The MPEG-7 Video Signature Tools for Content Identification

S. Paschalakis, *Member, IEEE*, K. Iwamoto, P. Brasnett, N. Sprljan, R. Oami, *Member, IEEE*,
T. Nomura, *Member, IEEE*, A. Yamada and M. Bober, *Member, IEEE*

Abstract—This paper presents the core technologies of the Video Signature Tools recently standardized by ISO/IEC MPEG as an amendment to the MPEG-7 Standard (ISO/IEC 15938). The Video Signature is a high-performance content fingerprint which is suitable for desktop-scale to web-scale deployment and provides high levels of robustness to common video editing operations and high temporal localization accuracy at extremely low false alarm rates, achieving a detection rate in the order of 96% at a false alarm rate in the order of five false matches per million comparisons. The applications of the Video Signature are numerous and include rights management and monetization, distribution management, usage monitoring, metadata association, and corporate or personal database management. In this paper we review the prior work in the field, explain the standardization process and status, and provide details and evaluation results for the Video Signature Tools.

Index Terms—content description, content identification, content localization, content fingerprint, MPEG, video fingerprint, video signature, video identification

I. INTRODUCTION

THE amount of video content which people generate and consume has been increasing at a stunning pace in recent years. According to figures released by the company itself, in 2010 two billion videos a day were watched on YouTube and hundreds of thousands of videos were uploaded daily, at a rate of 24 hours of content every minute [1]. Instantaneous access

to the vast volumes of multimedia content available on the web is something that we take for granted thanks to text-based search engines. Despite this, there are no standard tools which one can use to efficiently search for a copy of a specific piece of video content, possibly in an edited or modified version and embedded in a longer piece of video content, either on the web or in one's own personal database.

In recent years, video identification has been studied mainly in two different, but complementary, ways. Firstly, as a general problem of designing unique and robust fingerprints to allow the identification and localization of a video embedded in an unrelated longer video [2],[3]. Secondly, as the more specific problem of near-duplicate video clip detection in large databases [4],[5]. The latter investigations have focused on the design of complete retrieval systems, including fingerprint extraction and database indexing schemes, for the retrieval of short near-duplicate clips exhibiting a complete or significant temporal overlap. Such systems are not, by and large, designed for the identification and localization of a video embedded in a longer video, although there are exceptions [6]. This paper focuses on the first problem of designing video fingerprinting tools that achieve high levels of robustness to common video editing operations and are able to accurately detect and localize a piece of video content embedded in a longer piece of unrelated video content, with exceptionally low false alarm rates, in the order of only a few false alarms per million comparisons.

To attain an interoperable solution for video content identification, the Moving Picture Experts Group (MPEG), a working group of ISO/IEC, issued a worldwide open call for proposals in 2008 [7],[8]. The Video Signature, or Video Signature Tools to use its full name, has now been standardized and is the latest amendment to the MPEG-7 Standard, also known as ISO/IEC 15938 Multimedia Content Description Interface [9],[10].

The rest of this paper is organized as follows. Section II looks at related works in the field of interest. Section III presents the development process of the Standard, while Section IV looks at application scenarios for the Video Signature. Sections V and VI examine the Video Signature extraction and compression respectively. The matching and localization process used in the development of the Standard is presented in Section VII. The MPEG-7 evaluation methodology and results are presented in Section VIII, and a

S. Paschalakis is with Visual Atoms Ltd, The Surrey Technology Centre, 40 Occam Road, Guildford Surrey GU2 7YG, UK (tel: +44 (0)1483 688181, email: s.paschalakis@visualatoms.com). He was with Mitsubishi Electric R&D Centre Europe (MERCE).

K. Iwamoto is with NEC Corporation, 1753, Shimonumabe, Nakahara-Ku, Kawasaki, Kanagawa 211-8666, Japan (email: k-iwamoto@ay.jp.nec.com)

P. Brasnett is with Imagination Technologies Ltd, Imagination House, Home Park Estate, Kings Langley, Hertfordshire, WD4 8LZ, UK (email: paul.brasnett@imgtec.com). He was with MERCE.

N. Sprljan is with Maxeler (email: nsprljan@maxeler.com). He was with Visual Atoms Ltd and MERCE.

R. Oami is with NEC Corporation (email: r-oami@az.jp.nec.com)

T. Nomura is with NEC Corporation (email: t-nomura@da.jp.nec.com)

A. Yamada is with NEC Corporation (email: a-yamada@da.jp.nec.com)

M. Bober is with The University of Surrey and Visual Atoms Ltd (email: m.bober@visualatoms.com). He was with MERCE.

Copyright (c) 2011 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

brief comparative discussion based on recently reported results in video identification is presented in Section IX. Finally, Section X presents our concluding remarks.

II. RELATED WORKS

This section provides an overview of the key design choices in creating a video fingerprint and then examines a number of notable studies in the field, both for identification and localization of a video embedded in unrelated video content and for near-duplicate video clip detection.

Most video signatures rely on frame-level features. In turn, such features are usually keypoint-based [11],[12],[13],[14], block-based [2],[15],[16] or global [17],[18]. For keypoint-based approaches, SIFT features [14],[19],[20] or local descriptors around Harris interest points [12],[13],[21],[22] are a common design choice. Such approaches can attain increased robustness to content modifications, but also entail increased computational costs. Block-based approaches [2],[15],[16] involve the calculation of certain properties of pre-defined spatial regions in the frame. Block-based methods typically entail reduced computational costs, but are also less robust to transformations such as scaling and rotation. Finally, global features, such as the color histograms [17],[18] are the most computationally efficient, but typically have less discriminating power than keypoint-based and block-based descriptions and poor tolerance to global modifications, such as global color changes.

Another key design choice is whether a video signature will include temporal features or not. Spatial video signatures [2],[14] are a common design choice and are typically structured as a sequence of frame signatures. The underlying temporal relation between the frame signatures is then used only during the matching or retrieval process. In contrast to spatial video signatures, temporal video signatures actually contain temporal features, e.g. based on motion statistics [13], or by extraction of features jointly from groups of frames [12],[15]. Clearly, a video signature may have both a spatial component and a temporal component, which is the case with the MPEG-7 Video Signature, as will be seen later on.

Furthermore, a video signature may be viewed as temporally dense or sparse. Here, we use the term *dense* to refer to a video signature which describes every frame in the video, as in [2],[13]. Conversely, sparse video signatures operate at the keyframe level, with keyframes typically extracted at the shot level [11] or at fixed intervals [18].

Although the above can be viewed as the primary design choices, there are other choices which will have an impact on identification performance and computational efficiency, such as extraction from native resolution or sub-sampled frames, the use of intensity or color information, etc. All these factors need to be correctly balanced in the signature design. For example, keypoint-based frame signatures are typically employed in sparse video signatures [12] to ensure that their extraction and matching are computationally tractable.

Hampapur *et al.* studied the problem of identification of

video content embedded in longer video content in [2]. That early evaluation used only a few hundred copy queries 1.3 to 20.3 seconds in length, subjected to only a single modification from the original, namely a resolution change. In that evaluation, a block-based ordinal signature achieved the highest performance, with nearly 0% false positive and false negative rates for 10.3 second clips. The signature was calculated at each frame by dividing it into blocks, calculating the mean intensity of each block, sorting them in ascending order and then using the rank vector as the frame feature. However, for shorter queries, the best performance reported in [2] drops to only ~94% recall and precision. Furthermore, [2] does not report the matching segment localization accuracy.

A more recent evaluation of video signatures was carried out by Law-To *et al.* [3]. That evaluation was performed under two different scenarios. The first scenario used 72 5-second copy segments embedded in unrelated content and subjected to a single modification chosen from contrast change, cropping, blurring, letter-box, logo overlay or zoom. That experiment showed a block-based temporal ordinal signature, based on the mean intensities of frame blocks ranked in the temporal direction, achieving the highest performance, with perfect recall and precision. The second scenario used 60 copy segments between one and 10 seconds embedded in unrelated content and subjected to a random combination of modifications. In that experiment a keypoint-based signature achieved the highest identification performance, with 82% recall at 95% precision. That signature was based on keypoint temporal trajectory properties, calculated from a differential description of the region around Harris interest points identified in keyframes [13]. However, Law-To *et al.* note that their evaluation was limited by the size of the database, which contained only three hours of video and few queries.

