Robust Color Histogram Descriptors for Video Segment Retrieval and Identification

A. Müfit Ferman, *Member, IEEE*, A. Murat Tekalp, *Senior Member, IEEE*, and Rajiv Mehrotra

***Abstract*—Effective and efficient representation of color features of multiple video frames or pictures is an important yet challenging task for visual information management systems. Key frame-based methods to represent the color features of a group of frames (GoF) are highly dependent on the selection criterion of the representative frame(s), and may lead to unreliable results. In this paper, we present various histogram-based color descriptors to reliably capture and represent the color properties of multiple images or a GoF. One family of such descriptors, called alpha-trimmed average histograms, combine individual frame or image histograms using a specific filtering operation to generate robust color histograms that can eliminate the adverse effects of brightness/color variations, occlusion, and edit effects on the color representation. We show the efficacy of the alpha-trimmed average histograms for video segment retrieval applications, and illustrate how they consistently outperform key frame-based methods. Another color histogram descriptor that we introduce, called the intersection histogram, reflects the number of pixels of a given color that is common to all the frames in the GoF. We employ the intersection histogram to develop a fast and efficient algorithm for identification of the video segment to which a query frame belongs. The proposed color histogram descriptors have been included in the recently completed ISO standard MPEG-7 after extensive evaluation experiments.**

***Index Terms*—Color descriptors, image/video databases, MPEG-7, video segment retrieval.**

I. Introduction

As THE SIZE of multimedia databases increases, it becomes necessary to represent groups of frames or shots in a video with effective and efficient descriptors. A generic mechanism for segmenting video into groups of frames is through the popular shot-based representation model [1], [2]. Once the shot boundaries in a video sequence are identified, it is customary to describe the visual and color content of shots using key frames and key frame histograms, respectively [3]. Although the key frame color histogram is a very simple and, depending on how the key frames are chosen, computationally inexpensive descriptor of color content of a shot, the color description it provides varies significantly with the selection criterion. Some methods simply pick from every shot one or more frames in predetermined temporal locations (e.g., the first and/or last frame), while others employ color- and/or motion-based criteria for appropriate key frame selection [4]. In order to avoid the variations in the color description of a shot due to the inherent arbitrariness of key frame selection, a more favorable approach is to consider the color content of all the frames within a shot for color histogram computation. To this effect, one may consider the cumulative (average) color histogram as an appropriate choice [5]. However, the average color histogram becomes vulnerable to outlier frames within a shot. Examples of such outlier frames are those with brightness variations (e.g., when background lighting changes or a sudden flash occurs), edit effects (e.g., fades and dissolves), and text/graphics overlays (e.g., in sports video or news), some of which are depicted in Figs. 1(a)–3(a). Thus, it is desirable to develop reliable color descriptors for a group of frames that are representative of the actual color content of the collection of frames but also are unaffected by the presence of outlier frames, which may skew the color representation unfavorably.

In this paper, we present a set of robust color histogram descriptors for representation of the color content of a group of frames. One of the main advantages of the proposed descriptors is that, unlike key frame color histograms, they are uniquely defined (up to some parameters) for a given set of video frames, and thus provide a consistent standard feature set across different systems. This property also enables objective evaluation of the performance of the descriptors in retrieval applications. The proposed robust histogram descriptors have been accepted into the recently completed international standard, MPEG-7 (formally "Multimedia Content Description Interface"), after a series of vigorous evaluation (core) experiments [6], [7]. The rest of the paper is organized as follows. Section II presents a short description of the group-of-frames representation model and data structure for video sequences. The proposed color histogram descriptors, their properties, and implementation details are presented in Section III. Sections IV and V highlight two specific applications for the proposed descriptors; namely, video segment retrieval and identification. Section VI is dedicated to concluding remarks and future research.

II. Group-of-Frames Representation Model for Video Sequences

This section presents a general framework for efficient description and representation of video sequences. A video se-

Manuscript received March 15, 2000; revised February 5, 2002. This work was supported by the National Science Foundation SIUCRC grant, a New York Science and Technology Foundation grant to the Center for Electronic Imaging Systems at the University of Rochester, and a grant from Eastman Kodak Company. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tsuhan Chen.

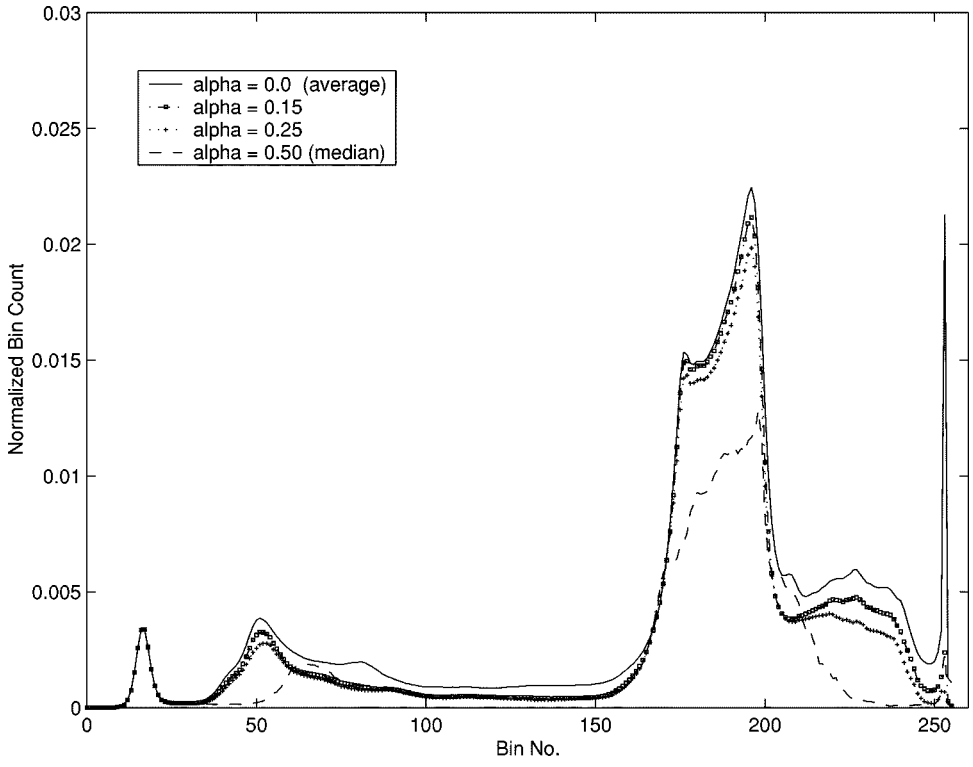
A. M. Ferman was with the Department of Electrical and Computer Engineering, University of Rochester, NY 14627-0126 USA. He is now with Sharp Laboratories of America, Inc., Camas, WA 98607-9489 USA (e-mail: mferman@sharplabs.com).

A. M. Tekalp is with the College of Engineering, Koç University, Istanbul, Turkey, and also with the Department of Electrical and Computer Engineering, University of Rochester, NY 14627-0126 USA (e-mail: tekalp@ece.rochester.edu).

