions. The ground truth objects are decided by having three human subjects rank the video objects for their local motion similarity to the query video objects. The similarity distance between two shapes is measured by computing the Euclidean distance on their local motion descriptors.

Retrieval performance results using the ACLM descriptor with various numbers of angular and circular segments is presented in Figure 1. Note that smaller ANMRR values represent a better retrieval performance. Employing a large number of angular and circular bins generally results in a better retrieval performance but with the cost of more bits required to represent the descriptor. The highest retrieval rates (i.e., lowest ANMRR) here are obtained by using 6 angular and 3 circular segments (ANMRR = 0.090) and 8 angular and 2 circular segments (ANMRR = 0.089).

Some query examples using 6 angular and 3 circular segments are presented in Tables 1 and 2. Note that the dimensions given in the parentheses are not the dimensions of the video objects, but the resolutions of the video sequences from which they are extracted. The dimensions of the video objects are different for each plane of the video object. One important point to note is that, because of the simple upsampling/downsampling methods used to obtain various resolutions of the same video objects, the different resolutions of the same objects are not likely to have exactly the same shapes. Thus, even though our descriptor is scale invariant, the query distances corresponding to the different resolutions of the same object may not be identical.

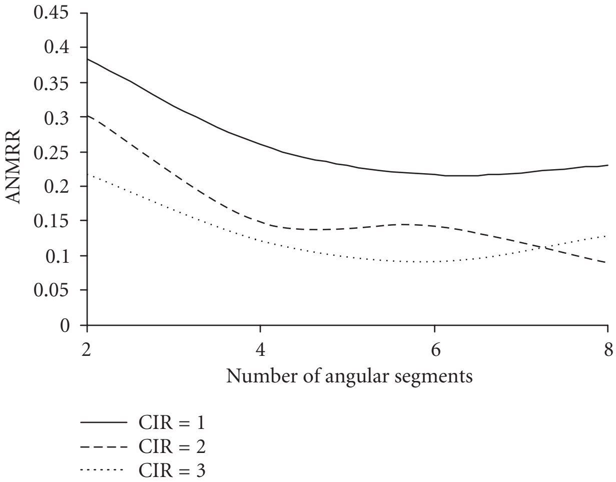

FIGURE 1: Retrieval results of the ACLM descriptor obtained by using various numbers of angular and circular (CIR) segments.

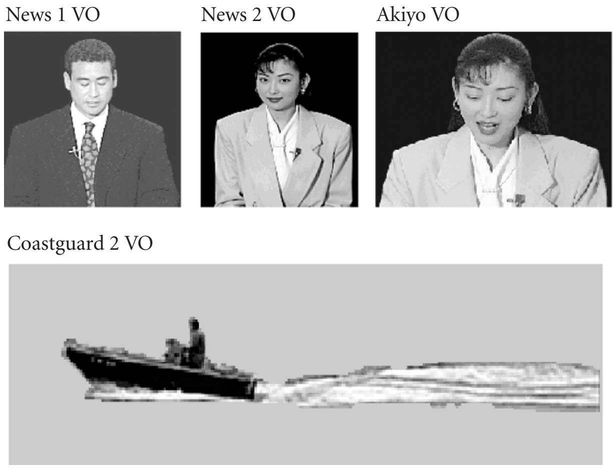

FIGURE 2: The video objects classified as being similar in terms of their local motion to the query video object News 1.

The first query, shown in Figure 2,[1] is a very low-motion anchorperson video object, News 1, which is coded in two different resolutions in our database. As presented in Table 1,

[1] The query and database items presented in this section are video objects, and the illustration given in the figures are some representative VOPs of these objects.

Table 1: Local motion retrieval results for the News 1 video object query.

Rank	Video object	Query distance
1	News 1 (360 × 240)	0.00
2	News 1 (180 × 120)	6.68
3	Akiyo (360 × 240)	11.07
4	News 2 (360 × 240)	12.55
5	Akiyo (180 × 120)	14.06
6	News 2 (180 × 120)	19.52
7	Coastguard 2 (352 × 288)	27.12
8	Coastguard 2 (176 × 144)	27.63
9	Coastguard 2 (528 × 432)	27.68

Table 2: Local motion retrieval results for the Hall Monitor 1 video object query.

