GOP Based Automatic Detection of Object-based Forgery in Advanced Video

Shunquan Tan* and Shengda Chen† and Bin Li‡
*Shenzhen Key Laboratory of Media Security,
College of Computer Science & Software Engineering, Shenzhen University, Guangdong Province, 518060 China
E-mail: tansq@szu.edu.cn
†School of Information Science and Technology, Sun Yat-sen University, Guangdong Province, 510006 China
E-mail: chshda@foxmail.com
‡Shenzhen Key Laboratory of Media Security,
College of Information Engineering, Shenzhen University, Guangdong Province, 518060 China
E-mail: libin@szu.edu.cn

Abstract—**Passive multimedia forensics has become an active topic in recent years. However, the research on video forensics, and especially on automatic detection of object-based video forgery is still in its infancy. In this paper, we develop an approach for automatic identification of object-based forged video encoded with advanced frameworks based on its GOP (Group Of Pictures) structure. The proposed approach contains two specific frame manipulation detectors for three categories of frames. GOP structures are used in the proposed approach to determine the sampling interval when extracting I frames or P/B frames in the training and testing procedure. In the construction of the frame manipulation detector, motion residuals are generated from the target video frame sequence. We regard the object-based forgery in video frames as image tampering in the motion residuals, and employ the feature extractors which are originally built for frequency domain image steganalysis to extract forensic features from the motion residuals. The experiments show that the proposed approach achieves excellent results.**

I. Introduction

With the wide availability of powerful media editing tools, it becomes much easier to tamper digital media without leaving any perceptible traces. This leads to an increasing concern about the trustworthiness of digital media contents [1] and there is a pressing need to develop effective forensic techniques to verify the authenticity, originality, and integrity of media contents.

While some efforts on video forensics have also been made in the last decades, most of the existing video forensic algorithms either expose the evidence of side-effects of forgery [2] or detect the so-called *frame-based forgery*, which refers to the manipulations that insert or delete frames [3]. Less attention has been paid to the forensics of *object-based forgery*, which adds new objects to a video scene or removes existing objects from it. We must emphasize that object-based forgery is a common video tampering method since the object added into or removed from a video is usually critical to the contents that video conveys. Therefore the attention paid to the forensics of object-based video forgery does not match its importance. In [4], Zhang et al. proposed a detection scheme of the forgery of the sole moving object in scene based on geometrical inconsistencies. In [5], Conotter et al. proposed a specific forensic method to detect the forgery of objects in ballistic motion based on physical inconsistencies. In [6], Chen et al. proposed a forensic method to detect the forgery of single moving object with absolutely static background relying on the statistical features of object contour. All of the above works are devoted to the manipulation of simplified scenes or specific objects. In [7], we proposed an algorithm which can not only identify object-based forged videos encoded with advanced frameworks but also detect the boundaries of the actual forged segments.

In this paper, we develop an approach for automatic identification of object-based forged video encoded with advanced video encoding standards based on its GOP (Group Of Pictures) structure. This paper is an extension of our work reported in [7]. The rest of this paper is organized as follows. In Sect. II, we analyze the characteristics of object-based forgery in advanced video and its impact to GOP structure. The proposed GOP based automatic detector of object-based forgery is described in Sect. III. We present experimental results in Sect. IV. Finally, the paper is concluded in Sect. V.

II. Object-based forgery and its impact to GOP

Fig. 1 gives the diagram of object-based video forgery procedure. Generally, videos are in compressed format. Therefore, when a pristine video undergoes some kinds of object-based forgery, the first step is to decompress it to a sequence of individual frames and each frame can be regarded as a still image. Then the frames in the selected segments of the sequence are tampered while the rest frames remain untouched. After all the manipulations are finished, the resulting frame sequence is re-compressed to generate a forged version. Please note that those untouched frames in a forged video do contain some artifacts introduced in re-compression though they do not have any perceptible difference compared to their original counterparts. Based on this scenario, we can make a

This work was supported in part by the NSFC (61332012, 61402295, 61572329), Guangdong NSF (2014A030313557), Shenzhen R&D Program (GJHZ20140418191518323, JCYJ20140418182819173).

