Open Access

Open Journal of Internet of Things (OJIOT)
Volume 7, Issue 1, 2021
http://www.ronpub.com/ojiot
ISSN 2364-7108

Video Source Forensics for IoT Devices Based on Convolutional Neural Networks

Dongzhu Rong[A], Yan Wang[A], Qindong Sun[A,B,]

[A]Shaanxi Key Laboratory of Network Computing and Security, Xi'an University of Technology, Xi'an, China,
[B]School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an, China, qdongsun@xjtu.edu.cn

Abstract

With the wide application of Internet of things devices and the rapid development of multimedia technology, digital video has become one of the important information dissemination carriers among Internet of things devices, and it has been widely used in many fields such as news media, digital forensics and so on. However, the current video editing technology is constantly developing and improving, which seriously threatens the integrity and authenticity of digital video. Therefore, the research on digital video forensics has a great significance. In this paper, a new video source passive forensics algorithm based on Convolutional Neural Networks(CNN) is proposed. CNN is used to classify the maximum information block of specified size in video I frame, and then the classification results are fused to determine the camera to which the video belongs. Experimental results show that the recognition algorithm proposed in this paper has a better performance than other methods in trems of accuracy and ROC curve. And our method still can have a good recognition effect even if a small number of I frames are used for recognition.

Type of Paper and Keywords

Short Communication: *Internet of Things Device Forensics, Digital Video Forensics, Video Feature Extraction, Convolutional Neural Networks, Digital Video Source Recognition*

1 Introduction

With the rapid development of Internet of Things in wearable devices, smart sensors and household appliances are affecting all aspects of our lives. The formal use of the fifth generation mobile network (5G) has brought several services to the Internet of Things, which are wide coverage, large connections and low latency network access. Facing the heterogeneous network access technology, the mobile Internet of Things data is characterized by mass, heterogeneity and dynamics. By 2022, some relevant researches show that there are about 1234 billion intelligent devices will be generated every year in the direction of Internet of Things network security[14]. The wide application of Internet of Things devices also makes the security of Internet of Things devices attract much attention. Especially in the research of digital forensics, the multimedia information of Internet of Things devices have a very important analytical significance.

Nowadays, multimedia technology is developing rapidly, the application range of digital video is becoming more and more extensive. For example, in order to obtain more evidence to identify the suspect, the law enforcement officers can match the obtained video evidence with the multiple suspects' smartphones. Specifically, you can extract the video frame of the video evidence, use the high-pass filter to obtain the high-

This paper is accepted at the *International Workshop on Very Large Internet of Things (VLIoT 2021)* in conjunction with the VLDB 2021 conference in Copenhagen, Denmark. The proceedings of VLIoT@VLDB 2021 are published in the Open Journal of Internet of Things (OJIOT) as special issue.

frequency noise frame, and input the noise frame into the convolutional neural network designed in this paper, which can accurately determine the phone of the suspect that the frame belongs to, and then obtain more evidence helps solve the case.

Based on the information of digital video, we directly use the digital video source forensics to identify and without any preprocessing of digital video in advance. Traditional video source forensics mainly compares the imaging traces of video frames (such as different CFA interpolation algorithms), sensor defects (such as random noise caused by non-uniformity of light response of photosensitive materials) and video post-processing (such as video frame compression) to determine the video source. Among them, the most widely used one is the sensor mode noise based on nonuniformity of light response proposed by Fridrich[5] which is caused by the defect of sensor manufacturing process and the different sensitivity of different COMS/CCD chips to light. Because this defect is unique for each image, the sensor mode noise can be used for video source identification.

The video source identification algorithm is more challenging than the general digital image source identification algorithm[8], which is mainly reflected in the following two aspects: on the one hand, because digital video is more compressible than digital image, it is mainly reflected in intra-frame coding and inter-frame prediction, which will cause more loss of noise information. The noise information used for source identification is reduced, and then reduce the recognition accuracy. On the second hand, most of the current video source identification methods are used to identify the central block of video frames, these methods do not consider the impact of the loss of pixels in the central area during video compression, which will also lead to a decrease in recognition accuracy.In this paper, an IOT video source recognition algorithm based on CNN is proposed. Our method makes full use of the high-frequency noise information in the largest information block in the video I frame, and inputs it into the CNN designed in this paper for classification. Finally, the source recognition results of the video are obtained by fusing the recognition results of each frame.

