ROBUST VIDEO HASHING BASED ON LOCAL FLUCTUATION PRESERVING FOR TRACKING DEEP FAKE VIDEOS

[1]*Lv Chen,* [1,*]*Dengpan Ye,* [2]*Yueyun Shang,* [1]*Jiaqing Huang*

[1] *Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, P.R. China*
[2] *School of Mathematics and Statistics, South Central University for Nationalities, P.R. China*

ABSTRACT

With the rapid development of deepfake techniques, massive face manipulation videos appeared on social networks. These deepfake videos not only violated the original video of the author's privacy, but also seriously threatened the security of the video database. Robust video hashing can map videos with similar visual content into similar hash codes, which is beneficial for tracking fake video material in social networks. In this paper, a robust video hashing algorithm based on local fluctuation preserving is proposed. The algorithm uses a shot segmentation model and local statistical descriptors, which is robust to many commonly-used digital operations and can accurately track the original version of these fake videos from a huge video database. An essential contribution is a shot segmentation model reconstruction from input video with image hashing and discrete wavelet transform, reaching initial data compression and against noise attack. In addition, as local statistical descriptors are content-based and local preserving features, the hash generated by local statistical descriptors can achieve good discrimination and ensure that the proposed hash has good tracking ability to fake videos from original videos.

Index Terms— Robust video hashing; Privacy protection; Tracking deepfake video; Key shot.

1. INTRODUCTION

Recently, fake video through digital operation, especially deep fake video, has become a hot topic. These manipulation videos are swapping the face of a person by the face of another person through various face manipulation techniques [1,2] such as FaceSwap, Face2Face, Deepfakes, Neural-Textures. Massive deepfake videos appeared in social net-works can fuel disinformation and reduce trust in media, which lead to pollution of a healthy online environment. Therefore, tracing the source of synthetic videos is of great value to protecting personal privacy and maintain copyright.

This research was funded by National Key Research Development Program of China (2019QY(Y)0206), and the National Natural Science Foundation of China NSFC (62072343, U1736211).
Corresponding author: Dengpan Ye (e-mail: yedp@whu.edu.cn).

Robust video hashing is a significant research subject in the fields of multimedia information and security. It has been widely applied in many applications [3,4,5,6] (video retrieval, video copying detection, etc.). Robustness and discrimination are two fundamental properties. Robustness means that the visually identical or similar videos have the same or similar hash value, and discrimination means the hash values of two videos with different visual content are very different. Specifically, for a given threshold T and a small positive number ε that is close to 0, the property of robustness and discrimination are described from the following formulation respectively:

$$P\{r[H(\mathbf{I}), H(\mathbf{I}_1)] < T \ \} \geq 1 - \varepsilon, \quad (1)$$

$$P\{r[H(\mathbf{I}), H(\mathbf{I}_2)] \geq T \ \} \geq 1 - \varepsilon. \quad (2)$$

Where $\mathbf{I}$ is the original visual multimedia including image and video, and $\mathbf{I}_1$ and $\mathbf{I}_2$ are visually similar and visually different versions of $\mathbf{I}$. $H(\mathbf{I})$ represents $\mathbf{I}$ corresponding multimedia hash. $P(\cdot)$ and $r(\cdot,\cdot)$ denote the probability of event occurrence and the metric function used for judging the similarity of two visual hashes, respectively.

1.1. Related work

For example, Yang et al. [7] designed a hash method by combines speeded up robust feature (SURF) and ordinal measure (OM) to form hash. The SURF-OM hashing can resist brightness adjustment and noise attack, but its discrimination is not desirable enough. Li and Monga [8] proposed a hash viewed video as a three-order tensor and then used low-rank tensor approximations (LRTA) to generate a video hash. The LRTA hashing can resist joint attack operations but weak to geometric attacks. Recently, Tang et al. [9] proposed a video hash with discrete cosine transform (DCT) and non-negative matrix factorization (NMF). This method uses dominant DCT coefficients to form feature matrices and then reduced with NMF into a compact hash. DCT-NMF hashing can resist noise attack and MPEG compression, but is fragile to video rotation. Khelifi and Bouridane [10] combine DCT and the discrete sine transform (DST) to design a novel hash. This approach computed the mean value of blocks in each frame to construct a three-