R. Mehrotra is with Eastman Kodak Company, Rochester, NY 14650-1816 USA (e-mail: rajiv.mehrotra@kodak.com).

Publisher Item Identifier S 1057-7149(02)04778-4.

(a)



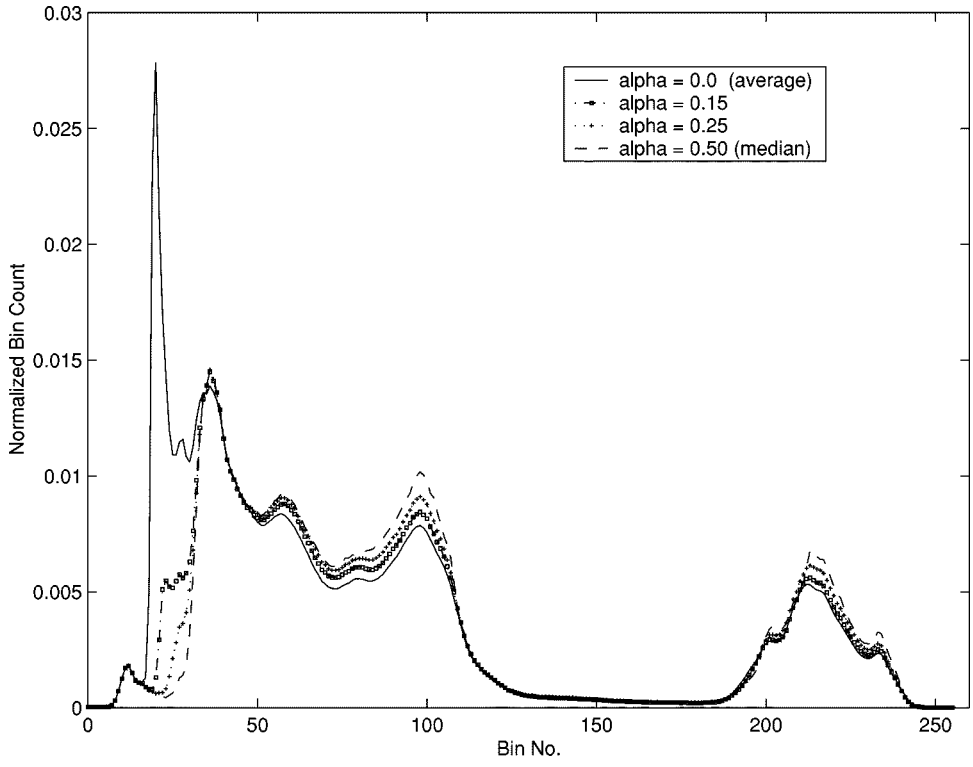
(b)

Fig. 1. (a) Sample shot from the MPEG-7 test sequence *lgerca_lisa_1*. Note how the camera roams freely over multiple areas and objects, and also the glare due to the poor quality of the camcorder. The corresponding alpha-trimmed average luminance histograms computed for different values of α are shown in (b). The effects of undesired color and brightness components on the GoF histogram are reduced by changing α.

quence is viewed as a set of "*groups of frames*," or GoFs, that are a collection of frames selected according to a certain criterion. The GoF framework is therefore more general and flexible than the popular shot-based approach, and well-suited for representing the sequential and hierarchical nature of video data. At the top of the hierarchy, the video stream is a single GoF; at the bottom, the smallest GoF is the frame. In between, shots, sub-shots, and groups of shots can all be considered as GoF instances. The resulting representation takes the form of an ordered tree structure, where each node corresponds to a distinct GoF. Several low-level (e.g., color, motion, and texture) and semantic (e.g., objects, events, concepts, and annotations) descriptors can be attached at each node of the tree. This general structure facilitates easy sharing of information between related components, and inherently supports access to data at various levels of semantic and syntactic abstraction. The GoF framework was first presented to MPEG-7 in response to a call for proposals for technology evaluation; since then, MPEG-7 has adopted a similar "segment" structure which shares many of the features of the proposed GoF model [8].

This paper does not address extraction or aggregation of GoFs. We note, however, that there exist several temporal segmentation methods in the literature, which can identify the shot boundaries and edit effects (fades, dissolves, etc.). The GoFs thus obtained can further be divided into sub-GoFs based on syntactic or semantic criteria, while high-level GoFs

(a)



(b)

Fig. 2. (a) Sample shot from the MPEG-7 test sequence *La Sombra de un Cipres es Alagarda* and (b) the corresponding set of alpha-trimmed average histograms. The average histogram exhibits a sharp peak at darker luminance values in the frames where the carriage occludes the camera. This effect is eliminated by choosing a different α for the GoF histogram.

can likewise be generated by merging lower level GoFs. In the rest of the paper, we concentrate on one of the syntactic visual descriptors, namely the GoF color histogram, that can be attached to various levels in the GoF hierarchy.

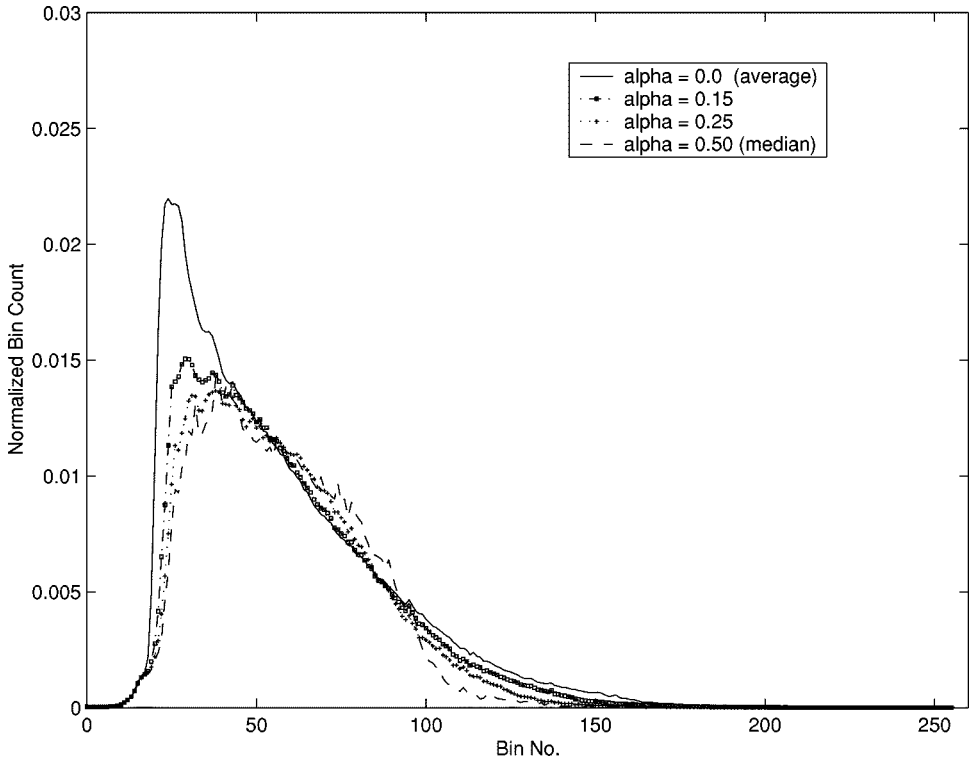
III. GoF Color Histogram Descriptors

In this section, we introduce a number of new histogram descriptors for GoF color representation. As noted in Section I, the simplest approach is to describe the color content of a GoF by the color histogram of a representative key frame (referred to as *key frame histogram*). First, we present a method to select the "optimum" key frame for the key frame histogram. Next, we define a family of *alpha-trimmed average histograms* as robust descriptors of color content of a GoF. Finally, we introduce the *intersection histogram* for the specific application of identification of the GoF in which a given query frame occurs. Here, the GoF color histogram definitions are provided for video sequences; however, the proposed descriptors can be directly extended to collections of still images, as long as a meaningful grouping or clustering of the images in a database is provided.

