ROBUST VIDEO FINGERPRINTING BASED ON VISUAL ATTENTION REGIONS

Xing Su[1], Tiejun Huang[2] , Wen Gao[2]

[1]Graduate School, Chinese Academy of Sciences, Beijing 100049, China
[2]Institute of Digital Media, Key Laboratory of Machine Perception, Peking University, Beijing 100871, China
xsu@jdl.ac.cn, {tjhuang, wgao}@pku.edu.cn

ABSTRACT

This paper presents a robust video fingerprinting based on visual attention regions. Video fingerprints, which are a set of short feature vectors, are unique to video clips and used for video identification. The performance of video fingerprinting is usually measured in terms of robustness and accuracy of identification. In our proposed approach, we extract video fingerprints using visual attention regions which remain the same for the perceptually same scenes with different types of distortions and different for different scenes. The experimental results show that the proposed video fingerprinting is effective for constructing video fingerprints that are robust against various content-preserving distortions and accurate in identifying different video clips.

Index Terms— Feature extraction, Video signal processing, Identification, Visual system, Machine vision.

1. INTRODUCTION

Video fingerprinting is a set of techniques including analyzing video content, reducing its unique characteristics to a set of short feature vectors that serve as "video fingerprints", and looking those fingerprints up in a video fingerprint database to determine the identity of the video clips. Robust video fingerprinting methods create compact bit-stream representation of video content [1], which can uniquely differentiate one video from another, so as to convenience video content identification applications such as filtering, copyright management, retrieval, automatic linking. The extracted fingerprints should satisfy the following three requirements [2]: (1) Robustness. The video fingerprints should be robust against various content-preserving distortions such as brightness change, resolution reduction, frame-rate reduction, logo overlay. (2) Accuracy of identification. If two video clips are two perceptually different, the video fingerprints extracted from them should be considerably different. (3) Database search efficiency. The video fingerprints should be convenient for efficient matching and be conducive to efficient database search.

Past work on video fingerprint extraction can be broadly categorized into two classes [3], viz. the class of methods based on a whole video clip and the class of methods based on individual video frames. In the former class, video fingerprints are derived from a whole video clip or a subset of frames which selected from the video sequence such as spatio-temporal transform coefficients [4], randomized first-order video summarization technique which summarizes a video by a small set of representative feature vectors [5]. One of the main drawbacks for this kind of methods is that the extracted video fingerprints cannot identify a portion of a distorted video clip which is shorter than the original one. In the latter class, video fingerprints are based on individual video frames. In order to extract video fingerprints, image feature extraction techniques are applied to individual video frames. Recently, more attention is placed on this kind of approaches. The following video fingerprinting are in this class: the differential block luminance [1], centroid of gradient orientations (CGO) [2], robust color histogram descriptors [6], the radial hash (RASH) [7] which uses radial projection of the image pixels, and so on. These video fingerprints can identify and retrieve a video clip with variable length. These video fingerprints are robust to various video processing including lossy compression, frame rate change, etc. But they are not robust against heavily video distortions through these distortions are perceptually the same.

This paper proposed a novel video fingerprinting following the latter class. But unlike the previous video fingerprinting, it is based on visual attention regions. The visual attention regions are invariant if the distorted video is content-preserving, even the video is distorted heavily. Often, the visual attention follows two mechanisms, bottom-up and top-down, corresponding to stimulus-driven and object-driven respectively [8]. Top-down mechanism is task-dependent and closely related to human brain, which is a much complex mechanism. Bottom-up models are mainly saliency based. Visual input is first decomposed into a set of topographic feature maps. Different spatial locations then compete for saliency within each map. All feature maps are combined into a unique saliency map. It does not require any top-down guidance. The attention regions, which can be represented in the saliency map, are the same in the perceptually same scenes and different in different scenes. They are robust against content-preserving distortions and are discriminative in different video clips. In the proposed approach, video fingerprints are extracted from these invariant and discriminative attention regions.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

The rest of this paper is organized as follows: Section 2 presents the proposed video fingerprinting extraction approach. Section 3 provides video matching methods and database strategy. Section 4 shows experimental results and performance evaluation and we conclude our work in section 5.

2. PROPOSED VIDEO FINGERPRINTING

The overall framework of the proposed video fingerprint extraction is showed in Fig.1.

