Extended Temporal Ordinal Measurement Using Spatially Normalized Mean for Video Copy Detection

HeungKyu Lee and June Kim

This letter proposes a robust feature extraction method using a spatially normalized mean for temporal ordinal measurement. Before computing a rank matrix from the mean values of non-overlapped blocks, each block mean is normalized so that it obeys the invariance property against linear additive and subtractive noise effects and is insensitive against multiplied and divided noise effects. Then, the temporal ordinal measures of spatially normalized mean values are computed for the feature matching. The performance of the proposed method showed about 95% accuracy in both precision and recall rates on various distortion environments, which represents the 2.7% higher performance on average compared to the temporal ordinal measurement.

Keywords: Video fingerprinting, video copy detection, content-based video retrieval.

I. Introduction

As a means to prevent users from uploading copyrighted digital video contents, video copy detection has recently received increasing attention as a file filtering application [1], [2]. One of the challenging issues in video copy detection is robust feature vector extraction.

The matching performance between motion-based and spatial ordinal-based feature vectors have been evaluated and compared in [3]. The spatial ordinal-based feature showed higher matching performance. It is immune to global changes that are introduced by the digitization and encoding process. In addition, it is very computationally efficient and appropriate to real-time service for a live video. Both motion-based and color-based features discard spatial information while they represent the relative change in the intensities or absolute color over time. Thus, they are also immune to global color changes, but their performances are worse than the ordinal-based feature. This is due to the fact that there are a number of shots in different parts of a video with the same color and motion information. This will cause false matching. This fact signifies that the use of spatial information is required. An edge-based feature represents the local spatial information while it is susceptible to global changes in brightness, blurring, and lossy compression because the edge information can be removed or changed by an encoder. Thus, the robust feature vectors that represent spatial and temporal information are required to increase the matching performance.

The temporal ordinal measurement [4] is known to have the best performance compared to other methods, even the spatio-temporal sequence matching [5]. However, the temporal ordinal measurement does not consider explicitly spatial normalization. Thus, to represent the distinct inter and intra relationships of a video, an extended temporal ordinal measurement is proposed.

II. Extended Temporal Ordinal Measurement

1. Intraframe and Interframe Normalization of a Video

Prior to the feature extraction procedure, we consider the normalized feature extraction domain. First, as an intraframe normalization method, a video frame is partitioned into $m \times n$ non-overlapped blocks so that it is invariant to the resolution of an input video. Second, the interframe normalization method is employed by subsampling video frames using a temporal mean method so that it is invariant against the encoding rate.

Manuscript received Nov. 27, 2009; revised Feb. 3, 2010; accepted Feb. 18, 2010.
This research was supported by Seokyeong University, Rep. of Korea, in 2009.
HeungKyu Lee (phone: +82 2 2262 5336, email: hklee@ispl.korea.ac.kr) is with the Department of Visual Information Processing, Korea University, Seoul, Rep. of Korea.
June Kim (email: june@skuniv.ac.kr) is with the Department of Information and Communication Engineering, Seokyeong University, Seoul, Rep. of Korea.
doi:10.4218/etrij.10.0209.0485

This makes the video frame rate fixed at F_R frames per second (fps), where F_R is smaller than the frame rate of an original video. As a result, the intraframe and interframe normalization method of a video can provide robustness against the scaling and frame rate change distortions. To avoid heavy computational costs, just the *Y* channel is selected for feature extraction. The interframe normalization is accomplished by computing a temporal mean, and then a video frame is partitioned into *m*×*n* blocks.

2. Feature Extraction Using Spatial Normalization of Block Intensity Mean Values

The proposed spatial normalization method begins by computing the block intensity mean values of *m*×*n* partitioned blocks on a subsampled frame. Let the mean values of partitioned blocks in *i*-th frame be $\mu_{i,K}=\{\mu_{i,1}, \mu_{i,2}, \cdots, \mu_{i,k}\}$, where k is the number of measurements in *m*×*n* (=*k*) partitioned blocks. First, the partitioned block mean can be normalized so that it obeys the invariance property against the multiplied and divided noise signals by dividing it into a global mean g_{μ_i} of block mean values $\mu_{i,K}$ as follows:

$$\mu'_{i,K} = \frac{\mu_{i,K}}{g_{\mu_i}}, \quad K = 1, 2, \cdots, k. \tag{1}$$

Let the frame pixels be $X = \{x_0, x_2, \cdots, x_{k-1}\}$, and the changed frame pixels by the constantly multiplying noise signal α be $\bar{X} = \{x_0 \times \alpha, x_1 \times \alpha, \cdots, x_{k-1} \times \alpha\}$. Then, the partitioned block mean $\mu_{i,K}$ of the original frame pixels is changed into $\alpha\mu_{i,K}$ by this noise signal. Thus, (1) can be rewritten with the changed block mean $\bar{\mu}_{i,K}$ and global mean $\bar{g}_{\mu_i}$ as

$$\mu'_{i,K} = \frac{\bar{\mu}_{i,K}}{\bar{g}_{\mu_i}} = \frac{\alpha\mu_{i,K}}{\alpha g_{\mu_i}} = \frac{\mu_{i,K}}{g_{\mu_i}}, \tag{2}$$

where the derived final equation in (2) is same as (1). It is the same with divided noise signals. Second, the normalized block mean can be normalized once again so that it is insensitive to the additive and subtractive noise signals. It is computed using the mean μ_{norm} and standard deviation σ_{norm} of normalized mean values $\mu'_{i,K}$ as

$$\mu''_{i,K} = \frac{\mu'_{i,K} - \mu_{\text{norm}}}{\sigma_{\text{norm}}}, \quad K = 1, 2, \cdots, k, \tag{3}$$

where *k*-dimensional feature vectors $\mu''_{i,K}$ are stored in the feature database. Then, the spatially normalized mean values using (3) are employed to compute rank matrices for feature matching in the next subsection.

3. Feature Matching Using Temporal Ordinal Measures

The feature matching procedure begins by computing temporal relative ordering (rank matrix) of spatially normalized mean values $\mu''_{i,K}$ on temporal frame sequences. The spatially partitioned block mean values $\mu''_{i,K}$, $i = 0, \cdots, N-1$, are sorted along the time series in ascending order, and its rank matrix $\pi_i[K]$ in each K index has [1×*N*] dimension where N is the length of query video sequences. Thus, the temporal similarity measure *TSM* between two video sequences is given by

$$TSM\left(V_{q,i}, V_{t(l),p+i}\right) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{C_N}\sum_{i=0}^{N-1}\left|\pi_{q,i}[k] - \pi_{t,p+i}[k]\right|\right), \tag{4}$$

where $\pi_{q,i}[k]$ is the rank matrix of [1×*N*] average matrix $\mu''_{q,i}[k]$ from query video sequences, $\pi_{t,p+i}[k]$ is the rank matrix of [1×*N*] average matrix $\mu''_{t,i}[k]$ from the target original video subsequence $V_t[p{:}p+N-1]$, where p is a clipped temporal frame index in an original video, and the normalization factor C_N is obtained by

$$C_N = \sum_{i=1}^{N}\left|N + 1 - 2\times i\right|, \tag{5}$$

where N is the number of measurements. In (4), we cannot see the clipped temporal frame index p. Thus, searching a similar video and temporal location is the same as finding a $V_t(l)$ video with a p value that has a minimum difference value, which should be lower than the predefined threshold value τ:

$$D = \arg\min_{p} TSM\left(V_{q,i}, V_{t(l),p+i}\right) < \tau, \tag{6}$$

where $0 \le p < M-N$, M is the length of an original video sequences, and l is the video index that are stored in the feature database.

III. Experimental Evaluations

The performance of the proposed method is evaluated using 240 movies wherein these movies have a variety of subsequences including sports, landscapes, action, and animation. The total length of the movies is approximately 320 h. Every query video is generated for experimental evaluations by using six types of modifications. The base modification of test type I is lossy compression to DivX 256 kbps, resized into CIF videos, and the frame rate changes from 25-30 fps to 15 fps. Test type II is a Gaussian-blurred set with a radius of 1 pixel that is re-sampled using the test type I videos. Test type III is a globally changed set in brightness (30%). Test type IV is a histogram equalized set. Test type V is a letter-box added set. The test type VI is a pillar-box added set.

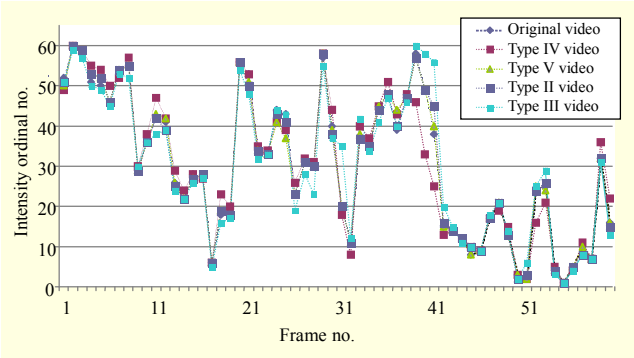

Fig. 1. Example of temporal ordinal ranking of spatially normalized mean feature vectors.

