ROBUST VIDEO SIGNATURE BASED ON ORDINAL MEASURE

Xian-Sheng Hua, Xian Chen, Hong-Jiang Zhang*

Microsoft Research Asia, Beijing 100080, China
xshua@microsoft.com, chenxian@tsinghua.org.cn, hjzhang@microsoft.com

ABSTRACT

In this paper we proposed a video signature based on ordinal measure of resampled video frames, which is robust to changing compression formats, compression ratios, frame sizes and frame rates. To effectively localize a short query video clip in a long target video through the proposed video signature, we developed a coarse-to-fine signature comparison scheme. In the coarse searching step, roughly matched positions are determined based on Sequence Shape Similarity, while in the fine searching step, dynamic programming is applied to handle similarity matching in the cases of losing frames and temporal editing processes are employed on the target video. Experiments showed that the proposed video signature has good robustness and uniqueness, which are the two essential properties of video signatures.

1. INTRODUCTION

Nowadays more and more digital videos are available on the web and in multimedia databases. Content-based video analysis is indispensable in order to efficiently manage and utilize these resources. Wide applications as video copy detection, video indexing, and video search engine require an effective (robust and unique) and compact description, i.e., video signature, of videos based on their content.

In order to be both efficient and effective, the video signature is required to have two essential properties, uniqueness and robustness. Uniqueness indicates the distinguishing capability of the video signature, which implies that videos with different content should have distinct signatures. While robustness indicates the capability of noise tolerance, which means that two videos with the same content should have identical or near the same signatures, even they are in different compression formats/ratios, frame sizes and/or frame rates.

Most existing video signature generation schemes adopted the following framework. First, the video is segmented into shots and each shot is represented by one or more key-frames [1]-[3]. The key-frames are then represented by certain high dimensional feature vectors (color histogram, edge distribution, texture, etc.). Finally the sequence of the key frames' features is taken as the signature of the whole video. The above methods have a primary disadvantage. Currently color histogram is the most widely used feature vector in both shot boundary detection and key frame extraction algorithms. However, it is known that two video clips with the same content but compressed in different formats/ratios may have distinct color characteristics [4], which makes the shot and key-frame based video signatures unreliable. Although some other approaches as the method in [5] do not depend on shot boundary detection and key-frame extraction, color histogram is still indispensable for comparing two video signatures. In fact, the robustness of these sorts of video signatures is not well investigated in literatures. Furthermore, little work is accomplished to discuss the distinguish capability (uniqueness) of video signatures.

In this paper, based on the analyses of the two essential properties of video signatures, we proposed a new robust video signature scheme. In this scheme, video is resampled at a uniform sampling rate, and the *ordinal measure* of the resampled video frames is employed as the signature. The resampling process is applied to handle the cases of changing frame rates, while the ordinal measure is robust to different compression formats/ratios and frame sizes. In addition, to make the signature tolerant to temporal editing, such as inserting in or cutting out a short clip, dynamic programming method is employed for locating a short query video clip in a long target video.

The rest of the paper is organized as follows. In Section 2 the generation scheme of the proposed video signature is introduced. Two types of sequence similarity measures are defined in Section 3. In Section 4, the approach for localizing a query video clip in a long video sequence is presented in detail. Experimental results are provided in Section 5, followed by conclusion remarks in Section 6.

2. SIGNATURE GENERATION

As we have mentioned, most existing video signatures employ feature vector extracted from each video frame, or key-frame of each shot. The key-frame based schemes are not robust to compression and resolution change, while the frame-by-frame based schemes are not robust to frame rate change, as well as that this type of signatures will be very large and has numerous redundant information. In our scheme, firstly the original video sequence is resampled at a uniform sampling rate, T_S fps (frames per second), thus the signature extracted from these sampled video frames is relative compact and capable of being tolerant to different compression formats/ratios, resolution changes and frame rate changes.

Then the ordinal measure is extracted and regarded as the feature vector of each sample frame, similar to that of [6]. Ordinal measure reflects the relative intensity distribution within an image, which was first proposed in [7] as a robust feature in image correspondence. The video frame is partitioned into $N = N_x \times N_y$ blocks and the average gray level in each block is computed. Then the set of average intensities is sorted in ascending order and the rank is assigned to each block (in this paper, N_x=N_y=3). The ranked $N_x \times N_y$ dimensional sequence, i.e., ordinal measure, is the inherent relative intensity distribution in a single frame, thus is naturally robust to the color degradation

* The work was performed when this author was visiting Microsoft Research Asia as an intern.

effect caused by different compression formats. Furthermore, ordinal measure is a very compact feature vector ($9 \times 0.5 = 4.5$ bytes/frame if we use 4 bits to represent number 0~8, and even 2.5 bytes/frame is enough if using proper coding method since there are only $9! = 362880 < 2^{20}$ possible combinations), thus is able to keep the whole video signature from being too large.