dimensional array and then used a signal calibration via DCT and DST to derive a hash. The DCT-DST hashing can be used for video content identification and authentication, but it is vulnerable to video rotation. Chen et al. [11] through low-rank and decomposition (LRD) extracted low-rank matrix from keyframes and compressed them with 2D-DWT to form a video hash. This method is resilient to some digital operations, but its discrimination should be improved.

In summary, most previous works do not reach desirable classification performance between robustness and discrimination. Focus on this issue, we exploit the shot segmentation model and local statistic descriptor to design a novel video hashing, which can make a good trade-off between robustness and discrimination. In addition, robust video hashing can map videos with similar content to similar hash code. Taking advantage of this feature, we explore the use of video hashing to trace the origin of fake videos that spread across social networks. Our contribution can be summarized as follows:

(1) We designed a shot segmentation model by combing robust image hashing and DWT. As robust image hashing can match the content of similar frames, which benefits to eliminates the redundant frames in input video and at the same time helps the proposed model to resist frame dropping and frame rate adjustment. In addition, exploiting LL-sub band of DWT coefficients to construct a segmentation model is also conducive to the proposed model anti-noise attack. Thus, hash generation from the shot segmentation model, good robustness of our hashing is achieved.

(2) We propose to use local linear embedding (LLE) and second-order statistic to design a local statistic descriptor. LLE can map points in high-dimensional space to low-dimensional embedding vectors, and second order statistics value reflects the local feature variation. Thus, the proposed local statistic descriptor has strong identification and local preserving ability for similar content. The use of the proposed local statistic descriptor can make our hashing good discrimination, and desirable fake material tracing capability in social networks.

(3) Experiments on popular databases such as Standard Benchmark, VOT2015, and FaceForesics++ verify the effectiveness of the proposed method. The results demonstrate that our hashing outperforms the existing well-known method in terms of robustness and discrimination, and can be used to track deepfake videos in social networks with high accuracy.

2. PROPOSED METHOD

To begin with, some simple pre-processing steps collectively reduce the length of extracted hash and increase its robustness to common video operations. (1) Input color video of three-channel is converted to grayscale **V**. (2) The pixels of all frames of **V** at the same position are orderly selected to form a tube, and then all pixel tubes are mapped to a moderate-length M by linear interpolation, together with smoothed by a low-pass filter with kernel size $1\times p$. (3) Each frame converted to $M\times M$ size using bi-cubic linear interpolation. Finally, we obtain a normalized video **V** with size $M\times M\times M$.

2.1. Shot Segmentation Model Construction

Robust image hashing is a novel technology of multimedia processing, and it maps an input image to a short and compact string. Due to the easy process of image hash, it has been widely used in many applications [12,13]. Define reference formula in detail [1,2].

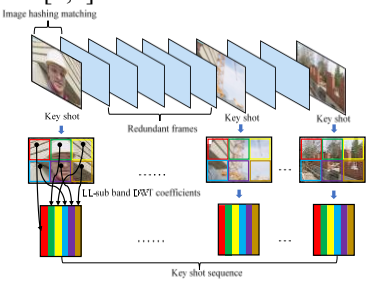


Fig.1. Diagram of the shot segmentation model.

