Fake-image detection with Robust Hashing

*

1st Miki Tanaka
Tokyo Metropoliltan University
Tokyo, Japan
tanaka-miki@ed.tmu.ac.jp

2nd Hitoshi Kiya
Tokyo Metropoliltan University
Tokyo, Japan
kiya@tmu.ac.jp

2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech) | 978-1-6654-1875-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/LIFETECH52111.2021.9391842

***Abstract*—In this paper, we investigate whether robust hashing has a possibility to robustly detect fake-images even when multiple manipulation techniques such as JPEG compression are applied to images for the first time. In an experiment, the proposed fake detection with robust hashing is demonstrated to outperform state-of-the-art one under the use of various datasets including fake images generated with GANs.**

***Index Terms*—fake images, GAN**

I. Introduction

Recent rapid advances in image manipulation tools and deep image synthesis techniques, such as Generative Adversarial Networks (GANs) have easily generated fake images. In addition, with the spread of SNS (social networking services), the existence of fake images has become a major threat to the credibility of the international community. Accordingly, detecting manipulated images has become an urgent issue [1].

Most forgery detection methods assume that images are generated by using a specific manipulation technique, and the methods aim to detect unique features caused by the manipulation technique such as checkerboard artifacts [2]–[5]. Actually tampered images are usually uploaded to SNS and image sharing services. SNS providers are known to process the uploaded images by resizing or compressing them into JPEG format [6]–[9]. Such manipulation may damage or lose the unique features of fake images. However, the influence of manipulations on images has not been discussed sufficiently when a number of manipulation techniques such as JPEG compression are applied at the same time. In this paper, we investigate the possibility that there is a method with robust hashing that has been proposed for image retrieval, and the proposed method with robust hashing is demonstrated to have a high fake-detection accuracy, even when multiple manipulation techniques are carried out.

II. Related work

A. Fake-image generation

Fake images are manually generated by using image editing tools such as Photoshop. Splicing, copy-move, and deletion are also carried out under the use of such a tool. Similarly, resizing, rotating, blurring, and changing the color of an image can be manually carried out.

In addition, recent rapid advances in deep image synthesis techniques such as GANs have automatically generated fake images. CycleGAN [10] and StarGAN [11] are typical image synthesis techniques with GANs. CycleGAN is a GAN that performs one-to-one transformations, e.g. changing apples to oranges, while StarGAN is a GAN that performs many-to-many transformations, such as changing a person's facial expression or hair color (see Figs.1 and 3). Furthermore, fake videos created using deep learning are called Deepfake, and various tampering methods have emerged, such as those using autoencoders, Face2Face [12], FaceSwap [13], and so on.

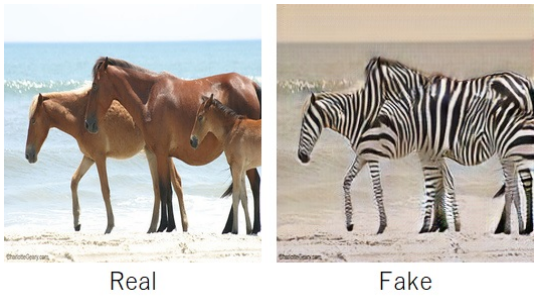

Fig. 1. Example Fake-images with CycleGAN

Real-world fake images may include the influence of a number of manipulation techniques such as image compression, resizing, copy-move at the same time, even if fake-images are generated by using GANs. Therefore, we have to consider such conditions for detecting real-world fake images.

B. Fake detection methods

Image tampering has a longer history than that of deep learning. Fragile watermarking [14], detection of double JPEG compression with a statistical method [15] [16], and use of PRNU (photo-response non-uniformity) patterns of each camera [17] [18] have been proposed to detect such tampers. However, most of them do not suppose to detect fake images generated with GANs. Moreover, they cannot detect the difference between fake images and just manipulated ones such as resized images, which are not fake images in general.

With the development of deep learning, fake detection methods with deep leaning have been studied so far. The methods with deep learning do not employ a reference image or the features of a reference image to detect tamper ones. The methods also assume that images are generated by using

978-1-6654-1875-1/21/$31.00 ©2021 IEEE

a specific manipulation technique to detect unique features caused by the manipulation technique.

There are several detection methods with deep learning for detecting fake images generated with an image editing tool as Photoshop. Some of them focus on detecting the boundary between tampered regions and an original image [19] [20] [21]. Besides, a detection method [22] enables us to train a model without tamper images.

Most detection methods with deep learning have been proposed to detect fake images generated by using GANs. An image classifier trained only with ProGAN was shown to be effective in detecting images generated by other GAN models [23]. Various studies have focused on detecting checkerboard artifacts caused in both of two processes: forward propagation of upsampling layers and backpropagation of convolutional layers [24]. In this work, the spectrum of images is used as an input image in order to capture the checkerboard artifacts.