A. Key Frame Histogram

Key frame histogram refers to the color histogram of a representative frame of a GoF that is selected according to some criterion [3], [9]. Although this approach is very simple and (often)

(a)



(b)

Fig. 3. (a) Fade-in sequence from the MPEG-7 test sequence *Pepa y Pepe*, and (b) the corresponding alpha-trimmed average histograms. By increasing α, it is possible to reduce the effects of the dark frames in the fade and to emphasize the actual color/luminance content observed over the group of frames.

computationally the least expensive, the color description provided by the resulting histogram varies significantly with the key frame selection criterion. In the following, we propose a particular key frame selection method to avoid this arbitrariness at the cost of computational complexity. For an arbitrary frame l in a given GoF and its histogram H_l, we define the average error E_{H_l} as

$$E_{H_l} = \frac{1}{M} \sum_i ||H_i - H_l||, \qquad i = 1, \ldots, M \qquad (1)$$

where M denotes the number of frames in the GoF. The histogram H_r that minimizes (1) is the *optimum* representative histogram for the given distance measure.

While it presents a consistent approach to the key frame selection problem, the above method is computationally demanding, as it requires an exhaustive search over all the frames in a GoF to determine which one minimizes the average error. For a GoF of duration M, a total of $M(M-1)$ histogram comparisons need to be carried out to determine the optimum histogram, which can become prohibitive very quickly. Note also that this procedure always finds a single key frame for a GoF; hence, it best suits those GoFs with more-or-less uniform color content.

B. Alpha-Trimmed Average Histograms

A more suitable color descriptor than the key frame histogram is one that represents the cumulative color information of all frames within a GoF. The most straightforward way to achieve this is to accumulate all pixel color values from all frames within a GoF into a single histogram. Proper normalization of this cumulative histogram then yields the *average histogram*. Each bin j in the average histogram for the kth GoF is defined as [5], [10]

$$AvgHist_k(j) = \frac{1}{M} \sum_{i=b_k}^{e_k} H_i(j), \qquad j = 1, \ldots, B \qquad (2)$$

where H_i denotes the histogram of the ith frame, b_k and e_k denote the start and end frames of the given GoF, $M = (e_k - b_k + 1)$ is the number of frames in the GoF, and B is the total number of bins in the histograms. A potential problem with using sample averages to represent each bin value in the GoF histogram is the sensitivity of the mean operator to outliers. Given M samples, any data point has weight M^{-1} in the computation of the mean [11], and a deviant data value can lead to a biased sample average that is not representative of the entire set.

One way to obtain a robust color descriptor is to replace the sample average with the *sample median*, which can efficiently eliminate the outliers present in the data . For the kth GoF, each bin in the *median histogram* is given by

$$MedHist_k(j) = \text{median}\{H_{b_k}(j), H_{b_k+1}(j), \ldots, H_{e_k-1}(j), H_{e_k}(j)\}. \quad (3)$$

To compute the value in every bin in $MedHist$, the ascending list of frame histogram values is constructed for the duration of the GoF, and the median of this list is assigned to the corresponding bin in $MedHist$. When M, the number of frames in the kth GoF, is even, the median is defined as the average of the two center values in the ordered list.

An alternative approach for computing the GoF histogram is to define a family of *alpha-trimmed average histograms*, which is generated using the *trimmed mean* operator [11]. An alpha-trimmed average histogram is obtained by sorting the array of frame histogram values for each bin in ascending order and averaging only the central members of the ordered array. Each bin j in the alpha-trimmed average histogram is computed by

$$\alpha TrimHist_k(j, \alpha) = \frac{1}{M - 2 \cdot \lfloor \alpha M \rfloor} \sum_{m=\lfloor \alpha M \rfloor + 1}^{M - \lfloor \alpha M \rfloor} \hat{h}_j(m) \quad (4)$$

where $\lfloor \alpha M \rfloor$ denotes the largest integer not greater than αM, and $\hat{h}_j$ is the *sorted* array of frame histogram values for the jth bin. The trimming parameter α, $0 \leq \alpha \leq 0.5$, controls the number of data points excluded from the average computation. The trimmed mean operator reduces the contribution of aberrant data points to the estimate by discarding an equal number of samples at each end of the sorted series. Note that when $\alpha = 0$, the resulting histogram is equivalent to the sample mean, while $\alpha = 0.5$ corresponds to the sample median (when M is odd).[1] Figs. 1(b)–3(b) depict, for each of the sample GoFs in Figs. 1(a)–3(a), the GoF (luminance) histograms computed using different values of α. It is evident from these figures that the effects of undesired luminance and chrominance variations within the GoF can be eliminated, and the true color content of the GoF accentuated, by modifying α. These figures also highlight the fact that for optimal performance, the value of α should be determined individually for each GoF. Generally, unless strong luminance and/or chrominance variations are observed throughout a GoF, the average histogram (i.e., $\alpha = 0$) can be used to provide a reliable representation of the GoF color content, with minimal computational overhead. Otherwise, a nonzero value for the trimming parameter should be adopted in (4) to reduce or eliminate the effects of these variations.

The variance σ_μ^2 of the mean values of the luminance/chrominance components for each frame f in a GoF—$\mu_{Lum}^{(f)}$ and $\mu_{Chr}^{(f)}$, respectively—provides a sufficient and convenient cue for determining the appropriate value of α. A large variance implies fluctuations in luminance or chrominance throughout the GoF, signaling the possible presence of outlier frames. It is, therefore, appropriate to determine the value of α as a function of, and proportionally to, σ_μ^2; i.e., $\alpha = F(\sigma_\mu^2)$. Typical choices for $F(.)$ may be a linear function, with an (empirically-defined) offset, or a step function. If computation of the trimming parameter individually for each GoF is not possible, the properties of the video content can be taken into account to (intuitively) determine the value of α on a more global level, i.e., for all the GoFs in the available video content. For example, α can be set to zero or a similarly low value for static programs such as talk shows, while for content with more activity (e.g., movies, music programs, etc.) the value of α should be larger.

A potential concern about the use of alpha-trimmed average histograms is the increased computational complexity for $\alpha \neq 0$, because sorting needs to be performed for each histogram bin. Using a common algorithm like *quicksort*, the total number of comparisons required to sort the histograms of a video segment with M frames is $B \times O(M \log M)$, where B is the number of histogram bins. However, fast methods that have been developed for median filtering can be adopted to reduce these computational requirements [11], [13].