It is obvious that the larger the resampling rate T_S is, the larger the size of the signature is, and the more precisely it is able to represent the original video, and vice versa. Therefore there is a trade-off in selecting a proper sampling rate, which is to be discussed with more details in Section 5.

3. SEQUENCE SIMILARITY

In this section, two similarity metrics for the proposed video signature are studied. One is the Sequence Shape Similarity (*SSS*), the other is the Real Sequence Similarity (*RSS*).

3.1. Sequence Shape Similarity

SSS intuitively measures the similarity of two video clips based on the "temporal shape" of their signatures. Figure 1 shows the curves of the first dimension of the ordinal measure sequence, extracted from the signatures of two videos with the same content but in different formats (MPEG1 and AVI). It can be seen that their "temporal shapes" are similar, although there are some small differences.

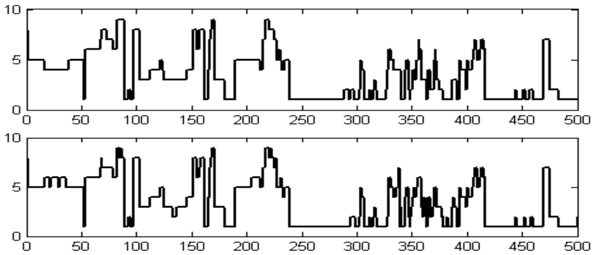


Fig.1. The curves of the ordinal measure (the first dimension) sequences of two video clips with the same content but different formats (MPEG1 and AVI).

Let $S_X = (X_1, X_2, \cdots, X_M)$ and $S_Y = (Y_1, Y_2, \cdots, Y_M)$ denote the signatures of two video clips with the same duration, their Sequence Shape Similarity (*SSS*) is defined by

$$SSS(S_X, S_Y) = \left(\sum_{i=1}^{M} I(d(X_i, Y_i) \le \varepsilon)\right) / M \qquad (1)$$

where $d(\cdot)$ is the distance metric defined on the ordinal measure (here L_1 distance is applied), and $I(x)$ is equal to 1 if x is true; otherwise, equal to zero. ε is a predefined distance threshold which makes the signature tolerant to the possible noises caused by the temporal resampling, compression or spatial up/down sampling. In our implementation, ε is set to 6.

3.2. Real Sequence Similarity

RSS is designated to measure two sequences' similarity during the fine searching phase (to be explained in detail in Section 4). Let $S_X = (X_1, X_2, \cdots, X_M)$, $S_Y = (Y_1, Y_2, \cdots, Y_N)$ denote the signatures of two sequences which may be in different length ($M \le N$). Due to losing frames or temporal editing, such as inserting in or cutting out short clips, dynamic programming [8] is applied to find the best match of these two sequences while the matching similarity, Real Sequence Similarity (*RSS*), is defined as

$$RSS(S_X, S_Y) = \left(\alpha N_{match} + \beta N_{miss} + \gamma N_{gap}\right) / M \qquad (2)$$

where N_{match}, N_{miss} and N_{gap} are the number of matched, mismatched and inserted (gap) elements for the best match, respectively. α, β and γ are predefined weights. In our experiments, α, β and γ are set to 1, 0 and -0.5, respectively.

4. SEQUENCE MATCHING

Let $S_X = (X_1, X_2, \cdots, X_M)$ and $S_Y = (Y_1, Y_2, \cdots, Y_N)$ denote the signatures of the query clip X and the target videos Y ($M \le N$), in which we want to locate the positions that X may appear in Y. The sequence matching scheme is composed of two phases, a coarse searching phase and a fine searching phase. In the coarse searching phase, rough positions of all possible matches are obtained by the following steps.

(1) Get the similarity curve by matching S_X along S_Y, and computing Sequence Shape Similarity at every step, i.e., generate the curve of $SSS(S_X, S_Y^{\,i})$, where $i \le N - M + 1$, and

$$S_Y^i = \left(Y_i, Y_{i+1}, \cdots, Y_{i+M-1}\right) \qquad (3)$$

(2) Threshold the curve at T_1. That is, only keep the similarity values which are above T_1 as candidate matches, and set all others to zero. In our implementation, T_1 is set to 0.5.