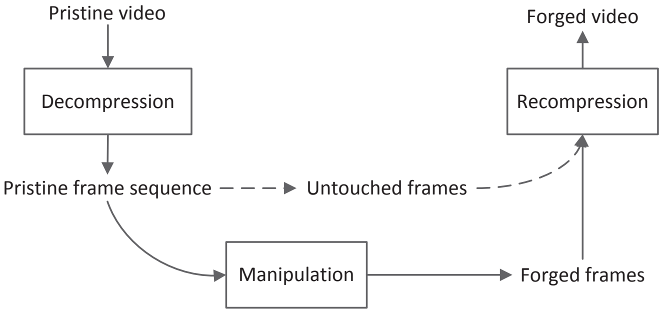


Fig. 1. The diagram of object-based video forgery procedure.

conclusion that there are two categories of frames in front of us if we want to figure out a suspicious video clip is forged or not. The first category is the **pristine frames**, namely the frames in a pristine compressed video stream which do not undergo any manipulation. The second category is the **double compressed frames**, namely the frames in a video stream which have undergone re-compression. In an innocent double-compressed video clip all the frames belong to "double compressed frames" category. However, in a forged video clip the frames can be further divided into two sub-categories: one is the **innocent double compressed frames** sub-category which denotes those frames do not contain forged contents, the other is the **forged frames** sub-category which denotes those frames have undergone tampering operations. An encoded video stream consists of a series of successive GOPs. Each GOP in turns contains three types of frames: I-frames (intra-coded frames), P-frames (predictive-coded frames) and B-frames (bipredictive-coded frames). An I-frame indicates the beginning of a GOP while P-frames and B-frames all rely on the I-frame in a GOP. All types of frames in a GOP exhibit strong correlations. In early video compression standards such as MPEG I/II, the structure of each GOP is fixed. However, in advanced frameworks such as H.264/MPEG-4, GOPs have much more flexible structures. The flexibility of the GOP structure in advanced video encoding frameworks presents great challenge to the forensics of object-based forgery.

However, in order to generate an object-based forged video without leaving any perceptible traces, object-based video forgery itself must be an elaborate and tedious operation. It usually requires some post-processing operations such as contour blurring, contrast adjusting, video in-painting and video layer fusion, which in turns inevitably alter some inherent statistical properties of the tampered pristine video. In this paper, we construct a GOP based automatic detector to detect the alteration of those inherent statistical properties, as detailed in Sect. III.

III. GOP based automatic detector of object-based forgery

The inherent statistical properties of a video can be divided into two categories: the *intra-frame inherent properties* which describe its spatial characteristics, and the *inter-frame inherent properties* which describe its temporal characteristics. The strong correlations among the neighboring frames in a video imply that each frame in a local temporal window comprises two parts: the motion part and the static part. The static part is identical to the basic anchor frame of the local temporal window, while the motion part is the motion residual relative to that anchor frame. For each frame in a video, its motion residual, as the principal part of the visual information presented by that frame, contains a substantial portion of the intra-frame properties of that frame. Furthermore, the motion residual also contains the inter-frame inherent properties of the corresponding frame since it represents the temporal changes from that frame to the basic anchor frame. Since the motion residuals contain both the intra-frame and inter-frame inherent properties of the corresponding frames, they become our primary analysis object. Actually, each GOP in an encoded video stream can be considered as a local temporal window. The I-frame in a GOP represents an anchor frame and the P-frames/B-frames in that GOP are actually the motion residuals of the corresponding I-frame. Since the flexible structure of GOPs in advanced video framework implies that GOP can not guarantee a fixed local temporal window structure, as a result in this paper we design a collusion operator based on a fixed local temporal window structure instead of using GOP structure directly. Denote a sequence of decompressed video frames of length N as

