A Near-Duplicate Video Retrieval Method Based on Zernike Moments

Tang-You Chang[1], *Shen-Chuan Tai*[1], *Guo-Shiang Lin*[2]

[1] Institute of Computer and Communication Engineering,
National Cheng Kung University
E-mail: e2490668@gmail.com, sctai@mail.ncku.edu.tw
[2] Department of Computer Science and Information Engineering,
Da-Yeh University
E-mail: khlin@mail.dyu.edu.tw

ABSTRACT

In this paper, a near-duplicate video retrieval method developed based on invariant features was proposed. After shot change detection, Zernike moments are extracted from each key-frame of videos as invariant features. We obtain the key-frame similarity by computing the difference of Zernike moments between key-frames of the query and test videos. To achieve near-duplicate video retrieval, each key-frame is considered as an individual sensor and then evaluating all of key-frames is considered as multiple sensors. The results of key-frames are fused to obtain a better performance of near-duplicate video retrieval. The experimental results show that the proposed method can not only find the relevant videos effectively but also resist to the possible modifications such as re-scaling and logo insertion.

Keywords: *near-duplicate video retrieval, Zernike moments, big data;*

1. INTRODUCTION

Due to remarkable advance of communication technology, the growth of online videos is huge in past few years. According to the report on the official YouTube website, there are more than 1 billion users and 300 hours of videos are uploaded to YouTube every minute. This means that how to manage the video content effectively becomes increasingly important. Therefore, some emerging services such as video sharing, video broadcasting, and video recommendation draw researchers' attention.

As for video sharing, most websites often allow their users to freely upload videos without any checking procedures. It is expected that some videos with the same or similar content can be found on Internet. Figure 1 illustrates these some cases of near-duplicate videos. Figure 1(a) is an original frame and Fig. 1(b), 1(c), 1(d) are three common duplicates of Fig. 1(a). These videos with similar content are called as near-duplicate videos (NDVs). To improve the efficiency of video management, it is an important issue to retrieve and manage these near-duplicate videos. The technique that retrieves near-duplicate videos is called Near-Duplicate Video Retrieval (NDVR). Though there are some existing methods [1-3], developing an efficient NDVR method is still challenging. Therefore, we aim at developing an effectively NDVR method.

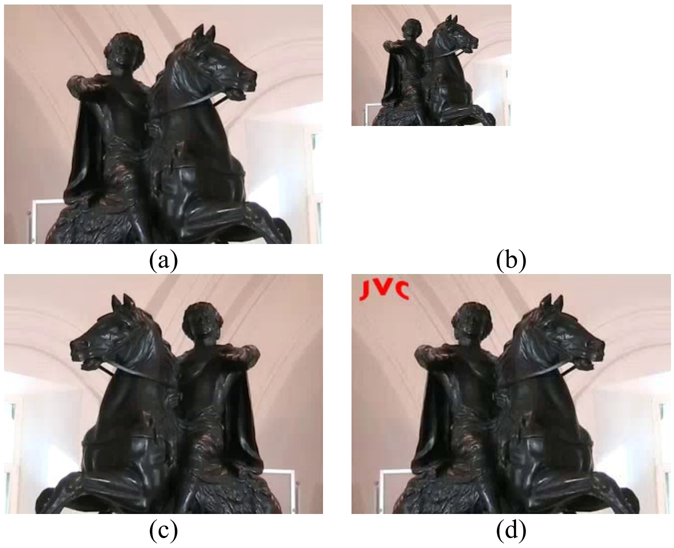

Fig. 1 Samples of near-duplicate videos: (a) original (b) rescaling (c) mirroring, and (d) logo insertion

2. Related Work

NDVR often contains three parts: video signature, similarity measure, and searching algorithm. Based on [2], existing video signatures can be classified into several classes according to what kinds of information are adopted.

