EXPLORING COMPRESSION EFFECTS FOR IMPROVED SOURCE CAMERA IDENTIFICATION USING STRONGLY COMPRESSED VIDEO

Wei-Hong Chuang, Hui Su, and Min Wu

University of Maryland, College Park, USA

ABSTRACT

This paper presents a study of the video compression effect on source camera identification based on the Photo-Response Non-Uniformity (PRNU). Specifically, the reliability of different types of frames in a compressed video is first investigated, which shows quantitatively that I-frames are more reliable than P-frames for PRNU estimation. Motivated by this observation, a new mechanism for estimating the reference PRNU and two mechanisms for estimating the test-video PRNU are proposed to achieve higher accuracy with fewer frames used. Experiments are performed to validate the effectiveness of the proposed mechanisms.

***Index Terms*—** Digital video forensics, source camera identification, photo-response non-uniformity, low-bit-rate video compression.

1. INTRODUCTION

Pocket-sized digital cameras and cell-phones with cameras have become popular and generated a large amount of digital images and videos. Compared to images, videos can capture more visual information, and therefore is an ideal format for recording rich and dynamic content.

Accompanying the growing importance of digital videos, concerns regarding their origin and authenticity have been raised and are receiving increasing attention. A systematic study of *digital video forensics* that answers different questions about a video's acquisition and processing history is important in order to establish the trustworthiness of digital videos. Several previous works on video forensics considered the identification of source devices and tampering operations. In [1], Chen *et al.* extended the source camera identification technique based on the Photo-Response Non-Uniformity (PRNU) [2] from image to video. McCloskey [3] proposed to take into account the influence of video content on the achievable performance of [1]. On tampering detection, Wang and Farid [4] demonstrated that frame insertion or deletion that are usually involved in video forgery form forensic traces and therefore can be detected. Luo *et al.* [5] showed that MPEG compression introduces different block artifacts into different types of frames, which can be used to detect video recompression.

In this paper, we examine the source camera identification problem, with a focus on *cell-phone cameras*. We focus on cell-phone cameras because more cell-phones are now equipped with the video recording capability, and we foresee that more videos will be generated by cell-phones in the future owing to their superior convenience. Previous works such as [6, 7] have developed and enhanced the methodology of source camera identification by means of the PRNU [2] which we will review shortly. These works considered the case when *still images* from the camera under investigation are used for PRNU estimation and matching. This methodology is extended in [1] to use *videos*, and the reported accuracy is promising when the test video is long enough. However, as also noticed in [3], the task of source camera identification using videos is more challenging than the image counterpart due to the degraded visual quality of videos. This problem is even more serious when we consider videos generated by cell-phone cameras that suffer from much stronger compression. Nevertheless, the rich temporal information in videos can help, if properly exploited, to achieve more accurate source camera identification.

As a video is composed of multiple frames, how each frame should be used to jointly estimate the PRNU deserves careful exploration. In this paper, we study the effect of video compression, and show that the reliability of frames for PRNU estimation can be considerably different, attributed to different levels of compression. We propose new mechanisms for PRNU estimation that leverage such a difference, and show that more accurate source camera identification can be achieved with fewer frames used.

2. PRNU FOR SOURCE CAMERA IDENTIFICATION

We review the basic principles of source camera identification based on PRNU. For a more detailed discussion, please refer to [2]. The manufacturing imperfections of charge-coupled device (CCD) and complementary metal-oxide semiconductor (CMOS) sensors result in slight variations of the sensitivity of sensors to the incident light. The pattern of sensitivity variation, commonly referred to as the Photo Response Non-Uniformity (PRNU) [2], can be seen as the "fingerprint" unique to individual imaging devices. It has been shown in [2] that, by applying a denoising filter on the image $\mathbf{F}$, the difference between $\mathbf{F}$ and its denoised version can be approximated by $\mathbf{V} = \mathbf{F}\mathbf{K}+\mathbf{M}$, where $\mathbf{V}$ is referred to as the noise residual, $\mathbf{K}$ is the PRNU pattern matrix that captures the variation pattern of sensor sensitivity, and $\mathbf{M}$ is the modeling noise that accommodates various noise sources, including shot noise,

Email contact: {whchuang, hsu, minwu}@umd.edu

dark current, read-out noise, quantization and compression noise, and the imperfection of the denoising filter. Please be informed that all multiplication operations throughout this paper are element-wise.