Alpha-trimmed average histograms provide an estimate of the color properties of all the frames within a GoF; hence, a natural way to determine the fidelity of these descriptors is in terms of the average error $E_{H_{GoF}}$. For a GoF of duration M, $E_{H_{GoF}}$ is defined as the average of the accumulated distances between each frame histogram H_i in the GoF and the GoF histogram H_{GoF}

$$E_{H_{GoF}} = \frac{1}{M} \sum_i ||H_i - H_{GoF}||. \quad (5)$$

Based on the distance norm used in computations, the average error E can take on various interpretations. For example

$$\text{If } ||H_i - H_{(.)}|| = \sum_j |H_i(j) - H_{(.)}(j)| \quad \text{then } E_{(.)} \equiv \text{mean absolute error } (MAE) \quad (6)$$

$$\text{If } ||H_i - H_{(.)}|| = \sum_j [H_i(j) - H_{(.)}(j)]^2 \quad \text{then } E_{(.)} \equiv \text{mean square error } (MSE). \quad (7)$$

It follows directly from statistics that when the error is computed using (6), the median histogram minimizes the error, while the minimum for (7) is attained with the average histogram; i.e.,

$$MAE_{\min} = \frac{1}{M} \sum_i \sum_j |H_i(j) - MedHist(j)| \quad (8)$$

$$MSE_{\min} = \frac{1}{M} \sum_i \sum_j [H_i(j) - AvgHist(j)]^2. \quad (9)$$

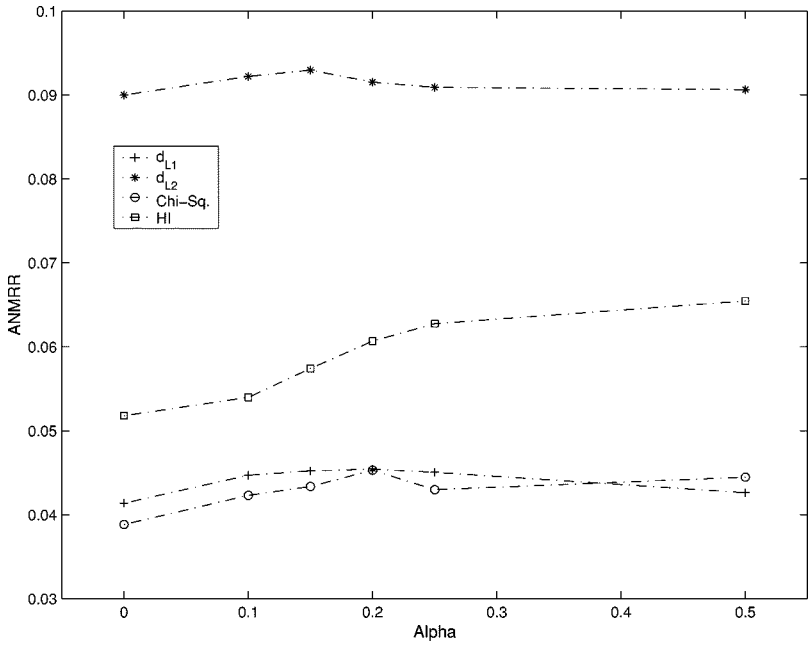
Comparison of $E_{H_{GoF}}$ and E_{H_r}, the minimum of (1), provides a consistent way to objectively assess the performance of key frame-based color representation methods and the proposed descriptors. Thus, for every GoF,

$$E_{H_r} \geq MAE_{\min} = E_{H_{med}} \quad (10)$$

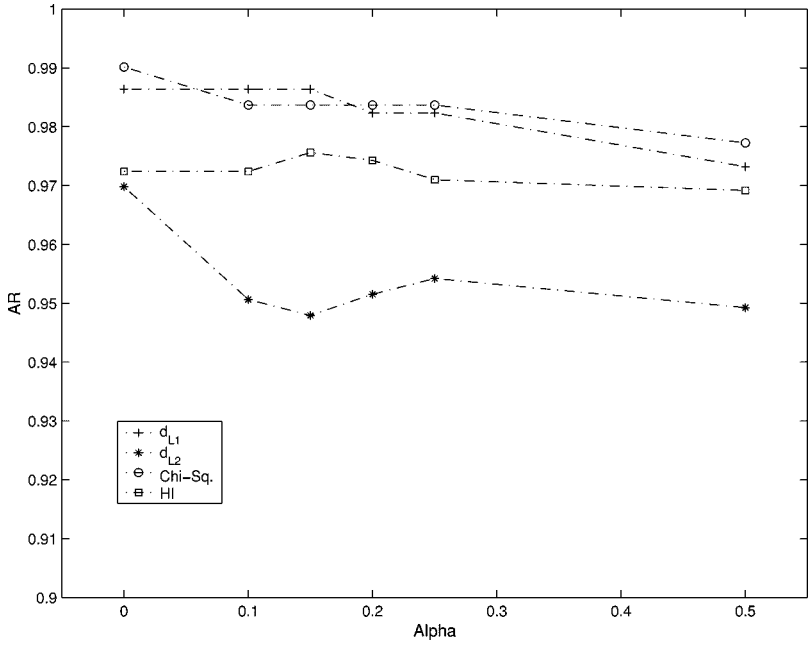
$$E_{H_r} \geq MSE_{\min} = E_{H_{avg}}. \quad (11)$$

The above equations imply that for any given GoF, the median and average histograms are better representatives of GoF color

[1]If M is even, (4) does not directly hold, and the median is defined as the average of the two center sample values.



(a)



(b)

Fig. 4. (a) $ANMRR$ and (b) average recall (AR) values obtained for video segment-to-segment matching experiments using various types of alpha-trimmed histograms and histogram distance measures.

content (in the MAE and MSE sense, respectively) than any single frame histogram selected from the GoF.

C. Intersection Histogram

Histogram intersection is a popular scalar-valued similarity metric for color-based indexing and object recognition [14]. It yields the number of pixels that have the same color in two images. In contrast, the intersection histogram [5] that we propose is itself a histogram, computed over the range of frames in a GoF. The value of the jth bin in the intersection histogram $IntHist_k$ of the kth GoF is determined by

$$IntHist_k(j) = \min_i \{H_i(j)\}. \tag{12}$$

Each bin value in $IntHist$ thus represents the number of pixels of a particular color that appear in all of the GoF frames.