Therefore, we can judge whether they are similar by the similarity of their image hash between two adjacent video frames. The first frame has defaulted as the key shot. Our proposed process can be expressed as follows:

$$\mathbf{V}_{match} = \{\mathbf{I}_1, H(\mathbf{I}_1,\mathbf{I}_2), H(\mathbf{I}_2,\mathbf{I}_3)\ldots, H(\mathbf{I}_{M-1},\mathbf{I}_M)\}, \quad (3)$$

$$H(\mathrm{X},\mathrm{Y}) = \begin{cases} Y, & if H(\mathrm{X},\mathrm{Y}) < T \\ null & otherwise \end{cases}. \quad (4)$$

Through above process, we obtain a video sequence and extract the first f shots to form $\mathbf{V}_{match}=\{\mathrm{I}_1, \mathrm{I}_2,\ldots \mathrm{I}_f\}$. Note that, some inter-frame attacks (randomly frame dropping and frame rate adjustment, etc.) will cause the video to lose some redundant visually similar frames. As robust image hashing perceives similar content, the key shot sequence selected through robust image hashing matching is not easily affected by above mention operations. To facilitate extraction of local features, each frame of $\mathbf{V}_{match}$ is divide into non-overlapping blocks with sized $b\times b$, and $B_{j,i}$ ($1\le i\le k, k=M/b$) is the i-th block index from left to right and top to bottom, and j-th is the frame index of $\mathbf{V}_{match}$. Then 2D-DWT is applied to $B_{j,i}$ and uses its LL-sub band coefficients to reconstructed DWT-based feature matrices. The reason is that LL-sub band DWT coefficients can resist minor modifications such as noise attacks, and together with preserve block content. The first n LL-sub band DWT coefficients of each block are picked up by zigzag scanning [14] algorithm to form an n-dimensional vector $\mathbf{d}_{j,i}=\{d_{j,i}\ (1),\ d_{j,i}\ (2),\ldots,d_{j,i}\ (n)\}(1\le j\le f, 1\le i\le k)$. Thus, stable DWT-based matrices **D** by arranging these vector $\mathbf{d}_{j,i}$.

$$\mathbf{D} = \begin{bmatrix} d_{j,1}(1) & d_{j,2}(1) & \dots & d_{j,k}(1) \\ d_{j,1}(2) & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ d_{j,1}(n) & \dots & \dots & d_{j,k}(n) \end{bmatrix}, j = 1,2,\dots,f. \quad (5)$$

Finally, DWT-based matrices are converted to normalized matrices to form a key shot sequence by next following rules:

$$v_{j,i}(z) = \frac{d_{j,i}(z) - u_{j,i}}{s_{j,i}}, (z = 1,2,\dots n, i = ,1,2,\dots k, j = ,1,2 \dots, f), \quad (6)$$

$$u_{j,i} = \frac{1}{n}\sum_{z=1}^{n} d_{j,i}(z), (i = 1,2,\dots k, j = 1,2,\dots f), \quad (7)$$

$$s_{j,i} = \sqrt{\frac{\sum_{z=1}^{n}[d_{j,i}(z) - u_{j,i}]^2}{n-1}}, (i = 1,2,\dots k, j = 1,2,\dots f). \quad (8)$$

$$\mathbf{V}_{key} = \begin{bmatrix} v_{j,1}(1) & v_{j,2}(1) & \dots & v_{j,k}(1) \\ v_{j,1}(2) & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ v_{j,1}(n) & \dots & \dots & v_{j,k}(n) \end{bmatrix}, j = 1,2,\dots,f. \quad (9)$$

2.2. Local statistical descriptor

This work, LLE [15] is considered to derive a local descriptor based on following considerations: (1) As LLE can map points in high-dimensional space to low-dimensional embedding vectors, we use LLE to constructed a local descriptor, which is a content-based feature and thus provides hash good discrimination. (2) In general, fake video is to modify the local area's original video. We combine LLE with second-order statistics to generate a local descriptor. The second-order statistics describe the data fluctuation, which strengthens the local preserving ability of LLE, and makes the hash coding of fake video material closer to the original. Therefore, it guarantees the excellent traceability of our hash.