Rank	Video object	Query distance
1	Hall Monitor 1 (360 × 240)	0.00
2	Hall Monitor 1 (540 × 360)	2.89
3	Hall Monitor 1 (180 × 120)	10.23
4	Hall Monitor 2 (180 × 120)	46.85
5	Hall Monitor 2 (360 × 240)	50.25
6	Hall Monitor 2 (540 × 360)	50.31
7	Fish 1 (352 × 240)	84.59
8	Stefan (176 × 144)	90.31
9	Stefan (352 × 244)	90.80

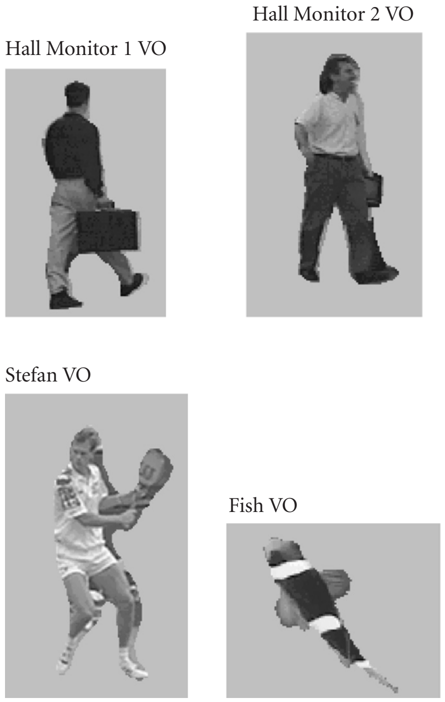

Figure 3: The video objects classified as being similar in terms of their local motion to the query video object Hall Monitor 1.

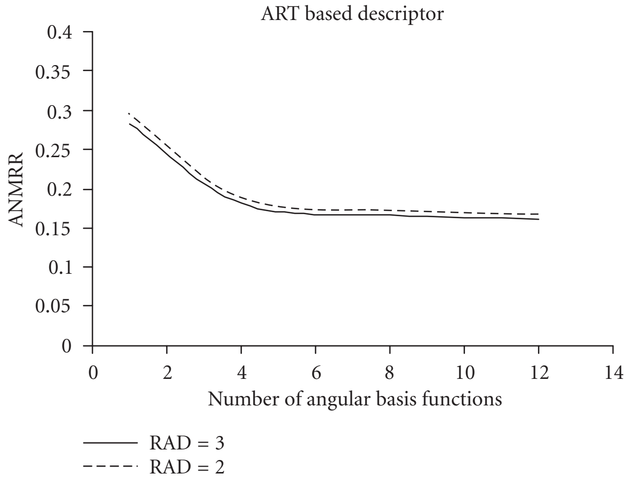

Figure 4: Retrieval results of the ART-based local motion descriptor obtained by employing different number of angular and radial (RAD) basis functions.

using the ACLM descriptor, the two different resolutions of the News 1 video object are retrieved as the first two items. The other highly ranked two anchorperson video objects, illustrated in Figure 2, are also very low in motion. The Coastguard video object, ranked 7th, 8th, and 9th, is also an object without any articulated parts (a boat object and its waves) and with moderate local motion. Our second query, Hall Monitor 1, is the video object of a walking man captured by a surveillance camera as shown in Figure 3. The query results for this object are presented in Table 2. The three different resolutions of the video object are ranked the highest, and another walking man video object from the same sequence, Hall Monitor 2, is ranked immediately after. The fish object, which has large moving fins and a tail as depicted in Figure 3, is ranked 6th. The different resolutions of a video object that contain a person playing tennis are ranked 8th and 9th. As can be seen from these query examples, the ACLM descriptor successfully classifies the local motion of the video objects.

