H-068

MPEG-7 Video Signature for Robust Video Identification

Kota Iwamoto Ryoma Oami Toshiyuki Nomura

Information and Media Research Labs., NEC Corporation

1. Introduction

With the proliferation of video content distribution on the Internet, illegal copying and distribution of copyrighted contents has become a major problem. In order to cope with video piracy, it is essential to identify duplicated video contents, possibly in edited or modified versions, from the vast amounts of contents available on the Internet. Video fingerprinting is a key technology for content identification. It extracts unique descriptors called fingerprints from contents to identify identical duplicates of videos or video scenes. Unlike watermarking technologies, they do not require any alternation to the content, and therefore can be used readily with all existing contents.

Video fingerprints can be extracted coarsely at segment-level, or more finely at frame-level. Segment-level descriptors [2] can be used for fast matching, however, they are not able to accurately detect the matching intervals. Frame-level descriptors are used for accurate detection. Conventional frame-level descriptors include color-based descriptors [3], local feature based descriptors [4] and spatial descriptors [5][6]. It has been reported that the ordinal measure descriptor [6], a type of spatial descriptor, is robust and also provides high extraction and matching speed, most suited for the purpose. The ordinal measure describes the relative intensity distribution in an image, as an ordered intensity ranks of the partitioned blocks. However, it has been reported that the ordinal measure is not robust to editing operations such as text/logo overlay [5], which commonly occurs in copied contents delivered through the Internet. Conventional descriptors are not designed to provide robustness to a wide variety of common editing operations.

We have developed a new frame-level descriptor, which has robustness to wide variety of common editing operations. Due to its high performance, it was accepted as a frame-level descriptor of a new ISO/IEC standard, ISO/IEC 15938-3/Amd. 4 – MPEG-7 Video Signature Tools – [1], which standardizes interoperable descriptor for robustly identifying video contents. The intensive evaluation test conducted in the standardization process showed that the standardized descriptor can robustly identify videos with various editing operations, such as text/logo overlay, camera capturing, severe compression, etc. even at an extremely low false acceptance rate. The frame-level descriptor is compact, thus enabling ultra-fast searching of contents on the Internet. In this paper, we will present the state-of-the-art frame-level descriptor of MPEG-7 Video Signature.

2. Frame-level Descriptor of MPEG-7 Video Signature

The frame-level descriptor of MPEG-7 Video Signature is composed of "frame signature" and "confidence" components. The frame signature is a descriptor representing the content of the frame, while the confidence is a measure of reliability of the frame signature. Figure 1 illustrates the extraction procedure of these components.

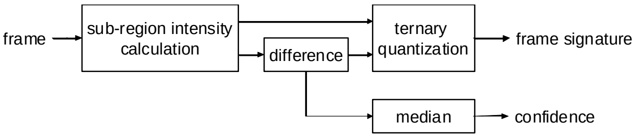

Figure 1: Extraction of frame-level descriptor.

2.1 Frame Signature Extraction

The frame signature represents quantized intensity and intensity differences of various sub-regions in a frame. It is composed of 380 dimensional vector of base-3 ternary values {0,1,2}, here denoted as $\mathbf{x} = \{x_1, x_2, \cdots, x_{380}\}$, extracted from pre-defined sub-region(s) associated with each element of the vector. The major 348 elements (*i*=33,…,380) are extracted by quantizing the difference between the average intensities of two sub-regions, and the remaining 32 elements (*i*=1,…,32) are extracted by quantizing the average intensity of a single sub-region.

Figure 2 illustrates samples of the sub-regions from which the frame signature is extracted (full specification of the sub-regions is described in [1]). They are configured at various scales, shapes and locations, imitating the human visual perception. These variations in the sub-regions provide uniqueness and robustness to the descriptor. Furthermore, they are sampled more densely at the center of the frame, where ROI (region of interest) is more likely to lie.

The ternary quantization is conducted as follows. Let $v1_i$ and $v2_i$ denote the average intensities of the sub-regions for element *i*, where $v2_i$=128 for elements *i*=1,…,32. The ternary value is calculated by,

$$x_i = \begin{cases} 2 & (\text{if} \quad v1_i - v2_i \ > \ th) \\ 1 & (\text{if} \quad |v1_i - v2_i| \ \le \ th) \\ 0 & (\text{if} \quad v1_i - v2_i \ < \ -th) \end{cases}, \tag{1}$$

where *th* is a threshold. The threshold *th* is not fixed, but determined adaptively for each frame, by considering the distribution of absolute differences $|v1_i\text{-}v2_i|$. It is chosen so that the distribution of the quantized value {0,1,2} across the vector becomes uniform, i.e. 1/3 each. This quantization strategy achieves robustness to changes in intensity ranges, while at the same time maximizing discriminability of the descriptor.

The extracted 380 dimensional ternary vector is encoded into 76 bytes representation. Each group of five consecutive elements is encoded into one byte value. The encoded value b_j (*j*=1,…,76) is calculated by the following equation.

$$b_j = 81 \times x_{5j-4} + 27 \times x_{5j-3} + 9 \times x_{5j-2} + 3 \times x_{5j-1} + x_{5j} \tag{2}$$

This encoding scheme achieves 20% size reduction compared with that of independently encoding each element with 2 bits representation.

