Comparative Evaluation of Color-Based Video Signatures in the Presence of Various Distortion Types

Aritz Sánchez de la Fuente, Patrick Ndjiki-Nya, Karsten Sühring, Tobias Hinz, Karsten Müller, and Thomas Wiegand

Abstract**—**The robustness of color-based signatures in the presence of a selection of representative distortions is investigated. Considered are five signatures that have been developed and evaluated within a new modular framework. Two signatures presented in this work are directly derived from histograms gathered from video frames. The other three signatures are based on temporal information by computing difference histograms between adjacent frames. In order to obtain objective and reproducible results, the evaluations are conducted based on several randomly assembled test sets. These test sets are extracted from a video repository that contains a wide range of broadcast content including documentaries, sports, news, movies, etc. Overall, the experimental results show the adequacy of color-histogram-based signatures for video fingerprinting applications and indicate which type of signature should be preferred in the presence of certain distortions.

Keywords**—**color histograms, robust hashing, video retrieval, video signature

I. Introduction

THE amount of accessible multimedia information is ever increasing. For the management of large-scale multimedia data, efficient organization of databases is required. One efficient management tool is video retrieval, which can also be used for a variety of other applications, such as content advertising, detection of duplicated video sequences and automatic detection of copyright violations.

In video retrieval, one or more videos are selected as query (query-by-example) and a list of similar videos from the database is returned in response. Recent approaches often use content-based video properties (e.g. color or texture features) for matching. Most frequently, a robust or perceptual video hash or a video signature is used for video identification. First approaches developed a cryptographic hash [1], which is bit-sensitive. Other approaches map the visual information onto a fixed-length hash code that remains practically unaltered in presence of distortions of the original signal [2]. Hence, robust hashing allows slightly altered (e.g. damaged) videos to be also recognized as being similar to the original unaltered video.

Radhakrishnan and Bauer [3] propose a robust signature based on projections of cropped coarse difference-images between adjacent frames onto randomly generated matrices. The output hash for each difference-image is of fixed length. A quantization using two reconstruction levels is afterwards applied using the median of the hash-elements as threshold. The final signature is the concatenation of all binary vectors. This description is highly robust against distortions introduced by video compression, spatial and brightness scaling, whereas being sensitive to other geometric and frame-rate conversions. The binarization step described above is also used by Coskun and Sankur [2]. They determine and apply one-bit quantization to the most significant coefficients of 3D DCT for a given video signal. In addition to blurring and noise robustness, this hash code is also robust to distortions introduced by operations including video compression as well as contrast and brightness manipulation.

In this paper, the robustness of color-based signatures to specific types of distortions is evaluated. The main advantage of these signatures is the very low complexity required for their extraction. For some distortion types this might be at the cost of retrieval performance. For other distortion types, color-based signatures might offer a very good efficiency-complexity trade-off. The signatures are evaluated in a new, modular framework that allows easy plug-in of new data and signatures.

The remainder of the paper is organized as follows. In section II, the framework and the proposed signatures are described. Section III summarizes the similarity measures used for each type of signature. The experimental results are presented in section IV. Finally, the conclusions are drawn and future research is discussed.

A. Sánchez de la Fuente is with the Image Processing Department, Fraunhofer Institute for Telecommunications Heinrich-Hertz-Institut, Berlin, Einsteinufer 37, D-10587 Germany (phone: +49-(0)30-31002-259; fax: +34-(0)30-3927200; e-mail: aritz.sanchez@hhi.fraunhofer.de).

P. Ndjiki-Nya, K. Sühring, T. Hinz, and K. Müller are with the Image Processing Department, Fraunhofer Institute for Telecommunications Heinrich-Hertz-Institut, Berlin, Einsteinufer 37, D-10243 Germany (e-mail: {ndjiki,suehring,hinz,kmueller}@hhi.de).

T. Wiegand is the Head of the Image Processing Department, Fraunhofer Institute for Telecommunications Heinrich-Hertz-Institut, Berlin, Einsteinufer 37, D-10243 Germany, and Head of the Image Communication Chair, Department of Telecommunication Systems, School of Electrical Engineering and Computer Sciences, Technical University of Berlin, Berlin, Einsteinufer 17, D-10587 Germany (e-mail: wiegand@hhi.fraunhofer.de).