The video-level global signature translates whole video into single signature. In [3], Huang et al. introduced a statistical model called Bounded Coordinate System (BCS). By combining the principal component analysis (PCA) and the Bi-distance Transformation (BDT), a video clip can be summarized as a coordinate system which records the dominating content-changing trends and ranges by directions and lengths. Since global signature is very compact, it usually reduces the content redundancy to bring various benefits in storage, management, computation, and retrieval. However, it ignores the local information like the object or region in video easily and exist doubt of representativeness.

Frame-level local signature extracts the local feature on individual frame like local key-point descriptors. Such usage in NDVR is closely related to near-duplicate image retrieval like [4]. Due to the higher computationally then video-level global feature, there exists two classes of frame level local signature usage. One of them is doing pairwise frame matching with local key-point descriptors. It costs lots of computation and need acceleration like pre-filtering.

To improve the time complexity issue of frame-level local signature, more studies tend to transform the local information into global signature. The Bag of Words (BoW) model is often applied for image classification in computer vision [5]. It seems the image as a document and the key-point descriptors are the words in the document. To convert the local descriptors of image into the visual words, each descriptor will be matched with "codebook" which has been trained by clustering algorithm like k-means clustering. A frame can be represented as a histogram of the occurrences of the visual words in that frame. The BoW model shows excellent precision and high efficiency in NDVR [6].

Spatio-temporal signature represents not only the frame information, but also the change between frames. Zobel and Hoad [7] used the shot information to represent the frame signature. It summarizes a video into a sequence of numbers, each representing a shot length, the change of color distribution between frames and the inter-frame change in spatial movement of the lightest and darkest pixels in each frame over time. However, all the codes only indicate the consecutive inter-frame difference within each individual sequence without carrying content information. Huang et al. [8] proposed a method which transforms a video stream into the one-dimensional Video Distance Trajectory (VDT), which preserves some content information due to the usage of the reference point. The difference of color histogram between frames will be calculated, and converted into a sequence of compact signatures called Linear Smoothing Functions (LSFs). LSF adopts compound probability to combine three independent video factors for effective segment similarity measure, which is then utilized to compute sequence similarity for NDVR.

Though the SIFT features are useful to find the corresponding frames, their computational complexity is too much and then limit their applicability of real applications. On the other hand, some modifications such as cropping and mirror reflection may be used to generate near-duplicate videos. Therefore, invariant features are necessary and important to develop an effective NDVR method.

3. Proposed Method

Image moments are scale invariant and have been used in many applications such as image retrieval [9],[11]. One important advantage of Zernike moments [10] is resistant to geometric rotation and noise. It is expected that this property of Zernike moments is useful for near-duplicate video retrieval. On the other hand, it is no doubt that the computational complexity of processing all of frames in a video is very high. To reduce the computational complexity, we only extract and examine the key-frames. Therefore, it motivates us to develop an efficient NDVR method based on Zernike moments. Figure 2 illustrate the proposed NDVR method.

Let an original video sequence be $\boldsymbol{I^O}$ and there are some key-frames $\boldsymbol{IK^O}$. A near-duplicate video sequence $\boldsymbol{I^C}$ with the same frame dimensions as $\boldsymbol{I^O}$ and there are some key-frames $\boldsymbol{IK^C}$. Similar to [12], we extract a set of features from $\boldsymbol{IK^O}$ and another set from $\boldsymbol{IK^C}$. Then, the similarity between $\boldsymbol{IK^O}$ and $\boldsymbol{IK^C}$ is measured by computing the distance between their features.