$$\mathbb{V} \triangleq \{F^{(1)}, F^{(2)}, \dots, F^{(N)}\}, N \in \mathbf{Z} \tag{1}$$

where $F^{(k)} = (F_{i,j}^{(k)}) \in \{0, \dots, 255\}^{n_1 \times n_2}$ represents the kth decompressed video frame, which is actually an 8-bit gray-scale still image of $n_1 \times n_2$. A collusion operation inside a temporal window of the target video frame sequence, which centered at frame $F^{(k)}$ with the window size of $L = 2 \times L_h + 1$ (L_h is the number of the left/right neighbors of $F^{(k)}$), is defined as follows:

$$\begin{aligned} C^{(k)} &= (C_{i,j}^{(k)}) \\ &= \mathfrak{C}[(F_{i,j}^{(k-L_h)}), \dots, (F_{i,j}^{(k)}), \dots, (F_{i,j}^{(k+L_h)})] \end{aligned} \tag{2}$$

where $C^{(k)}$ is the colluded result for $F^{(k)}$ and $\mathfrak{C}$. The collusion operator $\mathfrak{C}$ is actually an aggregate function that groups the pixels in the corresponding coordinates of every frames in the temporal window to generate $C_{i,j}^{(k)}$. The motion residual of $F^{(k)}$, is defined as:

$$R^{(k)} = |F^{(k)} - C^{(k)}| = (R_{i,j}^{(k)}) = (|F_{i,j}^{(k)} - C_{i,j}^{(k)}|) \tag{3}$$

where $|\bullet|$ denotes the absolute value.

The collusion operator used in our experiments is defined in Eq. (4), in which $\mathfrak{C}_{\text{MIN}}$ represent minimum collusion.

$$\mathfrak{C}_{\text{MIN}} \triangleq \min_{l \in [-L_h, L_h]} \{F_{i,j}^{(k+l)}\} \tag{4}$$

The resulting $R^{(k)}$ can be regarded as an 8-bit gray-scale still image.

From Eq. (3) we can see that with the introduction of motion residual, object-based video forgery turns into the

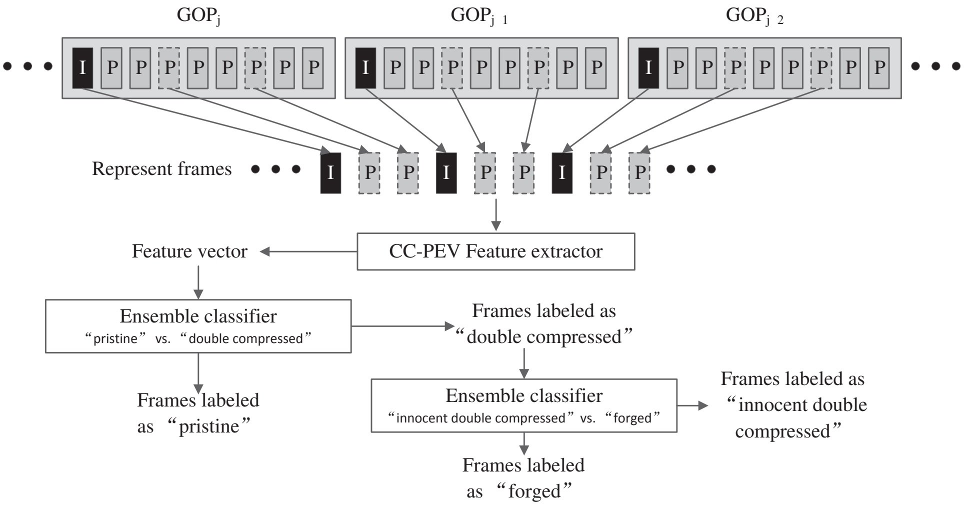

GOPj
GOPj 1
GOPj 2
I P P P P P P P P
Represent frames
I P P I P P I P P
CC-PEV Feature extractor
Feature vector
Ensemble classifier
"pristine" vs. "double compressed"
Frames labeled as "double compressed"
Frames labeled as "pristine"
Ensemble classifier
"innocent double compressed" vs. "forged"
Frames labeled as "innocent double compressed"
Frames labeled as "forged"