TRECVID [23],[24] has also been conducting content-based copy detection evaluations. TRECVID 2009 completed in March 2010 [25] and included video-only experiments, conducted on 938 copy queries and 469 non-copy queries, with a total duration of ~7.3 hours, exhibiting modifications including picture-in-picture, pattern insertion, re-encoding, frame dropping, cropping, closed captioning, etc. The primary performance measure used was minNDCR (minimum Normalized Detection Cost Rate) for two different profiles, “NOFA” (No False Alarm) and “Balanced”. minNDCR does not measure the copy localization accuracy, which was reported as secondary information. One of the best performing methods in TRECVID 2009 was proposed by Liu *et al.* [11]. That method utilizes sparse temporal sampling, representing each shot by a single keyframe, describes the frame contents using SIFT features, and uses locality sensitive hashing (LSH) for efficient indexing and query. Copy detection is based on a gradual keyframe-level query process, followed by video-level merging. The method achieved an average minNDCR of ~0.37 and ~0.38 for the NOFA and Balanced profiles respectively. For NOFA, this translates to a detection rate of ~63%. For Balanced, it is more difficult to determine the actual success

TABLE I SUMMARY OF SELECTED VIDEO IDENTIFICATION METHODS AND EXPERIMENTS – PART I

Method	Detecti on Type	Localization Type	Data	Modifications
MPEG-7 Video Sig.	Embedded segments	Fully integrated: accurate to 1sec	Independence: ~70,000 30-sec queries to 1,900 180-sec originals Robustness: ~70,000 30-sec queries to 545 180-sec originals Query matching segment length: 2 sec, 5 sec, 10 sec	Wide range manually applied and tested
Liu <i>et al.</i> [11]	Embedded segments	Not integrated: diagnostic only	~7.3 hours DB 938 copy queries, 469 non-copy queries, average length 32 sec	Wide range manually applied and tested
Hampapur <i>et al.</i> [2]	Embedded segments	Not reported	~2 hours DB Few hundred queries, 1.3 sec to 20.3 sec	Resolution reduction
Law-To <i>et al.</i> [3]	Embedded segments	Not reported	~3 hours DB (a) 9 15-min queries, each with 8 5-sec copy segments (b) 1 30-min query with 60 copy segments between 1 sec and 10 sec	Wide range manually applied and tested
Paisitkiangk rai <i>et al.</i> [6]	Embedded segments (coarse)	Not reported	12,790 web videos (24 text queries) 24 queries used in experiments	Those already naturally present in content
Shen <i>et al.</i> [4]	Whole clips	NA	~11,000 TV ads at ~60 sec each 20 query clips	Those already naturally present in content
Shang <i>et al.</i> [5]	Whole clips	NA	(a) 12,790 web videos (24 text queries) (b) ~50,000 web videos including (a) 24 queries	Those already naturally present in content
Sarkar <i>et al.</i> [18]	Whole clips	NA	~38,000 web videos main DB ~21,600 query videos with modifications (from 1,200 videos of DB)	Wide range manually applied and tested

TABLE II SUMMARY OF SELECTED VIDEO IDENTIFICATION METHODS AND EXPERIMENTS – PART II

Method	Performance	Speed
MPEG-7 Video Sig.	95.49% correct detection and localization at $\leq 5ppm$ false alarm (95.49% recall at $\geq 99.9995\%$ precision)	Matching/localization: $\geq 1,500$ matches/sec (pairwise matching) Coarse signature matching: ~0.0001ms (pairwise matching)
Liu <i>et al.</i> [11]	NOFA: minNDCR ~0.37 (~63% recall at 100% precision) Balanced: minNDCR ~0.38	Mean processing time per query between 100 sec and 200 sec
Hampapur <i>et al.</i> [2]	<i>Ordinal signature</i> : Combined recall/precision between ~94% and 100%	Not reported
Law-To <i>et al.</i> [3]	(a) <i>Temporal ordinal signature</i> : 100% recall at 100% precision (b) <i>Keypoint signature</i> : ~82% recall at ~95% precision	<i>Temporal ordinal</i> : ~40 min for ~15-min query in ~3 hour DB <i>Keypoint</i> : ~27 sec for ~15-min query in ~3 hour DB (DB index)
Paisitkiangk rai <i>et al.</i> [6]	MAP 0.896 (DB indexing), MAP 0.911 (pairwise matching) ~20% recall at ~96% precision (with DB indexing)	0.157 sec (with DB indexing), 3.72 sec (pairwise matching)
Shen <i>et al.</i> [4]	BCS: ~60% recall at ~90% precision ~20% recall at ~95% precision	50ms (with DB indexing)
Shang <i>et al.</i> [5]	(a) MAP 0.953, ~20% recall at ~98% precision (b) MAP 0.885, ~20% recall at ~98% precision	(a) 3.7ms (with DB indexing) (b) 17ms (with DB indexing)
Sarkar <i>et al.</i> [18]	MAP 0.93, ~10% recall at >99% precision	30ms (with DB indexing)

and false alarm rates. An average localization accuracy of ~85% was achieved for both profiles, with a mean processing time per query between 100 and 200 seconds.

In [4] a near-duplicate video clip detection system is presented, comprising two detection schemes. The first scheme uses the bounded coordinate system (BCS), a compact representation which ignores temporal information and summarizes each entire video via a single vector. The second scheme utilizes frame symbolization (FRAS), mapping each video to a sequence of symbols and taking temporal order into account. Database indexing is used to reduce the search space and improve matching speeds. The system was tested on ~11,000 TV commercials, each with duration of ~60 seconds. The results show BCS achieving the best performance, with ~60% recall for ~90% precision and an average search time of ~50ms. However, the ROC curves in [4] show that the recall

rate drops to ~20% for a precision of ~95%, indicating that the system would not achieve high recall rates at very high precision levels. Since the system of [4] is designed to match entire short clips, it does not address the issue of detecting and localizing a video segment inside a longer unrelated video.

In [5], another near-duplicate video clip detection system is presented, where videos are represented by a compact spatiotemporal feature relying on relative grey-level intensity distributions within a frame and the temporal structure along the frame sequence. For retrieval, an inverted file index structure is used to achieve high speeds. The system was evaluated on a database of 12,790 videos retrieved from the web using 24 different queries. A mean average precision (MAP) of 0.953 was achieved, with an average query speed of 3.7ms. To demonstrate the system's scalability, it was also evaluated on an extended database of ~50,000 web videos,

producing a MAP of 0.885, with an average query speed of 17ms. However, as with [4], the ROC curves in [5] indicate that the recall rate would drop to ~20% for a precision of ~98%. This system also does not address the issue of detecting a video segment inside a much longer video segment.

In [6], another near-duplicate video clip detection system is presented, which can also address the problem of detecting a video sequence embedded in a longer video sequence with a coarse granularity. The system samples video at a rate of 1 frame/sec and extracts signatures from 10-frame sequences with a 50% overlap. Each clip signature is the average of its frame descriptors, each of which is an ordinal spatial intensity distribution (OSID). The 10-second clip signatures strike a balance between frame-level signature approaches and video-level signature approaches. Multi-probe locality sensitive hashing (MPLSH) is used to index the clips. The system was tested on the database of [5] and achieved a MAP of 0.936 and 0.947, with and without MPLSH, at the clip signature level. At the video level, i.e. for queries comprising multiple clip signatures the system achieved 0.896 and 0.911 MAP with and without MPLSH respectively, with a retrieval speed of 0.157 and 3.72 seconds per query. However, the ROC curves in [6] also indicate that the recall rate of the system would drop to a level of ~20% for a precision of ~96%. To test the scalability of the system, it was tested on enlarged versions of the original database, to ~25,000 and ~50,000 videos, with little performance degradation.

In [18] a near-duplicate video detection system is presented which can detect temporally reduced videos, but its design does not support detection of video content embedded in unrelated video material. The system extracts frame-level features based on the MPEG-7 color layout descriptor [9]. In order to reduce the descriptor size and increase search speeds, k -means-based clustering is performed on the frame-level descriptors and the cluster centroids are used as the video fingerprint. A two-stage matching approach is then used: first k nearest neighbor matches are found using vector-quantized fingerprint representation, distance look-up tables and dataset pruning. Then, match candidates are confirmed or rejected using either a distance thresholding-based approach or a registration-based approach. Experimental evaluation was conducted on 38,000 web videos. A subset of 1,200 videos was used to generate 18 duplicates per video by applying modifications including Gaussian blurring, gamma correction, JPEG compression, frame cropping, etc. The system achieved a MAP of ~0.93 at a query time of ~30ms. However, an examination of the FA/FR curves in [26] shows that for higher precision values, of ~99% or higher, the recall rate would drop significantly and very rapidly, from ~70% to less than 10%. As the authors note themselves, the method will not be effective if the query contains portions of multiple videos.

Tables I and II provide a summary of the key video identification works examined above, and also provide a comparison with the MPEG-7 Video Signature, which will be discussed later in this paper.