For the traditional recognition algorithm based on PRNU or DL, they usually clip a fixed position in the video image, or resize the image to a fixed size for recognition. Such method cannot dynamically extract the position with the largest amount of information from the frame.The proposed algorithm solves this problem to a certain extent by dynamically selecting the largest information block.

In this paper, our contributions are as follows

- I frame information of video is made full use to identify video source, which improve the accuracy of video source identification.
- A new CNN structure is designed for video source recognition.

2 Related Work

The research of image source recognition technology is earlier than video source recognition, and video is composed of a series of frames. Therefore, image source recognition technology can be applied to video source recognition. In this section, we will summarize the video source identification algorithms separately from PRNU-based source identification algorithm and deep learning-based source identification algorithm.

Video source identification is mainly due to some inherent defects in the video collection stage, which leaves a unique fingerprint on the video content. Currently, PRNU (Light Response Nonuniformity) noise is the most widely used, which is used as fingerprint to identify the specific camera to which the image belongs. The camera source recognition based on PRNU is mainly caused by the defects in the sensor manufacturing process and the different sensitivity of different pixels to light[5]. The PRNU noise fingerprint can be expressed as:

$$K = \frac{\sum_{i=1}^{N} V_i F_i}{\sum_{i=1}^{N} F_i^2} \tag{1}$$

Here, V_i is the i-th residual noise and F_i is the i-th image. respectively. N is the total number of the test images.

Many researchers recognized the video source based on PRNU noise fingerprint, and achieved good results. For example, Chen et al.[4] proposed to estimate PRNU fingerprint from a single video frame for video source identification. Wei-Hong Chuang et al.[17] proposed a video source recognition algorithm based on PRNU, and they analyzed the influence of compression intensity on the accuracy of PRNU. Samet Taspinar[19] proposed a stable video source recognition algorithm based on PRNU. Firstly, the video stability is determined by judging whether the video frame is translated or rotated, then the video fingerprint is extracted, and finally the video source is identified by stabilizing the correlation between the video fingerprint and the camera fingerprint. Sai-Chung Law et al.[12] used PRNU to verify the video source in the video surveillance system, and they also analyzed the effects of video resolution, frame type and application scene on recognition accuracy and reliability.

Massimo Iuliani et al.[10] fused images and videos to identify the source based on PRNU, and they identified the source of digital video by using PRNU noise

fingerprint generated by still images taken by the same device and applied it to social network forensics. Enes Altinisik et al.[1] eliminated the filtering process applied to H.264 decoder to reduce blocking effect and improve the performance of fingerprint video source recognition based on PRNU noise. Irene Amerini et al.[2] in analyzed the source identification of videos shared on the internet, and identified the sources of videos shared on social networks by using PRNU noise to generate fingerprints. Kouokam et al.[11] put forward a PRNU noise fingerprint estimation method by using video frames, which is used to determine whether two videos from unknown sources come from the same device. Shaxun Chen et al.[6] used PRNU fingerprint to identify video source transmitted through wireless network, and it has a good recognition effect.

With the development of deep learning and the increase of video data sets in multimedia forensics applications, some researchers used deep learning to identify the video sources. For example, B. Hosler proposed a video source identification system based on deep learning, which uses CNN to identify video sources[7]. Due to the limitation of video data sets, the method of deep learning for video source recognition is still at the initial stage.

This paper proposed a video source recognition algorithm based on CNN. The I-frame information of video is made full use in our method. We extract the high-frequency noise in the frame through high-pass filter[16], and eliminate the scene information, then input it into CNN for classification. Finally, we fuse the classification results of each frame, in which the camera with the highest score in the classification results is the test video source camera.

3 APPROACH

This section introduces how our method to extracts the features of input video for video recognition.Figure 1 illustrates this process. The specific steps are as follows:

- All I frames are extracted from the test video, and the maximum information block with specified size is cut out from the I frames. Then, a high-pass filter is used on the maximum information block to obtain the sequence of high-frequency noise map of the video;
- Sequentially inputting the high-frequency noise image sequence into the trained CNN model to obtain the classification results of each image;
- Finally, we fuse the obtained classification result sequences, and the obtained result is the video source identification of the video source camera.