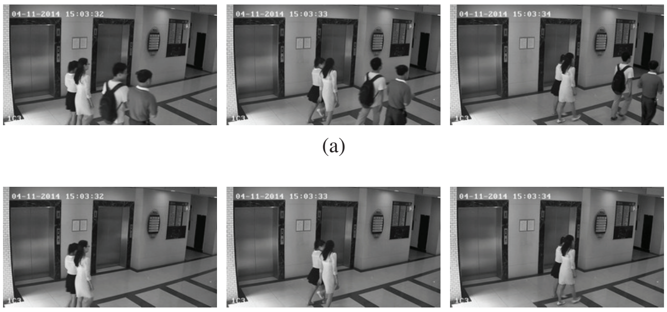
Fig. 2. The diagram of our proposed approach.

modification of the pixel values in the corresponding motion residuals. Therefore object-based video forgery can be regarded as image tampering in motion residuals. How to model the intra-frame and inter-frame inherent properties of a pristine video is the key issue of successful detection of object-based forgery in advanced video. We have already known that $R^{(k)}$ can be regarded as an 8-bit gray-scale still image. Since image forensics is the art of detecting image tampering/processing, we can use image forensic methods to detect tampering in motion residuals. The works in [8] have revealed that tampering/processing in still images can be modeled as image data hiding/steganography and state-of-the-art image steganalytic features can be used to detect them. Using motion residuals as intermediates, we can borrow some powerful statistical features from image steganalysis to model the alteration of the inherent properties of a video clip introduced by object-based forgery. Since in a video stream either the I-frames themselves or the motion vectors in the P-frames and B-frames are compressed using frequency-domain lossy compression scheme, we believe that the intra-frame and inter-frame inherent properties of the frames can be modeled by frequency-domain oriented feature sets. In this paper we employ the 548 dimensional CC-PEV frequency-domain image steganalytic feature set [9] which extracted from the motion residuals to model the intra-frame and inter-frame inherent properties contained in them. Ensemble classifier described in [10] is adopted in our work to construct the frame manipulation detector.

Since in each GOP, P-frames and B-frames all rely on the I-frame and they exhibit strong correlations, we do not need to extract CC-PEV features for all the frames in a GOP. For each GOP, we only select the leading I frame and extract several subsequent P/B frames with equal intervals. Their classification results represent the classification results of all the frames in the GOP structure. We call them the **represent frames** of a given GOP structure. Given a target video clip, the represent frames of every GOP structure in that video clip are selected and their corresponding motion residuals are constructed. Then a feature vector is extracted from each motion residual using the CC-PEV feature set. They comprise the input of the proposed frame manipulation detectors, which is made up of two ensemble classifiers. The first ensemble classifier is a "pristine" vs. "double compressed" classifier. Its task is to judge an input frame is "pristine" or "double compressed". Based on the output of the "pristine" vs. "double compressed" classifier, a simple majority strategy is adopted to pick out the "pristine" video clips. A target video clip will be classified as "pristine" if the majority of the frames it contains (more than fifty percent) are classified as "pristine", or else it is suspected to have undergone tampering operations. For those suspicious video clips, their corresponding feature vectors are fed to the second ensemble classifier, the "innocent double compressed" vs. "forged" classifier, whose task is to judge an input frame is "innocent double compressed" or "forged". A GOP structure is marked as "forged" if all the represent I frames and P/B frames are labeled as "forged" by the "innocent double compressed" vs. "forged" classifier. If there are at least one "forged" GOP structure in the suspicious video clip, it is considered to be a forged video clip, otherwise the suspicious video clips is indeed an innocent double-compressed one. For those forged video clips, successive GOP structures labeled as "forged" constitute the forged segments. Fig. 2 gives the diagram of our proposed approach. In Fig. 2, one leading I frame and two subsequent P frames are selected from each GOP structure to construct the represent frame sequence.

