MPEG-7 Visual Motion Descriptors

Sylvie Jeannin and Ajay Divakaran

***Abstract*—This paper describes tools and techniques for representing motion information in the context of MPEG-7 standardization for multimedia description interfaces. It first gives an overview of the current organization of the set of MPEG-7 motion descriptions, then illustrates this by presenting two of them, motion activity and motion trajectory, in more detail. It explains how to extract them from the content, how to express them in a compact way, and illustrates their use in concrete applications scenarios.**

***Index Terms*—Motion activity, motion representation, motion trajectory, MPEG-7, video indexing.**

I. Introduction

THE MOTION features of a video sequence provide the easiest access to its temporal dimension, and are hence of key significance in video indexing. Camera motion-estimation techniques, trajectory-matching techniques, and aggregated motion-vector histogram-based techniques form the core of past work in video indexing using motion characteristics[1],[2],[3] [6]–[14], [16], [18].

A huge volume of information is required to express the motion field of a typical video sequence, even when the motion field is coarse. Hence, the main aim of motion-based indexing is to capture essential motion characteristics from the motion field into concise and effective descriptors. Such indexing techniques span a wide range of computational complexity, since the motion fields can be sparse or dense, and the processing of the motion fields ranges from simple to complex. Motion-feature extraction in the MPEG-$1/2$ compressed domain has been popular because of the ease of extraction of the motion vectors from the compressed bitstream. However, since compressed domain motion vectors constitute a sparse motion field, they cannot be used for computing descriptors that require dense motion fields.

In combination with indexing based on still-image features such as color, texture, etc. (see [26] on color and texture visual descriptors), motion-based indexing has been shown to significantly improve the performance of similarity-based video retrieval systems [16]. Motion-based indexing is also useful by itself. It enables motion-based queries, useful in contexts where motion has a rich meaning. For instance, in tennis, the ball's trajectory allows discrimination between types of rallies (at the net or not), or in basketball, camera motion can help discriminate between the two teams by finding out whether the basket being attacked is on the right-hand side or on the left hand side. Motion descriptions can also be the basis for complementary functionalities, such as video hyperlinking based on trajectories [19], [21], refined browsing based on motion characteristics (detailed in Section III-C), or refinement of table of content [20], [21].

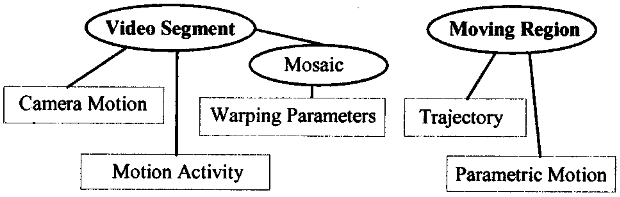

Fig. 1. MPEG-7 motion descriptions.

The rest of the paper is organized as follows. Section II provides an overview of the current set of MPEG-7 motion descriptors. In Sections III and IV, we illustrate this further by presenting two of these descriptors in more detail: Motion Activity in Section III and Motion Trajectory in Section IV. Section V concludes the paper.

II. Overview of Motion Descriptors

The motion descriptors that have currently been selected by MPEG-7 cover the range of complexity and functionality mentioned in the introduction, enabling MPEG-7 to support a broad range of applications. The over-arching guiding principle is to maintain simple extraction, simple matching, concise expression and effective characterization of motion characteristics.

They are organized as follows (see also Fig. 1) for a VideoSegment, i.e., a group of frames in a video.

1) Its overall activity or pace of motion, which often denotes its level of action, is captured by *Motion Activity*, further detailed in Section III of this paper.
2) The movement of the camera or of the virtual view-point on the scene is described by *Camera Motion*. This descriptor expresses which type(s) of camera motion is/are present in the segment among all possible camera motion types. It also gives their amplitude, and either their precise time localization in the segment or their relative importance in terms of duration (e.g., zoom in 30% of the time). Camera motions often mimic viewers' focus of attention, and can thus be used in certain contexts to discriminate events (e.g., actions in a basket game), or to refine tables of content [20]. More details can be found in [21], [15].
3) The global motions, which relate each frame to the segment's panorama (or Mosaic, largely studied in the literature [24], [25] and recently standardized by MPEG-4[4] [22] using the terminology "Sprite"), are

---

Manuscript received September 14, 2000; revised March 13, 2001.