The specific process is shown in the following figure.

3.1 Maximum information block extraction

According to the H.264 video coding standard, each frame of video is divided into several macro-blocks, and the noise information in the frame is seriously lost after being quantified by the encoder. In this paper, referring to the method proposed in [11], the video frame is divided into 8×8 blocks firstly, and then Discrete Cosine Transform (DCT) is used in each block. If the coefficient after DCT transformation only contains Direct Current(DC) component, this position will be marked as 0, otherwise it will be marked as 1. So that a labeled image A with the same size as the original image is obtained. Finally, we search a sub-image B which has a specified size and containing the most number of 1 by using sliding window from the image A .The time complexity of searching with sliding window is $o(n^4)$. Here, we use the integral graph to optimize our method. For a gray-scale image, the value of any point (x, y) in the integral image refers to the sum of gray-scale values of all points in the rectangular area from the upper left corner of the image to this point. By this way, the calculation of the area in the sliding window can be optimized as $o(1)$ when it slides every time, and the time complexity of the whole search can be reduced to $o(n^2)$.

3.2 Network Structure

This network structure combines the idea of [3] network construction and residual learning, and based on this, we use a large convolution kernel instead of pooling. Figure 2 shows the network structure proposed in this paper. The construction ideas mainly include the following aspects.

- Batch Normalization (BN) module is used in the network. The main reason for using BN module is that when min-batch is used to train neural networks in this paper, different batch data distributions are different, so the network must learn to adapt to different distributions in each iteration, which will greatly reduce the training speed of the network. Using BN method to standardize data processing can speed up the training process and improve the denoising performance.
- The network structure is designed in the form of bottleneck residual blocks connected in series. The main reason is that in the network training process based on random gradient descent, the multi-layer back propagation of error signals can easily

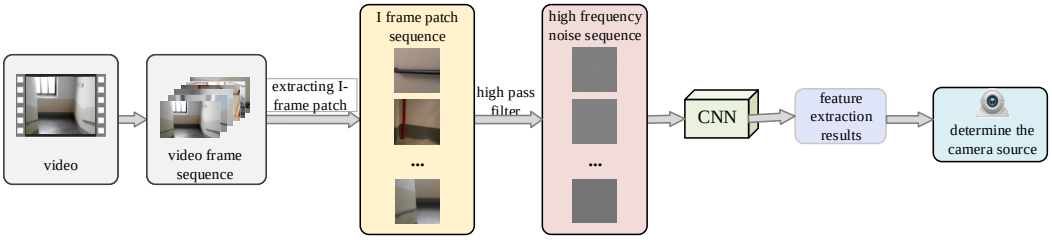

Figure 1: Overall process of video source identification

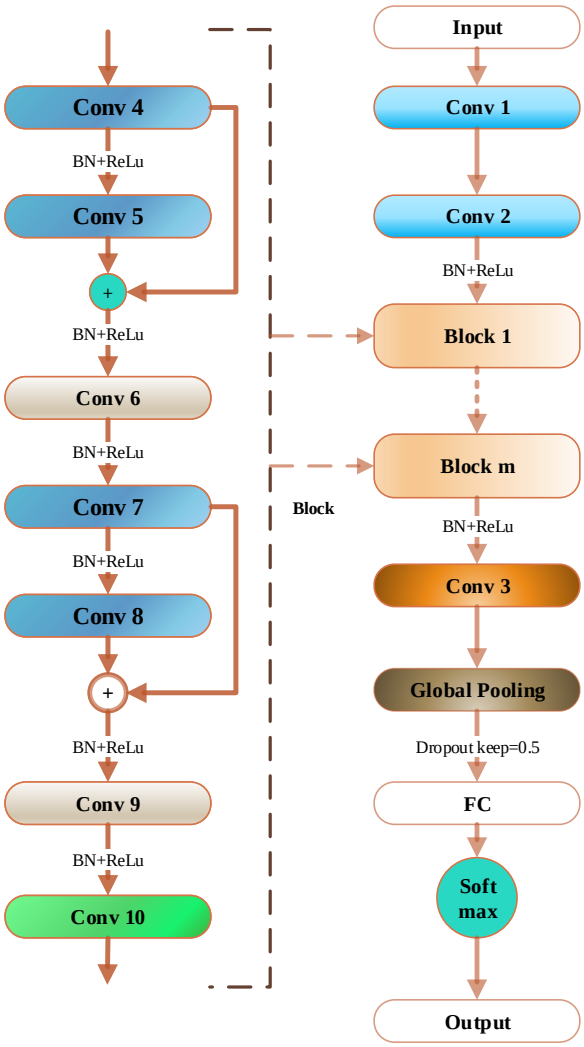

Figure 2: Network Structure

cause the phenomenon of "gradient dispersion" or "gradient explosion"[9], and when the model converges, with the increase of network depth, the training error does not decrease but increases.