The number of angular and radial functions of the ART descriptor determines how accurately the shape is represented. Considering that the video object shapes, different than trademark shapes for example, generally do not contain much detail, using a small number of basis functions to represent the shape maps would be sufficient and result in a more compact descriptor. Representation with a small number of basis functions also makes the descriptor more robust to the potential segmentation errors. The retrieval performance achieved by using different number of angular and radial functions is presented in Figure 4. As can be observed from the table, employing 4 angular and 2 radial basis functions offers a good trade-off between the retrieval performance (ANMRR = 0.181141) and the compactness of the descriptor.

5. CONCLUSIONS

In this paper, we proposed two local motion descriptors for the retrieval of video objects. As presented in Section 4, the ranking obtained by employing our descriptors closely matches with the human ranking. According to the ANMRR scores obtained, the ACLM descriptor offers a better retrieval rate than the ART-based descriptor. Given that each descriptor value is quantized to [0 255] range, ACLM descriptor requires 16 bytes and the ART-based descriptor requires 8 bytes to represent. ACLM descriptor is less computationally complex to extract. Nevertheless, if the ART coefficients of the video object is already computed and attached to the video objects as metadata for shape retrieval, then the extra computations required to extract the local motion descriptors based on the ART coefficients are minimal. Depending on the application, either of the proposed descriptors could be used for efficient video object retrieval by local motion.

REFERENCES

[1] A. Smolic, M. Hoeynck, and J.-R. Ohm, "Low-complexity global motion estimation from P-frame motion vectors for MPEG-7 applications," in *Proc. IEEE International Conference on Image Processing*, vol. 2, pp. 271–274, Vancouver, BC, Canada, September 2000.

[2] H. J. Zhang, Y. L. Chien, and S. W. Smoliar, "Video parsing and browsing using compressed data," *Multimedia Tools and Applications*, vol. 1, no. 1, pp. 89–111, 1995.

[3] R. R. Wang and T. Huang, "Fast camera analysis in MPEG domain," in *Proc. IEEE International Conference on Image Processing*, pp. 24–28, Kobe, Japan, October 1999.

[4] M. Shah, K. Rangarajan, and P. S. Tsai, "Motion trajectories," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 23, no. 4, pp. 1138–1150, 1993.

[5] R. Brunelli, O. Mich, and C. Modena, "A survey on video indexing," IRST Technical Report 9612-06, Istituto per la Ricerca Scientifica e Tecnologica, Trento, Italy, 1996.

[6] H. Sawhney, S. Ayer, and M. Gorkani, "Model-based 2D&3D dominant motion estimation for mosaicing and video representation," in *Proc. International Conf. on Computer Vision*, pp. 583–590, Boston, Mass, USA, June 1995.

[7] A. M. Ferman, B. Günsel, and A. M. Tekalp, "Motion and shape signatures for object-based indexing of MPEG-4 compressed video," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 4, pp. 2601–2604, Munich, Germany, April 1997.

[8] J. Little and Z. Gu, "Video retrieval by spatial and temporal structure of trajectories," in *SPIE Storage and Retrieval for Media Databases*, vol. 4315, San Jose, Calif, USA, January 2001.

[9] K. Rangarajan, W. Allen, and M. Shah, "Matching motion trajectories using scale-space," *Pattern Recognition*, vol. 26, no. 4, pp. 595–610, 1993.

[10] W. Chen and S. F. Chang, "Motion trajectory matching of video objects," in *SPIE Storage and Retrieval for Media Databases*, vol. 3972, pp. 544–553, San Jose, Calif, USA, January 2000.

[11] E. Ardizzone and M. La Cascia, "Automatic video database indexing and retrieval," *Multimedia Tools and Applications*, vol. 4, no. 1, pp. 29–56, 1997.

[12] S. F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "A fully automated content-based video search engine supporting spatiotemporal queries," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 602–615, 1998.

[13] Y. Deng and B. S. Manjunath, "NeTra-V: Toward an object based video representation," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 616–627, 1998.