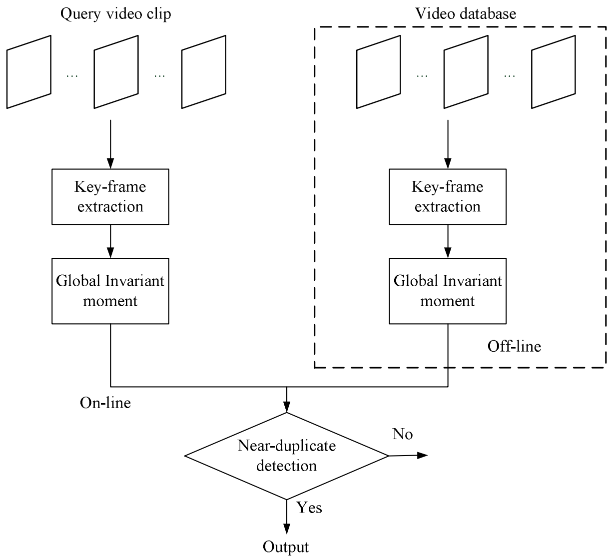

Fig. 2 Block diagram of the proposed NDVR method

3.1 Global Invariant Moment

Since the frame size may not be the same, we normalize the frames of the original video and the near-duplicate version before computing the Zernike moments. After normalization, we measure a Sobel edge map of each key-frame and then the edge map is used for computing Zernike moments.

First, we describe the $(i+j)$-order central moments as follows:

$$\phi_{ij} = \sum_x \sum_y \left(x-\bar{x}\right)^i \left(y-\bar{y}\right)^j f(x,y) \tag{1}$$

where $f(x,y)$ denotes the edge map and $(\bar{x},\bar{y})$ represents the centroid of the edge points. Then the moment φ_{ij} can be expressed as follows:

$$\varphi_{ij} = \frac{\phi_{ij}}{\phi_{00}^{1+(i+j)/2}} \tag{2}$$

Based on Eq. (2), six invariant moment M_i (i=1,2,…,6) can be expressed in the following:

$$M_1 = \frac{3}{\pi}(\varphi_{20} + \varphi_{02} - 1) \tag{3}$$

$$M_2 = \frac{9}{\pi^2}[(\varphi_{20} - \varphi_{02})^2 + 4\varphi_{11}{}^2] \tag{4}$$

$$M_3 = \frac{16}{\pi^2}[(\varphi_{30} - 3\varphi_{12})^2 + (3\varphi_{21} - \varphi_{03})^2] \tag{5}$$

$$M_4 = \frac{144}{\pi^2}[(\varphi_{30} + \varphi_{12})^2 + (\varphi_{21} + \varphi_{03})^2] \tag{6}$$

$$M_5 = \frac{13824}{\pi^4}\{(\varphi_{30} - 3\varphi_{12})(\varphi_{30} + \varphi_{12})[(\varphi_{30} + \varphi_{12})^2 - 3(\varphi_{21} + \varphi_{03})^2] + (3\varphi_{21} - \varphi_{03})(\varphi_{21} + \varphi_{03})[3(\varphi_{30} + \varphi_{12})^2 - (\varphi_{21} + \varphi_{03})^2]\} \tag{7}$$

$$M_6 = \frac{864}{\pi^3}\{(\varphi_{20} - \varphi_{02})[(\varphi_{30} + \varphi_{12})^2 - (\varphi_{21} + \varphi_{03})^2] + 4\varphi_{11}(\varphi_{30} + \varphi_{12})(\varphi_{21} + \varphi_{03})\} \tag{8}$$

According to Eqs. (3) to (8), six Zernike moments can be computed and then formed as a feature vector.

3.2 Similarity measurement

After measuring Zernike moments of each key-frame, we need to measure the similarity of two videos. Although Zernike moments had been used on image retrieval, they seem not robust to image scaling. According to our experience, the difference of Zernike moments between the original image and its near-duplicate is sensitive to image content. Unfortunately, image scaling is a very common process to generate near-duplicate videos. To reduce the impact of not only geometric scaling but also image content on near-duplicate video retrieval, a reference-based NDVR method is developed.