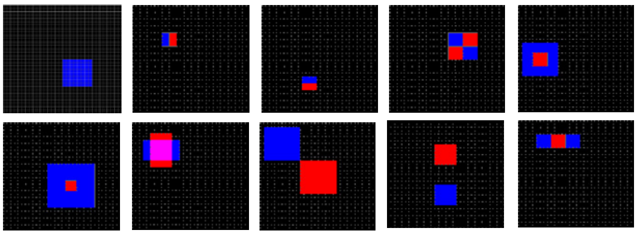

Figure 2: Sample of sub-regions in a frame. Blue and red regions represent the two sub-regions.

2.2 Confidence Extraction

The confidence component represents the complexity of the image content which the frame signature represents. It is calculated by taking the median value of the absolute differences $|v1_i\text{-}v2_i|$ of the elements i=33,…,380, and converting it to 1 byte (0-255) value. Low confidence value means that the intensity differences between sub-regions of the frame are small, representing a flat image with little content information. Thus, frame signature with low confidence value can be considered as less reliable. The confidence can be used during the matching to filter out unreliable false matches caused by flat images.

3. Matching of Frame-level Descriptor

Matching of frame signature and confidence is not specified as the normative part of MPEG-7 standard. Here, we explain the recommended method of matching.

Frame-by-frame matching is carried out between two sequences of frame signatures to detect matching intervals. The frame signatures between two frames $\mathbf{x}^1$ and $\mathbf{x}^2$ is matched by calculating the L1 distance between them, given as,

$$dist(\mathbf{x}^1, \mathbf{x}^2) = \sum_{i=1}^{380} \left| x_i^1 - x_i^2 \right| \quad (3)$$

For fast calculation, the distance can be computed in the encoded domain by using a look-up table. The distance of 1 byte encoded representation corresponding to 5 ternary elements can be pre-calculated as a look-up table. The distance can then be calculated by 76 look-ups and 76 additions, significantly improving the computation speed. A consecutive sequence where the distance is below a certain threshold is detected as a matching interval.

The confidence is used to filter out false matches caused by flat images. If the overall confidence of the detected matching interval is low, the match discarded as a false match.

4. Evalutation

Robustness to various editing operations is evaluated. The experiment is conducted using part of the MPEG-7 Video Signature dataset used in the standardization. The task is, given a 2 seconds query, to detect a matching interval, if any, embedded in a longer clip of approximately 3 minutes. First, an operational parameter is determined that achieves false acceptance rate of no more than 5ppm, using unrelated clips totaling 100 hours. Then using the parameter determined, the detection performance under various editing operations is tested. The tested editing operations are, camera capturing, text/logo overlay, severe compression, analog VCR recording, resolution reduction, monochrome conversion, brightness change, interlaced/progressive conversion, and frame-rate reduction. Queries were modified with these editing operations, with a total of more than 1600 queries for each editing operations.

The conventional ordinal measure descriptor [6] is also evaluated for comparison. The 3x3 block partitioning proposed in [6] was not able to achieve false acceptance rate of 5ppm. Therefore, we extended to 10x10 block partitioning, which can be expressed in a minimum of 66 bytes/frame, comparable with the 77 bytes/frame of MPEG-7 Video Signature.

Table 1 shows the detection rate under each editing operations. The MPEG-7 Video Signature achieves average detection rate of 96.4%, which is 12.2% higher than the ordinal measure. In particular, the MPEG-7 Video Signature achieves significant improvements on camera capturing and text/logo overlay, improving by 61.8% and 38.8% respectively.

Table 1: Detection rate at 5ppm false acceptance rate.

editing operation	ordinal measure	MPEG-7 Video Sig
camera capturing	28.29%	90.08%
text/logo overlay	45.02%	83.79%
severe compression	97.19%	99.45%
analog VCR rec.	90.75%	95.68%
resolution reduction	99.36%	99.72%
monochrome conv.	99.63%	100.0%
brightness change	98.04%	99.33%
I/P conversion	99.45%	99.63%
frame-rate reduction	99.76%	99.82%

5. Conclusion

We have presented the frame-level descriptor accepted in a new ISO/IEC standard, MPEG-7 Video Signature, for robust identification of video contents. It represents the intensity and intensity differences between various sub-regions in a frame, designed to provide uniqueness and robustness to the descriptor. The evaluation result shows that the MPEG-7 Video Signature achieves high-level of robustness to various editing operations, compared with the conventional ordinal measure descriptor.

References

[1] ISO/IEC 15938-3:2002/AMD 4:2010, Information Technology – Multimedia content description interface – Part 3: Visual, Amendment 4: Video signature tools.

[2] A. M. Ferman et al., “Robust Color Histogram Descriptors for Video Segment Retrieval and Identification”, IEEE Trans. on Image Processing, Vol. 11, No. 5, May 2002.

[3] E. Kasutani *et al*., “The MPEG-7 Color Layout Descriptor: A Compact Image Feature Description for High-speed Image/Video Segment Retrieval”, Proc. of ICIP2001, Oct. 2001.

[4] C. Chiu *et al*., "Efficient and Effective Video Copy Detection based on Spatiotemporal Analysis", Proc. ISM2007, Dec. 2007.

[5] K. Iwamoto *et al*., "Image Signature Robust to Caption Superimpostion for Video Sequence Identification", Proc. of ICIP2006, Oct. 2006.

[6] X.-S. Hua *et al*., "Robust Video Signature based on Ordinal Measure", Proc. of ICIP2004, Oct. 2004.