III. DEVELOPMENT OF THE STANDARD

A. Aim and Scope

The Video Signature was designed to find *identical* content, unlike previous MPEG-7 descriptors, such as Dominant Color and Scalable Color which were designed to find *semantically similar* content [9]. The scope of the Standard encompasses those aspects which are required for interoperability. There are four parts to the Standard, namely (i) the descriptor extraction and decoding, along with its DDL (Descriptor Definition Language [27]) and binary representation syntax [28], (ii) a reference software implementation and source code for the Video Signature Tools [29], (iii) the conditions and dataset for ensuring conformance to the Standard [30], and (iv) an exemplary pair-wise matching and localization scheme, as used during the MPEG-7 evaluation process [31]. Note that [31] is merely informative, as the Standard does not specify a normative matching, retrieval or database indexing scheme. Those options are left open to the application developers.

B. Requirements

The first step before the technical development of the Video Signature Tools was the definition of the requirements, which were subsequently used for the evaluation of the proposed technologies. A set of 11 requirements were identified, namely

- i) Uniqueness: A signature will describe a video uniquely.
- ii) Robustness to editing operations: The robustness of the Video Signature was tested against a large number of common editing operations. These are examined in more detail later.
- iii) Independence: The Video Signature shall achieve a false alarm rate of $\leq 5 \text{ ppm}$ (parts per million), i.e. no more than five false matches per million comparisons.
- iv) Fast matching: This refers to the pair-wise matching and localization speed, since the Standard does not specify any mandatory retrieval or database indexing scheme. The requirement was that it will be possible to match and localize at least 1,000 clip pairs per second on a desktop PC under the partial content matching query scenarios described later.
- v) Fast extraction: This was used as a secondary criterion in the MPEG-7 selection process, since large-scale signature extraction is typically an offline process. In practice, the extraction speed of the Video Signature from uncompressed video is ~900 frames/sec on a standard desktop PC.
- vi) Compactness: The compactness requirement was that the Video Signature shall not exceed 30,720 bits/sec of content at 30 frames/sec, i.e. 1,024 bits per frame on average.
- vii) Non-alteration of the content: This ensured that the Video Signature shall not require any content modifications, making it suitable for immediate use with all existing content.
- viii) Self-containment of the signatures: This requirement ensured that no access to the video content from which a signature is extracted shall be necessary for matching, making it possible for applications to operate at the signature level.
- ix) Coding independence: This requirement ensured that the extraction of the signatures shall be independent of the encoding of the video content.

TABLE III KEY NOTATIONS

Symbol	Description	Symbol	Description
u	Frame signature element	ThA_j	Ternarization threshold for averages category A_j
x	Ternarized frame signature element	ThD_j	Ternarization threshold for differences category D_j
$\mathbf{x}$	Vector of all 380 ternarized frame signature elements	Thd	Single Jaccard distance threshold (Matching Stage 1)
$\mathbf{w}$	Set of 5 fine frame-level signature words	Thd_c	Composite Jaccard distance threshold (Matching Stage 1)
c	Fine frame-level signature confidence	Thx_H	Frame signature L1 distance threshold for Hough transform (Matching Stage 2)
$\mathbf{s}$	Fine frame-level signature	Thx_G	Frame signature L1 distance threshold for matching interval growing (Matching Stage 3)
$\mathbf{b}$	Coarse segment-level signature bag-of-words	ThT	Matching segment duration threshold (Matching Stage 3)
$\mathbf{b}$	Set of 5 coarse segment-level signature bags-of-words	Thc_1	Frame-level confidence threshold (Matching Stage 3)
v	Video Signature for a video	Thc_2	Matching segment confidence threshold (Matching Stage 3)
KP / PP	Key Picture / Predicted Picture		
GOP	Group of Pictures		

x) Partial matching: This requirement ensured that it will be possible to detect a duplicated video segment embedded within a longer segment of different content.

xi) Accurate temporal localization of duplicated and embedded content: This requirement specified that the start point/duration of duplicated and embedded content shall be identified within one second/two seconds of the actual start point/duration for a match to be deemed successful.

These requirements were designed to ensure that the Standard will be applicable to many different applications. Where hard limits are imposed, e.g. a false alarm rate of $\leq 5ppm$ or a maximum descriptor size, those were agreed after deliberations within the MPEG-7 group on the appropriate limits that would maximize the benefit of the Standard in different applications and products.

IV. APPLICATIONS

The Video Signature has a wide range of potential applications. Below we list a few of the main applications, although this is not meant to be an exhaustive list.

i) Rights management and monetization: For content owners, the aim is the detection of possible copyright infringement or content monetization online. For content consumers, the aim is to identify the copyright owner, if any, to avoid infringement or ensure proper attribution.

ii) Distribution management: An organization may operate a video fingerprint database for all its sensitive content and automatically identify and stop accidental transmission of such content via email, unauthorized copy to external device, etc.

iii) Usage monitoring: The aim here is to track and record statistics such as distribution and frequency of content usage. For example, an advertising agency may wish to check whether its material is distributed or broadcast as expected.

iv) Video content-based linking: In a similar fashion to text, the video content in a web page can be used to infer association with other web pages.

v) Database management and de-duplication: For high-volume content creators and owners, such as studios or archives, as well as for personal video libraries.

The development process of the MPEG-7 Video Signature Tools was guided by the requirements presented in Section III, which were designed to include a wide range of applications.

Naturally, different applications will weigh those requirements differently. For example, certain applications, such as personal database management, may not require a false alarm rate of $5ppm$ (parts per million) or less. However, the MPEG-7 Video Signature descriptor delivers this false alarm rate while achieving a detection rate in the order of 96%, which is higher than the recall rates typically reported in the literature for significantly lower precision rates. Clearly, if a higher recall rate is required, this can be achieved by relaxing the false alarm rate requirement. For applications such as near-duplicate video clip detection in large databases, the accurate localization requirement will not apply. For such applications, the localization aspect of the matching function may be disabled, which will also result in significantly improved matching speed profiles. Furthermore, in all applications which involve a known database, suitable database indexing may be used for the improvement of retrieval speeds.

As an International Standard, the MPEG-7 Video Signature Tools bring two main benefits to the different systems and applications. Firstly, MPEG follows a systematic peer-reviewed evaluation process, leading to the adoption of the best technologies from various proposals. Secondly, the Video Signature Tools enable interoperability, i.e. they allow different users and systems to talk to each other in terms of descriptors rather than in terms of videos. This is achieved in a number of ways. The syntax of the Video Signature is both unambiguous and flexible. Reference software is provided as part of the Standard, to aid in the development of compliant products and services. Developers may also generate their own implementations and then test their conformance according to the conformance conditions and dataset, which also form part of the Standard. The advantages of interoperability are different for different parties in the content creator to consumer chain. For example, for content creators, standardized description tools will allow them to generate standardized content descriptions to register their content in different databases, while users will be able to search different databases using the same standardized descriptors.

V. VIDEO SIGNATURE EXTRACTION

The Video Signature comprises two parts: (i) fine signatures extracted from individual video frames and (ii) coarse signatures extracted from sets of frame-level signatures based

Fig. 1. Organization of the frame signature $\mathbf{x}$, words $\mathbf{w}$ and confidence value c into the fine signature $\mathbf{s}$.

on a “bag-of-words” approach. To aid in the understanding of the Video Signature extraction, compression and matching, a list of the main symbols used is presented in Table III.

A. Fine Signature

The fine signatures are extracted from 32×32 pixel 8-bit luminance information, derived by the block averaging of the luminance channel of a frame. Each fine signature comprises (a) a set of local features, termed the “frame signature”, (b) a small representative subset of the local features, organized into different “words” of the frame signature, and (c) a global “frame confidence” measure. This is illustrated in Fig. 1.

The computation of the local features involves the calculation of local average intensities and differences at various levels of granularity in the frame. Each local feature is termed an element or dimension of the frame signature. In total there are 380 such elements in a frame signature, 32 averages and 348 differences. Collectively, the frame signature elements are designed to capture the local intensity content and intensity interrelations at different regions and scales of the luminance channel while being very simple to compute. A sample of these elements is shown in Fig. 2, and all other average and difference elements are calculated in a similar fashion. A detailed description of all 380 elements is not provided here, but can be found in [28].

The frame signature elements are divided into different categories according to characteristics such as the element type (average or difference) and pattern type. Overall, there are two categories of average elements, A_1 and A_2 , with 20 and 12 elements respectively (32 average elements in total), and eight categories of difference elements, $D_1 \dots D_8$, with 116, 25, 36, 30, 62, 9, 50 and 20 elements respectively (348 difference elements in total). This categorization is significant for the next step, which is the ternarization of the elements.