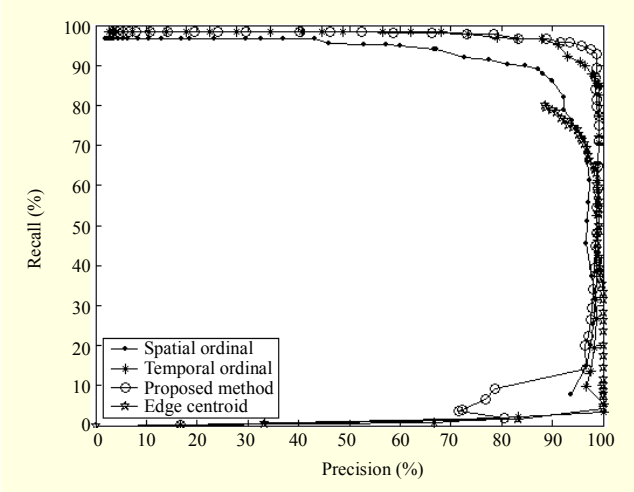

Fig. 2. Performance comparison using average precision and recall.

Test type sets II to VI are re-sampled from the test type I set. Total 1440 (=240×6) query videos with 50 seconds query length are used for matching performance evaluations.

First, we evaluated the robustness of the proposed spatial normalization method under some distortion environments. Figure 1 describes an example of temporal ordinal ranking of spatially-normalized mean-feature vectors. The four transformations include test types II to V. It showed that the proposed method's results are very similar between the original video and distorted videos.

From this result, we evaluated the extended temporal ordinal measurement using proposed feature vectors and compared it to the spatial ordinal measure, edge-centroid features [6], and temporal ordinal measures [4]. We used the 2×2 partitioned blocks from the interframe normalized video at 2 fps for homogeneous experimental environments and comparisons. The performance evaluations are analyzed using the precision and recall rate [5]. As shown in Fig. 2, the result showed that the proposed method has the best performance compared to other methods. In Fig. 2, the ideal curve must pass through (100, 100), that is, 100% precision and 100% recall. The proposed method showed similar results with the temporal ordinal measure [4] in test types III and IV, but also improvements in other test sets. Averaging precision and recall rates shows that the proposed method is highly robust against lossy compression, global change in brightness, blurring, histogram equalization, resizing, frame rate change, and combined distortions. The proposed method showed the best performance at threshold τ =0.3, where the precision rate is 95.68%, and the recall rate is 94.92%. Meanwhile, as a second rank, the temporal ordinal measure showed that its precision rate is 92.92%, and its recall rate is 92.27% at threshold τ =0.29. From this result, the proposed method accomplished the improvement about 2.7% with respect to both precision and recall rates, and it proved experimentally that the spatio-temporal feature vectors are more robust than spatial or temporal features.

IV. Conclusion

When a video is transformed by re-encoding tasks, the frame pixel information and its feature vectors are changed. To cope with this, robust feature vector selection insensitive to these modifications is very important. From this point of view, the extended temporal ordinal feature vectors were the salient ones to make the matching system robust against various distortions by transforming the partitioned block mean into spatially normalized mean and by using temporal information.

References

[1] J. Oostveen, T. Kalker, and J. Haitsma, "Feature Extraction and a Database Strategy for Video Fingerprinting," *Proc. Int. Conf. Recent Adv. Vis. Inf. Syst.*, 2002, pp. 117-128.

[2] J. Yuan et al., "Fast and Robust Short Video Clip Search Using an Index Structure," *Proc. the 6th ACM SIGMM*, 2004, pp. 61-68.

[3] A. Hampapur, K.H. Hyun, and R. Bolle, "Comparison of Sequence Matching Techniques for Video Copy Detection," *SPIE Proc.*, vol. 4676, 2002, pp. 194-201.

[4] L. Chen, and F.W.M. Stentiford, "Video Sequence Matching Based on Temporal Ordinal Measurement," *Pattern Recognition Lett.*, vol. 29, 2008, pp.1824-1831.

[5] C. Kim and B. Vasudev, "Spatiotemporal Sequence Matching for Efficient Video Copy Detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no.1, Jan. 2005, pp.127-132.

[6] S. Lee and C.D. Yoo, "Robust Video Fingerprinting for Content-Based Video Identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no.7, 2008, pp. 983-988.