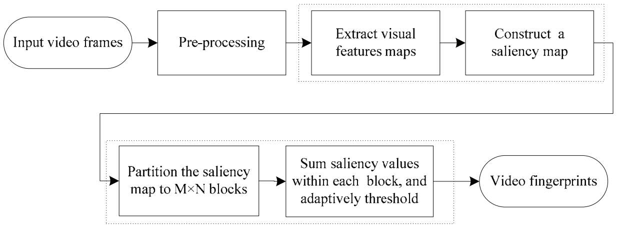
Input video frames
Pre-processing
Extract visual features maps
Construct a saliency map
Partition the saliency map to M×N blocks
Sum saliency values within each block, and adaptively threshold
Video fingerprints

Fig. 1. The framework of video fingerprint extraction

In the first step, the input video frames, which is decoded from the video files, is pre-processed. These frames are not proper for extracting video fingerprints directly, for they may be decoded from different video files and have different size, different quality and so on. To eliminate these differences among these frames, we pre-process the input video frames before extracting features. All the input video frames are resized to CIF (352×288). Then they are smoothed, which makes the edges and orientation features more stable and reduce the influence of noises. In the second step, visual feature maps, such as color maps, intensity maps, orientation maps, are computed from the input frame. In the third step, in order to eliminate modality-dependent amplitude differences, these feature maps are normalized. Then, these normalized feature maps are combined to a unique saliency map. The process of constructing saliency map is denoted by the dashed rectangle in Fig.1. In the fourth step, the saliency map is partitioned into a grid of N rows and M columns, resulted in M×N blocks. Finally, the average saliency value of each block is calculated. All the values of the blocks are adaptively quantized as video fingerprints. Another dashed rectangle denotes the process of extracting video fingerprint from the saliency map in Fig.1.

2.1. Saliency Map

Visual attention is one of the most important characteristics in human visual system. Most current computational attention models emphasize on bottom-up mechanism. Bottom-up mechanism is stimulus-driven, which is closely to the mechanism in human sensory system. The bottom-up models are mainly saliency-based. Simple multi-scale feature maps detect local spatial discontinuities in intensity, color, orientation, and are combined into a unique saliency map [9]. Itti and Koch proposed an applied system for detecting human attention regions [8]. The experiments in their paper show that the system works well. We follow their work to extract the saliency map with three main steps: (1) Compute the visual feature maps such as color maps, intensity maps and orientation maps. (2) Normalize these feature maps. (3) Combine these feature maps to a unique saliency map. Each pixel value of saliency map represents the weight for the attention regions. If a location has a larger saliency value in the saliency map, it is more attention-getting. We can quantize the values of the saliency map and get a coarse representation of the saliency map, which is used for video fingerprint extraction.

In the first step, we follow the Itti's approach [8]. Each feature map is computed by a set of linear "center-surround" operations. Typical visual neurons are most sensitive in a small region of the visual space (the center), while stimuli presented in a broader, weaker antagonistic region concentric with the center (the surround) inhibit the neuronal response. Center-surround is implemented in the model as the difference between fine and coarse scales. So, a set of topographic feature maps are decomposed from the visual input. These feature maps are intensity maps, color maps and orientation maps.

All these feature maps are from different visual modalities and with unrelated dynamic ranges. Before they are combined into a unique saliency map, they are normalized in order to eliminate such influence. We use iterative localized normalization [9], which relies on simulating local competition between neighboring salient locations in the feature map. It is closely to computation mechanism in human vision system. In the iterative localized normalization, each feature map is then iteratively convolved by a large 2D "difference of Gaussians" (DoG) filter, and negative results are set to zero after each iteration.

$$DoG(x,y) = \frac{c_{ex}^2}{2\pi\sigma_{ex}^2} e^{-\frac{x^2+y^2}{2\sigma_{ex}^2}} - \frac{c_{inh}^2}{2\pi\sigma_{inh}^2} e^{-\frac{x^2+y^2}{2\sigma_{inh}^2}} \qquad \textbf{(1)}$$

In our implementation, σ_{ex} =2% and σ_{inh} =25% of the input image width, c_{ex} =0.5 and c_{inh} =1.5. We denote this iterative localized normalization with $N(.)$.

After all the feature maps are normalized, they are combined into a unique saliency map. The intensity, color and orientation feature maps are combined respectively with Eq.(2). Then these three maps are normalized and combined to a saliency map with the same method.

$$S = \frac{1}{m}\sum_{i=1}^{m} N(X_i) \qquad \textbf{(2)}$$

In Eq.(2), X_i represents a feature map, $N(.)$ is normalization function, S is the combined map.