More specifically, the frame signature comprises ternarized elements, i.e. quantized to three levels. We found that ternarization produced improved results compared to binarization, especially for the difference elements, avoiding element value oscillation in flat frame regions. For the average element categories A_1 and A_2 , this ternarization proceeds as follows. Let u_i , $i \in \{0, \dots, N_j^A - 1\}$, be an average element of category A_j , $j \in \{1, 2\}$, containing a total of N_j^A elements. The ternarized element x_i is calculated as

$$x_i = \begin{cases} 2 & \text{if } (u_i - 128) > ThA_j \\ 1 & \text{if } |u_i - 128| \leq ThA_j \\ 0 & \text{if } (u_i - 128) < -ThA_j \end{cases} \quad (1)$$

The threshold ThA_j is not fixed but adaptive, re-calculated for each frame and for each category A_j as the 33.3% percentile rank of the absolute values $|u_i - 128|$. That is, for each category A_j with N_j^A elements we calculate $|u_i - 128| \forall i \in \{0, \dots, N_j^A - 1\}$ and sort the results in ascending order. The threshold ThA_j is then the n^{th} element of the sorted list, where $n = \lfloor 0.3333 \cdot N_j^A \rfloor$. For the difference element categories $D_1 \dots D_8$ the ternarization proceeds in a similar fashion. Let u_i , $i \in \{0, \dots, N_j^D - 1\}$, be an element of category D_j , $j \in \{1, \dots, 8\}$, with a total of N_j^D elements. The ternarized element x_i is calculated as

$$x_i = \begin{cases} 2 & \text{if } (u_i) > ThD_j \\ 1 & \text{if } |u_i| \leq ThD_j \\ 0 & \text{if } (u_i) < -ThD_j \end{cases} \quad (2)$$

The threshold ThD_j is again adaptive, re-calculated for each frame and for each category D_j as the 33.3% percentile rank of the absolute values $|u_i|$, as described above. The aim of making the ternarization thresholds frame-adaptive and category-adaptive is to achieve a more uniform distribution of the frame signature elements across the three quantization bins. Making the thresholds frame-adaptive avoids the situation of a frame signature with diminished information content, e.g. when a video frame is too bright/dark or of very poor contrast, while making them category adaptive prevents the statistics of a certain type of element influencing the quantization of other elements. The vector $\mathbf{x}$ of all 380 ternarized elements forms the frame signature.

In the context of the Video Signature, a “word” refers to a compact representation of the complete frame signature and is a small ordered subset of elements of vector $\mathbf{x}$, i.e. a simple projection from a 380-dimensional space to a lower-dimensional space. For two video frames, the distance between two corresponding words is an approximation of the distance between the complete frame signatures. The Video Signature utilizes Q Ψ -dimensional words with $Q = 5$ and $\Psi = 5$, i.e. five different projections from the original 380-dimensional space of the frame signature to a 5-dimensional space. The values $Q = 5$ and $\Psi = 5$ were found through experimentation to provide a good tradeoff between compactness and discriminative power. Collectively, the five words $\mathbf{w}$ provide a good and compact representation of the complete frame signature and are used at a later stage for the formation of the coarse signature.

The global frame confidence value is an 8-bit integer value calculated from the 348 difference elements of categories $D_1 \dots D_8$. More specifically, we place all 348 absolute values of difference elements in a single list and sort in ascending order. Then, the 174th element of the sorted list, denoted by M , is selected as the median value and the global frame signature confidence c is calculated as

$$c = \min(\lfloor M \cdot 8 \rfloor, 255) \quad (3)$$

A low confidence measure implies that the frame is a flat

Fig. 2. Representative frame signature elements: (a) One A1 element, $\{a\}$, the mean luminance of the pixels in region a , (b) One A2 element, $\{a\}$, (c) Four D1 elements, $\{a\}-\{b\}$, $\{c\}-\{d\}$, $\{e\}-\{f\}$ and $\{g\}-\{h\}$, (d) Two D2 elements, $\{a\}-\{b\}$ and $\{c\}-\{d\}$, (e) Two D3 elements, $\{a\}-\{b\}$ and $\{c\}-\{d\}$, (f) Two D4 elements, $\{a\}-\{b\}$ and $\{c\}-\{d\}$, (g) Two D5 elements, $\{a\}-\{b\}$ and $\{c\}-\{d\}$, (h) Two D6 elements, $\{a\}-\{b\}$ and $\{c\}-\{d\}$ – note that the regions overlap, (i) Two D7 elements, $\{a\}-\{b\}$ and $\{c\}-\{d\}$, and (j) Four D8 elements, $\{a\}-\{b\}$, $\{c\}-\{d\}$, $\{e\}-\{f\}$ and $\{g\}-\{h\}$.

Fig. 3. Extraction pattern of coarse signatures $\mathbf{b}$ from the fine signatures $\mathbf{s}$.

image with little or no information in the difference elements. This information is subsequently used in the matching process.

The “frame confidence” c , the “words” $\mathbf{w}$ and the “frame signature” $\mathbf{x}$ make up the fine signature $\mathbf{s}$ of the Video Signature. It is worth noting that, in the binary representation, the ternary elements of the frame signature and the words are not represented by two bits each but are encoded. More specifically, each group of five consecutive elements is encoded into an 8-bit value, resulting in a 20% size reduction. Thus, the fine signature described here is quite compact at only 656 bits of storage. In practice, as part of the Video Signature, the fine signatures require less space for storage as temporal redundancies can be exploited, as will be seen later.

B. Coarse Signature

The coarse signatures are extracted from sets of fine signatures based on a “bag-of-words” approach [14],[32],[33],[34].

Fig. 4. Organization of coarse signatures $\mathbf{b}$ and fine signatures $\mathbf{s}$ in the Video Signature bitstream for a video with M frames.

The bag-of-words representation is extracted for temporal segments of 90 consecutive frames. As seen earlier, each fine signature contains five words, i.e. five subsets of the complete frame signature. For each of these five words, the values that it takes over the 90 frame sequence are plotted into a 243-bin histogram (since we have five ternary values, $3^5 = 243$). Thus, five histograms h_k , $k \in \{1, \dots, 5\}$, are generated, one for each of the five subsets of the complete frame signature for each word. Then, each histogram is binarized, by setting the value of each bin to 1 if it is greater than 0, and leaving as 0 otherwise. This gives rise to five binary occurrence histograms b_k , $k \in \{1, \dots, 5\}$, which become the coarse segment signature $\mathbf{b}$ for the 90-frame segment. In the Video Signature coarse signatures are generated for 90-frame segments but with a 45-frame overlap, e.g. if a first coarse signature is extracted for frames m to $m+89$, then the second coarse signature will be extracted for frames $m+45$ to $m+134$, etc. This is illustrated in Fig. 3. Each coarse signature requires 1215 bits for storage which, given the coarse signature overlap, results in 810 bits/sec at 30 frames/sec.

C. Video Signature Organization

At the bitstream level, the Video Signature of a piece of video content contains all the coarse signatures first, followed by the fine signatures, as illustrated in Fig. 4. This arrangement facilitates efficient bitstream access for a coarse-

Fig. 5. Alignment of coarse signature temporal segments with compression temporal segments.

Fig. 6. Temporal segment transformed into a prediction difference matrix. Ternary "0" shown as white, ternary "1" as gray, and ternary "2" as black.

to-fine matching process, which is examined in Section VII.

VI. VIDEO SIGNATURE COMPRESSION

The frame signatures contain a considerable amount of temporal redundancy which is eliminated through a specially designed lossless compression scheme. The compression is applied only to groups of 45 consecutive frame signatures, aligned with the coarse signature temporal segments, as illustrated in Fig. 5. Compression is not applied to the coarse signatures or the fine signature words or frame confidences.

Every compressed temporal segment starts with a Key Picture (KP) frame signature. A KP frame signature is coded in a non-predictive way. The other frame signatures of the compressed temporal segment are Predicted Picture (PP) frame signatures, each predicted from the temporally preceding frame signature. A Group of Pictures (GOP) is defined as a set of frame signatures between two KPs, including the temporally first KP and all PPs before the next KP. A GOP therefore consists of a KP and zero or more PPs. The method of selecting the lengths of GOPs needs to balance maximization of the compression performance against flexible random access to specific frame signatures. The criterion that is used in this work consists of counting the number of elements that remain unchanged between the current and reference frame signatures and comparing that number to a predefined constant. If this number is lower than the constant, i.e. very few elements remain unchanged, this indicates a low correlation between the current and reference signatures, and the current signature is encoded as a KP frame signature.

Frame signatures are transformed by modulo-3 subtraction between corresponding elements in the current and reference frame signature, $\tilde{x}_{i,m} = (x_{i,m-1} - x_{i,m}) \pmod{3}$ where i and m are the element and frame index respectively. Such

Fig. 7. The structure of a coded GOP.