For source camera identification using output images, it is usually assumed that N images taken by the camera under investigation are available for PRNU estimation. When the modeling noise $\mathbf{M}$ is assumed as white Gaussian with per-pixel variance identical across all the images, a maximum-likelihood estimate of $\mathbf{K}$ can be derived as:

$$\hat{\mathbf{K}} = \frac{\sum_{i=1}^{N} \mathbf{V}_i \mathbf{F}_i}{\sum_{i=1}^{N} (\mathbf{F}_i)^2}, \tag{1}$$

where $\mathbf{V}_i$ and $\mathbf{F}_i$ are the ith noise residual and ith image, respectively [2].

The typical setting of source camera identification assumes the camera under investigation is available. To match test images against this camera, a training procedure is performed first to obtain a *reference PRNU*. Ideal training images are those with smooth content and high yet unsaturated luminance. Then a PRNU estimate from the test image is calculated using Eq. (1) and compared against the reference PRNU. A popular sub-optimal similar metric between two PRNU matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ is the Normalized Cross-Correlation (NCC) given by $\mathrm{NCC}(\mathbf{S}_1, \mathbf{S}_2) = \frac{(\mathbf{S}_1 - \bar{\mathbf{S}}_1) \otimes (\mathbf{S}_2 - \bar{\mathbf{S}}_2)}{\|\mathbf{S}_1 - \bar{\mathbf{S}}_1\| \|\mathbf{S}_2 - \bar{\mathbf{S}}_2\|}$, where $\otimes$ denotes the dot product, and $\bar{\mathbf{S}}_1$ and $\bar{\mathbf{S}}_2$ are the average value of $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively. A correlation matrix $\mathbf{C}$ can be obtained where $C(i, j)$ is the NCC value between $\mathbf{S}_1$ and $\mathbf{S}_2$ when $\mathbf{S}_2$ is shifted by (i, j). Another PRNU similarity metric that compensates for the camera-specific NCC range is called the Peak to Correlation Energy (PCE), defined as $\mathrm{PCE}(\mathbf{S}_1, \mathbf{S}_2) = \frac{(n - |\mathcal{N}_{\mathrm{peak}}|) C_{\max}^2}{\sum_{(i,j) \notin \mathcal{N}_{\mathrm{peak}}} C(i,j)^2}$, where $C_{\max} = \max_{i,j} C(i, j)$, $\mathcal{N}_{\mathrm{peak}}$ is a small neighborhood surrounding the shift corresponding to $C_{\max}$, and n is the size of $\mathbf{S}_2$. PCE characterizes if the maximum correlation is much higher than the average correlation, or in other words, if there is a peak in the correlation matrix. We adopt the PCE metric in this paper.

PRNU-based source device identification using output videos has been studied in previous works [1] and [3]. Particularly, in [1], PRNU is utilized to determine if two video clips come from the same source camcorder. The main idea is to treat each frame as one image in a video consisting of N frames, and then apply Eq. (1) to obtain an estimate based on the multiple frames, *i.e.*, the entire video. It is advised in [1] that each frame be treated equally mainly to reduce the complexity of implementation. The authors reported that source camcorder can be identified as long as the video is sufficiently long. In [3], the method described above is examined with special attention to the influence of video content. It was observed that edges can be mistaken as noise by the denoising filter, which is further amplified if frames in the video are highly correlated. It is proposed in [3] to assign higher weights to pixels in smooth areas to alleviate this problem, which actually shares a similar spirit with other image-based PRNU estimation techniques such as [7].

Table 1. Cell-Phone Cameras Used in Our Experiment

Index	Model	Format	Resolution
1	RIM Blackberry 9530	3GP	480 × 352
2	Sony Ericsson W705a	MP4	320 × 240
3	Motorola Cliq	3GP	352 × 288
4,5	Apple iPhone 4 (×2)	MOV	568 × 320

3. COMPRESSION EFFECT ON PRNU ESTIMATION

Most of cell-phone cameras today support low-bit-rate video coding standards MPEG-4 AVC/H. 264. The typical resolution ranges from 320×240 to 480×352 pixels, and the bit rate may vary between 300 to 1000 kbps. Such strongly-compressed videos are generated in order to meet a more stringent storage-space constraint and to reduce the transmission effort. Strong compression may lower the accuracy of PRNU estimation, as it creates blocking artifacts and coarsely quantized intensity levels, and eliminates a significant amount of content detail that carry the PRNU-induced noise.