To detect fake videos called DeepFake, a number of detection methods have been investigated so far. Some methods attempt to detect failures in the generation of fake videos, in terms of poorly generated eyes and teeth [25], the frequency of blinking as a feature [26], and the correctness of facial landmarks [27] or head posture [28]. However, all of these methods have been pointed out to have problems in the robustness against the difference between training datasets and test data [1]. In addition, the conventional methods have not considered the robustness against the combination of various manipulations such as the combination of resizing and DeepFake.

III. Proposed Method with Robust Hashing

A. Overview

Figure2 shows an overview of the proposed method. In the framework, robust hash value is computed from easy reference image by using a robust hash method, and stored in a database. Similar to reference images, a robust hash value is computed from a query one by using the same hash method. The hash value of the query is compared with those stored the database. Finally, the query image is judged whether it is real or fake in accordance with the distance between two hash values.

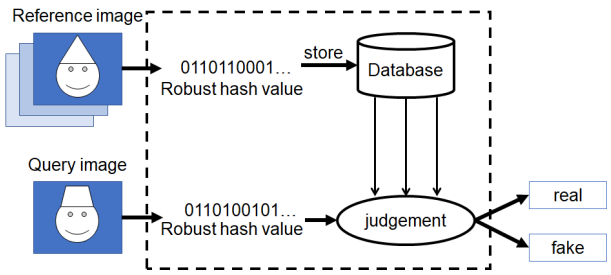

Fig. 2. Overview of proposed method

B. Fake detection with Robust Hashing

Various robust hashing methods have been proposed to retrieval similar images to a query one [29], [30]. In this paper, we apply the robust hashing method proposed by Li et al [29] for applying it to fake-image detection. This robust hashing enables us to robustly retrieve images, and has the following properties.

- Resizing images to 128×128 pixels prior to feature extraction.
- Performing 5×5-Gaussian low-pass filtering with a standard deviation of 1.
- Using rich features extracted from spatial and chromatic characteristics.
- Outputting a bit string with a length of 120 bits as a hash value.

In the method, the similarity is evaluated in accordance with the hamming distance between the hash string of a query image and that of each image in a database.

Let vectors $\boldsymbol{u} = \{u_1, u_2, \ldots, u_n\}$ and $\boldsymbol{q} = \{q_1, q_2, \ldots, q_n\}$, $u_i,\ q_i \in \{0, 1\}$ be the hash strings of reference image U and query image Q, respectively. The hamming distance $d_H(\boldsymbol{u}, \boldsymbol{q})$ between U and Q is given by:

$$d_H(\boldsymbol{u}, \boldsymbol{q}) \triangleq \sum_{i=1}^{n} \delta(u_i, q_i) \tag{1}$$

where

$$\delta(u_i, q_i) = \begin{cases} 0, & u_i = q_i \\ 1, & u_i \neq q_i \end{cases} . \tag{2}$$

To apply this similarity to fake-image detection, we introduce a threshold d as follows.

$$\begin{cases} Q \in \mathbb{U}', \min\limits_{u \neq q, u \in \mathbb{U}} (d_H(\boldsymbol{u}, \boldsymbol{q})) < d \\ Q \notin \mathbb{U}', \min\limits_{u \neq q, u \in \mathbb{U}} (d_H(\boldsymbol{u}, \boldsymbol{q})) \geq d \end{cases} \tag{3}$$

where $\mathbb{U}$ is a set of reference images and $\mathbb{U}'$ is the an of images generated with image manipulations from $\mathbb{U}$, which does not include fake images. According to eq. (3), Q is judged whether it is a fake image or not.

IV. Experiment Results

The proposed fake-image detection with robust hashing was experimentally evaluated in terms of accuracy and robustness against image manipulations.

A. Experiment setup

In the experiment, four fake-image datasets: Image Manipulation Dataset [31], UADFV [26], CycleGAN [10], and StarGAN [11] were used. The details of datasets are shown in Table I (see Figs. 1 and 3). The datasets consist of pairs of a fake-image and the original one. JPEG compression with a quantization parameter of $Q_J = 80$ was applied to all query images. $d = 3$ was selected as threshold d in accordance with the EER (Equal error rate) performance.

As one of the state-of-the-art fake detection methods, Wang's method [23] was compared with the proposed one. Wang's method was proposed for detecting images generated by using CNNs including various GAN models, where a classifier is trained by using ProGAN.