- We use convolution kernel to replace pooling. The main purpose is that pooling layer does not have trainable parameters, which belongs to simple down-sampling, it will inevitably lead to the loss of useful information. However, convolution layer has trainable parameters, it also has the function of feature fusion on the basis of down-sampling, which can restrain the information loss in the process of down-sampling.

  We build the network according to the above idea. The input of the network is a $511 \times 511 \times 3$ high-frequency noise image. Two 3×3 convolution kernels are used in the front end of the network instead of 5×5 convolution kernels. And the residual blocks on the right side of the figure are used for stacking in the middle part of the network. Each residual block reduces the resolution of the input image by half, the number of channels of the input image is doubled after every two residual blocks. The output channel of the middle part of the network is $2 \times 2 \times 128$. Finally, we use the maximum pooling layer to transform the picture into vector, which is sent to the full connection layer to take the output result of softmax function as the final output.

3.3 Training

In order to eliminate the interference of scene information in network performance testing, all frames participating in network training and testing are eliminated by using wavelet-based high-pass filter in advance.

When training the network, all the data come from I frames. The construction process of the training data sets is as follows: Firstly, FFMPEG is used to extract I frames from the 1080P video in the training sets, and the label of each I frame is the camera to which the video

belongs. Then, we use a high-pass filter to eliminate the scene information. A training set $G_n\{nosie_n : label_n\}$ is obtained, which contains n images with high-frequency noise from I frames, and each image has its corresponding label, which is the source camera to which it belongs, and the label is encoded by one-hot.

After preprocessing the training data sets, we start to train the network. The input size of the network is $511 \times 511 \times 3$, and each input data of the network is randomly cut from the original frame. The loss function of the network uses standard cross entropy, the optimizer selects Adma, the initial learning rate is set to be 0.001, and the learning rate is reduced to 0.5 times of the original one for every 30 epochs. The total number of epochs trained by the network is set to be 100 times.The loss function uses the cross entropy function.

3.4 Testing

The input of the network is an image I with a specified size, and the output of the network is an m-dimensional prediction vector $f = G(I)$, where $G(I)$ represents prediction by neural network, and $\sum_i^m f_i = 1$. The identification process of a single test video is as follows:

- Firstly, we use FFMPEG to divide the test video into n I frames, which are called sequences I_n;
- Then, we cut I_n into a fixed-size sequence by using the maximum information block extraction algorithm. And in order to obtain a high-frequency noise map sequence $I_n^{'}$, we use a high-pass filter to extract high-frequency noise from the fixed-size frame sequence;
- Finally, the sequence $I_n^{'}$ is be entered into the network, and we fuse the results according to the following formula.

$$f^{'} = \frac{1}{n}\sum_{i=0}^{n} G(I_i^{'}) \tag{2}$$

Each component in the one-dimensional vector $f^{'}$ represents the probability that the video belongs to a certain device, and the source camera of the video is the one with the highest probability.

4 EXPERIMENT

In order to verify the performance and the effectiveness of the proposed model, the following series of experiments are made:

- We evaluate the influence of different types of frames on the accuracy of camera source recognition in this network.
- We compare the performance of the proposed algorithm with several popular algorithms;

4.1 Data Set

To verify the performance of the method proposed in this paper, we processes some parts of the data in the VISION [18]data set. This processing will generate a data set for testing the algorithm. The resolution of all videos in this data set is 1080P, and all videos are re-coded into MP4 by ffmpeg. The video code rate is consistent with the original video. This new data set contains six cameras, totaling 114 videos. Sixty percent of them are used to train the network models, with a total of 66 videos; Forty percent are used to test the performance of the model, with a total of 48 videos.The models of these six cameras are iPhone5c,Xperia Z1 Compact, HuaWei P9 Lite, iPhone6Plus, RedmiNote3, OnePlus A3000.