II. Signature Framework

A. Overview

The new framework pursues three objectives: the complete video retrieval process with all corresponding phases; a separate operation of the individual phases; and a modular system in which data and signatures can be easily added or removed. The first problem is addressed by covering a large number of aspects a fingerprinting task might involve: wide variety of data, including altered and damaged videos, feature extraction and matching, and benchmarking of the system. As depicted in Fig. 1, video processing is divided into five independent modules.

The first stage as depicted in Fig. 1 generates test sets of a predefined number of random clips extracted from the video repository (further explained in section IV.A). The distorted versions of all the query clips are added to each test set during the second phase (see section IV.B). Next, the signatures are extracted from the query videos as well as from the sequences contained in the test sets. Matching of the query clips with the test sets is then conducted. Finally, the matching results are analyzed (see section IV.D) to assess the efficiency of the signatures.

B. Integrated Signatures

Five signatures have been implemented in this work. They all derive from the color histograms of the sequence frames (see Fig. 2). The histograms are 256-bin wide and the HSV color-space is used for their generation. As explained in [4], the reliability of color information to characterize visual content makes it one of the most widely used features in retrieval tasks. The simple representation provided by color histograms and the ease of their comparison (e.g. Euclidean distance) is the other main reason for the chosen feature in the present descriptions. Such descriptions do not properly represent grayscale content, which is therefore out of scope. This paper investigates the robustness of color-based signatures to typical video distortions.

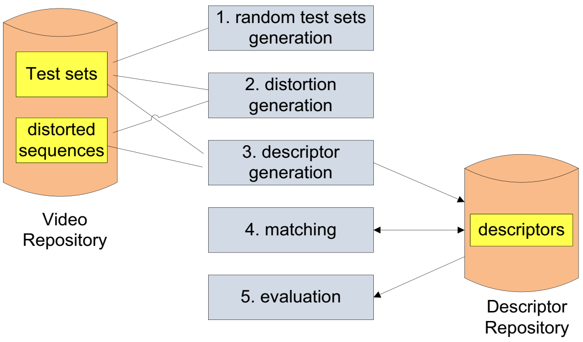

Fig. 1 Operation chain of the proposed framework

1) The SCMean Signature

The first signature is calculated as the normalized color histogram temporally averaged along the sequence frames. The average population $\bar{p}_m$ of the m-th bin on the final histogram is calculated as follows:

$$\bar{p}_m = \frac{1}{WH}\frac{1}{N}\sum_{j=1}^{N} p_m^j \quad , \quad m = 1,2,\ldots,M, \tag{1}$$

where M is the total number of bins. The population p_m^j of the m-th bin of the j-th frame (j = 1,2,…,N) of the sequence is averaged, i.e. summed up and normalized by N, W (width) and H (height) (1). Hence, the description is a decimal vector of 256 elements and is, referred to as *SCMean*.

This signature needs a single vector to represent the overall color content of the whole sequence. This computational saving may imply some detrimental loss of detail. Dissimilar videos are expected to have different color content, hence producing different descriptions. However, such a signature may be highly sensitive to color degradations.

2) The SCDiffHist Signature

The *SCDiffHist* signature characterizes the change in color between frames and is inspired by the color-shift signature described in [4], where the signature consists of the scalar distances between normalized color histograms of adjacent frames. For the *SCDiffHist* signature, the absolute unnormalized difference histogram, d^j, between neighboring frames j and (j+1) is gathered.

$$d_m^j = \left| p_m^j - p_m^{j+1} \right| \quad , \quad m = 1,2,\ldots,M, \quad j = 1,2,\ldots,N-1 \tag{2}$$

The normalized mean histogram of these difference histograms, d^j, is obtained as a decimal description with 256 elements.

$$\bar{d}_m = \frac{1}{2WH}\frac{1}{N-1}\sum_{j=1}^{N-1} d_m^j \quad , \quad m = 1,2,\ldots,M \tag{3}$$

Note that the largest possible distance between two histograms is now twice the dimensions of the frame. For this reason, the normalization factor is doubled (2WH).