S. Jeannin is with Philips Research USA, Briarcliff Manor, NY 10510 USA (e-mail: sylvie.jeannin@philips.com).

A. Divakaran is with Mitsubishi Electric Research Laboratories, Murray Hill Laboratory, Murray Hill, NJ 07974 USA (e-mail: ajayd@merl.com).

Publisher Item Identifier S 1051-8215(01)04989-8.

[1] Virage system [Online]. Available: http://www.virage.com

[2] QBIC system [Online]. Available: http://www.qbic.almaden.ibm.com

[3] VideoQ system [Online]. Available: http://www.ctr.columbia.edu/videoq/

[4] MPEG-7 CD [Online]. Available: http://www.cselt.it/mpeg/

captured by the *Warping Parameters*. These parameters are the coefficients of the parametric motion model used to define the panorama-to-frame relation (translational, pan-and-zoom, affine, perspective, or parabolic) [2]. Their transmission enables functionalities related to mosaics to be provided at the user end, while avoiding their complex computation. For a MovingRegion (i.e., an arbitrary-shaped spatiotemporal entity).

4) Its coarse displacement, namely successive positions in time of one selected representative point for the region such as its center of mass, is captured by *Motion Trajectory*, further detailed in Section IV of this paper.
5) More detailed information about its displacement is captured by *Parametric Motion*. This descriptor uses the same motion model and syntax as the *Warping Parameters* described above, and enables to retrieve objects of similar motions undergoing rotations or deformations, that are not captured by *Motion Trajectory*.

In the following, we illustrate this further by presenting two of the descriptors in more detail, one relating to segment motion, the other one to object motion. For each of them, we highlight the notions that are common to all motion descriptions, and give details on the concepts involved: the syntax, the algorithms needed to use them, and some applications and conditions of usage.

III. MOTION ACTIVITY

A human watching a video or animation sequence perceives it as being a slow sequence, fast-paced sequence, action sequence, etc. The activity descriptor [2] captures this intuitive notion of "intensity of action" or "pace of action" in a video segment. Examples of high "activity" include scenes such as "goal scoring in a soccer match" or "a high-speed car chase," while examples of low action shots are "news reader shot," "an interview scene," etc. Video content, in general, spans the gamut from high to low activity. This descriptor enables us to accurately express the activity of a given video sequence/shot and comprehensively covers the aforementioned gamut. It can be used in diverse applications such as content re-purposing, surveillance, fast browsing, video abstracting, video editing, content-based querying, etc.

A. Description

To more efficiently enable the applications mentioned above, we need additional attributes of motion activity. Thus, motion activity includes the following attributes.

Intensity of Activity—This is expressed by an integer lying in the range [1]–[5]. A high value of intensity indicates high activity while a low value of intensity indicates low activity.

Direction of Activity—While a video shot may have several objects with differing activity, we can often identify a dominant direction. The Direction parameter expresses the dominant direction of the activity if any. It is expressed by a three-bit integer that has a value corresponding to any of eight equally spaced directions.

Spatial distribution of Activity—The spatial distribution of activity [1] indicates whether the activity is spread across many regions or restricted to one large region. It is an indication of the number and size of "active" regions in a frame. For example, a talking head sequence would have one large active region, while an aerial shot of a busy street would have many small active regions. The spatial distribution parameter is expressed by three integers using a total of 16 bits.

TABLE I
QUANTIZATION THRESHOLDS FOR MPEG-1 VIDEO

Activity Value	Range of σ (Std. Dev. of motion vector magnitude)
1	$0 \le \sigma < 3.9$
2	$3.9 \le \sigma < 10.7$
3	$10.7 \le \sigma < 17.1$
4	$17.1 \le \sigma < 32$
5	$32 \le \sigma$

Temporal Distribution of Activity—The temporal distribution of activity expresses the variation of activity over the duration of the video segment per shot, i.e., whether the activity is sustained throughout the duration of the sequence or is confined to a part of the duration. The temporal distribution parameter is expressed by five 6-bit integers.

In the following section, we detail the extraction process for the intensity attribute of the descriptor. Note that it is the only essential attribute of the descriptor. The other three are optional. Thus, the descriptor can range from 3 to 52 bits in size.

B. Extraction

Since the intensity of motion activity is a subjective measure of the pace or action in a video segment, we first need to construct a ground truth. We can then assess the performance of automatically computed measures of intensity of motion activity by finding their fidelity to the ground truth (see [3], [5] for detailed investigations of this topic).