(3) The local maximums of the thresholded curve are identified as the coarsely matched locations.

Figure 2 shows a similarity curve of matching a 29-second query clip against an 11-minute TV program containing the query clip. It can be seen from the figure that there exists one match for the query clip at around (resampled) frame 4300 with the similarity of about 0.95.

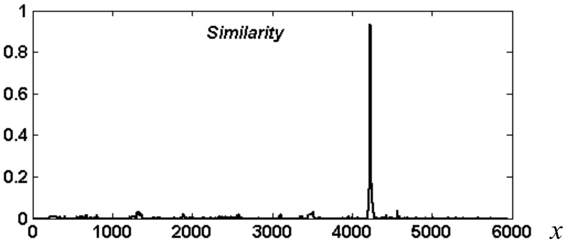


Fig.2. Similarity curve from matching a short clip against a long video sequence when $T_s = 10$ fps and $\varepsilon = 6$ (x: frame number).

Considering there may exist some temporal editing as cutting or inserting frames or small clips, dynamic programming method is applied in the fine searching phase, as following steps.

(1) A sub-sequence ($2M$ in length) centered at each coarsely matched location is extracted from S_Y, denoted by S_Y^{*}.

(2) Use the Needleman/Wunsch [8] method to obtain the best match for S_X and S_Y^{*}. If *RSS* of the best match is above a threshold T_2, the matched location is found. In our experiments, T_2 is set to 0.6.

In order to roughly estimate the uniqueness of the proposed scheme, the probability of two random signatures that will be determined as similar is estimated. Assume that the elements in the two random signatures S_X and S_Y are independent. According to Equation (1), we then have

$$Prob(SSS(S_X,S_Y)\ge\eta)=Prob\left(\sum_{i=1}^{M} I(d(X_i,Y_i)\le\varepsilon)\ge\eta\right) \quad (4)$$

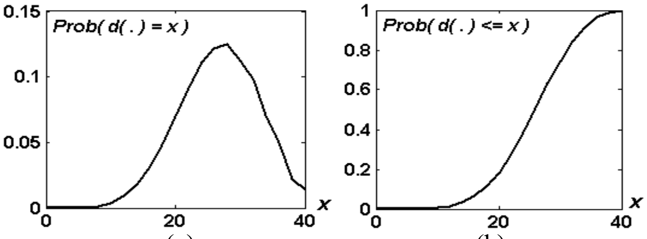
As $X_1, X_2, \cdots, X_M, Y_1, Y_2, \cdots, Y_M$ are assumed independent, $Prob(d(X_i,Y_i)\le\varepsilon)$ are identical for all $1\le i\le M$. For simplicity, we denote it as P_ε. Accordingly, Equation (4) can be rewritten as

$$Prob(SSS(S_X,S_Y)\ge\eta)=(P_\varepsilon)^{\eta M} \quad (5)$$

To obtain the value of P_ε for each ε, we exhaustively calculate all possible distances of any two 9-dimensioanl ordinal measures (there are $(9!)^2$ cases in total). Figure 3(a) shows the distribution curve of $Prob(d(X_i,Y_i)=x)$, while (b) is the integral curve of (a), i.e., $Prob(d(X_i,Y_i)\le x)$. From this distribution, we can obtain

$$P_\varepsilon\approx 0.0006 \quad (6)$$

where $\varepsilon=6$. Therefore, the probability of two random signatures will be identified as similar can be estimated by $(0.0006)^{\eta M}$, which is a very small value.



(a) (b)

Fig.3. Distance possibility (a) and distribution (b) of two random ordinal measures, in which x denotes the value of distance.

The above estimation roughly shows the uniqueness of the proposed video signature, although the independence assumption is strong. More accurate estimation can be obtained by better modeling the distribution of ordinal measure within a video sequence, as well as taking the dynamic programming based matching into consideration. These will be our future works.

5. EXPERIMENTAL RESULTS

Experiments for the proposed video signature scheme consist of three parts, parameter selection, uniqueness test, and robust test.

5.1. Parameter Selection

In this sub-section, firstly we investigate the selection of appropriate resampling rate in the proposed video signature scheme. Figure 4 is the matching curves of searching a 5-second MTV segment (AVI format, 15fps) in the original 4-minute MTV (MPEG1 format, 29.97fps), under different temporal resampling rate T_S. Note that the frame rates and compression formats of the segment and the original video are different. From the three similarity curves it can be seen that the "shape" almost remains the same but the value of the peak reduces as T_S decreases. The reason is that, when T_S decreases, the signature of the query clip becomes shorter and less precise due to the error caused by resampling.