The intersection histogram is characteristically different from the family of alpha-trimmed average histograms because it provides the "least common" color traits of the given group of frames, rather than an estimate of the color distribution. This distinct property of the intersection histogram makes it appropriate for fast identification of the GoF in which a given query

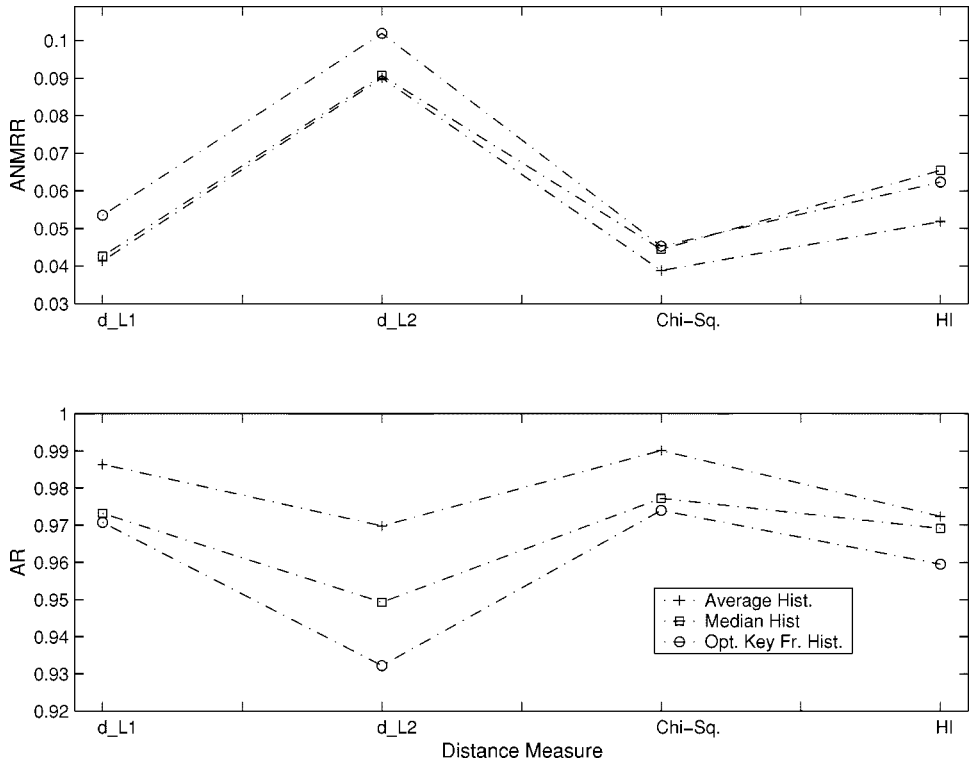


Fig. 5. (Top) $ANMRR$ and (bottom) AR values obtained in video segment retrieval experiments using average, median, and key frame histograms. The optimal key frame histograms are those that minimize (1) for $L1$ distance norm.

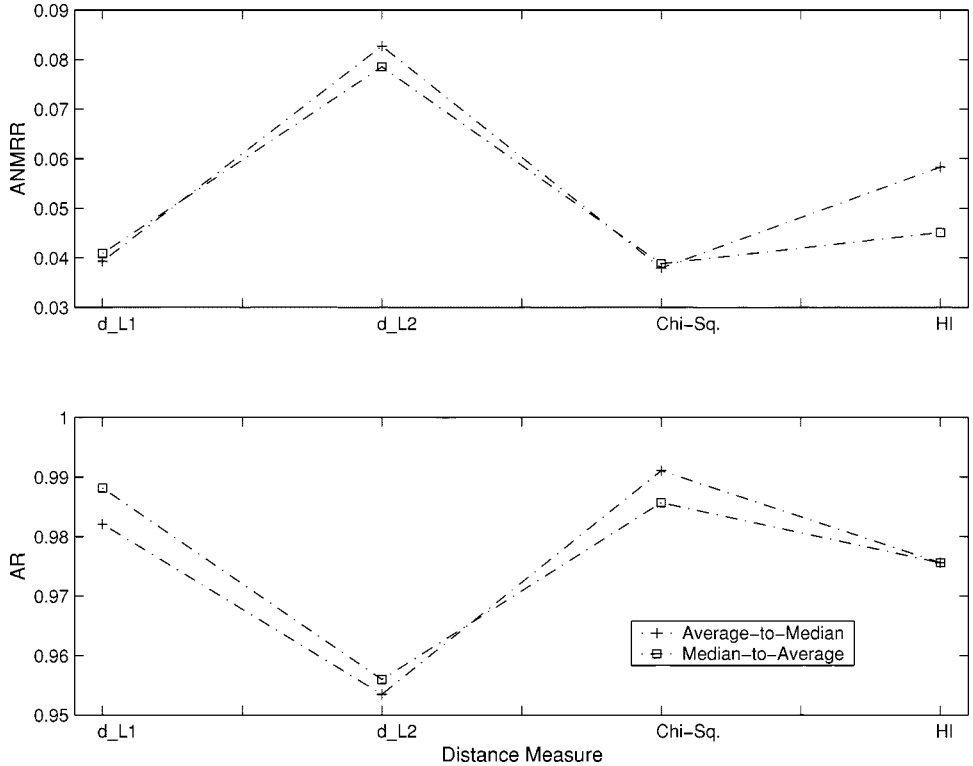


Fig. 6. Cross-retrieval results for different types of GoF histograms. The top figure depicts the $ANMRR$ values obtained when the average histogram is compared to the median histograms for retrieval and vice versa. The corresponding AR values are shown in the bottom figure.

image occurs. Given any GoF k and its intersection histogram, $IntHist_k$, it follows by definition that

$$H_i^k(j) \geq IntHist_k(j), \qquad \forall i \in GoF_k,\ j = 1, \ldots, B \tag{13}$$

where H_i^k denotes the histogram of the ith frame in the kth GoF. This property can be directly used to determine the specific GoF that a frame belongs to. Given a query frame f, it is possible to eliminate the GoFs in the database that do not contain f using (13), since the bin-wise differences $H_f(j) - IntHist_k(j)$ must

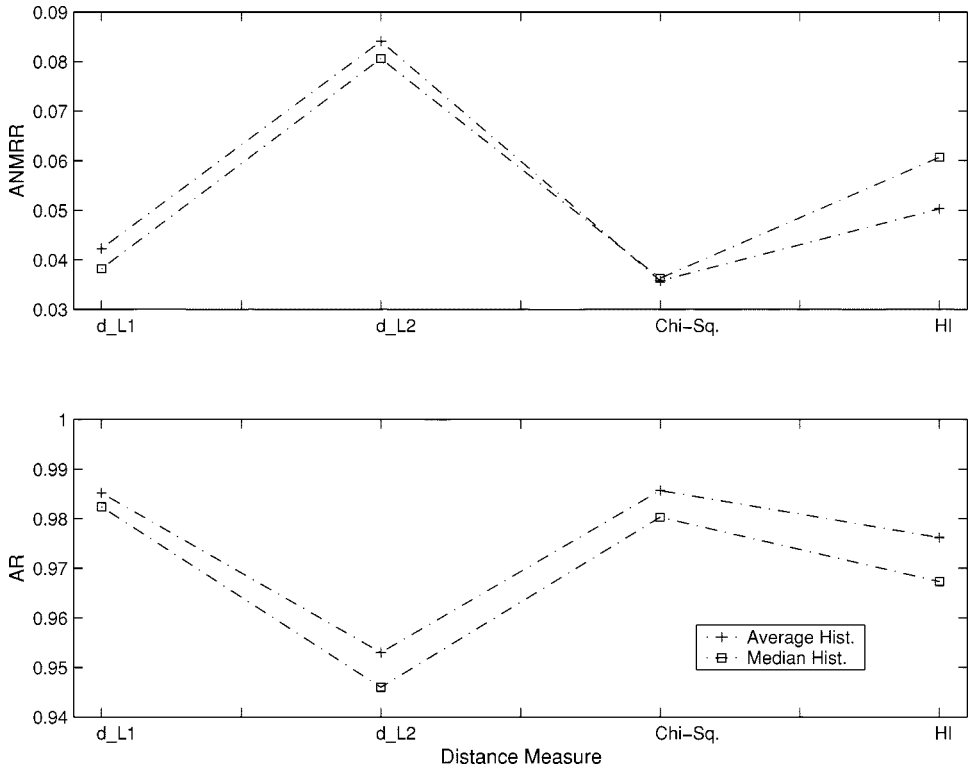


Fig. 7. (Top) $ANMRR$ and (bottom) AR values obtained for the frame-to-video segment matching experiments, where a frame histogram is compared against a GoF histogram during retrieval.