3) The SCComp Signature

The first- and second-order moments of the difference histograms are used for the third signature. Two different criteria are considered for the calculation of moments. On the one hand, the average $\left(\bar{c}_{time,x}, \bar{c}_{time,y}\right)$ of the centers of mass of every difference histogram (4) and their standard deviations $\left(\sigma_{time,x}, \sigma_{time,y}\right)$ (5) are estimated, both in x- and y-axis.

$$\bar{c}_{time,x} = \frac{1}{N-1}\sum_{j=1}^{N-1} c_{j,x} \ , \tag{4}$$

$$\sigma_{time,x} = \frac{1}{N-1}\left(\sum_{j=1}^{N-1}\left(\bar{c}_{time,x} - c_{j,x}\right)^2\right)^{1/2} , \tag{5}$$

where $c_{j,x}$ denotes the coordinate in x-direction of the center of mass of the j-th difference histogram, and N refers to the number of frames of the sequence.

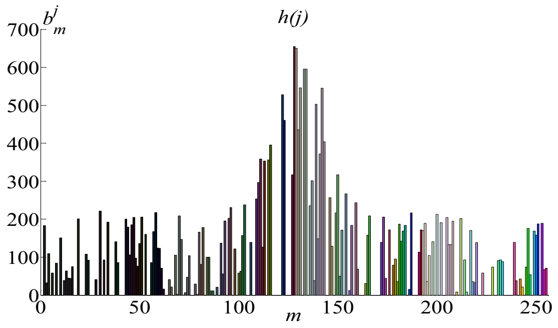

Fig. 2 Color histogram *h(j)* of frame *j*

On the other hand, the center of mass $\left(\bar{c}_{spatial,x}, \bar{c}_{spatial,y}\right)$ and the standard deviations $\left(\sigma_{spatial,x}, \sigma_{spatial,y}\right)$ in the mean difference histogram are calculated (6), (7), and (8). Here, the y coordinate of the centers of mass determined is coincident for both methods $\left(\bar{c}_{time,y} = \bar{c}_{spatial,y} = \bar{c}_y\right)$. The y moments are normalized by W and H.

$$\bar{c}_{spatial,x} = \frac{\sum_{m=1}^{M} m\bar{p}_m}{\sum_{m=1}^{M} \bar{p}_m}, \tag{6}$$

$$\bar{c}_{spatial,y} = \frac{1}{2WH} \sum_{m=1}^{M} \bar{p}_m , \tag{7}$$

$$\sigma_{spatial,x} = \frac{1}{M}\left(\sum_{m=1}^{M}\left(\bar{c}_{spatial,x} - m\right)^2\right)^{1/2}, \tag{8}$$

where m corresponds to the bin number of the mean difference histogram, and M denotes the number of coefficients of a histogram.

All the equations for the y coordinate are similarly obtained with adjusted (4), (5), and (8), with the additional normalization factor $2WH$.

The generated signature *SCComp* has only 7 decimal elements, as in (9):

$$\mathbf{y} = [\bar{c}_y, \bar{c}_{time,x}, \bar{c}_{spatial,x}, \sigma_{time,y}, \sigma_{time,x}, \sigma_{spatial,y}, \sigma_{spatial,x}] \tag{9}$$

SCComp inherits the potential weaknesses of *SCMean* and *SCDiffHist*. Despite the advantages it features in terms of its compactness, this signature describes the video signal very coarsely.

4) *The SCHistBin Signature*

The signature *SCHistBin* is a binary variation of the *SCDiffHist* signature. It is inspired by [2], [3], where the purpose is to obtain a higher degree of robustness. Here, the unnormalized mean difference histogram is quantized with two reconstruction levels, using the median of the bin populations as threshold. Any bin-value higher than the median is set to 1. The others are set to 0.

Binarization is expected to yield higher robustness to homogeneous color changes in the temporal direction compared to *SCDiffHist*. A higher degree of robustness is also reported in [2], [3].