4.2 Evaluation Indexes

In order to evaluate the performance of CSI-CNN network model, we use ACC, ROC curve and AUC as evaluation indexes, which can be defined as follows:

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \tag{3}$$

In ROC curve, the abscissa is false positive rate (FPR) and the ordinate is true rate(TPR). FPR and TPR can be calculated as follows.

$$FPR = \frac{FP}{FP + TN} \tag{4}$$

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

Here, TP represents the number of samples that are actually positive examples and are judged as positive examples by the classifier, that is, the number of samples that the image belongs to a certain camera and the model classifies as belonging to the camera. FP represents the number of samples that are actually negative examples and judged as positive examples by the classifier, that is, the number of samples whose images do not belong to a certain camera but are classified by the model as belonging to the camera. FN indicates the number of samples that are actually positive examples but judged as

Table 1: Accuracy on different types of frames

Block Size	MAX-I	Center-I	Center-I+P
128×128	0.855	0.853	0.841
256×256	0.913	0.865	0.856
512×512	0.948	0.923	0.884

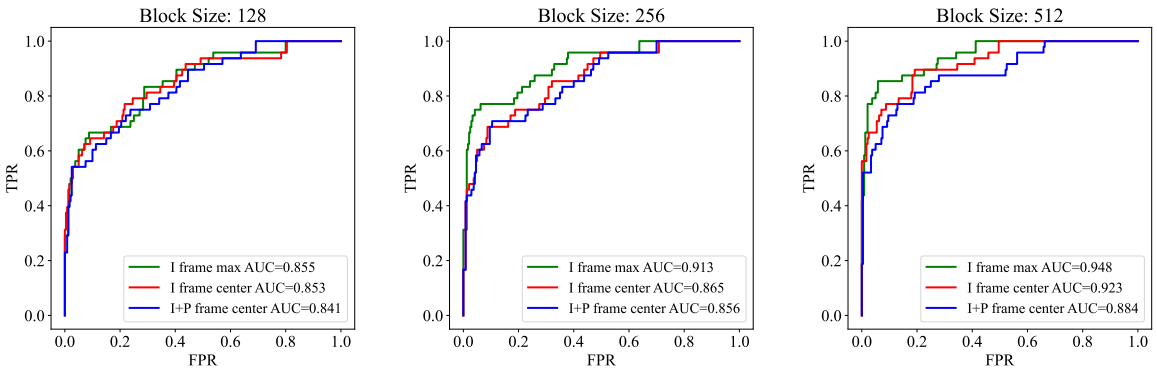

Figure 3: The ROC curve on different frame types

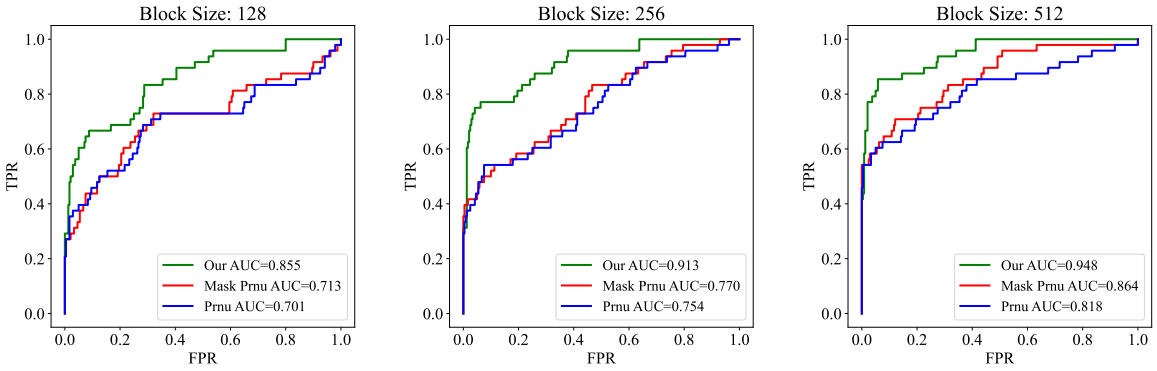

Figure 4: The ROC curve on different frame algorithms

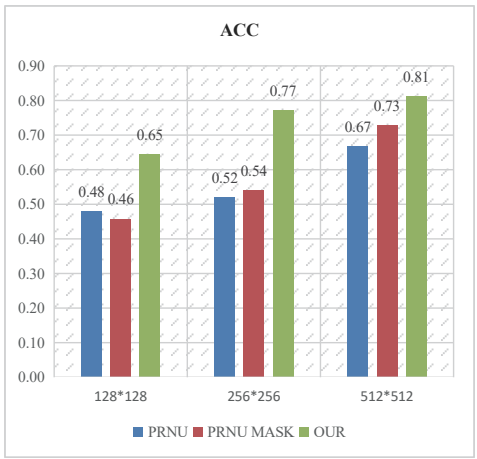

Figure 5: Accuracy on different algorithms

negative examples by the classifier, that is, the number of samples that the image does not belong to a certain camera and the model classifies as not belonging to the camera. TN represents the number of samples that are actually negative cases and judged as negative cases by the classifier, that is, the number of samples that the image belongs to a certain camera and the model classifies as not belonging to the camera. AUC is the area under ROC. It is equivalent to Mann–Whitney U test and can be calculated as follows[15]:

$$AUC = \frac{\sum_{i \in p} rank_i - \frac{(M+1)M}{2}}{MN} \tag{6}$$

Among them, M represents the number of positive samples, and N represents the number of negative samples. p represents a positive sample. $rank_i$ represents the descending rank of i in the sample set.

4.3 Experiment and Discussion

In order to determine the influence of different frame types on the accuracy of the model, we use three combinations of frames to verify. The test data types used respectively are as follows:

- I frame center block;
- I frame maximum information block;
- I frame center block and P frame center block.

At the same time, in order to prove the robustness of the experimental results, we test these three types of frames under three different resolutions in this paper.