Construction of Ground Truth: Our ground truth data set consists of 637 video segments chosen from the MPEG-7 test set. We attempt to maximize the diversity of content while maintaining a manageable size. We use human subjects to rate the intensity of motion activity of each of the segments on a five notch scale *viz.*: 1) very low intensity; 2) low intensity; 3) medium intensity; 4) high intensity; and 5) very high intensity. We then average the ratings across the subjects to arrive at the ground truth for each of the video segments.

Automatic Measurement of Motion-Activity Intensity: With compressed video, motion vectors provide the easiest approach to the gross motion characteristics of the video segment. Since motion-vector magnitude is an indication of the magnitude of the motion itself, it is natural to use statistical properties of the motion-vector magnitude of macroblocks to measure intensity of motion activity. We find that both the standard deviation and the average of the motion-vector magnitude reasonably match the ground truth after proper quantization. However, we find that the standard deviation of the motion-vector magnitude provides a slightly better approximation of the ground truth, and hence use quantized standard deviation of the motion-vector magnitude to compute the intensity of motion activity (see Table I for the thresholds used for quantization of the standard deviation).

Fig. 2. Video browsing—extraction of the ten most “active” video segments in a news program.

Note that these thresholds can be normalized linearly with respect to frame size.[4] The threshold values have been obtained by optimizing the fidelity to the ground truth. However, they work well for all kinds of content and encoding conditions because of the diversity of the ground truth data set.

C. Usage and Applications

Video Browsing: The motion-activity intensity descriptor enables selection of the video segments of a program based on intensity of motion activity. Fig. 2 illustrates the retrieval of the ten highest motion-activity intensity segments of a news program from Spanish TV, item V3 of the MPEG-7 test -set.

In this case, and with other news programs, such retrieval allows us to immediately locate sports segments. The descriptor can similarly help locate other desired parts of the program such as the “head and shoulders” segments.

Content-Based Querying of Video Databases: Another application is as a simple first stage of content filtering for content-based querying applications [4]. We can use motion activity to separate the high and low motion parts of the video sequence and then use color descriptors within each part. Since intensity of motion activity is a scalar, it is exceedingly simple to match. Hence, it lends itself well to use as a first stage content filter. For example, while looking for the news anchor, it would be useful to narrow down the search by filtering out all the shots of higher or lower action from the search space. Thus, on the news footage from Spanish TV mentioned earlier, we are able to narrow down the search space from 1200 to 100 using the motion-activity intensity before using a color descriptor over the reduced search space. We are thus able to speed up the content-based matching by a factor of four or more, compared to using the color descriptor on the entire search space.

D. Further Possibilities and Discussion

The aforementioned applications, such as video summarization, surveillance, video editing, content-repurposing, etc., can all capitalize on the gross motion categorization obtained with the motion-activity descriptor.

In short, the motion-activity descriptor is easy to extract, compact (hence easy to match), and enables a useful gross motion-based clustering of video content. Furthermore, it is robust to encoder parameter changes and other sources of noise.

IV. MOTION TRAJECTORY

Motion trajectory describes the displacements of objects in time, objects being defined as spatio-temporal regions whose trajectories are important for the given application.

It is a fairly high-level description, as the object position at each instant is given by one representative point. Such simplicity is relevant, as humans perceive objects motion at a high level [17]. We have also verified through Core Experiments that such high-level data are rich enough to enable most relevant applications (video hyperlinking [19], motion-based querying [23], etc).

A. Description

The trajectory model is a first- or second-order piecewise approximation along time, for each spatial dimension.

The core of the description is a set of *keypoints*, representing the successive spatio-temporal positions of the described object (positions of one representative point of the object, such as its center of mass). They are defined by their coordinates in space (2-D or 3-D) and time.

Additionally, interpolating parameters can be added to specify nonlinear interpolations between keypoints, using a second-order function of time [(3) and (4)]. By default, in the absence of interpolating data, linear interpolation is used [(1) and (2)].

First-order approximation (default interpolation)

$$f(t) = f_a + v_a(t - t_a) \tag{1}$$

with

$$v_a = \frac{f_b - f_a}{t_b - t_a}. \tag{2}$$

Second-order approximation

$$f(t) = f_a + v_a(t - t_a) + \frac{1}{2}a_a(t - t_a)^2 \tag{3}$$

with

$$v_a = \frac{f_b - f_a}{t_b - t_a} - \frac{1}{2}a_a(t_b - t_a) \tag{4}$$

where

a_a	interpolating parameter;
v_a, a_a	object's speed and acceleration, respectively, considered constant on $[t_a, t_b]$;
f_a, f_b	positions at times t_a and t_b.