Secondly, selection of appropriate distance threshold ε is studied. The same query and target videos clips above are used. When T_S is fixed to 10 fps, the matching curves under different distance thresholds are shown in Figure 5. We can see that as the threshold increases, the similarity curve becomes noisier, thus it is more difficult to locate the correct position of the query clip in the target video sequence.

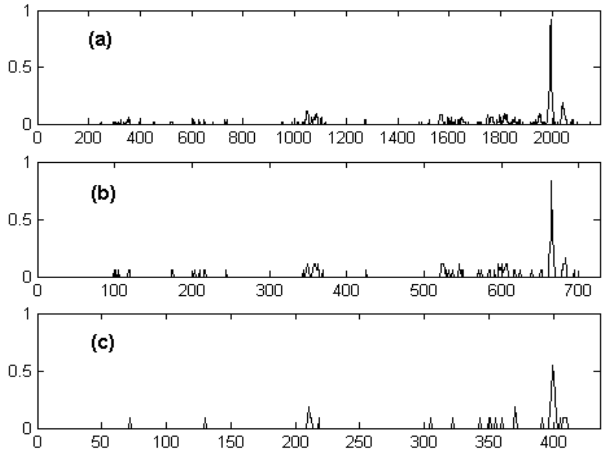


Fig.4 Similarity curve of different sampling rate: (a) 10 fps (b) 4 fps (c) 2 fps. Distance threshold is set to 6 in all cases.

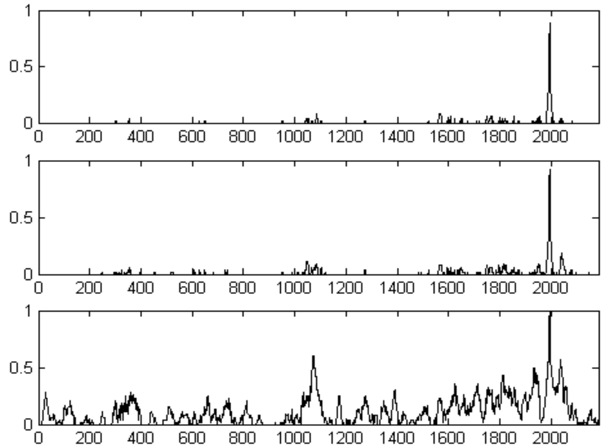


Fig.5. Similarity curves of different thresholds: (a) ε = 4 (b) ε = 6 (c) ε = 12. Sampling rate is set to 0.10s/sample in all cases.

Through a number of experiments, though the overall performance is not quite sensitive to the above two parameters, it is found that ε = 6 and T_S = 10 fps can yield relative better results.

5.2. Uniqueness Test

To test the uniqueness property of the proposed video signature, a 6-hour TV program is divided into 72 segments (5 minutes each), compressed in different formats or ratios (AVE, MPEG1, Intel YUV, MS-RLE, Cinepak, MPEG2, MPEG4 or WMV), and resampled at different frame rates (15~29.97 fps) and resolutions (100×75~640×480). Then each segment is taken as the query clip to search similar sub-sequences on the original TV program.

According to the experiment, all query segments are correctly located in the original video (with very small offset, as illustrated in Table 1), and no false alarm exists. Table 1 shows the average offset of the located positions and the ground-truths, as well as the average similarity of the query clips and the located clips, and the average maximal similarity between the query clips and other sub-sequences in the video except the located ones. These results show that the correlation between different clips but with the same content is far much larger than that of clips with different content.

Tab.1. Uniqueness test

$\triangle T$	*S*	*S’*
0.055 second	0.959	0.106

Note: •*T* is the average offset between the detected locations and the ground-truths; *S* is the average similarity between the query clips and the located clips; *S’* is the average maximal similarity between the query clips and other sub-sequences in the video except the located ones.

5.3. Robustness Test

Actually the uniqueness test above also shows the robustness of the proposed video signature. Here we present another experiment which is more focused on robustness. Another 2-hour TV program (MPEG1, 29.97 fps) is used in this test, from which we try to locate several commercial segments. Three commercial segments, denoted by *Com1*(25s), *Com2*(29s) and *Com3*(23s) are selected from the original video and taken as query clips, which are recompressed in AVI, Cinepak, MS-RLE, or Intel YUV format in distinct compression qualities, and some of them are resampled at 15 fps, as illustrated in Table 2. Totally there are 4, 2, and 1 occurrences for *Com1*, *Com2* and *Com3* in the original video, respectively. From the results it can be seen that all commercials clips are precisely located without any false alarm. It is observed that the similarity values of *Com1* are generally smaller than that of *Com2* or *Com3*. This is because *Com1* contains lots of actions thus temporal resampling process brings more errors in this case.