Specifically, LLE is applied to each matrix of $\mathbf{V}_{key}$, and $\mathbf{Y}_j=[y^j_1, y^j_2,\dots, y^j_N]$, $(1\le j\le f, 1\le i\le N)$ are the matrices of low-dimensional embedding vectors, and each vector with d elements. Then we calculate the variance statistic of $y^j_i (1\le j\le f, 1\le i\le N)$ and concatenate them for a local statistical descriptor vector $\mathbf{x}$ with size $q=f\times N$ by the next equation:

$$\mathbf{x} = [\sigma_1(1), \dots, \sigma_1(N), \sigma_2(1) \dots, \sigma_2(N), \sigma_j(1), \dots, \sigma_j(N)]. \quad (10)$$

$$\sigma_j(i) = \frac{1}{d}\sum_{l=1}^{d} \left| y_i^j(l) - u_j(i) \right|^2, (j = 1,2,\dots f, i = 1,2,\dots N). \quad (11)$$

$$u_j(i) = \frac{1}{d}\sum_{l=1}^{d} y_i^j(l), (j = 1,2,\dots f, i = 1,2,\dots N) \quad (12)$$

where $y^j_i(l)$ is the l-th element of y^j_i. Finally, we calculate the mean value of $\mathbf{x}$ and quantify it to form a compact hash.

$$h_i = \begin{cases} 1, & if\ x(i) > \mathbf{x}_{mean} \\ 0 & otherwise \end{cases}, (i = 1,2,\dots,q), \quad (13)$$

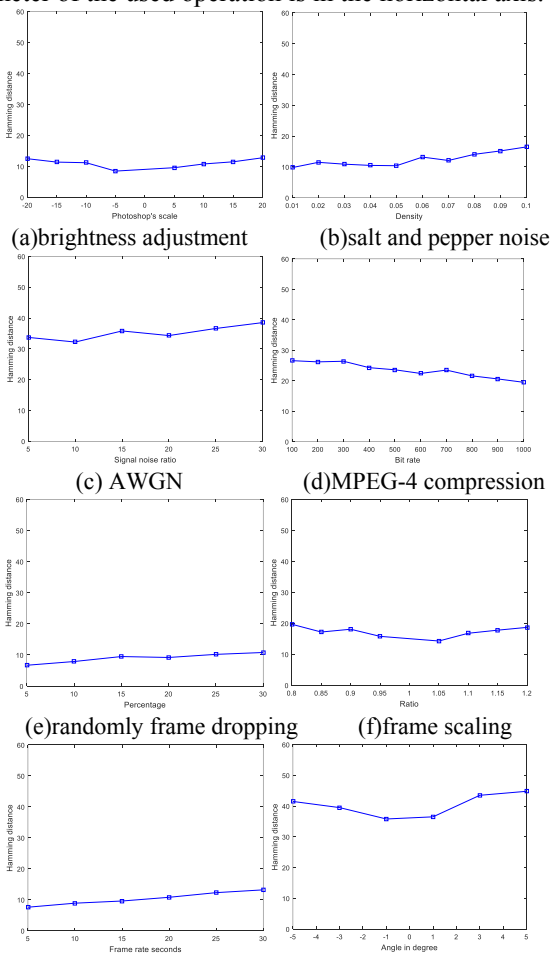
where $x(i)$ and $\mathbf{x}_{mean}$ are the element and mean of $\mathbf{x}$.

3. EXPERIMENTAL

The used parameters of our hashing are as follows: the normalized video is 256×256×256; Gaussian low-pass filter with kernel size is 1×20; key shot f is 4 and non-overlapping block size is 8×8; the first 64 LL-sub band DWT coefficients in the zigzag are adopted; the dimension of embedding vector and the number of nearest neighbors are 40 and 15. Therefore, proposed hash length is 256bits. Namely, M=256; p=20; f=4; b=8; n=64; N=40; d=15; q=256. Note that, the robust image hashing algorithm used TD hashing [16], and parameter setting according to their recommended values.

3.1. Robustness and Discrimination

Two popular video databases (standard benchmark[11], VOT 2015[17]) are conducted to robustness and discrimination experiments. All 20 videos from standard benchmark and all 60 videos from VOT2015 form an original video dataset. For the robustness test, we extract 80 videos and their visually similar videos hashes and calculate the Hamming distance between them. The mean value of Hamming distances under the detailed parameters of digital operations shows in Fig.2, where the vertical axis is the Hamming distance and the parameter of the used operation is in the horizontal axis.