Fig. 3 shows a trajectory obtained with this representation, instantiated by the parameters shown in bold (x_i, t_i, a).

Description Properties: The description has properties in terms of simplicity of use, scalability, flexibility, and compactness, which are typical of MPEG-7 descriptors.

1) *Simplicity of Use:* Targeted functionalities can be easily implemented based on the description. There is direct and independent access to positions, speeds, and accelerations, in each dimension (see also Section IV-C).
2) *Scalability:* Keypoints data can be used by itself; we can consider interpolation parameters as a refinement layer.

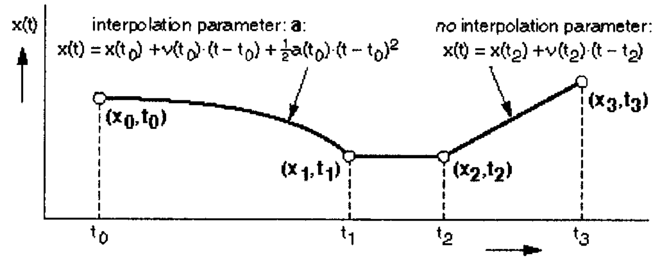

Fig. 3. Example of trajectory representation (one dimension).

3) *Flexibility:* The description can be adapted to fit various content and/or application: There is no constraint on numbers or temporal instances of keypoints. This enables systems to choose a total number of keypoints that provides the global precision/compactness that they require, and to choose keypoints separated by variable time intervals that match the local trajectory's smoothness.
4) *Variable Size/Compactness:* The description can be very compact and the tradeoff between size and approximation precision can be chosen depending on the context. The description size is roughly proportional to the number N of keypoints, and depends on the image size and on the coordinates resolution (around 150*N bits per keypoint for CIF images give a very good precision). For typical content and applications, 2 keypoints/s give good approximations.

B. Extraction

As we aim at describing the motion of a given object or region, we suppose that we know where it is located in the visual data, spatially and temporally. In MPEG-4, *alpha channels* specify spatio-temporal localizations of objects [22]: we assume that such an input is available.

If it is not the case, a segmentation, either in the image or in the motion-vector domain [9], may be performed. Segmentation processes are however complex. Besides, leaving the object definition to a nonspecified segmentation process, whose quality is unknown, is not rigorous. Different techniques may give different segmentation results, leading to different trajectories.

Extraction of Keypoints Data: Alpha-channels give direct access to spatio-temporal positions, using the following process.

1) Keypoints' time instants are initialized as successive frames time instants (at this step one keypoint per frame).
2) Then, for each time instant:
    a) the keypoint positions (x, y) are instantiated as the coordinates of the object's center of mass;
    b) interpolating parameters are calculated as the local second-order derivatives of the positions.

Selection of Keypoints and Functions: The description can be further adapted to the application by subsampling the obtained set of keypoints (i.e., merging adjacent approximation intervals) using the flexibility of the descriptor in terms of number of keypoints and time intervals length. Appropriate algorithms are not standardized by MPEG-7, and should be chosen and optimized based on the constraints of the application. For instance:

1) if the main constraints are memory and processing time, e.g., the indexing is done on the fly on the output of a tracking system, one can apply a sequential procedure. The system starts with an approximation interval which is the union of the first two intervals: $[t_0, t_1]U[t_1, t_2]$, made of the three first points. It sequentially adds points to this interval until the interpolation error exceeds a given threshold. This interval is kept, and the process repeats, from the next three instants;
2) if the main constraint is the size of the description, as in a database dealing with large amounts of data, the indexing being done off-line, it is more appropriate to apply a recursive procedure. The system starts with one interval containing all the points, and splits the interval into two at the position where the interpolation error is maximum. The splitting process is recursively repeated on each two new intervals either until the desired description size is reached, or until the error drops below a given threshold;
3) if the main constraint is simplicity, keypoints can be selected using regular time intervals sampling.

Complexity: Though the extraction is not normative, its complexity was analyzed and had to be reasonable. The complexity of the first two selection methods described in Section IV-B depends on the number of both the initial and final trajectory points [23]. If the time interval between keypoints is fixed, there is no complexity at all in the extraction process.