always be nonnegative if $f \in k$. The fast search method involves computation of bin-wise differences between H_f and the intersection histograms $IntHist_k$ of each GoF k in the database

$$D = H_f(j) - IntHist_k(j), \qquad j = 1, \ldots, B. \tag{14}$$

If $D < 0$ for any bin j, then the GoF is immediately rejected as a candidate. Otherwise, a *match coefficient* $C_{f,k}$ is computed as

$$C_{f,k} = 1 - \frac{1}{N_f} \sum_j D \tag{15}$$

where N_f denotes the total number of pixels in frame f. If f belongs to the kth GoF, then the cumulative binwise differences $\sum_j D$ is expected to be small. Hence, a larger $C_{f,k}$ value in (15) implies a higher likelihood that $f \in GoF_k$. The obtained $C_{f,k}$ values can be rank-ordered to yield the list of GoFs to which the query frame may belong.

IV. Video Segment Retrieval With Alpha-Trimmed Average Histograms

In this section, we present the results of color-based video segment retrieval experiments, where three different methods were tested. First, the retrieval performances of the alpha-trimmed average histograms were determined for different values of α. The second experiment was designed to compare the retrieval performance of the proposed color histogram descriptors with that of key frame-based methods. Finally, *cross-retrieval* experiments, where the color histogram descriptors of the query GoF and the database elements are different, were performed to determine the cross-platform compatibility of the proposed descriptors.

A. GoF Database, Query Set, and Ground Truth

The video retrieval experiments were conducted on 1544 GoFs extracted from eight video sequences (over 328 000 video frames) picked out of the MPEG-7 content set [15]. The selected video data, which comprised sitcoms, a TV movie, a documentary, a TV news program, a game show, a music/variety show, excerpts from basketball and soccer games, a bicycle race, and edited home videos, provides a heterogeneous set drawn from a variety of content domains. The GoF boundaries were first identified with the automatic shot boundary detection tools described in [16] and subsequently refined manually by a human operator. Each resulting GoF is a shot, a dissolve, a fade, or a wipe; blank frames and segments where the actual segment bounds were hard to determine were excluded from the test set. Six different alpha-trimmed average histogram descriptors with $\alpha = \{0, 0.10, 0.15, 0.20, 0.25, 0.5\}$ were computed for every GoF. Additionally, the histogram of a representative key frame was identified for each GoF, in order to evaluate the performance of key frame-based video segment retrieval methods. The selected key frame histogram is the one that minimizes E_{H_r} for the $L1$ distance measure in (1); hence, it is *optimal* with respect to the chosen error criterion. All histogram computations were carried out in the YCbCr color space [17], with the luminance component coarsely quantized to eight uniformly spaced bins to reduce sensitivity to variations in lightness. Quantization of the chrominance channels, on the other hand, was nonuniform, because the distribution of pixels in the chrominance space peaks around the mid-range values [18]. Each chrominance channel was quantized to 12 bins, yielding a total of 1152 bins for each histogram. A total of

Fig. 8. Example results of the GoF identification experiments. The actual GoF that the input frame belongs to is framed by a rectangle. Note how a small group of GoFs [i.e., the last two retrieved items in (a) and (b)] are picked up as candidates in each query, even though they are clearly irrelevant. The total number of pixels in each of these GoF histograms is very low, and, as a result, they satisfy (13) for a large number of query frames.

31 query GoFs and the corresponding ground truth data were then picked from the GoF database, and used in the retrievals experiments. The queries range from almost identical static GoFs to dynamic scenes and edit effects/transitions.

B. Histogram Similarity Measures

In order to determine whether any trends would emerge with respect to different histogram types and similarity measures, four different histogram similarity measures were used in the retrieval experiments. A short description of each histogram similarity measure is provided below. In the following definitions, H_a and H_b denote two histograms with B bins, and

$$C_1 = \sqrt{\frac{N_{H_b}}{N_{H_a}}}, \qquad C_2 = \frac{1}{C_1}$$
$$N_{H_a} = \sum_{j=1}^{B} H_a(j), \qquad N_{H_b} = \sum_{j=1}^{B} H_b(j).$$

1) The $L1$ Distance Measure considers the absolute bin-wise differences between histograms H_a and H_b

$$d_{L_1}(H_a, H_b) = \sum_{j=1}^{B} |C_1 \cdot H_a(j) - C_2 \cdot H_b(j)| \qquad (16)$$

d_{L_1} is normalized by $2 \cdot \sqrt{N_{H_a} \cdot N_{H_b}}$ to the range [0, 1].

2) The $L2$ Distance Measure is proportional to the accumulated squared bin differences

$$d_{L_2}(H_a, H_b) = \sqrt{\sum_{j=1}^{B} [C_1 \cdot H_a(j) - C_2 \cdot H_b(j)]^2} \qquad (17)$$

d_{L_2} is normalized by $\sqrt{2 \cdot N_{H_a} \cdot N_{H_b}}$ to the range [0, 1].

3) The Chi-Square Test is used to compare two binned data sets and to determine if they are drawn from the same distribution function

$$\chi^2(H_a, H_b) = \sum_{j=1}^{B} \frac{[C_1 \cdot H_a(j) - C_2 \cdot H_b(j)]^2}{H_a(j) + H_b(j)} \qquad (18)$$

χ^2 is normalized by $N_{H_a} + N_{H_b}$ to the range [0, 1].

4) Histogram Intersection determines the number of pixels that share the same color in the two histograms

$$HI(H_a, H_b) = \sum_{j=1}^{B} \min[H_a(j), H_b(j)]. \quad (19)$$

HI is normalized by $\min(N_{H_a}, N_{H_b})$ to yield a *match value* in the range [0, 1]. Note that when $N_{H_a} = N_{H_b}$, histogram intersection is equivalent to the complement of the $L1$ distance measure [14].

C. Evaluation of Retrieval Performance

In this section, we present the performance measures used to evaluate the retrieval results obtained using the GoF color histogram descriptors. Given a query set and the corresponding ground truth data, we adopt the measures developed by the MPEG Video Group for evaluation of MPEG-7 core experiments. These measures are designed to determine how many of the correct GoFs are retrieved and how high their rankings are among the retrievals.