Fig. 8. Run-length coding of vectorized prediction matrix. "v" refers to variable number of bits, while "1" refers to length of one bit.

transformed frame signatures make up a prediction difference matrix, as illustrated in Fig. 6, which is scanned into a 1-dimensional vector. This vectorization step is done by concatenating the columns of the prediction difference matrix which exploits the fact that within one GOP some frame signature elements are likely to stay constant, which leads to longer zero runs. The structure of a GOP as coded (composed of a KP and PPs transformed into a prediction difference matrix) is depicted in Fig. 7.

The modulo-3 difference operation between ternary elements of the signatures results in the three ternary values, "0", "1" and "2". The vectorized prediction matrix is thus composed of these three ternary symbols, which are coded differently. Ternary "1" and "2" are coded with one bit each, binary "0" for ternary "1" and binary "1" for ternary "2". After each ternary symbol, a run-length codeword for a run of zeros is inserted. If no zeros follow, a codeword for zero length is used. Ternary "0" is implicitly encoded, since after each non-zero ternary symbol a zero run must follow, unless the non-zero ternary symbol is last in the vectorized prediction matrix. This coding scheme for the vectorized prediction difference matrix is depicted in Fig. 8.

For the selection of the entropy coding method two aspects need to be considered, namely complexity and the source statistics. Since the presented compression method is aiming for low-complexity implementations, we chose fixed codeword variable length entropy coding, as it allows for parsing of a bitstream and decoding the codewords by using a set of very simple algorithmic steps. More specifically we used Exponential-Golomb codes, also known as Exp-Golomb, which offer a good approximation of the ideal code for power-law distributions of form $p_\omega \sim \beta \omega^{-\alpha}$, where p_ω is the probability of unsigned integer symbol ω , with constants

$\beta > 0$ and $\alpha > 0$. These codes are a good fit for power-law distributions that have long tails, i.e. probabilities p_ω are relatively large for large ω , which was experimentally confirmed to be the case for our source data [35]. Exp-Golomb codes are parameterized by a non-negative order number γ , for which the codewords are of length

$$l_\omega = 1 + 2 \lfloor \log_2(\omega + 2^\gamma) \rfloor - \gamma \quad (4)$$

When selecting the order number for the Exp-Golomb coding it was experimentally derived that $\gamma = 2$ provides the best fit to the probability distribution of the zero run-lengths.

The decoding consists of inverting the operations performed at the encoder: first Exp-Golomb and run-length decoding are performed and the resulting ternary symbols are put into a prediction difference matrix in the vectorization scan order. Then, frame signatures within each GOP are processed in order of increasing frame index m starting from the KP, and then all PPs are reconstructed by computing $x_{i,m} = (x_{i,m-1} - \tilde{x}_{i,m}) \pmod{3}$. Finally, each temporal segment, potentially composed of multiple GOPs, is put in its correct temporal position in the Video Signature. It should be noted that, in a similar fashion to video coding standards like MPEG-2 [36] and MPEG-4 AVC [37], it is the decoding process of the Video Signature that forms the normative part of the Standard. Users are free to implement their encoder in different ways, provided that the resultant stream can be decoded by the normative decoder.

Experiments on the MPEG-7 dataset showed that this scheme achieves a mean compression ratio of $\sim 23\%$ for the frame signature blocks, which translates to a mean compression ratio of $\sim 27\%$ for the complete Video Signature. In comparative tests, this scheme outperformed three generic compression algorithms, namely bzip2, Prediction by Partial Match (PPM)+Arithmetic Coding, and LZMA [38]. It is also worth noting that, by aligning the compressed temporal segments with the coarse signature temporal segments, and by leaving the coarse signatures uncompressed, only a small fraction of the compressed segments require decompression with the coarse-to-fine matching process described next.

VII. VIDEO SIGNATURE MATCHING AND LOCALIZATION

The Standard does not specify a mandatory matching or retrieval scheme. For informative purposes, this section describes the pair-wise matching and localization method that was used during the MPEG-7 evaluation process. The matching between two Video Signatures v^1 and v^2 is carried out in three stages. The first stage uses the coarse signatures to identify candidate matching segments. The second stage uses the fine signatures to identify candidate parameters of frame rate ratio and temporal offset between the candidate matching segments. The third stage performs frame-by-frame matching to determine candidate matching intervals using the fine signatures, assesses the quality of each match, and selects a best matching interval between v^1 and v^2 .

A. Stage 1: Coarse Signature Matching

In the first matching stage, all of the coarse signature temporal segments of Video Signature v^1 are compared with all of the temporal segments of Video Signature v^2 . For two segments f^1 and f^2 , their similarity is assessed by comparing the histograms b_k^1 and $b_k^2 \forall k \in \{1, \dots, 5\}$ and merging the results. The distance between b_k^1 and b_k^2 is measured by the Jaccard distance metric given by

$$d_k = 1 - \frac{b_k^1 \cap b_k^2}{b_k^1 \cup b_k^2} \quad (5)$$

This measures the distance of the segments f^1 and f^2 as a function of the distinct words they have in common and all the distinct words that they contain jointly. Since each coarse signature $\mathbf{b}$ contains $Q = 5$ bags of words, we have five Jaccard distances. These distances are fused to give the composite distance d_c as

$$d_c = \sum_{k=0}^{Q-1} d_k \quad (6)$$

A decision on the similarity of the segments is reached by thresholding each of the Jaccard distances d_k and the composite distance d_c . More specifically, the segments f^1 and f^2 are passed to the second stage of matching if more than half of the d_k Jaccard distances are less than a threshold Thd and the composite distance d_c is less than another threshold Thd_c , otherwise they are declared not matching and the matching process ends.

B. Stage 2: Temporal Parameter Estimation

For the segment pairs passed to this stage, a Hough transform is used to identify candidate parameters of temporal offset and frame rate ratio between the segments of each candidate pair. These are linear properties and can therefore be estimated using two strongly corresponding frame pairs.

First, the L1 distance between the frame signature elements of all frame pairs between the two segments are calculated. The frame pairs whose distance is smaller than a threshold Thx_H are selected as strongly corresponding frame pairs. Then, two strongly corresponding frame pairs are selected to calculate the temporal offset and frame rate ratio, and a vote is cast to the calculated parameters in the Hough space. The voting is done for all possible combinations of two strongly corresponding frame pairs. Finally, the temporal parameter sets with a high response in the Hough space are selected as candidate parameter sets, and are passed to the third stage of matching. If no parameter set in the Hough space satisfies the selection criteria, the segment pairs are declared not matching and the matching process ends.

C. Stage3: Localization and Verification

In the third stage of matching, the matching interval, i.e. the start and end position of the match between the two segments, is determined by temporal interval growing based on a frame-by-frame matching using the frame signatures.

First, the estimated temporal offset is used to determine the

Fig. 9. The MPEG-7 query scenarios, referred to as direct or partial matching, with copy segment lengths of two, five and 10 seconds. Segments in red indicate unrelated non-reference content.

initial temporal matching position. Then, using the estimated frame rate ratio, the temporal interval is extended frame-by-frame towards both temporal directions by calculating the L1 distance between the frame signature elements of corresponding frames. The temporal extension stops when the frame signature L1 distance exceeds a threshold Thx_G . If the length of the resultant matching interval is shorter than a given minimum duration ThT , the matching interval is eliminated as a non-match. Otherwise, the frame confidence element associated with each frame in the matching interval is checked to verify the match. The overall confidence of the matching interval is calculated as the ratio of the number of frames which have a frame confidence that is higher than a threshold Thc_1 to the total number of frames in the interval. If the overall confidence is below a level Thc_2 , the matching interval is eliminated as a false match caused by frames with low information content.

This process is carried out for all of the candidate temporal parameter sets, thus generating multiple candidate matching intervals. The best interval is finally selected as a function of the candidate intervals' L1 distances and lengths.

The reference software contains a complete implementation of this matching and localization method and provides default values for all thresholds. However, different values, or different matching approaches, may be used in different applications to achieve the desired operational characteristics.

VIII. MPEG-7 EVALUATION METHODOLOGY AND RESULTS

A. Dataset

The dataset used in the MPEG-7 experiments comprises footage of various types, such as film, news, documentary, cartoons, sport, home video, etc. The independence tests were conducted on $\sim 70,000$ 30-second queries, with a total duration of ~ 570 hours, and 1,900 3-minute original, non-matching clips, unrelated to each other, with a total duration of ~ 95 hours. Thus, a total of over 120 million video comparisons were made in order to derive the required false alarm

operating settings. The robustness tests were conducted on $\sim 70,000$ 30-second queries, carrying a wide range of content modifications and with a total duration of ~ 600 hours, and 545 3-minute reference clips, with a total duration of ~ 27 hours, in order to derive the success rates of the Video Signature.