C. Usage

As for most MPEG-7 descriptions, the genericity of the description allows different usages.

Similarity Matching: A simple and flexible trajectory distance measure $M(T1, T2)$ is a linear weighting of distances between object positions M_P, speeds M_S, and accelerations M_A. It is given by the generic expression

$$M(T1,T2) = \frac{w_P M_P(T1,T2) + w_S M_S(T1,T2) + w_A M_A(T1,T2)}{w_P + w_S + w_A}$$

with

$$M_P(T1,T2) = \sum_i \frac{(x1_i - x2_i)^2 + (y1_i - y2_i)^2 + (z1_i - z2_i)^2}{\Delta t_i}$$

and with similar definitions for M_S and M_A. i is the index of the time interval and Δt_i is the duration of the ith time interval. The matching strategy is adapted to the application or to the user's interests by adjusting the weights (w_P, w_S, w_A).

When combined with other descriptors to retrieve similar video segments, trajectory descriptions have been shown to significantly improve retrieval performance [16].

Higher Level Queries: As our syntax gives direct and independent access to natural notions (positions, speeds and accelerations), many higher level queries can be easily processed. For instance:

1) for "*show me objects passing near this area*," useful for surveillance of a building with protected area, we simply test the keypoints' spatial positions values;

2) for "*show me objects moving faster than speed_limit*," useful for road monitoring, we calculate the speeds from the positions and time, and compare them to speed_limit.

Results on high-level queries can be found in [15].

Further Usages: Video Hyperlinking: A further example of trajectory usage is video hyperlinking. Rich interactive viewing experiences can be built if some predefined objects in videos contain links to other multimedia contents. Using several trajectories to represent the evolution of a few contour points of an object in time enables this efficiently [19]. It provides the required precision on the object spatio-temporal localization (an error of 20% is acceptable for hyperlinking) for a reasonable cost in terms of bit rate and processing complexity.

V. CONCLUSION

As shown in this paper, MPEG-7 provides a complete, coherent, and useful collection of motion descriptors that capture the different aspects of motion in videos with a broad range of precision. The basic objective is to provide useful concise descriptors that are easy to extract and match. We detailed in this paper how motion activity and motion trajectory both meet this objective. More generally, the results of the Core Experiments conducted in the MPEG-7 Visual group since March 1999 (some being given in [3], [19], [23]) show that this objective has been satisfied for all selected descriptors. Altogether, they facilitate a wide variety of applications besides the most obvious *viz.* the content-based querying of multimedia databases.

ACKNOWLEDGMENT

For motion activity, the authors would like to thank V.V. Vinod for initiating this work, K. Peker for useful suggestions and for generating some of the data, as well as the whole "Motion Activity Group" comprising V. V. Vinod, H. Sun, H. Kim, C.-S. Park, D.-G. Sim, N.-K. Kim, B. S. Manjunath, X. Sun, H. Shin, G. Rattinassababady, D. Manoranjan, L.-S. Fong, and T. C. Ling. For motion trajectory, the authors would like to thank A. She, T. Naveen, and A. Tabatabai for their help in defining the descriptor syntax, as well as the whole "Motion Trajectory Group" comprising O. Hori, T. Kaneko, J. Kim, K.-W. Lee, and K. Masukura.

Finally, the authors acknowledge the work of all MPEG members who contributed to the definition of all other motion descriptions.

REFERENCES

[1] A. Divakaran and H. Sun, "A descriptor for spatial distribution of motion activity," in *Proc. SPIE Conf. Storage and Retrieval from Image and Video Databases*, San Jose, CA, Jan. 2000, pp. 24–28.

[2] *MPEG-7 Visual Part of the XM 4.0*, ISO/IEC MPEG99/W3068, Dec. 1999.

[3] A. Divakaran, H. Sun, H. Kim, C. S. Park, B. S. Manjunath, X. Sun, H. Shin, and V. V. Vinod *et al.*, "Report on the MPEG-7 Core Experiment on the Motion Activity Feature," ISO/IEC MPEG99/m5030, Melbourne, Australia, Oct. 2000.

[4] A. Divakaran, A. Vetro, K. Asai, and H. Nishikawa, "Video browsing system based on compressed domain feature extraction," *IEEE Trans. Consumer Electron.*, Aug. 2000.