[14] ISO/IEC JTC1/SC29/WG11, ISO/IEC 15938-3:2002, "Information technology—Multimedia content description interface—Part 3: Visual," 2002.

[15] M. Abdel-Mottaleb, N. Dimitrova, L. Agnihotri, et al., "MPEG-7: A content description standard beyond compression," in *IEEE 42nd Midwest Symposium on Circuits and System*, vol. 2, pp. 770–777, Las Cruces, NM, USA, August 1999.

[16] G. Stalidis, N. Maglaveras, A. Dimitriadis, and C. Pappas, "Modeling of cardiac motion using wavelets: comparison with Fourier-based models," in *Proc. IEEE Computers in Cardiology*, pp. 733–736, 1998.

[17] H. Miyamori and S. Iisaku, "Video annotation for content-based retrieval using human behavior analysis and domain knowledge," in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 320–325, Grenoble, France, March 2000.

[18] J. Hoey and J. Little, "Representation and recognition of complex human motion," in *Proc. IEEE Computer Vision and Pattern Recognition*, vol. 1, pp. 752–759, Hilton Head, SC, USA, June 2000.

[19] H. Fujiyoshi and A. J. Lipton, "Real-time human motion analysis by image skeletonization," in *Proc. IEEE Workshop on Applications of Computer Vision*, pp. 15–21, Princeton, NJ, USA, October 1998.

[20] J. Little and J. E. Boyd, "Recognizing people by their gait: The shape of motion," *Videre: Journal of Computer Vision Research*, vol. 1, no. 2, pp. 2–32, 1998.

[21] B. Heisele and C. Woehler, "Motion-based recognition of pedestrians," in *Proc. IEEE International Conference on Pattern Recognition*, vol. 2, pp. 1325–1330, Brisbane, Australia, August 1998.

[22] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 781–796, 2000.

[23] "Z-Cam from 3DV Systems," 2002, http://www.3dvSystems.com.

[24] "Electronic perception technology from Canesta," 2002, http://www.canesta.com.

[25] ISO/IEC JTC1/SC29/WG11, "Coding of audio-visual objects: Video," 1999.

[26] D. Zhong and S. F. Chang, "AMOS: An active system for MPEG-4 video object segmentation," in *Proc. IEEE International Conference on Image Processing*, pp. 4–7, Chicago, Ill, USA, October 1998.

[27] H. T. Nguyen, M. Worring, and A. Dev, "Detection of moving objects in video using a robust motion similarity measure," *IEEE Trans. Image Processing*, vol. 9, no. 1, pp. 137–141, 2000.

[28] ISO/IEC JTC1/SC29/WG11, "Description of core experiments for MPEG-7 color/texture descriptors," doc no. N3090, December 1999.

Berna Erol received the B.S. degree from the Istanbul Technical University, Istanbul, Turkey, in 1994, and the M.A.S. and Ph.D. degrees from the Department of Electrical and Computer Engineering, the University of British Columbia, Vancouver, Canada, in 1998 and 2002, respectively. Since September 2001, she has been with the Multimedia Document Analysis Group at Ricoh California Research Center as a Research Scientist. Her research interests include multimedia signal processing and communications, image and video compression, object-based video representations, and content-based retrieval and analysis. She has coauthored more than twenty journal papers, conference papers, and book chapters.

Faouzi Kossentini received the B.S., M.S., and Ph.D. degrees from the Georgia Institute of Technology, Atlanta, USA, in 1989, 1990, and 1994, respectively. He is presently an Associate Professor in the Department of Electrical and Computer Engineering at the University of British Columbia, where he is involved in research in the areas of signal processing, communications, and multimedia, and more specifically in subband/wavelet image transformation, quantization, audiovisual signal compression and coding, channel error resilience, joint source and channel coding, image and video communication, and image analysis. He has coauthored more than one hundred and thirty journal papers, conference papers, and book chapters. Dr. Kossentini is a senior member of the IEEE. He has served as a Vice General Chair of the ICIP-2000, and he has also served as an Associate Editor of the IEEE Transactions on Image Processing and the IEEE Transactions on Multimedia.