5) *The SCMeanHaar Signature*

The last signature corresponds to the definition of MPEG-7's GoF/GoP descriptor [5]. The unnormalized mean color histogram is wavelet-transformed, using the Haar Transform, as defined in the MPEG-7 specification [6]. In this paper, it is referred to as *SCMeanHaar*.

III. Similarity Measures

For the *SCMean*, *SCDiffHist*, and *SCMeanHaar* signatures, the ℓ_2 norm (Euclidean distance) is used as similarity measure, D, (10) between two signatures $\boldsymbol{a}$ and $\boldsymbol{b}$.

$$D = \left(\sum_{i=1}^{M} \left|a_i - b_i\right|^2\right)^{1/2} \tag{10}$$

For the *SCComp* signature, the similarity between two sequences $(\boldsymbol{a},\boldsymbol{b})$ is measured via area comparison.

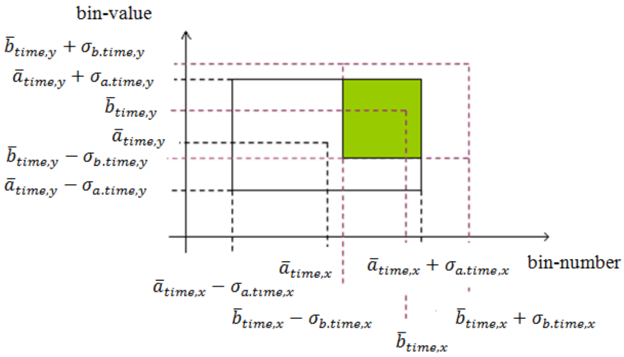
bin-value
$\bar{b}_{time,y} + \sigma_{b.time,y}$
$\bar{a}_{time,y} + \sigma_{a.time,y}$
$\bar{b}_{time,y}$
$\bar{a}_{time,y}$
$\bar{b}_{time,y} - \sigma_{b.time,y}$
$\bar{a}_{time,y} - \sigma_{a.time,y}$
$\bar{a}_{time,x}$
$\bar{a}_{time,x} + \sigma_{a.time,x}$
bin-number
$\bar{a}_{time,x} - \sigma_{a.time,x}$
$\bar{b}_{time,x} - \sigma_{b.time,x}$
$\bar{b}_{time,x} + \sigma_{b.time,x}$
$\bar{b}_{time,x}$

Fig. 3 Intersection areas of "temporal moment rectangles" (analog for the spatial moments)

As depicted in Fig. 3, the center of mass and the standard deviations define the center and the margins of a rectangle. The intersection of the areas described by the rectangles of two signatures is calculated and normalized with respect to the largest area (11), both for the temporal $\left(\phi_{area.time}\right)$ and spatial $\left(\phi_{area.spatial}\right)$ moments. The mean of $\phi_{area.time}$ and $\phi_{area.spatial}$ is subtracted from 1, so that a perfect match has a distance $D = 0$ (12).

$$\phi_{area} = \frac{S_a \cap S_b}{\max(S_a, S_b)} \tag{11}$$

$$D = 1 - \frac{\phi_{area.time} + \phi_{area.spatial}}{2} \tag{12}$$

As *SCHistBin* is a binary signature, the normalized Hamming distance (13) is used for similarity estimation. That is, the number of deviating bits is normalized by the length M of the signatures.

$$D=\frac{1}{M}\sum_{a_i\neq b_i}1 \;\;,\;\; i\in[1,M] \;\;,\;\; a_i\in\boldsymbol{a} \;\;,\;\; b_i\in\boldsymbol{b} \qquad (13)$$

IV. Evaluation

A. Data set

The developed signatures are tested on a large variety of video content, encoded in H.264/MPEG4-AVC format at QCIF (176 pixels x 144 pixels) resolution. The data is organized in shots or tracks. The latter are defined as a sequence of frames that corresponds to an uninterrupted camera operation [4].

