SCIENCE CHINA
Information Sciences

• LETTER •

May 2016, Vol. 59 059101:1–059101:3
doi: 10.1007/s11432-016-5528-6

Spherical torus-based video hashing for near-duplicate video detection

Xiushan NIE[1,4], Yane CHAI[1], Ju LIU[2*], Jiande SUN[2] & Yilong YIN[3]

[1]*School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan* 250014*, China;*
[2]*School of Information Science and Engineering, Shandong University, Jinan* 250100*, China;*
[3]*School of Computer Science and Technology, Shandong University, Jinan* 250100*, China;*
[4]*Digital Media Technology Key Laboratory of Shandong Province, Shandong University of Finance and Economics, Jinan* 250014*, China*

Received November 16, 2015; accepted December 28, 2015; published online March 2, 2016

Citation Nie X S, Chai Y E, Liu J, et al. Spherical torus-based video hashing for near-duplicate video detection. Sci China Inf Sci, 2016, 59(5): 059101, doi: 10.1007/s11432-016-5528-6

Dear editor,

With the rapid development of multimedia technologies, users can easily generate and share multiple videos through the Internet. Similarly, numerous illegal and useless near-duplicate videos generated through simple reformatting, transformation, and editing appear on the web. These near-duplicate videos inconvenience users when they are surfing the Internet and are considered a copyright infringement. Robust video hashing, which is also called video fingerprinting, does not require access to the video contents at the time of creation and can be used to detect existing contents. In general, a robust video hash is a short digest extracted from a video; however, it is robust to content-preserving attacks such as noise, logo addition, and contrast change. That is, similar to a human fingerprint that identifies a specific person, video hashing can classify video contents by extracting and comparing short digests. As such, video hashing is a feasible means to detect near-duplicate videos.

Frame-based hashing methods are the primary type of video hashing [1,2], in which each frame or key frame of a video is regarded as an image. These methods employ advanced image hashing techniques and concatenate image hashes as the final hash. However, compared with still images, a video is a time sequence. As such, spatiotemporal information plays an important role in video content analysis. Therefore, fusing the spatial and temporal information of videos is important in content representation. Several spatiotemporal mechanisms have been developed in recent years [3–6]. Generally, these methods primarily consider the relation and structure along the axis direction in 3D space, in which contents may change only slightly because the objects in a short video clip remain in nearly the same position in adjacent frames. As we all know, as the number of content changes is small, the amount of information is also assumed to decrease. Consequently, existing spatiotemporal methods cannot capture rich information. Moreover, users usually focus on the center of an image where visual saliency commonly appears, and then gradually look at the entire image. Thus, the ring partition performed on an image along the direction of the radius is close to the perception of the human visual system.

In this letter, inspired by the ring partition in an image [7], we propose a spherical torus based video

* Corresponding author (email: juliu@sdu.edu.cn)

The authors declare that they have no conflict of interest.

hashing scheme to overcome the limitations of existing methods. We combine the spatial and temporal information using spherical tori, where the spatial and temporal information are fused along the tangent planes of a sphere; it can capture more content changes in temporal evolution than the existing fusion method. To handle all videos similarly and enhance the security of the proposed method, we first experimentally produce a normalized video, and then select the locations of the overlapping sub-cubes based on a key and spatial sampling. Second, each video cube is partitioned into different spherical tori, which are then projected onto a spatiotemporal image that contains the spatial and temporal information of the corresponding video. Finally, hashes are generated by applying non-negative matrix factorization (NMF) to the spatiotemporal image.

Spherical torus generation. In each overlapping video cube, the entire cube is partitioned into a few spheres according to different radii. Then, all spherical tori are projected onto an image according to a certain strategy. The primary advantage of using spherical tori is that it is helpful in obtaining substantial spatiotemporal information along the tangent planes of a sphere. For example, when two videos are similar in scene or object but different in content, rich spatiotemporal information can provide a stronger distinguishing ability. The procedures for generating the spherical torus and the spatiotemporal image are described as follows.