B. Query Conditions

The Video Signature Tools were evaluated under two query conditions, i.e. (i) direct content matching and (ii) partial content matching. With direct content matching, the whole query clip matches with a part of the original clip. With partial content matching, only a part of the query clip matches with a part of the original clip. Each of the two query types has three query scenarios in order to evaluate performance at different temporal granularities. The query scenarios correspond to durations D of the segment to be matched, where $D = 2, 5$ and 10 seconds. In the case of partial content matching, D indicates the minimum durations of the segment to be matched; the total duration of the query clip is 30 seconds. The different query scenarios are illustrated in Fig. 9.

C. Evaluation Criteria

All candidate technologies proposed to MPEG were evaluated by setting limits for three requirements, namely (i) independence (false alarm rate of $\leq 5ppm$), (ii) matching speed (at least 1,000 clip pairs per second under the partial content matching scenarios), and (iii) descriptor size (no more than 30,720 bits/sec of content at 30 frames/sec), and then deciding on which technology is adopted into the Standard based on the performance for the robustness requirement, expressed as the mean success rate. For a given modification, the success rate, R , is defined as $R = C/T$ where C is the number of correct matches found and T is the number of videos that match.

D. Independence Tests

In the independence test, each clip in a database of 1,900 3-minute clips unrelated to each other was divided into six 30-second clips, each of which was used to produce the required six queries, to a total of $\sim 70,000$ queries. Each query clip was then compared against all the unrelated 3-minute clips, to determine the operational settings for the $\leq 5ppm$ false alarm rate. Thus, the false alarm rate was determined based on approximately 120 million clip comparisons.

E. Robustness Tests

In the robustness test, a database of 545 3-minute clips of various types was used, and each clip generated the six required query clips. Each of these queries was then subjected to one of nine modification categories, with each category having between one and three intensity levels, giving 22 different modifications and a collection of over 70,000 query clips. The modified query clips were then compared with the original clips in order to produce the detection success rates at the operational parameters determined in the independence test. The complete list of content modifications is shown in

TABLE IV MODIFICATIONS IN THE MPEG-7 ROBUSTNESS TESTS

Modification Type	Modification Level		
	Heavy	Medium	Light
(TLO) Text/logo overlay	30% of screen area	20% of screen area	10% of screen area
(CIF) Compression at CIF resolution	64 kbps	256 kbps	512 kbps
(RR) Resolution reduction from SD	–	QCIF	CIF
(FR) Frame-rate reduction from 25/30fps	4 fps	5 fps	15 fps
(CAM) Camera capture at SD resolution	10% extra border	5% extra border	0% extra border
(VCR) Analog VCR recording/recapture	3 times	2 times	1 time
(MON) Color to monochrome	–	–	$I = 0.299R + 0.587G + 0.114B$
(BR) Brightness change (additive)	+36	–18	+9
(IP) Interlaced/progressive conversion	–	–	$P \rightarrow I \rightarrow P$ or $I \rightarrow P$

TABLE V VIDEO SIGNATURE SUCCESS RATES (%) IN THE MPEG-7 ROBUSTNESS TESTS

Modification	Direct Content Matching									Partial Content Matching								
	D = 2 sec			D = 5 sec			D = 10 sec			D = 2 sec			D = 5 sec			D = 10 sec		
	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L
TLO	79.63	89.54	98.90	87.16	93.03	99.45	88.62	93.21	100.00	66.97	71.93	78.35	65.50	75.05	78.35	65.50	73.76	97.95
CIF	99.27	99.63	99.63	99.82	100.00	99.82	99.63	100.00	100.00	96.88	99.44	98.53	97.30	99.08	98.53	97.98	99.08	98.90
RR	–	99.27	99.82	–	99.63	99.82	–	99.82	100.00	–	98.90	98.90	–	98.90	98.90	–	94.50	98.53
FR	99.63	99.27	99.63	99.63	99.82	99.82	100.00	100.00	100.00	81.18	86.11	96.34	81.89	86.51	99.44	93.73	94.62	99.72
CAM	79.27	92.27	92.29	83.67	93.55	95.05	91.01	95.21	95.60	49.17	90.61	91.01	77.25	92.82	93.94	86.42	93.55	94.50
VCR	94.42	94.81	95.78	96.45	96.85	96.88	96.59	97.04	97.24	93.63	94.07	95.96	95.68	96.11	96.70	96.25	96.48	96.69
MON	–	–	99.82	–	–	100.00	–	–	99.82	–	–	99.08	–	–	99.27	–	–	99.63
BR	100.00	98.53	99.63	100.0	99.45	99.82	100.00	100.00	99.82	98.72	98.35	98.90	98.72	98.72	98.72	98.72	98.72	99.45
IP	–	–	99.82	–	–	100.00	–	–	99.63	–	–	98.90	–	–	99.27	–	–	99.27

Modification Levels: H:Heavy M:Medium L:Light

TABLE VI OVERALL SUCCESS RATES OF VIDEO SIGNATURE

Matching Scenario		Average Success Rate (%)
Direct	2 sec	96.65
	5 sec	97.76
	10 sec	98.22
Partial	2 sec	91.73
	5 sec	93.43
	10 sec	95.12
All		95.49

TABLE VII MATCHING SPEED RESULTS

Matching Scenario		Matches Per Second
Partial	2 sec	980
	5 sec	1618
	10 sec	2023
All		1540

Table IV. Some modifications, such as text/logo overlay, are straightforward, while others introduce a more complex combination of changes. For example, camera capture at SD resolution results in a resolution change, introduction of a border, and a significant change in color, while frame rate reduction entails temporal interpolation at 4 frames/sec. The MPEG-7 evaluation procedure prevented any proposed method from anticipating a modification or otherwise using certain characteristics of any given modification, e.g. no proposed method could rely on the fact that frame rate

reduction was happening at known frame rates, or the locations of the text/logo overlays.

In the MPEG-7 evaluation methodology, accurate localization was an integral part of successful matching, i.e. a detection was deemed successful only if it adhered to strict localization conditions. For the direct content matching, a match was correctly detected when the detected start point was within one second of the actual start point. For partial content matching, correct detection required that (i) the detected start point in the original clip be within one second from the true start point, (ii) the detected start point in the query clip be within one second of the actual start point, and (iii) the detected duration of the matching part be within two seconds of the actual duration of the matching part.

The success rates of the Video Signature are shown in detail in Table V, and the average success rates are summarized in Table VI. The overall success rate of the Video Signature is 95.49%. As can be seen from Table V, the detection performance varies with different content modifications, strength levels, query conditions, and the length D of the segment to be matched. The results show that, as expected, direct content matching is a simpler problem than partial content matching. In all cases, the length of the original video is of no importance, as the matching process is designed to “scan” the original video in order to detect any matching segment. The results also indicate that the length D of the segment to be matched is of high importance, with performance increasing as D increases. This is expected, since the descriptors of longer segments will, in general, carry more

information, making their identification and localization less error-prone. The durations $D = 2, 5$ and 10 seconds used here are much lower than the average matching clip durations typically reported in the literature, and were deliberately chosen to test the performance limits of the Video Signature.

F. Speed Tests

The extraction speed of the Video Signature descriptor from uncompressed video is ~ 900 frames/sec on a standard PC.

The matching speed of the Video Signature was measured by querying a set of 1,700 3-minute clips with 300 30-second query clips from the partial content matching tests, i.e. three sets of 100 query clips with $D = 2, 5$ and 10. The average matching speed for each of the three sets, and the overall average, is shown in Table VII. The overall matching speed achieved was just over 1,500 matches per second on an Intel Xeon X5460 (single core implementation), running at 3.16 GHz and with 8GB of RAM.

Table VII shows that the matching speed increases as the length D of the matching segment increases. Since comparing a 30-second clip to a 3 minute clip to identify and localize a common segment of length D is more difficult and error-prone as D becomes smaller, this results in an increased activation rate of Stages 2 and 3 of the matching and localization process described in Section VII. This entails more complex processing than Stage 1, and increased overall processing times. Stage 1 of matching, i.e. the pair-wise matching of coarse segment-level signatures, is deterministic in terms of speed, and its speed is in the order of 0.0001ms per coarse signature pair. It should again be noted that all matching speed figures refer to pair-wise matching. No database indexing algorithm was used in our experiments to speed up the matching process, in order to obtain an accurate picture of the detection performance of the Video Signature descriptor itself.

G. Storage Costs

In its uncompressed form, the fine frame-level signature is quite compact, requiring only 656 bits of storage. The coarse segment-level signature is also very compact, requiring only 1215 bits of storage. Thus the complete uncompressed Video Signature storage cost is 683 bits/frame, or 20,490 bits/sec at 30 frames/sec, i.e. ~ 9 MB per hour of video at 30 frames/sec. In its compressed form, the complete Video Signature storage cost is, on average, 184 bits/frame, or 5,532 bits/sec at 30 frames/sec, i.e. ~ 2.5 MB per hour of video at 30 frames/sec.