2.2. The Proposed Video Fingerprints

The attention regions can be represented by the saliency map. But the dimension of the saliency map is too high for video fingerprints. We obtain a coarse representation (Q) of the saliency map by partitioning the saliency map (S) into $M \times N$ blocks (M =12, N =8 in our implementation) and averaging the saliency value in the blocks of size $W_x \times W_y$, as it is showed in Eq.(3).

$$Q(m,n) = \frac{1}{W_x W_y} \sum_{i=(m-1)W_x}^{mW_x} \sum_{j=(n-1)W_y}^{nW_y} S(x,y) \qquad (3)$$

$$m = 1,2,\dots,M; n = 1,2,\dots,N$$

Then the coarse representation (Q) of the saliency map (S) is adaptively quantized to a binary vector. The threshold is obtained by the average of the maximum and minimum of Q.

$$Threshold = (\max\{Q(i,j)\} - \min\{Q(i,j)\})/2 \qquad (4)$$

The binary vector is the proposed video fingerprint. The size of it is $M \times N$ bits.

3. FINGERPRINT MATCHING AND DATABASE STRATEGY

Considering the huge amount of video databases potentially for video fingerprint extraction and comparison, an effective fingerprint matching approach and an efficient video fingerprint database strategy are proposed to speed up the progress of matching and retrieval.

In the proposed video fingerprinting, the video fingerprint is a set of binary data. To perform video fingerprint matching between two different fingerprints, both with length N, the normalized Hamming distance is taken.

$$H(B_1,B_2) = \frac{1}{N}\sum_{N} B_1 \otimes B_2 \qquad (5)$$

In Eq.(5), $\otimes$ is the exclusive OR (XOR) operator. If the Hamming distance $H(B_1,B_2)$ is less than some pre-defined threshold the video fingerprints are classified as 'matching', and vice versa. In our video fingerprint matching, the main operator is just OR (XOR) that runs fast and easily in computer, which can remarkably speed up the matching.

Given K fingerprints of a query video clip, a single video fingerprint of a frame is not sufficient for a reliable matching. Minimal unit, which is composed of a fixed durations of video segment, mostly 2 seconds, 5 seconds, or 10 seconds long, is minimal matching unit. A sliding-window, which is larger than the minimal unit, is used for a query video clip. We use coarse to fine method to speed up the whole retrieval. At coarse level, the interval between two sliding-windows is large. Then a candidate video clip set is obtained, which likely contains the correct matching for the query video clip. In the set we match them in fine level. The interval between two slip-windows, which is set smaller than that in coarse level, is used to retrieval the most similar one that is matched best.

In the case of that the number of video fingerprints is too huge, we propose to use an efficient database strategy. We create a lookup table (LUT) [10] for all possible video fingerprints, and we let the entries of the table point to the position where the video fingerprint is extracted, which is for looking up the source of a video fingerprint. The information of the pointer includes the name of the video clip, the position of the frame, and the description of the video clip. In this way one video fingerprint may be associated with multiple pointers to video clips and positions. The multiple pointers are stored in a linked list. Given a query clip, we extract video fingerprints of video frames and lookup them in the table. To the video fingerprint of each query video frame, which is matched in the video fingerprint database, a vote is given for the candidate clips and the corresponding positions. The video clip in the candidate set, which wins the most vote and the candidate frames are nearly consecutive, is the matching one. In this video database strategy the extracted video fingerprints do not have to match the video fingerprints in the fingerprint database one by one which is inefficient and time-consuming. The proposed method costs less and is efficient in video fingerprinting

In our implementation, we sampled 4 frames per second. The video fingerprints are calculated for the sampled frames. Extracting video fingerprints from each video frame is time-consuming and not necessary, for there is little different information in the adjacent frames. Sampling 4 frames per second is enough for the proposed video fingerprinting, which can speed up the video fingerprint extraction of the whole video clip.