(a)brightness adjustment (b)salt and pepper noise

(c) AWGN (d)MPEG-4 compression

(e)randomly frame dropping (f)frame scaling

(g)frame rate adjustment (h)video rotation

Fig.2 Hamming distances under digital operations.

For the discrimination test, we extracted hashes for each original video, calculated Hamming distances between its hash and the other 79 hashes. Thus, the pair of 80×79/2=3160 Hamming distances are obtained. Fig.3. showed the minimum Hamming distance between two different videos is 71, and the maximum is 177. The correct detection rate of similar videos is equivalent to robustness and, the false identification rate of different videos is comparable to discrimination under different thresholds listed in Table1.

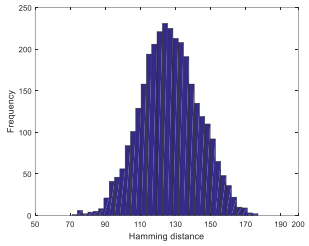


Fig.3. Hash distance distribution of different videos.

Table1: Percentage of correct detection under different thresholds.

T	Correct detection rate of similar videos	False identification rate of different videos
70	99.79%	0
75	99.82%	5.20×10^{-6}
80	99.84%	6.22×10^{-6}
85	99.89%	3.23×10^{-5}
90	99.95%	5.54×10^{-5}

3.2 Performance Comparisons

In this section, we compared some well-known robust video hashing algorithms, including SURF-OM hashing[7], DCT-DST hashing[10] and DCT-NMF hashing[9]. The test videos are the same as the section 3.1. All videos resized to 256×256×256, other parameters and hash similarity setting adopted as their reported. The performance was evaluated using the Receiver Operating Characteristics (ROC) curves [8]. Note that, video hashing of ROC much more nearby the left-top indicates a more desirable balance between robustness and discrimination. To conduct quantitative analysis, the area under the ROC curve (AUC) is taken. SURF-OM hashing is 0.9803; DCT-DST hashing is 0.9967; DCT-NMF hashing is 0.9978; Proposed hashing is 0.9989.

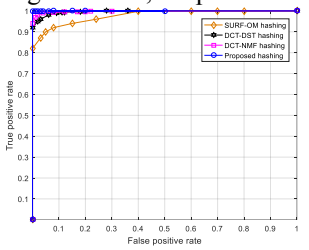


Fig.4. ROC curve among comparison algorithms.

3.3 Tracking deepfake videos

The block diagram of the tracking deep video application through perceptual hashing is shown in Fig.5. A suspected deepfake video downloaded from social networks. Then generate the corresponding video hash through the video hash algorithm and make a request, according to the match between the hash value of the request video and the hash value in the open database. Finds video hashes with similar hash values and returns the corresponding video metadata. Finally, the user can trace the source of the suspected video from the social network based on the video's metadata. The experiment used a well-known deepfake video database called FaceForesics++[18].

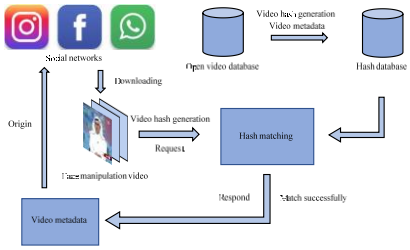


Fig.5. The block diagram of the tracking deepfake video system.

In this experiment, 1000 requests are considered to obtain the average. The curve closest to the upper righthand corner of the graph indicates the best performance. The average recall/precision curves [7] for origin system using the perceptual video hashes are shown in Fig.6.

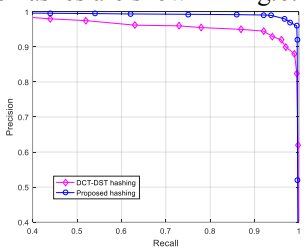


Fig.6. The average recall/precision curves among comparison hashing.

In addition, the accuracy of different typical operations is separately recorded and listed in Table 2. It can be intuitively seen from Fig.6. The precision of 4 manipulation operations when recall amount is close to 1 indicates that our algorithm can be used for tracking deepfake videos with high accuracy.