H. Syntax Flexibility

An important decision in designing the robustness experiments, and the Video Signature itself, was that robustness to certain modifications should be achieved through the flexible syntax of the Video Signature rather than by the core descriptor. Thus, to achieve good robustness to modifications such as picture-in-picture of arbitrary size, or black bar insertion of arbitrary size, with or without aspect ratio changes, one efficient approach, employed here and in other works [11],[39], is to first detect such modifications and

then describe only the spatial regions of interest. As defined in the Standard, a Video Signature for a given video may carry a descriptor for the full spatial extent of a video, or for one or more spatial sub-regions, or for the full spatial extent and for one or more sub-regions, making it possible to address modifications such as picture-in-picture or black bar insertion. A number of methods exist in the literature for picture-in-picture detection and black bar detection [11],[39], which may be used in conjunction with the Video Signature, but are beyond the scope of the Standard.

IX. DISCUSSION

Clearly, a direct comparison between the methods reported in the literature is not straightforward for a number of reasons. While some works address the identification of entire short clips, others investigate the identification of a video segment embedded in a much longer unrelated video. In the latter case, localization may be treated as integral to correct identification, may be reported only as auxiliary information, or may not be reported at all. While some investigations place strong emphasis on balancing the false positives and false negatives in a system, others aim at extremely low false alarm rates. This, in turn, results in diverging performance assessment measures. Furthermore, some investigations focus on the evaluation of entire retrieval systems, including an appropriate database indexing component to maximize query speeds with minimum degradation in identification performance, while others assess the performance of video fingerprints through pair-wise matching experiments. Finally, a like-for-like comparison between the techniques reported in the literature is not straightforward because of differences in the dataset sizes and compositions, the nature of the content modifications, and the number of queries on which results are reported.

Based on the results presented in Tables I and II, but keeping the above factors in mind, we believe that the MPEG-7 Video Signature compares favorably to recently reported methods. Tested on approximately one hundred and forty thousand queries, and with a wide range of content modifications, the MPEG-7 Video Signature achieved an average success rate of 95.49% with a false alarm rate no more than $5ppm$, i.e. with a precision ≈ 1 . This performance relates to both the identification and the accurate localization of a video segment embedded in a longer unrelated video segment. In terms of matching speed, the method presented in Section VII is, understandably, significantly slower than near-duplicate video clip retrieval techniques, which are not designed for accurate localization of content embedded in longer unrelated content. However, the method described in Section VII does not form part of the Standard, and is only used to assess the performance of the Video Signature. Different applications may use any matching and localization scheme, or employ any kind of suitable database indexing scheme, while still being conformant with the Standard. In terms of storage costs, the uncompressed/compressed Video Signature storage cost is very economic at 683/184 bits per

frame. Clearly, global descriptors extracted at the clip level, or descriptors which utilize temporal down-sampling of the content, may be more economic, but will also fail to produce the detection and localization accuracy achieved by the Video Signature.

X. SUMMARY AND CONCLUSIONS

The MPEG-7 Video Signature achieves high levels of robustness to common video editing operations, and is able to accurately detect and localize a piece of video content embedded in a longer piece of unrelated video content, with exceptionally low false alarm rates. The reference software developed during the standardization process is a useful tool for researchers and implementers and contains all the necessary routines for the Video Signature extraction, encoding, decoding and matching. Another essential resource for implementers is the conformance part of the Standard, which provides sample video data, sample Video Signature metadata extracted by the reference software, and the conditions that must be satisfied for any implementation to be considered conformant to the Standard.

REFERENCES

- [1] http://www.youtube.com/t/fact_sheet
- [2] A. Hampapur, K. Hyun, and R. Bolle, "Comparison of sequence matching techniques for video copy detection", *Conference on Storage and Retrieval for Media Databases*, pp. 194–201, 2002.
- [3] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford, "Video Copy Detection: A Comparative Study", *In Proc. 6th ACM International Conference on Image and Video Retrieval*, Amsterdam, NL, July 9–11 2007, pp. 371–378.
- [4] H. T. Shen, X. Zhou, Z. Huang, J. Shao, and X. Zhou, "UQLIPS: A Real-time Near-duplicate Video Clip Detection System", *In Proc. 33rd International Conference on Very Large Data Bases*, Vienna, AT, September 23–28 2007, pp. 1374–1377.
- [5] L. Shang, L. Yang, F. Wang, K.-P. Chan, and X.-S. Hua, "Real-time Large Scale Near-duplicate Web Video Retrieval", *In Proc. 2010 ACM International Conference on Multimedia*, Firenze, IT, October 25–29 2010, pp. 531–540.
- [6] S. Paisitkriangkrai, T. Mei, J. Zhang, X.-S. Hua, "Scalable Clip-based Near-duplicate Video Detection with Ordinal Measure", *In Proc. 2010 ACM International Conference on Image and Video Retrieval*, Xi'an, CN, July 5–7 2010, pp. 121–128.
- [7] ISO/IEC JTC 1/SC 29/WG 11, "MPEG Press Release", Hannover, DE, July 2008.
- [8] ISO/IEC JTC1/SC29/WG11, "Updated Call for Proposals on Video Signature Tools", Busan, KR, October 2008.
- [9] ISO/IEC 15938-3 2002, Information Technology – Multimedia content description interface – Part 3: Visual.
- [10] T. Sikora, "The MPEG-7 Visual Standard for Content Description – An Overview", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, June 2001.
- [11] Z. Liu, T. Liu, B. Shahraray, "AT&T Research at TRECVID 2009 Content-based Copy Detection", *2009 TRECVID Workshop*, Gaithersburg, MD, USA, November 16–17 2009. (<http://www-nlpir.nist.gov/projects/tvpubs/tv9.papers/att.pdf>)
- [12] A. Joly, O. Buisson, and C. Frelicot, "Content-based copy detection using distortion-based probabilistic similarity search", *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 293–306, February 2007.
- [13] J. Law-To, O. Buisson, V. Gouet-Brunet, and N. Boujemaa, "Robust voting algorithm based on labels of behavior for video copy detection", *In proc. 14th ACM International Conference on Multimedia*, Santa Barbara, CA, USA, October 23–27 2006, pp. 835–844.
- [14] C. Chiu, C. Yang, and C. Chen, "Efficient and Effective Video Copy Detection Based on Spatiotemporal Analysis", *In Proc. 9th IEEE International Symposium on Multimedia*, Taichung, TW, December 10–12 2007, pp. 202–209.
- [15] L. Chen and F. Stentiford, "Video Sequence Matching based on Temporal Ordinal Measurement", *Pattern Recognition Letters*, vol. 29, no. 13, pp. 1824–1831, October 2008.
- [16] M. Yeh and K. Cheng, "A Compact, Effective Descriptor for Video Copy Detection", *In proc. 17th ACM International Conference on Multimedia*, Beijing, CN, October 19–23 2009, pp. 633–636.
- [17] M. Bertini, A. D. Bimbo, and W. Nunziati, "Video clip matching using MPEG-7 descriptors and edit distance", *In Proc. 5th ACM International Conference on Image and Video Retrieval*, Tempe, AZ, USA, July 13–15 2006, pp. 133–142.
- [18] A. Sarkar, V. Singh, P. Ghosh, B. S. Manjunath, and A. Singh, "Efficient and Robust Detection of Duplicate Videos in a Large Database", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 6, pp. 870–885, June 2010.
- [19] D. G. Lowe, "Object Recognition from Local Scale-Invariant Features", *In proc. 7th International Conference on Computer Vision*, Kerkyra, GR, September 20–27 1999, vol. 2, pp. 1150–1157.
- [20] X. Wu, A. G. Hauptmann, and C. Ngo, "Practical Elimination of Near Duplicates from Web Video Search", *In Proc. 15th ACM International Conference on Multimedia*, Augsburg, DE, September 2007, pp. 218–227.
- [21] K. Mikolajczyk, C. Schmid, "A Performance Evaluation of Local Descriptors", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, October 2005.
- [22] E. Maani, S. A. Tsafaris, and A. K. Katsaggelos, "Local Feature Extraction for Video Copy Detection in a Database", *In Proc. 15th IEEE International Conference on Image Processing*, San Diego, CA, USA, October 12–15 2008, pp. 1716–1719.
- [23] <http://trecvid.nist.gov/>
- [24] A. F. Smeaton, P. Over, W. Kraaij, "Evaluation campaigns and TRECVID", *In Proc. 8th ACM International Workshop on Multimedia Information Retrieval*, California, USA, October 26–27 2006, pp. 321–330.
- [25] W. Kraaij, G. Awad, P. Over, "TRECVID-2009 Content-based Copy Detection Task Overview", *2009 TRECVID Workshop*, Gaithersburg, MD, USA, November 16–17 2009. (<http://www-nlpir.nist.gov/projects/tvpubs/tv9.slides/tv9.cd.slides.pdf>)
- [26] A. Sarkar, V. Singh, P. Ghosh, B. S. Manjunath, and A. Singh, "Duplicate Video Detection: Comparison of Proposed Distance Function with Dynamic Time Warping and Dependence of Detection Accuracy on Keyframe Selection", Univ. California, Santa Barbara, CA, VRL Tech. Rep., May 2008 [Online]. Available: <http://vision.ece.ucsb.edu/publications/Sarkar-tech-report-DTW-09.pdf>
- [27] ISO/IEC JTC 1/SC 29/WG 11, "Introducing MPEG-7 DDL – an Overview", Nice, FR, October 2005.
- [28] ISO/IEC 15938-3:2002/AMD 4:2010, Information Technology – Multimedia content description interface – Part 3: Visual, Amendment 4: Video signature tools.
- [29] ISO/IEC 15938-6:2003/AMD 4:2011, Information Technology – Multimedia content description interface – Part 6: Reference Software, Amendment 4: Reference software for video signature tools.
- [30] ISO/IEC 15938-7:2003/AMD 6:2011, Information Technology – Multimedia content description interface – Part 7: Conformance testing, Amendment 6: Conformance testing for video signature tools.
- [31] ISO/IEC 15938-8:2002/AMD 6:2011, Information Technology – Multimedia content description interface – Part 8: Extraction and use of MPEG-7 descriptors, Amendment 6: Extraction and matching of video signature tools.
- [32] D. D. Lewis, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval", *In proc. 10th European Conference on Machine Learning*, Chemnitz, DE, April 21–24 1998, pp. 4–15.
- [33] P. Scovanner, S. Ali, and M. Shah, "A 3-Dimensional SIFT Descriptor and its Application to Action Recognition", *In proc. 15th International Conference on Multimedia*, Augsburg, DE, September 24–29 2007, pp. 357–360.
- [34] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos", *In proc. 9th International Conference on Computer Vision*, Nice, FR, October 13–16 2003, pp. 1470–1477.