4. PERFORMANCE EVALUATION

The performance of the proposed video fingerprinting is evaluated on a video clip set collected from internet, TV broadcasting, DVD and VCD. The total length of the set is over 100 hours. Evaluation of video fingerprinting contains two aspects: (1) Independence test. It is test for accuracy of identification. The perceptually different video clips should be distinguished by video fingerprints. If the perceptually different video clips are detected as the same ones, we call this case 'false alarm'. (2) Robustness test. A video fingerprint should be robust against various content-preserving distortions. If the video clips, which are from the same source and are obtained by content-preserving distorted, are detected as different ones, we call this case ‘missed detection’. The probability of missed detection (P_{md}) and false alarm (P_{fa}) is expressed as Eq.(6).

$$P_{md} = \frac{\text{\#missed detection}}{\text{\#total test}}$$

$$P_{fa} = \frac{\text{\# false alarm}}{\text{\#total test}} \qquad \textbf{(6)}$$

In our experiments, the following distortions are selected for robustness test: (1) Resolution reduction to CIF (2) Resolution reduction to QCIF (3) Frame-rate reduction to 4fps (4) Frame-rate reduction to 5fps (5) Frame-rate reduction to 15fps (6) Change to gray scale (7) Logo overlay by 5% (8) Logo overlay by 20% (9) Logo overlay by 30%. These modifications are used to estimate the probability of missed detection.

Video distortions	CGO[2]	Our method
Resolution CIF	1.02%	0.16%
Resolution QCIF	1.67%	0.10%
Frame-rate 4fps	1.01%	1.78%
Frame-rate 5fps	0.97%	1.30%
Frame-rate 15fps	0.47%	0.18%
Color to gray	0.86%	1.30%
Logo overlay by 5%	18.8%	0.42%
Logo overlay by 20%	34.9%	0.49%
Logo overlay by 30%	43.9%	0.59%

Table 1. The probability of incorrect detection ($P_{md} = P_{fa}$)

The performance of video fingerprinting is evaluated by the probability of incorrect detection P_{md} when $P_{md} = P_{fa}$. P_{md} and P_{fa} indicate the performance of robustness and accuracy of identification. Table 1 presents the probability of incorrect detection in our proposed method and CGO [2]. The results show that our method performs well in accuracy of identification and robustness to video distortions. In the case of Logo overlay, the information of orientations in logo area is different from the original video frame, so the CGO doesn’t perform well. But our method is also robust in this case, for it is based on attention regions which are got from multi-feature maps.

5. CONCLUSION

In this paper, we proposed a robust video fingerprinting based on visual attention regions which are obtained in bottom-up mechanism and avoid the effect from the top level of Human Visual System (HVS). The experiment results show that the saliency map is effective for constructing robust and accurate video fingerprints for identification. We also proposed an effective video fingerprint matching method and efficient database strategy. The future work is to propose a video fingerprinting with improved robustness and accuracy of identification and combine it with other features to get a high accurate video fingerprinting solution.

6. ACKNOWLEDGMENT

This work was supported by the National Key Technology R&D Program [2006BAH02A10 and 2006BAH02A13] of China.

7. REFERENCES

[1] R. Radhakrishnan, and C. Bauer, "Robust video fingerprints based on subspace embedding," *Proc. ICASSP 2008*, pp.2245-2248, Apr.2008.

[2] S. Lee, C. D. Yoo, “Robust Video Fingerprinting for Content-Based Video Identification,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 18, no. 7, pp.983 – 988. Jul. 2008.

[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. Oct. 2007.

[4] B. Coskun, B. Sankur, N. Memon, “Spatio–Temporal Transform Based Video Hashing,” *IEEE Trans. Multimedia*, vol.8, no.6, pp.1190 – 1208, Dec. 2006.

[5] S. S. Cheung and A. Zakhor, “Efficient Video Similarity Measurement with Video Signature,” *IEEE Trans. Circuits and Systems for Video Technology,* vol.13, no.1, pp.59-74, Jan.2003.

[6] A. M. Ferman, A. M. Tekalp, and R. Mehrotra, “Robust Color Histogram Descriptors for Video Segment Retrieval and Identification,” *IEEE Trans. Image Processing*, vol. 11, no. 5, pp. 497 – 508, May. 2002.

[7] C. D. Roover, C. D. Vleeschouwer, F. Lefebvre, and B. Macq, “Robust video hashing based on radial projections of key frames,” *IEEE Trans. Signal Processing*, vol. 53, no. 10, pp.4020 – 4037, 2005.

[8] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.20, no.11, pp.1254 – 1259, Nov. 1998.

[9] L. Itti and C. Koch, “A Comparison of Feature Combination Strategies for Saliency-Based Visual Attention Systems,” *Proc. SPIE, human vision and electronic imaging IV (HVEI'99)*, vol. 3644, pp. 473-482, Jan.1999.

[10] J. Oostveen, T. Kalker, and J. Haitsma, “Feature Extraction and a Database Strategy for Video Fingerprinting,” *5th Int. Conf. on Visual Information Systems*, vol. LNCS 2314, pp. 117–128, 2002.