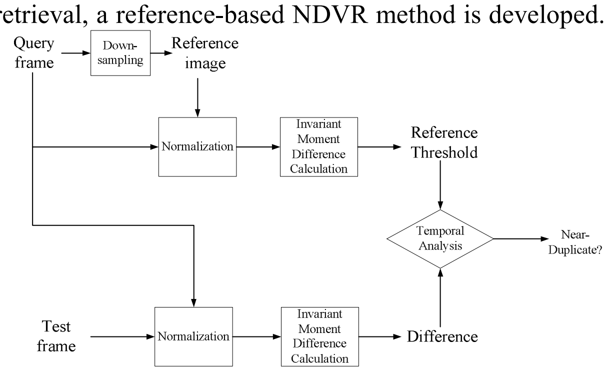

Fig. 3 An illustration of reference-based NDVR

To reduce the impact of image content on near-duplicate video retrieval, a reference-based approach is adopted. This means that we calculate and consider the difference of Zernike moments between the original frame and its quarter version as a reference value.

For the feature f_m^Q of each key-frame in the query video v^Q, its reference value can be computed. To search near-duplicate frames, the feature vector $[f_{i1}^T, f_{i2}^T, ..., f_{iN}^T]$ of one key-frame of the test video v_i^T can have the corresponding difference of Zernike moments with respect to f_m^Q in v^Q. If the difference between f_m^Q and f_{in}^T is lower than the reference value of f_m^Q, the video frame is considered as a near-duplicate. Then a binary sequence $[S_1^Q, S_2^Q, ..., S_M^Q]$ can be obtained.

3.3 Temporal analysis

In addition to spatial analysis for each key-frame, we also consider the results of similarity measurement from all of key-frames to increase the performance of the proposed method. Here we consider each key-frame as a sensor and whether one key-frame is a near-duplicate as the output of the sensor. It is expected that determining whether a key-frame is a near-duplicate is a binary hypothesis testing problem. Then the final decision can be obtained by combining all of output of sensors based on multiple-sensor fusion concept [13].

We define the state of two hypothesis H_0 and H_1 for each sensor, where H_0 represents that v_i^T is not the NDV of v^Q and H_1 means v_i^T is the NDV of v^Q. Since determining whether a key-frame is a near-duplicate is a binary hypothesis testing problem, we examine all of key-frames and then a binary sequence $[S_1^Q, S_2^Q, ..., S_M^Q]$ can be obtained below:

$$S_i^Q = \begin{cases} -1 & \text{if } H_0 \text{ is declared} \\ +1 & \text{if } H_1 \text{ is declared} \end{cases} \tag{9}$$

Based on the Bayesian criterion, the fused decision D can be expressed as follows:

$$D = \text{sgn}\left(-Q + \sum_{i=1}^{N_S}\left(\omega_i^Q \cdot S_i^Q\right)\right) \tag{10}$$

where Q is a control factor. The weight ω_i^Q is expressed as follows:

$$\omega_i^Q = \delta\left(S_i^Q - 1\right) \cdot \log\frac{1-P_{FN}}{P_{FP}} + \delta\left(S_i^Q + 1\right) \cdot \log\frac{1-P_{FP}}{P_{FN}} \tag{11}$$

where $\text{sgn}(\cdot)$ denotes a sign function ($\text{sgn}(x)=1$ for x>0 and $\text{sgn}(x)=-1$ for x<0); $\delta(x)$ is the impulse function ($\delta(x)=1$ for $x=0$ and $\delta(x)=0$ for $x \neq 0$). P_{FP} is the number negative case labeled as positive; P_{FN} refer to the number of the positive case which is incorrectly labeled as negative. Finally we can decide if the v_i^T is the NDV of v^Q by D:

$$\begin{cases} D < 0 & \text{if } H_0 \text{ is declared} \\ D \geq 0 & \text{if } H_1 \text{ is declared} \end{cases} \quad (12)$$

Both P_{FP_i} and P_{FP_i} can be estimated from the training data. The control factor Q can be used to adjust the precision of the proposed method.