We take an empirical approach to understand the impact of compression on PRNU estimation, in particular, if different frames have different reliability for PRNU estimation. As it is a non-trivial task to calculate the frame quality without the uncompressed video for reference, we judge the frame reliability in terms of their correlation with the reference PRNUs. We collect 5 recently-released cell-phones with video recording capability as listed in Table 1. Twenty videos that contain indoor and outdoor scenes of 30 seconds are taken with each camera. Interestingly, we find that all frames are either I- or P-frames, and no B-frame is found. We obtain the reference PRNUs of all these cameras according to the procedure recommended in Sec. 4. The sequence of frame type of each video can be represented as $\{I, P_1, P_2, P_3, P_4, \ldots, I, P_1, P_2, P_3, P_4, \ldots, I, \ldots\}$. The PRNU of each test video can be estimated with the subset of frames corresponding to the same symbol (*i.e.*, the same offset from I-frames).

For each camera, the PCE value averaging over 20 videos with matched against reference PRNU is shown in Fig. 1. The PCE value is much higher (about twice) when the PRNU is estimated using I-frames, but the difference in PCE between different subsets of P-frames is not obvious. That is, the PRNU extracted from I-frames are more correlated to the reference PRNU than those from P-frames, which implies that I-frames are more reliable than P-frames for PRNU estimation. In the meantime, the average PCE values associated with P_1, P_2, P_3, and P_4 have similar values of $31.8, 30.6, 32.7, 32.0$, respectively, indicating that P-frames with different offsets have similar reliability for PRNU estimation.

4. REFERENCE PRNU ESTIMATION

In order to perform robust matching between the reference PRNU and the PRNU from test videos, it is crucial to obtain reliable reference PRNUs in the training process. As compression poses a critical impact on PRNU estimation as shown

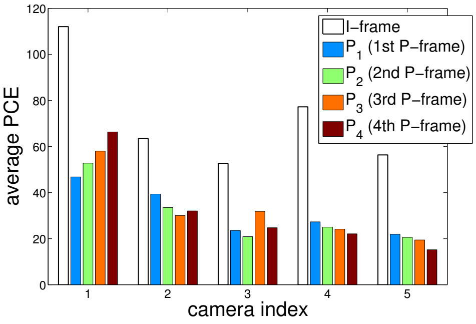

Fig. 1. Average PCE for different offsets from I-frames.

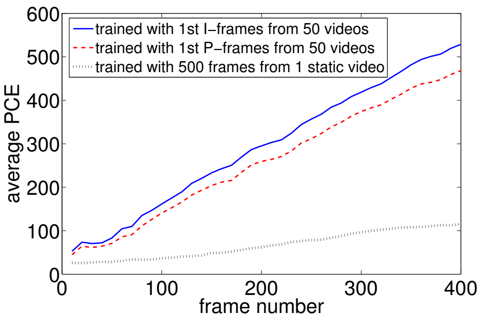

Fig. 2. Comparison of different mechanisms for reference PRNU estimation, in terms of the achievable PCE value for different test-video frame numbers. Blackberry 9530 is used.

in Sec. 3, it is reasonable to favor I-frames if enough I-frames are available. Besides, since the compression under our consideration is strong, various noise sources may be dominated by compression noise that is highly content-dependent. One should avoid the use of videos with (nearly) static content otherwise the overall modeling noise associated with different frames in a video will be unfavorably correlated and cannot be easily removed through frame averaging.

These observations motivate us to use multiple short videos, instead of one long video, to obtain the reference PRNU. Specifically, a total of N short videos (shorter than 1 second) that contain smooth and bright scenes are first collected, and then the first frame of each video will be used to jointly estimate the reference PRNU. Since practically the first frame in each video is an I-frame, there are as many I-frames as the number of training videos available for reference PRNU estimation. Moreover, because these I-frames are from different videos, it is expectable that they will have lower correlation with one another.