Figure 3 shows the ROC curves obtained by testing three different frame combinations at three different separation rates. It can be seen from Figure 3 that with the increase of block size, the AUC value of test data gradually increases.And with the increase of available data, the performance of classifier will gradually improve. Using three different rates of separation, test the accuracy of three different combinations of frames. From Table 1, we can see that when using the largest information block of I frame for classification, we can get the highest accuracy . Therefore, in the next experiment, we use I frame maximum information block for classification in our algorithm.

When evaluating the model proposed in this paper, we compare our method with the classical PRNU model[13] and the PRNU model based on MASK[11]. We use these three methods to test the performance of the model under different resolutions

Figure 4 describe the ROC curves of the comparison algorithm and our algorithm under different resolutions. Figure 5 describes the accuracy histogram of the comparison algorithm and our algorithm under different resolutions. From experimental data, it can be seen that the performance of the proposed method is significantly better than the two mainstream algorithms.

5 Conclusion

In this paper, we aimed at to improve the security of multimedia video information in mass Internet of Things devices.We analyzed the characteristics of video stream, and established the video source forensics network model.Through training the model, we obtained the optimal parameters for extracting noise features. And compared with the existing video source forensics algorithms, the results show that our proposed network model can well identify the test video source.

In the follow-up work, we will do further research in the following aspects. We will analyze the timing information of video stream and fuse the timing information features to identify video sources; And the recognition method in this paper will be extended to short video forensics on social network platform.

Acknowledgements

The research presented in this paper is supported in part by the National Natural Science Foundation(No.:U20B2050), The Youth Innovation Team of Shaanxi Universities, the Innovation Project of Shaanxi Provincial Department of Education(No.:17.JF023).

References

[1] E. Altinisik, K. Tasdemir, and H. T. Sencar, "Extracting prnu noise from h.264 coded videos," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018.

[2] I. Amerini, R. Caldelli, A. D. Mastio, A. D. Fuccia, C. Molinari, and A. P. Rizzo, "Dealing with video source identification in social networks," *Signal Processing: Image Communication*, pp. 1–7, 2017.

[3] M. Boroumand and J. Fridrich, "Deep learning for detecting processing history of images," *Electronic Imaging*, vol. 2018, no. 7, pp. 213–1–213–9, 2018.

[4] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, "Source digital camcorder identification using sensor photo response non-uniformity," in *Conference on Security, Steganography, and Watermarking of Multimedia Contents IX*, 2007.