- [35] N. Sprljan, P. Brasnett, and S. Paschalakis, "Compressed Signature for Video Identification", *In Proc. 28th Picture Coding Symposium*, Nagoya, JP, December 7-10 2010, pp. 454-457
- [36] ISO/IEC 13818-2 2000, Information Technology – Generic coding of moving pictures and associated audio information: Video.
- [37] ISO/IEC 14496-10 2005, Information Technology – Coding of audio-visual objects – Part 10: Advanced Video Coding.
- [38] D. Salomon, "Data Compression: The Complete Reference (4th Edition)", Springer, pp. 139,241,851, 2007.
- [39] O. Orhan, J. Liu, J. Hochreiter, J. Poock, "University of Central Florida at TRECVID 2008: Content Based Copy Detection and Surveillance Event Detection", *2008 TRECVID Workshop*, Gaithersburg, MD, USA, November 17-18 2008.

Stavros Paschalakis (M'01) received the B.Eng. degree in Computer Systems Engineering (1st Class Hons.) in 1996 and the Ph.D. degree in Electronic Engineering in 2002, both from the University of Kent, UK. He was a Research Associate at the University of Kent between 1997 and 1999. After receiving his Ph.D., he joined Mitsubishi Electric R&D Centre Europe in the UK (MERCE-UK) as a Research Engineer, and later as R&D Manager and Multimedia Technologies Team Leader. In 2011 he co-founded Visual Atoms Ltd, a UK-based independent R&D and technology consulting company. Dr. Paschalakis is the author of over 40 technical papers and patents. His research interests include image and video processing, video coding, multimedia databases, computer vision, and hardware implementation of vision algorithms. He is a project editor of the MPEG-7 Video Signature Tools Standard and a chair of the MPEG-7 Visual group.

Kota Iwamoto received the B.E. degree and the M.E. degree in Electronics, Information and Communication Engineering from Waseda University in 2001 and 2003, respectively. He joined NEC Corporation in 2003 and is now an Assistant Manager of Information and Media Processing Laboratories. His research interests include image processing, image/video indexing and retrieval, image classification, and their applications. He was involved in the activities of the ISO/IEC JTC 1/SC 29/WG 11 (known as MPEG) for the development of MPEG-7 Video Signature Tools standard as an ISO project editor.

Paul Brasnett received the MEng. degree in Mathematics for Intelligent Systems in 2002 and the PhD degree in Electrical and Electronic Engineering in 2006, both from the University of Bristol, UK. In 2005 he joined Mitsubishi Electric R&D Centre Europe in the UK as a Research Engineer then as a Senior Research Engineer. In 2010 he joined Imagination Technologies as a Senior Research Engineer in the PowerVR Research Group. From 2007-2010 he was co-chair of the MPEG-7 Visual group. Dr. Brasnett is an author of over 20 peer reviewed technical papers and patents. His research interests include image and video processing, image and video coding, computer vision, machine learning and algorithms for general purpose GPU computing.

Nikola Sprljan received the B.Eng. degree in Electrical Engineering at the Faculty of Electrical Engineering and Computing at the University of Zagreb, Croatia, in 2001 and the Ph.D. degree in Electronic Engineering in 2006, at Queen Mary, University of London, UK. He joined Mitsubishi Electric R&D Centre Europe in the UK (MERCE-UK) in 2006 as a Research Engineer. In 2011 he joined Visual Atoms Ltd as a Senior Research Engineer and then he joined Maxeler as Acceleration Architect. His research interests include image and video representation, media compression, streaming and robust transmission, as well as multimedia databases and retrieval.

Ryoma Oami received the B.S. degree and the M.S. degree from the University of Tokyo in 1993 and 1995, respectively. He joined NEC Corporation in 1995 and is now a Principal Researcher of Information and Media Processing Laboratories. His research interests in image and video processing and analysis for understanding visual contents. He has joined the standardization activities in ISO/IEC JTC 1/SC 29/WG 11 (MPEG) since 2006, and contributed to developing several MPEG-7 and MPEG-A standards as an ISO project editor. He is a member of the Institute of Electrical and Electronics Engineer (IEEE).

Toshiyuki Nomura received the B.S. degree and the M.S. degree from Nagoya University in 1990 and 1992, respectively. He joined NEC Corporation in 1992, where he is a Research Engineer in Media and Information Research Laboratories. His research interests include multimedia signal processing such as coding, analysis and understanding. He was involved in the activities of the Audio Subgroup, ISO/IEC JTC1/SC29/WG11 (known as MPEG/Audio) for international standardization of speech and audio coding. Mr. Nomura is a member of the Institute of Electrical and Electronics Engineer (IEEE). He received the 1998 Shinohara Memorial Academic Encouragement Award from IEICE, the 2005 Ohm Technology Award. He is a project editor of ISO/IEC 14496-3:2001 and ISO/IEC 14496-5:2000.

Akio Yamada received the Ph.D. degree in information electronics science from Nagoya University, Japan in 1993. He joined NEC Corporation in 1993 and is now a Senior Research Manager of Information and Media Processing Laboratories, NEC Corp. His research interests include visual communication systems, multimedia signal processing such as coding, analysis and understanding, and various media service development. Since 1999, he has been a participant of the MPEG and JPEG Group, where he has been active in the development of MPEG-7, MPEG-A and JPSearch Standards as an ISO project editor. Dr. Yamada received the ITE Niwa-Takayanagi Best Paper Award and IEICE Yong Researcher Award in 1994 and 1996, respectively.

Mirosław Bober (M'97) is a professor of video processing at the University of Surrey, UK. In 2011 he co-founded Visual Atoms Ltd, a UK-based independent R&D and technology consulting company. Between 1997 and 2011 he was with Mitsubishi Electric R&D Centre Europe in the UK (MERCE-UK) as Chief Scientist and later as General Manager. Before joining MERCE-UK, he was with the University of Surrey in the UK, as a research scientist and then as lecturer and leader of the Image Communication and Multimedia Systems Group. Dr. Bober received an M.Sc. degree (with distinction) in Electrical Engineering from the AGH University of Science and Technology, Poland, in 1990. Subsequently he received an M.Sc. degree (with distinction) in Signal Processing and Artificial Intelligence in 1991 and a Ph.D. in 1995, both from the University of Surrey, UK. His research interests include various aspects of image processing, computer vision and pattern recognition. He has been involved in the development of MPEG-7 for over 10 years, chairing the MPEG-7 visual group. Dr. Bober is an inventor of over 70 patents, and author of over 80 refereed publications, including three books and book chapters.