We compare this mechanism of reference PRNU estimation with two alternatives: 1) using the first P-frames (*i.e.*, the second frame in a video) from multiple videos and 2) using a long video with static content. We refer to these three mechanisms as $\mathcal{M}_I$, $\mathcal{M}_P$, and $\mathcal{M}_L$, respectively. For $\mathcal{M}_I$ and $\mathcal{M}_P$, 50 short videos are used to estimate the reference PRNU. For $\mathcal{M}_L$, a long video with 500 static frames is used to estimate the reference PRNU. In Fig. 2, we show for the three mechanisms the PCE values averaging over 20 test videos with respect to different frame numbers from the test video. One can see that $\mathcal{M}_I$ is consistently superior to $\mathcal{M}_P$, which increases as more frames from the test video are used. On the other hand, estimating the reference PRNU using a long but static video is much less effective. If the reference PRNU is obtained in such a way, then even if much more frames in the test video are used, then correlation between the test-video PRNU and the reference PRNU is still much smaller.

5. EFFICIENT PRNU MATCHING BY FRAME REORDERING AND WEIGHTING

We have shown that I-frames extracted from videos are more reliable than P-frames for PRNU estimation. Nevertheless, the average PCE value when all frames are used is 300.9, much higher than the average PCE value of 72.3 if only I-frames are used. It is therefore reasonable to use all the frames in a video to obtain a PRNU estimate, and this is in line with the conclusion made in [1]. Two issues, however, need to be addressed more carefully. First, using all the frames in a video can be prohibitively time-consuming, since all frames have to go through a denoising process with non-negligible complexity to extract the frame-wise PRNU. Besides, since I-frames and P-frames have distinct reliability, they should be treated differently when combined for PRNU estimation.

To address the first issue, if the number of frames that can be processed in PRNU estimation is limited, a reasonable choice is to first use more reliable frames, *i.e.*, I-frames. This is feasible in terms of video decoding complexity since I-frames are at the beginning of the Group of Picture (GOP) and can be easily located. In this paper, we assume that information required to decode the subsequent P-frames are stored after an I-frame is completely decoded, so that the decoding of P-frames can be performed without re-decoding the I-frames. For the second issue, by allowing the ith frame has its modeling noise variance of σ_i^2, we can generalize Eq. (1) as $(\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\mathbf{V}_i\mathbf{F}_i)/(\sum_{i=1}^{N} \frac{1}{\sigma_i^2}(\mathbf{F}_i)^2)$, which indicates that a frame should be assigned a weight inversely proportional to its modeling noise variance. We assume that all I-frames have the same modeling noise variance of σ_I^2, and all P-frames have the same modeling noise variance of σ_P^2. Since videos generated by cell-phones are strongly compressed, σ_I^2 and σ_P^2 are mainly determined by the level of compression noise, and therefore should be directly related to the signal-to-noise ratio (SNR) of each frame type. Estimating the SNR using only the compressed video is in general a difficult task [8]; in this paper, we arbitrarily take $\sigma_P^2 = 2\sigma_I^2$, or equivalently assign weights 2 and 1 to I-frames and P-frames, respectively. Please be reminded that this setting is merely to demonstrate that proper weighting may improve PRNU estimation.

We compare the sequential frame parsing (*i.e.*, reading frames from the beginning of the video in a sequential manner), the proposed frame reordering mechanism with equal weights, and the proposed frame reordering mechanism with the 2 : 1 weights. Fig. 3 shows the PCE values for these

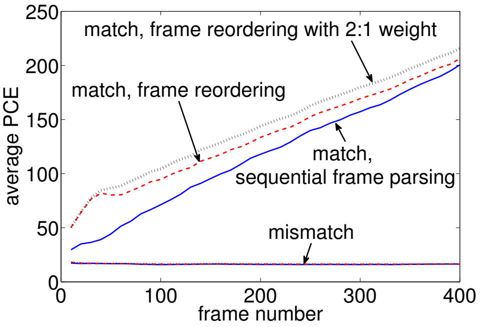

Fig. 3. Average PCE value w.r.t. different number of frame.