978-1-6654-1875-1/21/$31.00 ©2021 IEEE

TABLE I
DATASETS

dataset	Fake-image generation	real	fake
		No. of images	
Image Manipulation Dataset [31]	copy-move	48	48
UADFV [26]	face swap	49	49
CycleGAN [10]	GAN	1320	1320
StarGAN [11]	GAN	1999	1999

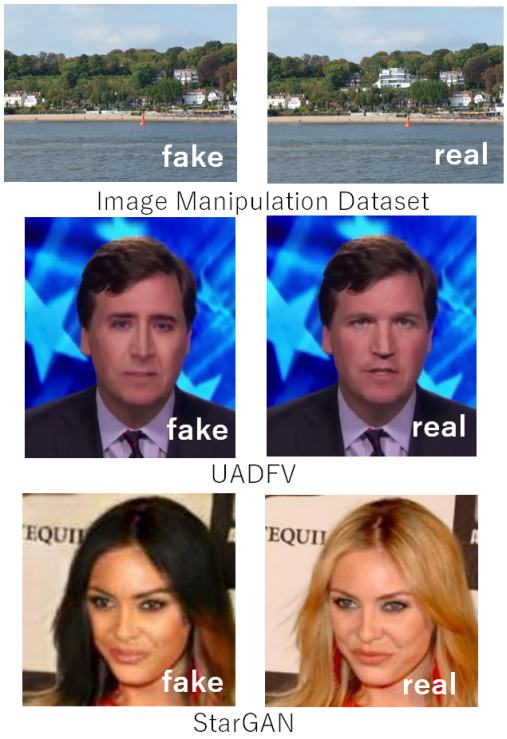

Fig. 3. Example of datasets

The performance of fake-image detection was evaluated by using AP (Average Precision) and Accuracy (fake), given by,

$$Accuracy\ (fake) = \frac{N_{tn}}{N_{Qf}} \tag{4}$$

where N_{Qf} is the number of fake query images, and N_{tn} is the number of fake query ones that are correctly judged as fake images.

B. Results without additional manipulation

Table II shows experimental results under the use of the two detection methods. From the table, it is shown that the proposed method had a higher performance than Wang's method in terms of both AP and Acc (fake). In addition, the performance of Wang's method heavily decreased when using the image manipulation and UADFV datasets. The reason is that Wang's method focuses on detecting fake images generated by using CNNs. The image manipulation dataset does not consist of images generated with GANs. In addition, although UADFV consists of images generated by using DeepFake, they have the influence of video compression.

TABLE II
COMPARISON WITH WANG'S METHOD

Dataset	Wang's method [23]		proposed	
	AP	Acc (fake)	AP	Acc (fake)
Image Manipulation Dataset	0.5185	0.0000	0.9760	0.8750
UADFV	0.5707	0.0000	0.8801	0.7083
CycleGAN	0.9768	0.5939	1.0000	1.0000
StarGAN	0.9594	0.5918	1.0000	1.0000

C. Results with additional manipulation

JPEG compression with $Q_J = 70$, resizing with a scale factor of 0.5, copy-move or splicing was applied to query images. Therefore, when query images were fake ones, the fake query ones included the effects of two manipulations at the same time.

Table III shows experimental results under the additional manipulation, where 50 fake images generated by using CycleGAN, in which horses were converted to zebras, were used (see Fig.1). The proposed method was confirmed to still maintain a high accuracy even under the additional manipulation. In contrast, Wang's method suffered from the influence of the addition manipulation. In particular, for splicing and resizing, Wang's method was affected by these operations. That is why the method assume that fake images are generated by using CNNs, to detect unique features caused by using CNNs. However, splicing and resizing don't depend on CNNs, although CycleGAN includes CNNs.

TABLE III
COMPARISON WITH WANG'S METHOD UNDER ADDITIONAL MANIPULATION (DATASET: CYCLEGAN)

additional manipulation	Wang's method [23]		proposed	
	AP	Acc (fake)	AP	Acc (fake)
None	0.9833	0.6200	0.9941	1.0000
JPEG($Q_J = 70$)	0.9670	0.6000	0.9922	0.9800
resize (0.5)	0.8264	0.2400	0.9793	1.0000
copy-move	0.9781	0.6000	1.0000	1.0000
splicing	0.9666	0.4800	0.9992	1.0000

V. CONCLUSION

In this paper, we proposed a novel fake-image detection method with robust hashing for the first time. Although various robust hashing methods have been proposed to retrieve similar images to a query one so far, a robust hashing method proposed by Li et al was applied to various datasets including fake images generated with GANs. In the experiment, the proposed method was demonstrated not only to outperform a state-of-the-art but also to be robust against the combination of image manipulations.

978-1-6654-1875-1/21/$31.00 ©2021 IEEE

REFERENCES

[1] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[2] Y. Sugawara, S. Shiota, and H. Kiya, "Super-resolution using convolutional neural networks without any checkerboard artifacts," in *Proc. of IEEE International Conference on Image Processing*, 2018, pp. 66–70.