100 query shots are matched against 20 randomly generated test sets. The query tracks were cleared beforehand, i.e. trivial sequences were removed (e.g. too short clips, static videos). The query clips represent a wide range of visual content (Fig. 4): documentaries, sports, news, real/animated films, advertisements, and talk shows, among others. Each test set contains 2200 sequences, among which 1200 are distorted versions of the query tracks. A ground-truth set consists of a query track and distorted versions of it. For the retrieval experiment, each query shot has 12 distorted versions. The ground-truth size is thus always 13. However, there might be other visually similar sequences that are not listed as ground-truth. This problem could be addressed by annotating the content of the video database accordingly. However, this is impractical due to its size: 770 000 tracks. This is the major motivation for drawing random subsets from the video repository for evaluation. Reliability of the results is achieved by repeating the subset evaluation sufficiently often. The number of test sets used is expected to provide the required generality of the results, independently of the analyzed data.

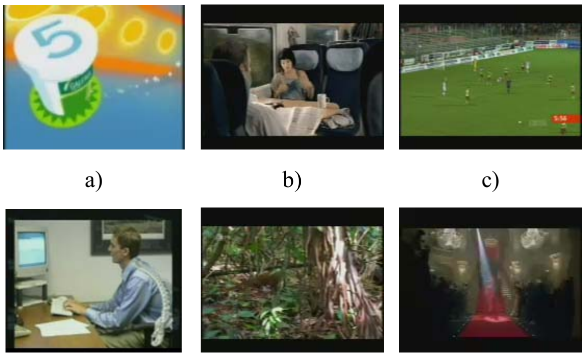
a) b) c)
d) e) f)

Fig. 4 Example query shots

B. Tested distortions

Some exemplary visual alterations affecting color content are tested. They are: *additive white Gaussian noise* (*AWGN*), with zero mean and a standard deviation of 50; *salt & pepper noise* (multicolor), with a 5% probability of occurrence; an additive cosine in time of amplitude 25.5 (of a maximum of 255) and period 50 frames; and color inversion. The latter is chosen, in order to show the limitations of the proposed signatures. All distorted video signals are encoded with three different quantization steps: 16, 40, and 48. Example images are shown in Fig. 5.

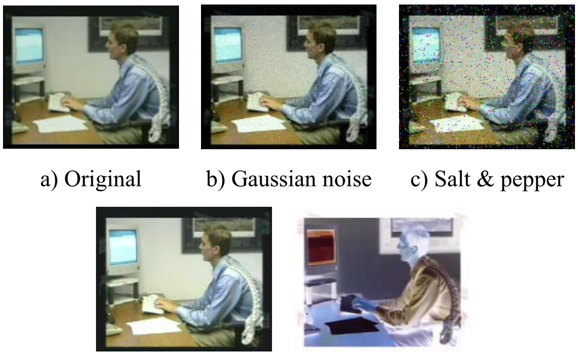
a) Original b) Gaussian noise c) Salt & pepper
d) Cosine distortion e) Color inversion

Fig. 5 Snapshots of evaluated distortions

C. Efficiency measure

As a measure of efficiency, the *average retrieval rate* (*ARR*) [5] of each signature is considered. The parameter is calculated along the 20 test sets and for every query track. This value can be expressed as the ratio between the number of correctly recovered sequences and the total number of sequences stored in the ground-truth sets:

$$ARR=\frac{1}{L}\sum_{i=1}^{L}ARR_i\,,$$
$$ARR_i=\frac{1}{NQ}\sum_{q=1}^{NQ}\frac{\#correct.matches(q)}{NG(q)}\,, \qquad (14)$$

where L denotes the number of test sets considered (20 in this case), ARR_i refers to the average retrieval rate for every query shot and for the i-th test set, NQ is the number of query tracks used (100), and $NG(q)$ expresses the number of sequences that belong to the ground-truth set of query q (in the present experiment, $NG(q)=12, \forall q$, as the query is ignored).

The threshold $K(q)$ represents the highest rank in the list of results that a ground-truth sequence of query q can achieve and is defined in [5] as

$$K(q)=\min\{4NG(q), 2\max\{NG(q),\forall q\}\}\,. \qquad (15)$$

Every ground-truth sequence with a rank higher than $K(q)$ is classified as missed, else it is a true positive match. For this experiment, $K(q) = 24$.