[5] 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.

[6] S. Chen, A. Pande, Z. Kai, and P. Mohapatra, "Live video forensics: Source identification in lossy wireless networks," in *IEEE*, 2014, pp. 28–39.

[7] B. Hosler, O. Mayer, B. Bayar, X. Zhao, and M. C. Stamm, "A video camera model identification system using deep learning and fusion," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[8] K. Houria, Y. Zaz, and T. Mantoro, "Hybrid watermark fingerprint algorithm for h.264 compressed video authentication," in *2018 International Conference on Computing,*

Engineering, and Design (ICCED), 2018, pp. 250–253.

[9] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[10] M. Iuliani, M. Fontani, D. Shullani, and A. Piva, "Hybrid reference-based video source identification," *Sensors*, vol. 19, no. 3, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/3/649

[11] E. K. Kouokam and A. E. Dirik, "Prnu-based source device attribution for youtube videos," *Digital Investigation*, vol. 29, pp. 91–100, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1742287618304377

[12] S. C. Law and N. F. Law, "Prnu-based source identification for network video surveillance system," in *2018 IEEE International Conference on Consumer Electronics (ICCE)*, 2018.

[13] C.-T. Li, "Source camera identification using enhanced sensor pattern noise," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 2, pp. 280–287, 2010.

[14] S. Li, K.-K. R. Choo, Q. Sun, W. J. Buchanan, and J. Cao, "Iot forensics: Amazon echo as a use case," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6487–6497, 2019.

[15] S. J. Mason and N. E. Graham, "Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation," *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, vol. 128, no. 584, pp. 2145–2166, 2002.

[16] M. K. Mihcak, I. Kozintsev, and K. Ramchandran, "Spatially adaptive statistical modeling of wavelet image coefficients and its application to denoising," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 6. IEEE, 1999, pp. 3253–3256.

[17] C. Park, "Exploring compression effects for improved source camera identification using strongly compressed video wei-hong chuang , hui su , and min wu," 2011.

[18] D. Shullani, M. Fontani, M. Iuliani, O. A. Shaya, and A. Piva, "Vision: a video and image dataset for source identification," *Eurasip Journal on Information Security*, vol. 2017, no. 1, p. 15, 2017.

[19] S. Taspinar, M. Mohanty, and N. Memon, "Source camera attribution using stabilized video," in *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2016, pp. 1–6.

APPENDICES A

It is mainly used to introduce video coding knowledge.H.264 is a new digital video compression format after MPEG4, which is proposed by ISO and ITU. In the H.264 compression standard, frames I, P, and B are used to represent the transmitted video picture.

- I frame, also known as intra-frame encoded frame, is an independent frame with all its own information. It can be decoded independently without referring to other images, and can be simply understood as a static picture.The first frame in the video sequence is always frame I, because it's the key frame.

- P frame, also known as interframe prediction coding frame, requires reference to the previous I frame to encode. Represents the difference between the current frame and the previous frame, which may be frame I or frame P. Decoding requires the previous cached image to be superimposed with the differences defined in this frame to generate the final image. P frames generally take up fewer bits of data than I frames, but the disadvantage is that P frames are very sensitive to transmission errors due to their complex dependence on previous P and I reference frames.

- B frame is also known as bidirectional predictive coding frame, that is, B frame records the difference between this frame and before and after the frame. That is to say, to decode B frame, not only the previous cached picture should be obtained, but also the later picture should be decoded, and the final picture should be obtained through the superposition of the before and after pictures and the data of this frame. B frame compression rate is high, but higher requirements for decoding performance.

AUTHOR BIOGRAPHIES

Dongzhu Rong received the bachelor degree from Xi'an Aeronautical University and is now studying for his master's degree at Xi'an University of Technology. His research interests include digital media forensics, network security and machine learning.

Yan Wang received her Master degree from Xi'an University of Technology, China. She is currently a Ph.D. student at Xi'an University of Technology, China. Her research interests include digital media forensics, network security and machine learning.

Qindong Sun received his Ph.D. degree in School of Electronic and Information Engineering from the Xi'an Jiaotong University, China. He is currently a professor at the School of Cyber Science and Engineering of Xi'an Jiaotong University. His research interests include network security, online social networks and internet of things.