Let the number of ground truth GoFs for a query q be $ng(q)$. The number of correctly retrieved items in the top K retrievals are denoted by $nr(q)$, with $K = \min\{4 \times ng(q), 2 \times GTM\}$, and $GTM = \max\{ng(q)\}$ over all defined queries Q. The number of misses for query q is given by $M(q) = ng(q) - nr(q)$, and the *recall* value by $R(q) = nr(q)/ng(q)$. Each of the $ng(q)$ ground truth items in the top K retrievals are then assigned a *rank value* $r(i)$, $i = 1, \ldots, ng(q)$; a rank of $K + 1$ is assigned to each of the $M(q)$ missed GoFs. The *average retrieval rank* and the *modified retrieval rank* for query q are then computed as

$$ARR(q) = \sum_{i=1}^{ng(q)} \frac{r(i)}{ng(q)}$$

and

$$MRR(q) = ARR(q) - \frac{ng(q)}{2} - 0.5$$

respectively. $MRR(q)$ is normalized to the range [0, 1] to yield the *normalized* $MRR(q)$, or $NMRR(q)$

$$NMRR(q) = \frac{MRR(q)}{K - \frac{ng(q)}{2} + 0.5}. \quad (20)$$

The average of $NMRR(q)$ and $R(q)$ over the set of all available queries Q yield the overall retrieval performance of the given method

$$ANMRR = \frac{1}{Q} \sum_{q=1}^{Q} NMRR(q) \quad (21)$$

$$AR = \frac{1}{Q} \sum_{q=1}^{Q} R(q). \quad (22)$$

A low value of $ANMRR$ denotes a high retrieval rate (small number of misses) with the relevant items ranked at the top. On the other hand, $ANMRR = 1$ represents the worst case, with none of the relevant items in the database present in the top K retrievals. For average recall AR, on the other hand, higher values imply better retrieval performance.

D. Video Segment-to-Segment Matching

The first set of experiments were carried out using different types of alpha-trimmed average histograms for video segment retrieval. Fig. 4 illustrates the effects of varying α on the retrieval performance of the alpha-trimmed average histograms. Overall, the performance of the proposed histogram descriptors remains very high for different α values, especially when the $L1$ and χ^2 similarity measures are employed. It is worthwhile to note that Fig. 4 reflects the case where the same α value has been adopted for every available GoF in the database. This approach is not representative of the optimal application scenario, where the value of α is determined individually for each GoF, or at least for each program or content type separately.

Fig. 5 compares the performance of the average and median histograms against the (optimum) key frame histogram. Except for a few instances when histogram intersection is used, the family of alpha-trimmed average histograms consistently outperforms the key frame histogram for both performance criteria. Our experiments strongly favor the key frame histograms, because they were optimally identified for each GoF after an exhaustive search. It is thus expected that the performance of key frame-based retrieval will deteriorate when the key frame selection technique is altered.

When individual queries are considered, it is evident that certain GoF histogram types are more appropriate for retrieval of specific types of GoFs. Generally, it is more suitable to select $\alpha > 0$ on transition-type GoFs, in cases of occlusion by a large object, or when gradual variations are observed in the background and/or luminance characteristics. The performances of the alpha-trimmed average histograms and key frame histograms are on par in cases where 1) camera remains stationary and object movement is not too significant and 2) there is object/camera movement, but the scene is dominated by the background object.

Another important retrieval case to consider is when the query GoF is described with a different type color histogram descriptor than that of the items in the search database. This scenario was simulated using average and median histogram as query and database descriptors, respectively (and vice versa). The results, shown in Fig. 6, reveal that there is almost no loss in retrieval performance when different kinds of alpha-trimmed average histograms are used to represent the query and database GoFs in retrieval applications.

E. Frame-to-Video Segment Matching

The final retrieval application we considered for GoF histogram descriptors is retrieval of video segments based on their similarity to a query image or frame. This application is identical to the cross-retrieval application discussed in the previous section when the query is described simply by a (key) frame histogram. To test the performance of the GoF histogram descriptors within this context, the key frame histograms for the 31 query GoFs were compared against the average and median histograms using the four distance measures. The retrieval results, presented in Fig. 7, show that the average and median histograms also perform very well in this scenario. The results

clearly illustrate that the proposed descriptors are fully compatible with most existing databases, where multi-frame color representation is achieved with key frame histograms (or features thereof).

V. GoF Identification Using Intersection Histograms

As noted in Section III-C, a significant application area for intersection histograms is identification of the particular GoF to which a query image belongs. It is possible to carry out this search in a very fast and efficient manner by using (14). The method is especially useful for large databases because, unlike regular similarity-based search methods (which assign a similarity value to every item in the database), it retrieves only a small number of candidate items from the database, and the correct GoF is guaranteed to be among the retrievals (if the query frame does, in fact, appear in the database). The search method can be very useful in stock footage, sports, and news archives, as well as home video collections, where an individual frame is frequently captured from a video stream and utilized as a still frame in printing, publishing, etc. We express the performance of the proposed search method quantitatively by the average number of retrieved items per query and the ranking of the correct GoF among the retrievals.

The proposed method was tested using the key frames picked from the 31 query GoFs. An average of 16 candidate GoFs were retrieved for each query; in 28 of the 31 queries, the correct GoF was retrieved as the top-ranked candidate. For the other three cases, the correct GoF was ranked second (in five retrieved items), third (out of 12 retrieved items), and, in the worst case, sixth (out of 20 retrieved items). Fig. 8 depicts the results obtained for several queries. Note that specific GoFs which contain a very small number of total pixels are retrieved in a large number of the queries (observe the last two candidate GoFs for the examples shown in Fig. 8). The values of the match coefficient C [see (15)] for these GoFs are very low, however, and these items can be readily eliminated as potential candidates by thresholding C. The average number of retrievals is reduced to 6 (without any loss in retrieval performance) when a simple threshold value of 0.05 is used to prune the candidate GoFs.

VI. Conclusions

In this paper, we have presented various types of histogram-based descriptors for joint color content representation of multiple frames. The family of alpha-trimmed average histograms, which include the average and median histograms as special cases, provide a robust set of color descriptors that can eliminate the effects of aberrant frames within a GoF. This set of descriptors consistently outperforms key frame-based representations for video segment retrieval applications. The intersection histogram, on the other hand, reflects the number of pixels of a given color that is common to all the frames in the GoF, and can be employed in specific applications to yield efficient and reliable results. The proposed descriptors are also appropriate for other tasks within a video management system, such as aggregation of GoFs based on color similarity for semantic scene reconstruction, and representative frame selection for visual summarization. These applications are currently under investigation.

References

[1] G. Davenport, T. A. Smith, and N. Pincever, "Cinematic primitives for multimedia," *IEEE Comput. Graph. Applicat.*, vol. 11, pp. 67–74, July 1991.

[2] A. Hampapur, "Designing video data management systems," Ph.D. dissertation, Dept. Comput. Sci. Eng., Univ. Michigan, Ann Arbor, 1995.