Table2: Precision of different typical manipulations

	FaceSwap	Face2Face	DeepFake	N-Textures
[10]	0.9772	0.9740	0.9245	0.9138
Our	0.9922	0.9918	0.9854	0.9555

4. CONCLUSION

Different from previous work, this study proposes to extend robust video hashing technology to a new traceability scenario. Extensive experiments have verified that our hash algorithm is robust to commonly-used digital operations and maintains desirable discrimination capability, and our hash can be used to tracking deepfake videos in social networks.

5. REFERENCES

[1] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Nießner, Face2face: Real-time face capture and reenactment of RGB videos. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2387-2395, 2016.

[2] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3207-3216, 2020.

[3] X. Nie, W. Jing, C. Cui, C. J. Zhang, L. Zhu and Y. Yin, Joint Multi-View Hashing for Large-Scale Near-Duplicate Video Retrieval, *IEEE Transactions on Knowledge and Data Engineering*, vol.32, no.10, pp. 1951-1965, Oct. 2020.

[4] D. Ye, Z. Wei, X. Ding, Robert H. Deng, Scalable Content Authentication in H.264/SVC Video Using Perceptual Hashing based on Dempster-Shafer theory. *International Journal of Computational Intelligence Systems,* vol.5, no.5, pp.777-787, 2012.

[5] L. Shen, R. Hong, H. Zhang, et al. Video Retrieval with Similarity-Preserving Deep Temporal Hashing. *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol.15, no.4, pp.1-16, 2019.

[6] Y. Wang, X. Nie, Y. Shi, X. Zhou and Y. Yin, Attention-Based Video Hashing for Large-Scale Video Retrieval, in *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 3, pp. 491-502, 2021.

[7] G. Yang, N. Chen and Q. Jiang, A robust hashing algorithm based on SURF for video copy detection, *Computers & Security*, vol.31, no.1, pp.33-39, 2012.

[8] M. Li and V. Monga, Robust video hashing via m ultilinear on subspace projections, *IEEE Transaction s on Image Processing*, vol.21, no.10, pp.4397-440 9, 2012.

[9] Z. Tang, L. Chen, H. Yao, et al. Video Hashing with DCT and NMF. *The Computer Journal*, vol.63, no.7, pp 1017-1030, 2020.

[10] F. Khelifi, A. Bouridane. Perceptual video hashing for content identification and authentication. *IEEE Transactions on Circuits and Systems for Video Technology*, vol.29, no.1, pp.50-67, 2017.

[11] L. Chen, D. Ye, S. Jiang. High Accuracy Perceptual Video Hashing via Low-Rank Decomposition and DWT, *In Proceedings of the International Conference on Multimedia Modeling*. Springer, Cham, pp.802-812, 2020.

[12] C. Qin, M. Sun, C.C. Chang, Perceptual hashing for color images based on hybrid extraction of structural features. *Signal processing*, vol.142, pp.194-205, 2018.

[13] Z. Huang, S. Liu, Robustness and discrimination oriented hashing combining texture and invariant vector distance. *In Proceedings of the 26th ACM international conference on Multimedia*, pp. 1389-1397, 2018.

[14] Z. Tang, H. Lao, X. Zhang, K. Liu, Robust image hashing via DCT and LLE. *Computers & Security*, vol.62, pp.133-148, 2016.

[15] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding. *Science*, vol. 290 no. 5500, pp. 2323-2326, 2000.

[16] Z. Tang, L. Chen, X. Zhang, S. Zhang, Robust Image Hashing with Tensor Decomposition, *IEEE Transactions on Knowledge and Data Engineering*, vol.31, no.3, pp.549-560, 2019.

[17] VOT visual object tracking, https://www.votchallenge.net/vot2015/dataset.html, accessed October 18, 2018.

[18] Rossler, D. Cozzolino, L. Verdoliva, C. Ries, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images. *In Proceedings of the IEEE International Conference on Computer Vision*, pp. 1-11, 2019.