4. Experimental Results

In this paper, we focus on some common operations, rescaling, mirroring, and logo insertion, for generating NDVs. On the other hand, we utilize three performance measurements, *precision*, *recall*, and *accuracy* rates, to evaluate the proposed method [14]. The performance measurements are expressed below:

$$precision = \frac{P_{TP}}{P_{TP} + P_{FN}}$$

$$recall = \frac{P_{TP}}{P_{TP} + P_{FP}}$$

$$accuracy = \frac{P_{TP} + P_{TN}}{P_{TP} + P_{FP} + P_{TN} + P_{FN}}$$

where P_{FP} and P_{FN} denote the numbers of false and miss detection; P_{TP} refers to the number of the positive cases which is correctly labeled; P_{TN} represents number of negative cases correctly labeled. In the following experiments, we first analyze the ability of Zernike moments to decide the parameter in Eq. (11) and then evaluate the efficacy of the proposed NDVR method.

A. Analysis of Zernike moments

We select 200 pictures from Google website to analyze the ability of Zernike moments. Ten selected images are used to generate near-duplicate versions by using the mirroring, logo insertion and rescaling operations. Since the edge information of each key-frame, we analyze the impact of the threshold in the Sobel filter on NDVR.

Table 1 shows P_{FP} and P_{FN} while different thresholds in the Sobel filter. As we can see in Table 1, the minimal value of P_{FP} occurs when threshold = 2; when the threshold = 1, we can get the minimum of P_{FN}. To consider simultaneously P_{FP} and P_{FN}, we set the threshold value as 1.To prevent the zero problem of log function, we set P_{FN} a small value and rewrite Eq. (11) as follows:

$$\omega_i^Q = \delta(S_i^Q - 1) \cdot 9.1415 + \delta(S_i^Q + 1) \cdot 2.7104 \quad (13)$$

B. Performance of NDVR

After the parameters are measured, 200 test videos are download form TRECVID website [15] for NDVR test.

Table 1 P_{FP} and P_{FN} while different thresholds in the Sobel filter

Threshold	1	2	3	4	5	6
P_{FP}	6.65%	3.1%	3.3%	12.1%	20.8%	24.2%
P_{FN}	0%	30%	30%	25%	25%	25%

Ten video clips are selected to produce near-duplicated videos by using three operations, mirroring, logo insertion, and rescaling. We analyze the correlation between Q and the efficiency of the proposed method and Fig. 4 illustrate the result.

As we can see in Fig. 4, there is a cross point of *precision* and *recall* cruve at Q = 5. For the further analysis, the accuracy with each value of Q are also calculated and shown in Table 2. The highest value of accuracy rates can be obtained when Q = 8. Due to the consideration between *precision* & *recall* and accuracy rates, Q = 5 or Q = 7 can be selected in the proposed method.

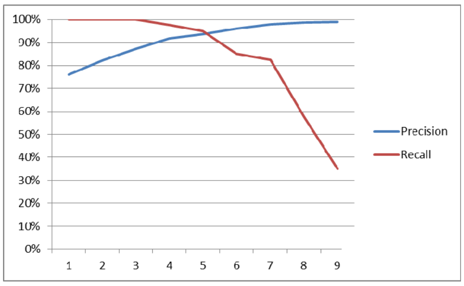

Fig. 4 Correlaction between Q and *precision* & *recall* rates

Table.2 Accuracy rates of the proposed method

Q	1	2	3	4	5
Accuracy	68.9%	78.7%	85.8%	91.3%	93.5%
Q	6	7	8	9	
Accuracy	96.4%	98.0%	98.5%	98.5%	

5. Conclusions

In this paper, we present a NDVR method developed based on Zernike moments. After shot change detection, Zernike moments are computed from each key-frame of videos as invariant features. We measure the key-frame similarity by computing the difference of Zernike moments between key-frames of the query and test videos. To achieve near-duplicate video retrieval, each key-frame is considered as an individual sensor and then evaluating all of key-frames is considered as multiple sensors. The results of examining key-frames are fused to obtain a better performance of near-duplicate video retrieval. To evaluate the proposed method, we collect a lot of near-duplicate videos. The accuracy rate of the proposed method can be up to 98%. Experimental results show that the proposed method can achieve near-duplicate video retrieval.