Another experiment illustrates the reason for adopting the dynamic programming method. We insert a 5-second video segment (randomly chosen from the above 2-hour TV program) into every occurrences of *Com1*, *Com2* and *Com3* at random positions, and then search them in the resulted video sequence using the query clips. If we only take the coarse searching phase and use *SSS* as similarity measure, the average peak value of the similarity curve at the matched positions is only 0.642. But after the fine searching phase, the actual similarity between the query and its matches is 0.874, which is more close to the real case. Besides, the locations determined by dynamic programming method are more precise that those of determined by *SSS* measure only.

6. CONCLUSION

In this paper we have proposed a video signature scheme, and investigated its two essential features: uniqueness and robustness. In the scheme, the original video is resampled, and then each sample is represented by its ordinal measure. The ordinal measure sequence is then taken as the video signature. To make the signature be more robust to various kinds of variations, such as format, compression ratio, frame rate and editing, two sequence similarity measures, Sequence Shape Similarity (*SSS*) and Real Sequence Similarity (*RSS*), are defined. To locate a query video clip in a long target video, a coarse-to-fine signature comparison scheme is presented, in which we first got the rough matching positions from the SSS curve, and then adopted the dynamic programming method to precisely locate the query clip in the target video. Experiments have showed that our approach has good uniqueness property, as well as it is robust to changing of video formats, compression ratios, frame rates, resolution and a certain quantity of temporal editing. Future work would be to construct more precise theoretical model to prove and improve the robustness and uniqueness of the video signature, as well as to test it on a large-scale video database.

Tab.2. Search commercial clips in TV program

Query Clips (format/fps)	**Matching Location (second)**	**Similarity**	**Recall**	**Precision**
Com1 29.97fps Intel YUV	a) 889.2s ~ 897.3s b) 2868.8s~2876.9s c) 4677.0s~4685.1s d) 6227.6s~6235.7s	0.790 0.889 0.790 0.704	100%	100%
Com1 15fps Cinepak	a) 889.2s ~ 897.2s b) 2868.7s~2876.7s c) 4677.0s~4685.0s d) 6227.6s~6235.6s	0.838 0.700 0.825 0.725	100%	100%
Com2 29.97fps Intel YUV	a) 1991.4s~2001.7s b) 4950.6s~4960.9s	1.000 0.971	100%	100%
Com2 15fps MS-RLE	a) 1991.4s~2001.6s b) 4950.6s~4960.8s	0.951 0.980	100%	100%
Com3 29.97fps Cinepak	a) 2051.5s~2080.8s	0.997	100%	100%
Com3 15fps MS-RLE	a) 2051.4s~2080.7s	0.980	100%	100%

7. REFERENCES

[1] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, “Adaptive key frame extraction using unsupervised clustering,” in *Proc. IEEE ICIP’98*, vol. 1, pp. 866-870, Oct. 1998.

[2] W. Wolf, "Key frame selection by motion analysis," in *Proc. ICASSP 96*, Vol. II, pp. 12281231, 1996.

[3] H.S. Chang, S. Sull and S.U. Lee, "Efficient video indexing scheme for content-based retrieval," in *IEEE Trans. Circuits Syst. Video Technol.*, Dec 1999.

[4] A. Hampapur and R. M. Bolle, “Comparison of distance measures for video copy detection.,” in *Proc. of Int. Conf. on Multimedia and Expo*, Aug. 2001.

[5] S.-C. Cheung and A. Zakhor, “Estimation of web video multiplicity,” in *Proc. SPIE—Internet Imaging*, vol. 3964, San Jose, CA, pp. 34–36, Jan. 2000.

[6] R. Mohan, “Video sequence matching,” in *Proceedings of the International Conference on Audio, Speech and Signal Processing, IEEE Signal Processing Society*, 1998.

[7] D. Bhat and S. Nayar, “Ordinal measures for image correspondence,” in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20 Issue: 4, pp. 415–423, April 1998.

[8] S. B. Needleman, and C. D. Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” *Journal of Molecular Biology*, 48, pp. 443-453, 1970.