[3] P. Aigrain, H. J. Zhang, and D. Petkovic, "Content-based representation and retrieval of visual media: A state-of-the-art review," *Multimed. Tools Applicat.*, vol. 3, no. 3, pp. 3–26, 1996.

[4] H. J. Zhang, J. Wu, D. Zhong, and S. W. Somaliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recognit.*, vol. 30, no. 4, pp. 643–658, Apr. 1997.

[5] A. M. Ferman and A. M. Tekalp, "Multiscale content extraction and representaton for video indexing," in *Proc. Multimedia Storage and Archival Syst. II*, vol. SPIE-3229, Nov. 1997, pp. 23–31.

[6] A. M. Ferman, S. Krishnamachari, A. M. Tekalp, M. Abdel-Mottaleb, and R. Mehrotra, "Group-of-frames/pictures color histogram descriptors for multimedia applications," in *Proc. ICIP 2000*, vol. 1, Vancouver, BC, Canada, 2000, pp. 65–68.

[7] *ISO/IEC Std. 15 938-Part 3: Information Technology—Multimedia Content Description Interface: Visual, 2002*.

[8] *ISO/IEC Std. 15 938-Part 5: Information Technology—Multimedia Content Description Interface: Multimedia Description Schemes, 2002*.

[9] F. Idris and S. Panchanathan, "Review of image and video indexing techniques," *J. Vis. Commun. Image Represent.*, vol. 8, no. 2, pp. 146–166, 1997.

[10] D. Zhong, H. J. Zhang, and S.-F. Chang, "Clustering methods for video browsing and annotation," in *Proc. Storage and Retrieval for Image and Video Databases IV*, vol. SPIE-2670, 1996, pp. 239–246.

[11] J. B. Bednar and T. L. Watt, "Alpha trimmed means and their relationship to median filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 145–153, Feb. 1984.

[12] G. R. Arce, N. C. Gallagher, and T. A. Nodes, "Median filters: Theory for one or two dimensional filters," in *Advances in Computer Vision and Image Processing*, T. S. Huang, Ed, CT: JAI Press, 1986.

[13] E. Ataman, V. K. Aatre, and K. M. Wong, "A fast method for real-time median filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 415–421, Aug. 1980.

[14] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 11, pp. 11–32, 1991.

[15] MPEG Requirements Group. (1998, Oct.) Description of MPEG-7 content set. ISO/IEC JTC1/SC29/WG11/N2467, Atlantic City. [Online]. Available: http://mpeg.telecomitalialab.com/working_documents.htm.

[16] A. M. Ferman and A. M. Tekalp, "Efficient filtering and clustering methods for temporal video segmentation and visual summarization," *J. Vis. Commun. Image Represent.*, vol. 9, no. 4, pp. 336–351, Dec. 1998.

[17] F. W. Billmeyer and M. Saltzman, *Principles of Color Technology*, 2nd ed. New York: Wiley, 1981.

[18] X. Wan and C.-C. Kuo, "Color distribution analysis and quantization for image retrieval," in *Proc. Storage and Retrieval for Image and Video Databases IV*, vol. SPIE-2670, 1996, pp. 8–16.

A. Müfit Ferman (S'91–M'00) received the B.Sc. and M.Sc. degrees in electronics and communication engineering from Istanbul Technical University, Istanbul, Turkey, in 1993 and 1995, and the M.S. and Ph.D. degrees in electrical engineering from the University of Rochester, Rochester, NY, in 1997 and 2001, respectively.

From February 1994 to August 1995, he was a Research and Teaching Assistant with the Department of Electronics and Communication Engineering at Istanbul Technical University. In September 1995, he joined the Department of Electrical and Computer Engineering at the University of Rochester as a Graduate Assistant. Since August 2000, he has been with Sharp Laboratories of America, Inc., Camas, WA. His research interests include content-based video indexing and retrieval, visual information management systems, data mining, and pattern recognition applications.

Dr. Ferman is a member of the ACM and SPIE.

A. Murat Tekalp (S'80–M'82–SM'91) received B.S. degrees in electrical engineering and mathematics from Bogazici University, Istanbul, Turkey, in 1980, with the highest honors, and the M.S. and Ph.D. degrees in electrical, computer, and systems engineering from Rensselaer Polytechnic Institute (RPI), Troy, NY, in 1982 and 1984, respectively.

From December 1984 to August 1987, he was a Research Scientist and then a Senior Research Scientist with Eastman Kodak Company, Rochester, NY. He joined the Electrical and Computer Engineering Department, University of Rochester, Rochester, NY, in September 1987, where he is currently a Distinguished University Professor. He is also affiliated with the Koç University, Istanbul. His current research interests are in digital image and video processing, including image/video restoration, video segmentation, object tracking, video coding, content-based video description, and secure media. He has served as an associate editor for the *Journal of Multidimensional Systems and Signal Processing* (1994–1999). He was an area editor for the journals *Graphical Models* and *Image Processing* (1995–1998). He was also on the editorial board of the *Journal of Visual Communication and Image Representation* (1995–1999). He is the Editor-in-Chief of the *EURASIP Journal on Image Communication*. He authored *Digital Video Processing* (Englewood Cliffs, NJ: Prentice-Hall, 1995). He holds five U.S. patents. His group contributed technology to the ISO/IEC MPEG-4 and MPEG-7 standards.

Dr. Tekalp received the NSF Research Initiation Award in 1988 and was named as Distinguished Lecturer by IEEE Signal Processing Society in 1998. He has chaired the IEEE Signal Processing Society Technical Committee on Image and Multidimensional Signal Processing (January 1996–December 1997) and is a founding member of the Technical Committee on Multimedia Signal Processing. He has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING (1990–1992), IEEE TRANSACTIONS ON IMAGE PROCESSING (1994–1996). He was appointed as the Technical Program Chair for the 1991 IEEE Signal Processing Society Workshop on Image and Multidimensional Signal Processing, and the Special Sessions Chair for the 1995 IEEE International Conference on Image Processing and Technical Program Co-Chair for IEEE ICASSP 2000 in Istanbul. He is the founder and first Chairman of the Rochester Chapter of the IEEE Signal Processing Society. He was elected as the Chair of the Rochester Section of IEEE in 1994–1995. He is the General Chair of IEEE International Conference on Image Processing to be held in Rochester, NY, in September 2002.

Rajiv Mehrotra is the Program Manager for the Media Asset Management Program of the Entertainment Imaging Division of Kodak and the Technology Manager for Media Asset Management R&D program of Kodak R&D Labs, Rochester, NY. Prior to joining Kodak, he held faculty positions at the University of South Florida, Tampa, the University of Kentucky, Lexington, and the University of Missouri, Kansas City. He is a co-editor of the book *The Handbook of Multimedia Information Management* (Englewood Cliffs, NJ: Prentice-Hall, 1997).

Dr. Mehrotra was co-editor of a special issue of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING ON MULTIMEDIA INFORMATION SYSTEM (August 1993). He also co-edited a special issue of *IEEE Computer* on image database management (December 1989). He is on the editorial boards of IEEE MULTIMEDIA AND PATTERN RECOGNITION and has served on the program/organizing committees of several international conferences.