[5] K. A. Peker and A. Divakaran, "Automatic measurement of intensity of motion activity of video segments," in *Proc. SPIE Conf. Storage and Retrieval from Multimedia Databases*, San Jose, CA, Jan. 2001.

[6] A. Akutsu, Y. Tonomura, H. Hashimoto, and Y. Ohba *et al.*, "Video indexing using motion vectors," in *Proc. Visual Communications and Image Processing*, vol. 1818, 1992, pp. 1522–1530.

[7] E. Ardizzone *et al.*, "Video indexing using MPEG motion compensation vectors," *Proc. IEEE Int. Conf. Multimedia Computing and Systems*, 1999.

[8] J. W. Davis, "Recognizing Movement Using Motion Histograms," MIT Media Lab., Perceptual Comput. Section, Cambridge, MA, Tech. Rep. 487, Apr. 1998.

[9] N. Dimitrova and F. Golshani, "Motion recovery for video content analysis," *ACM Trans. Inform. Syst.*, vol. 13, no. 4, pp. 408–439, Oct. 1995.

[10] V. Kobla, D. Doermann, K. I. Lin, and C. Faloutsous, "Compressed domain video indexing techniques using DCT and motion vector information in MPEG video," in *Proc. SPIE Conf. Storage and Retrieval for Image and Video Databases V*, vol. 3022, 1997, pp. 200–211.

[11] E. Sahouria, "Video indexing based on object motion," M.S. thesis, Dept. of EECS, UC Berkeley, Berkeley, CA.

[12] Y. P. Tan, S. R. Kulkarni, and P. Ramadge, "Rapid estimation of camera motion from compressed video with application to video annotation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, pp. 133–146, Feb. 2000.

[13] Y. T. Tse and R. L. Baker, "Camera zoom/pan estimation and compensation fore video compression," in *Proc. SPIE Conf. Image Processing Algorithms and Techniques II*, Boston, MA, 1991, pp. 468–479.

[14] W. Wolf, "Key frame selection by motion analysis," in *Proc. ICASSP '96*, vol. II, 1996, pp. 1228–1231.

[15] S. Jeannin and R. Jasinschi *et al.*, "Motion descriptors for content-based video representation," *Signal Processing: Image Commun. J.*, vol. 16, no. 1–2, pp. 59–85, Sept. 2000.

[16] S. F. Chang and W. Chen *et al.*, "A fully automated content-based video search engine supporting multi-objects spatio-temporal queries," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 8, pp. 602–615, Sept. 1998.

[17] D. Marr, *Vision*. San Francisco, CA: Freeman, 1982.

[18] W. Niblack, D. B. Ponceleon, and D. Petkovic *et al.*, "Updates to the QBIC system," in *Proc. IS&T SPIE, Storage and Retrieval for Image and Video Databases VI*, vol. 3312, San Jose, CA, 1998, pp. 150–161.

[19] O. Hori and T. Kaneko, "Results of Spatio-Temporal Region DS Core/Validation Experiment," ISO/IEC MPEG99/m5414, Maui, HI, Dec. 1999.

[20] J. Llach and P. Salembier, "Analysis of video sequences: Table of contents and index creation," in *Proc. Int. Workshop Very Low Bit-Rate Video Coding (VLBV'99)*, Kyoto, Japan, Oct. 29–30, 1999, pp. 52–56.

[21] S. Jeannin and B. Mory, "Video motion representation for improved content access," *IEEE Trans. Consumer Electron.*, vol. 46, no. 3, pp. 645–655, Aug. 2000.

[22] *Information Technology—Coding of Audio-Visual Objects*, Standard MPEG4, Version 1, ISO/IEC 14 496.

[23] K. W. Lee, W. S. You, and J. Kim, "*Quantitative Analysis for the Motion Trajectory Descriptor in MPEG-7*," ISO/IEC MPEG99/m5400, Dec. 1999.

[24] R. Szeliski, "Image mosaicing for tele-reality applications," in *Proc. IEEE Workshop Applications of Computer Vision*, Dec. 1994, pp. 44–53.

[25] M. Irani and P. Anandan *et al.*, "Efficient representation of video sequences and their application," *Signal Processing: Image Commun.*, vol. 8, pp. 327–351, 1996.

[26] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada *et al.*, "Color and texture descriptors," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 11, pp. 703–715, June 2001.