In the future, by re-design the equation of proposed similarity measure, we could do the NDVD which not only find the duplicate video in database but also discover the duplicate clip in video. Furthermore, we may combine the proposed method with SURF feature matching technology to increase the accuracy rate and analyze the possible operations for generating near-duplicate videos.

Acknowledgment

This research was supported by the Ministry of Science and Technology, Taiwan, under the grant of MOST 103-2221-E-212-004-MY2.

References

[1] X. Wu., W. Zhao, and C.-W. Ngo, "Near-duplicate key-frame retrieval with visual key-words and semantic context," In Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR'07), pp. 162–169, 2007

[2] J. J. Liu, Z. Huang, H. Cai, H. T. Shen, "Near-Duplicate Video Retrieval: Current Research and Future Trends," ACM Computing Surveys (CSUR), Vol. 45, No. 4, 2013.

[3] Z. Huang, H. T. Shen, J. Shao, X. F. Zhou, and B. Cui, "Bounded coordinate system indexing for real-time video clip search, " ACM Trans. Inf. Syst. Vol. 27, No. 3, pp. 17–33, 2009

[4] O. Chum, and J. Matas, "Large-scale discovery of spatially related images," IEEE Trans. Pattern Anal. Mach. Intell. Vol. 32, No. 2, pp. 371–377, 2010.

[5] J. Sivic, and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," In Proc. the 9th International Conference on Computer Vision (ICCV'03), pp. 1470–1477, 2003.

[6] H. S. Min, J. Y. Choi, W. D. Neve, and Y. M. Ro, "Near-Duplicate Video clip detection using model-free semantic concept detection and adaptive semantic distance measurement," IEEE Trans. Circuits and Systems for Video Technology, Vol. 22, No. 8, pp. 1174–1187,2012.

[7] J. Zobel and T. C. Hoad, "Detection of video sequences using compact signatures," ACM Trans. Inf. Syst. Vol. 24, No. 1, pp. 1–50, 2006.

[8] Z. Huang, H. T. Shen, J. Shao, B. Cui, and X. Zhou, "Practical online near-duplicate subsequence detection for continuous video streams," IEEE Trans. Multimedia vol. 12, No. 5, pp. 386–398, 2010.

[9] M. K. Hu, "Visual Pattern Recognition by Moment Invariants," IRE Trans. Info. Theory, vol. IT-8, pp.179–187, 1962.

[10] W. Y. Kim, Y. S. Kim, "A region-based shape descriptor using Zernike moments", Signal Processing: Image Communication, Vol. 16, pp. 95-102, 2000.

[11] S. K. Hwang, W. Y. Kim, "A novel approach to the fast computation of Zernike moments," Pattern Recognition, Vol. 35, Issue 12, pp. 2905-2911, 2002.

[12] Y. X. Peng and C. W. Ngo, "Clip-based similarity measure for query-dependent clip retrieval and video summarization," IEEE Trans. Circuits and Systems for Video Technology, Vol. 16, pp. 612–627, 2006.

[13] G.-S. Lin, M.-K. Chang, Y.-J. Chang, and C.-H. Yeh, "A Gender Classification Scheme Based on Multi-Region Feature Extraction and Information Fusion for Unconstrained Images," accepted by *Multimedia Tools and Applications*, 2015.

[14] T.-Y. Chang, S.-C. Tai, and G.-S. Lin, "A passive multi-purpose scheme based on periodicity analysis of CFA artifacts for image forensics," Journal of Visual Communication and Image Representation, Vol. 25, No. 6, pp.1289–pp.1298, Aug. 2014.

[15] TRECVID Guidelines [Online]. Available: http://trecvid.nist.gov/