D. *Experimental results*

Fig. 6 and Fig. 7 show the retrieval rates for each distortion type and each quantization step (QP). The standard deviation of the *ARR* along the test sets is very low (under 0.46%) and is, therefore, not displayed. From the results, it is obvious that *SCComp* always has very low retrieval rates (below 30%). This is caused by a high loss of description accuracy. The highest rates are achieved by *SCMeanHaar* and *SCMean*, closely followed by *SCHistBin*. *SCMean* shows a homogeneous response to the different types of distortion.

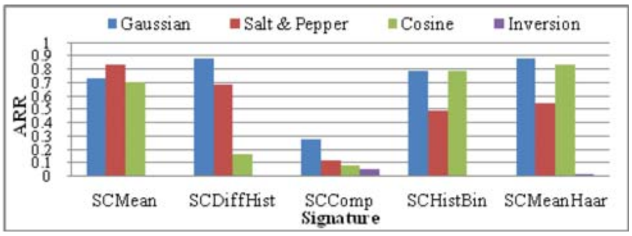

Fig. 6 ARR for different distortion classes

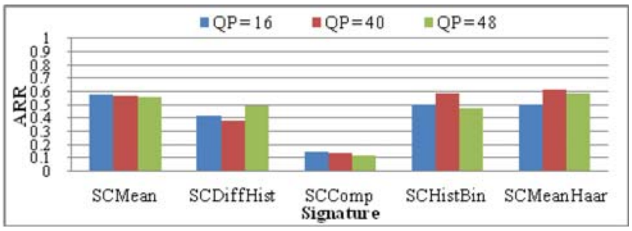

Fig. 7 ARR for different Quantization Parameters (QP)

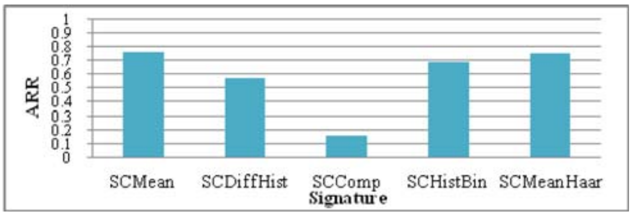

Fig. 8 Overall ARR without color inversion

For all investigated signatures, the retrieval rates are very similar for different quantizations. Hence, all color-based signatures are relatively invariant to coding quantization.

Among the tested distortions, the highest *ARR* scores are obtained for Gaussian noise for almost every signature. Here, *SCDiffHist* and *SCMeanHaar* are very robust. For salt & pepper noise, *SCMean* signature achieves over 80% with the other signatures achieving lower *ARR* values. For cosine distortion, *SCMeanHaar* achieves almost 85% *ARR* followed by the other signatures. Not surprisingly, color inversion leads to very bad scores, as it contains complementary color distribution. This attack might be overcome if image-difference were performed previous to histogram generation. Hence, Fig. 8 displays the overall retrieval rates for the three typical distortions, where color inversion is not considered. The results lie between 70% and 75% for *SCMean*, *SCDHistBin*, and *SCMeanHaar*. *SCComp* is the worst color-based signature. Although it is the only description that includes temporal information, it shows too low discrimination capability.

Overall, the most efficient signature appears to be *SCMean*. Most of the signatures proposed are appropriate for identifying distorted video signals of identical content in applications, where the considered set of distortions occurs.

E. *Size of the signatures*

The size of the descriptions is independent of the length of the sequences and of the dimensions of the image, as listed in TABLE I for each of the signature types.

TABLE I
SIGNATURE SIZE IN BYTES

SCMean	SCDiffHist	SCComp	SCHistBin	SCMean-Haar
2052	2052	61	34	2052

Considering that the average length of the sequences tested is 4.22 s and the required disk space for the coded bitstream of a single track is around 100 kbytes, the values of the table can be considered quite small. Three signatures are about 50 times smaller than the original video and two signatures around 1700 and 3000 times smaller.