[3] Y. Sugawara, S. Shiota, and H. Kiya, "Checkerboard artifacts free convolutional neural networks," *APSIPA Transactions on Signal and Information Processing*, vol. 8, p. e9, 2019.

[4] Y. Kinoshita and H. Kiya, "Fixed smooth convolutional layer for avoiding checkerboard artifacts in cnns," in *Proc. in IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 3712–3716.

[5] T. Osakabe, M. Tanaka, Y. Kinoshita, and H. Kiya, "Cyclegan without checkerboard artifacts for counter-forensics of fake-image detection," *arXive preprint arXive:2012.00287*, 2020. [Online]. Available: https://arxiv.org/abs/2012.00287

[6] T. Chuman, K. Iida, W. Sirichotedumrong, and H. Kiya, "Image manipulation specifications on social networking services for encryption-then-compression systems," *IEICE Transactions on Information and Systems*, vol. E102.D, no. 1, pp. 11–18, 2019.

[7] T. Chuman, K. Kurihara, and H. Kiya, "Security evaluation for block scrambling-based etc systems against extended jigsaw puzzle solver attacks," in *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 229–234.

[8] W. Sirichotedumrong and H. Kiya, "Grayscale-based block scrambling image encryption using ycbcr color space for encryption-then-compression systems," *APSIPA Transactions on Signal and Information Processing*, vol. 8, p. e7, 2019.

[9] T. Chuman, W. Sirichotedumrong, and H. Kiya, "Encryption-then-compression systems using grayscale-based image encryption for jpeg images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 6, pp. 1515–1525, 2019.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. of IEEE International Conference on Computer Vision*, Oct 2017.

[11] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[12] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[13] Y. Nirkin, I. Masi, A. Tran Tuan, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," in *Proc. of IEEE International Conference on Automatic Face Gesture Recognition*, 2018, pp. 98–105.

[14] A. T. S. Ho, X. Zhu, J. Shen, and P. Marziliano, "Fragile watermarking based on encoding of the zeroes of the z-transform," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 3, pp. 567–569, 2008.

[15] G. Zhenzhen, N. Shaozhang, and H. Hongli, "Tamper detection method for clipped double jpeg compression image," in *Proc. of International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2015, pp. 185–188.

[16] T. Bianchi and A. Piva, "Detection of nonaligned double jpeg compression based on integer periodicity maps," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 842–848, 2012.

[17] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, "Determining image origin and integrity using sensor noise," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 74–90, 2008.

[18] G. Chierchia, G. Poggi, C. Sansone, and L. Verdoliva, "A bayesian-mrf approach for prnu-based image forgery detection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 4, pp. 554–567, 2014.

[19] Y. Rao and J. Ni, "A deep learning approach to detection of splicing and copy-move forgeries in images," in *Pros. of IEEE International Workshop on Information Forensics and Security*, 2016, pp. 1–6.

[20] J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj, and B. S. Manjunath, "Exploiting spatial structure for localizing manipulated image regions," in *Proc. of IEEE International Conference on Computer Vision*, Oct 2017.

[21] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Pros. of learning rich features for image manipulation detection," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[22] M. Huh, A. Liu, A. Owens, and A. A. Efros, "Pros. of fighting fake news: Image splice detection via learned self-consistency," in *Proc. of European Conference on Computer Vision*, September 2018.

[23] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

[24] X. Zhang, S. Karaman, and S. Chang, "Detecting and simulating artifacts in gan fake images," in *Proc. of IEEE International Workshop on Information Forensics and Security*, 2019, pp. 1–6.

[25] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proc. of IEEE Winter Applications of Computer Vision Workshops*, 2019, pp. 83–92.

[26] Y. Li, M. Chang, and S. Lyu, "In ictu oculi: Exposing ai created fake videos by detecting eye blinking," in *Proc. of IEEE International Workshop on Information Forensics and Security*, 2018, pp. 1–7.

[27] X. Yang, Y. Li, H. Qi, and S. Lyu, "Exposing gan-synthesized faces using landmark locations," in *Proc. of ACM Workshop on Information Hiding and Multimedia Security*, 2019, p. 113–118.

[28] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 8261–8265.

[29] Y. N. Li, P. Wang, and Y. T. Su, "Robust image hashing based on selective quaternion invariance," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2396–2400, 2015.

[30] K. Iida and H. Kiya, "Robust image identification with dc coefficients for double-compressed jpeg images," *IEICE Transactions on Information and Systems*, vol. E102.D, no. 1, pp. 2–10, 2019.

[31] "Image manipulation dataset," https://www5.cs.fau.de/research/data/image-manipulation/.

978-1-6654-1875-1/21/$31.00 ©2021 IEEE