The size of the video cube is $Q \times Q \times Q$. n is the total number of spherical tori in the video cube (including the minimum sphere that can be considered a spherical torus), U_k is the set of all pixels in the kth spherical torus ($k = 1, 2, \ldots, n$), and Y_k is the set of luminance values of all pixels in U_k ($k = 1, 2, \ldots, n$). Given that each spherical torus is used as a column of the spatiotemporal image, we only use the pixels in the inscribed sphere of the video cube and divide such sphere into spherical tori with equal volume.

The spherical torus can be determined by calculating the radius of the sphere as well as the distance between each pixel and video cube center. Each pixel of the spherical torus is identified using two adjacent radii, except for the innermost radius. r_k ($k = 1, 2, \ldots, n$) is assumed as the kth radius; hence, the values of these radii are arranged in ascending order. Therefore, r_1 and r_n are the radii of the minimum and maximum spheres, respectively. Evidently, $r_n = \lfloor Q/2 \rfloor$ for the video cube whose size is $Q \times Q \times Q$, where $\lfloor \rfloor$ indicates rounding down. To calculate the values of the other radii, we use the following formulas to determine the volume of inscribed sphere V and the average volume of each spherical torus A:

$$V = \frac{4}{3}\pi r_n^3, \quad A = \lfloor V/n \rfloor . \tag{1}$$

Assuming $r_0 = 0$, the different radii r_k ($k = 1, 2, \ldots, n$) can be calculated using the following formula:

$$r_k = \sqrt[3]{\frac{3A}{4\pi} + r_{k-1}^3}. \tag{2}$$

Let $p(x, y, z)$ be the pixel value in the yth row, xth column, and zth frame of the video ($1 < x, y, z < Q$), and (x_c, y_c, z_c) be the coordinates of the video cube center. If Q is an even number, then $x_c = Q/2 + 0.5$, $y_c = Q/2 + 0.5$ and $z_c = Q/2 + 0.5$; otherwise, $x_c = (Q + 1)/2$, $y_c = (Q + 1)/2$ and $z_c = (Q + 1)/2$. Therefore, the Euclidean distance between pixel $p(x, y, z)$ and video cube center (x_c, y_c, z_c) can be expressed as follows:

$$d = \sqrt{(x - x_c)^2 + (y - y_c)^2 + (z - z_c)^2}. \tag{3}$$

After obtaining all the radii as well as the distance between the pixel and the video cube center, we divide the pixels into n sets using the following formulas:

$$U_1 = \{\, p(x, y) \,|d \leqslant r_1 \,\}, \tag{4}$$

$$U_k = \{p(x, y) \ | r_{k-1} < d \leqslant r_k\}. \tag{5}$$

Thus, a set of luminance values Y_k ($k = 1, 2, \ldots, n$) in the kth spherical torus is obtained. The data in this set are then rearranged in ascending order to generate a new vector θ_k.

Considering that the video is digital and the pixel coordinates are discrete, the number of pixels in each set Y_k is not necessarily equal to A. Hence, through linear interpolation, θ_k is mapped onto a new vector $\boldsymbol{X}_k$, which is A in size. Finally, the spatiotemporal image $\boldsymbol{X}$ is formed by combining the new vectors as follows:

$$\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n]. \tag{6}$$

Hash generation. Each column of the spatiotemporal image, which consists of pixel luminance values, is a high-dimensional vector. To achieve a compact representation of the image, NMF is used to reduce matrix dimension. For an arbitrary non-negative matrix $\boldsymbol{X} = \{x_{ij}\}_{M \times N}$, NMF can identify non-negative matrices $\boldsymbol{B} \in \Re^{M \times R}$ and $\boldsymbol{C} \in \Re^{R \times N}$, which satisfy the following formula:

$$\boldsymbol{X} \approx \boldsymbol{B}C, \tag{7}$$

where R is the rank of NMF, and it is set to 2 in this study, matrix $\boldsymbol{B}$ is the base matrix, and