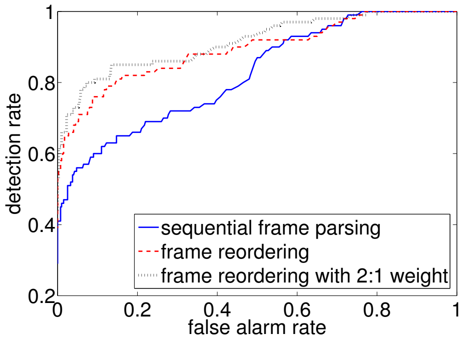

Fig. 4. ROC curve with 100 frames for PRNU estimation.

three mechanisms, averaging over totally 100 videos from 5 cameras. One can see that 1) with more frames, the difference between the match and mismatch cases becomes more obvious; 2) the frame reordering mechanism significantly increases the PCE values, especially when the frame number is smaller; 3) for all the frame numbers, the 2 : 1 weights assigned to I-frames and P-frames create additional increase in PCE. Note that these two mechanisms do not increase the PCE in the mismatch case.

We also compare these mechanisms in terms of their source camera identification accuracy. The Receiver Operating Characteristic (ROC) curves for the three mechanisms for two frame numbers 100 and 300 are shown in Fig. 4 and 5, where the horizontal axis is the false alarm rate and the vertical axis is the detection rate. One can see that with an increased number of frames, the accuracy is improved for all the three mechanisms. Frame reordering increases the accuracy especially for a smaller number of frames, and further improvement can be obtained by assigning higher weights to more reliable frames. It is also noteworthy that frame ordering and unequal weighting have a complimentary nature: the former is advantageous if only a limited number of frames can be processed, while the latter is more useful if more frames are available.

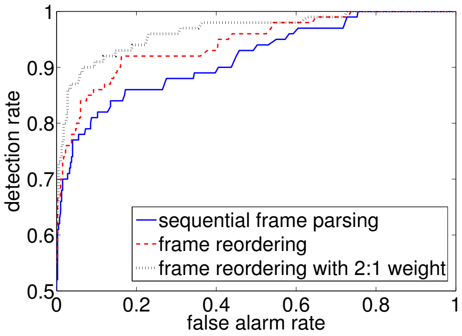

Fig. 5. ROC curve with 300 frames for PRNU estimation.

6. CONCLUSION

In this paper, we explore the impact of compression on source camera identification using the Photo-Response Non-Uniformity (PRNU) extracted from compressed videos. We consider videos generated by cell-phone cameras, which are strongly compressed to reduce the storage and transmission requirement. Although the authors in [1] stated that each frame in a video should be treated equally, we find that different frame types (I and P) actually have different levels of reliability for PRNU estimation. Motivated by this observation, we propose an effective mechanism for estimating the reference PRNU pattern. Moreover, we show that by reordering and weighting the frames in a video according to their reliability, we can achieve more accurate source camera identification with fewer frames used.

7. REFERENCES

[1] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, "Source digital camcorder identification using sensor photo-response nonuniformity," in *Proc. of SPIE Electronic Imaging*, Photonics West, Jan. 2007, pp. 1G–1H.

[2] J. Fridrich, "Digital image forensics," *IEEE Signal Proc. Magazine*, vol. 26, no. 2, pp. 26 –37, March 2009.

[3] S. McCloskey, "Confidence weighting for sensor fingerprinting," in *Proc. of IEEE Conf. on Com. Vision and Patt. Rec. Workshops*, 2008.

[4] W. Wang and H. Farid, "Exposing digital forgeries in video by detecting double mpeg compression," in *Proc. of ACM Multimedia and Security Workshop*, Geneva, Switzerland, 2006.

[5] W. Luo, M. Wu, and J. Huang, "Mpeg recompression detection based on block artifacts," in *Proc. of SPIE Conf. on Security, Forensics, Steganography, and Watermarking of Multimedia Contents*, January 2008, vol. 6819.

[6] J. Lukas, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," *IEEE Trans. on Infor. Security and Forensics*, vol. 1, pp. 205–214, Jan. 2006.

[7] C. T. Li, "Source camera identification using enhanced sensor pattern noise," *IEEE Trans. on Infor. Forensics and Security*, vol. 5, no. 2, pp. 280–287, June 2010.

[8] T. Brandao and M. P. Queluz, "No-reference PSNR estimation algorithm for H.264 encoded video sequences," in *Proc. of 16th Euro. Sig. Proc. Conf.*, Lausanne, Switzerland, Aug. 2008.