F. *Complexity of the signature extraction and matching*

For a sequence of N frames, width W and height H, approximate computational complexity orders of the signature extraction and matching are summarized in TABLE II. The term WH refers to the M-bin histogram generation from the decoded frames. On the one hand, for every signature the extraction times highly depend on W, H, N, and M. On the other hand, the matching is always constant for *SCComp* and depends only on M when any of the other four signature types are matched.

TABLE II
COMPUTATIONAL COMPLEXITY OF EXTRACTION AND MATCHING

Signature	Extraction	Matching
SCMean	$O(WH + NM)$	$O(M)$
SCDiffHist	$O(WH + NM)$	$O(M)$
SCComp	$O(WH + NM)$	$O(1)$
SCHistBin	$O(WH + NM)$	$O(M)$
SCMeanHaar	$O(WH + NM)$	$O(M)$

TABLE III lists the signature extraction and matching speeds (on a CPU of 2.66 GHz and 8 GB of RAM, with QCIF resolution, M = 256 bins, and a frame rate of 25 fps). The results show a very similar time-efficiency during the extraction for every signature type, which is validated by the complexity orders given by TABLE II. Due to the slightly higher time consumption introduced by the Haar Transform, only *SCMeanHaar* extracts less than 13 seconds of video per second, whereas the other four types of fingerprints reach a speed over 14.4 and under 14.8 seconds of video per second.

Regarding the matching stage, the signatures compared by the Euclidean distance are the slowest, closely followed by *SCHistBin* (Hamming distance), and the fastest matches occur for *SCComp* (almost twice as fast as the slowest matching fingerprints).

TABLE III
COMPUTATIONAL SPEED OF EXTRACTION AND MATCHING

Signature	Extraction seconds of video / second	Matching pairs of signatures / second
SCMean	14.43	20438
SCDiffHist	14.77	20438
SCComp	14.64	38261
SCHistBin	14.68	21256
SCMeanHaar	12.96	20438

V. Conclusions and Future Work

Five signatures based on color histograms have been tested in the presence of various distortion types. The results show that limiting the descriptions to histogram moments causes very high loss of description accuracy, leading to inefficient video retrieval. The use of difference histograms does not seem to improve the robustness of the signatures in comparison to the original histograms. However, the difference histogram performs better, when binarization of the bin population is operated. Size reduction and lower matching complexity are additionally achieved. The mean color histogram shows similar results for different distortions. Therefore, this signature is considered as the steadiest and most satisfactory overall. Another advantage of all signatures is their compactness. Using a number of random subsets of the database, content-independence of the evaluation results is achieved.

Future work will focus on the improvement of the presented signatures as well as on further investigations on complexity reduction. Furthermore, other damage classes (e.g. geometric attacks, severe alterations on color content, or distortions in monochrome sequences) will be investigated. For such distortions, some other feature or a more robust version of the presented signatures may be necessary.

References

[1] B. Preneel, R. Govaerts, and J. Vandewalle, "Cryptographic hash functions", Proceedings of the 3rd Symposium on State and Progress of Research in Cryptography, W. Wolfowicz (ed.), Fondazione Ugo Bordoni, pp. 161-171, 1993.

[2] B. Coskun, and B. Sankur, “Robust Video Hash Extraction”, EUSIPCO 2004, pp. 2295-2298, Vienna, April 2004.

[3] R. Radhakrishnan, and C. Bauer, “Content-Based Video Signatures based on Projections of Difference Images”, IEEE 9th Workshop on Multimedia Signal Processing, pp. 341-344, Greece, 1-3 Oct. 2007.

[4] T.C. Hoad, and J. Zobel, “Detection of Video Sequences Using Compact Signatures”, ACM Transactions on Information Systems (TOIS), Volume 24, Issue 1, pp. 1-50, January 2006.

[5] B.S. Manjunath, P. Salembier, and T. Sikora, “Introduction to MPEG-7: Multimedia Content Description Interface”, John Wiley & Sons, Inc., New York, NY, 2002.

[6] A. Yamada, M. Pickering, S. Jeannin, L. Cieplinski, J.R. Ohm, and M. Kim, “MPEG-7 Visual Part of eXperimentation Model Version 10.0”, ISO/IEC JTC1/SC29/WG11/N4063, Singapore, March 2001.