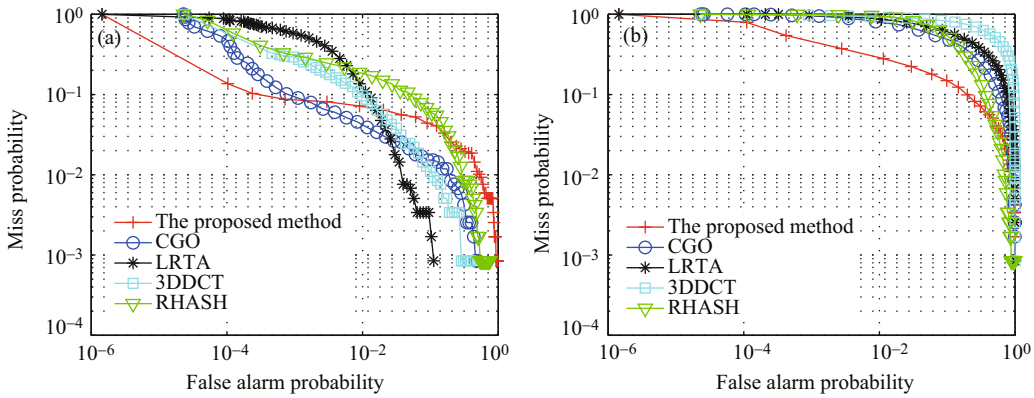

Figure 1 (Color online) The performance (ROC (Log)) under different modifications. (a) Letterbox; (b) noise+blur+pip+caption insertion.

matrix $\boldsymbol{C}$ is the coefficient matrix. NMF is also adopted to reduce the dimension of the spatiotemporal image $\boldsymbol{X}$ and to concatenate all columns in coefficient matrix $\boldsymbol{C}$ to acquire the desired hash vector of the video cube; the final hash vector is obtained by concatenating the hash vectors of all the video cubes. Obviously, the final hash is a vector with real-values. Therefore, the distance between the different hash vectors can be computed by Euclidean distance. Some existing video hashing may quantify the real-value vector into binary via secret keys, such as median quantification in [2]. Obviously, the common strategy of quantification can be definitely applied in the proposed method.

Experiments. The videos used in the experiments were downloaded from CC_WEB_VIDEO (vireo.cs.cityu.edu.hk/webvideo) and the OV (www.openvideo.org) data sets. We compared the proposed method with those based on CGO (frame-based) [1], LRTA (spatiotemporal-based) [6], 3DDCT (spatiotemporal-based) [3] and RHASH (frame-based hashing) [2] methods. The receiver operating characteristic (ROC) curve was used to evaluate the performance of the different algorithms which are shown in Figure 1. We can see that the performance of the proposed method is superior to the other methods for most modifications. The other experimental results are shown in the supporting information.

Conclusion. In this study, a video hashing algorithm was proposed based on spherical torus and NMF. In the proposed approach, a spatiotemporal image was established by generating spherical torus that linked spatiotemporal characteristics with one another. NMF was then employed to reduce the dimensionality. The main contribution of the proposed method was that the spherical torus partition was used to capture richer video-content spatiotemporal information, which can lead to better performance in near-duplicate video detection.

Acknowledgements This work was supported by the NSFC Joint Fund with Guangdong under Key Project (Grant No. U1201258), National Natural Science Foundation of China (Grant No. 61573219, 61101162), Shandong Natural Science Funds for Distinguished Young Scholar (Grant No. JQ201316) and National Science Foundation of Shandong Province (Grant No. ZR2014FM012).

Supporting information The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Lee S, Yoo C. Robust video fingerprinting for content-based video identification. IEEE Trans Circuits Syst Video Tech, 2008, 18: 983–988
2 Roover C D, Vleeschouwer C D, Lefebvre F, et al. Robust video hashing based on radial projections of key frames. IEEE Trans Signal Process, 2005, 53: 4020–4037
3 Coskun B, Sanku B, Memon N. Spatio-temporal transform based video hashing. IEEE Trans Multimedia, 2006, 8: 1190–1208
4 Wei Z K, Zhao Y, Zhu C, et al. Frame fusion for video copy detection. IEEE Trans Circuits Syst Video Tech, 2011, 21: 15–28
5 Nie X S, Liu J, Sun J D, et al. Robust video hashing based on representative-dispersive frames. Sci China Inf Sci, 2013, 56: 068104
6 Li M, Monga V. Robust video hashing via multilinear subspace projections. IEEE Trans Image Process, 2012, 21: 4397–4409
7 Tang Z J, Zhang S Q, Zhang S C. Robust perceptual image hashing based on ring partition and NMF. IEEE Trans Knowl Data Eng, 2014, 26: 711–724