IV. Experiments

We test the proposed algorithm on SYSU-OBJFORG dataset (will be publicly available in the near future) where all of the video clips are extracted from primitive video footages of several static commercial surveillance cameras. It contains 100 pristine video clips and 100 forged video clips generated



(a)

(b)

Fig. 3. (a) Representative frames of a pristine video clip. (b) Representative frame of the corresponding forged version of the pristine video clip shown by Fig. 3(a) in which two walking men were erased from the scene.

Our proposed method					
P/B frames selected in one GOP	Pristine clips	Forged clips	Pristine frames	Double compressed frames	Forged frames
All	99.40%	**100%**	99.93%	**95.52%**	**83.31%**
1	**100%**	92.20%	99.90%	93.91%	80.95%
2	**100%**	96.00%	**99.96%**	95.00%	81.28%
3	**100%**	95.20%	99.95%	94.81%	79.48%
CHEN-6D					
—	78.2%	75.2%	—	—	—

TABLE I
EXPERIMENTAL RESULTS. THE BEST RESULTS IN EACH COLUMN ARE BOLD.

from the corresponding pristine ones. Every forged video clips contain one or two forged segments which lasting from one to five seconds. The object-based forgery in those segments includes adding/erasing moving figures and changing the positions of the figures in the scene. Fig. 3(a) and Fig. 3(b) show one of the pristine video clips and its corresponding forged version. There are totally about 11,000 pristine frames selected to undergo object-based forgery manipulations. According to the best knowledge of the authors, SYSU-OBJFORG is the largest object-based forged video database ever reported in the literature. Since the research on video forensics is still in its infancy, it is difficult to find a public large-scaled forged video dataset. Besides SYSU-OBJFORG, the extended SULFA provided by Bestagini et al. [1] which contains 20 pristine/forged video sequences is publicly available. However, it cannot be used in our experiments since the videos in it are provided in uncompressed raw YUV-format frame sequences. That violates the motivation of our proposed approach, namely in practical videos are in compressed format.

In our experiments, 50% of the video clips are randomly selected from the "pristine" group. They constitute the training set along with their corresponding forged versions. The rest 50% video clips are for testing. All the experiments are repeated 10 times and the average results are reported. The performance of the proposed approach is shown in Table I. The first column in Table I denotes the number of the P/B frames selected with equal intervals. "All" means that all of the subsequent P/B frames in a given GOP are used in the training and testing procedure. All the data in the rest cells are the classification accuracy of the corresponding target. From Table I we can see that our proposed approach can achieve excellent performance when used to classify pristine and forged clips. The increment of the the number of the P/B frames selected with equal intervals does not affect largely on the performance of identifying the category of a given video frame. With the consideration of efficiency, taking two P/B frames alongside with the leading I frame per GOP structure is the best choice. It is harder to pick out forged frames from their double-compressed counterparts. However, our proposed approach can still achieves average more than 80% accuracy. The forensic method proposed by Chen *et al.* [6] (CHEN-6D for short) are selected to present a contrast. CHEN-6D can only be used to detect forged video clips. From this criterion we can see that the superior performance of our method compared with CHEN-6D is quite apparent.

V. Concluding Remarks

In this paper, we developed an approach for automatic identification of object-based forged video which encoded with advanced frameworks based on its GOP (Group Of Pictures) structure. The experiments show that the proposed approach achieves excellent results in SYSU-OBJFORG, the largest object-based forged video database ever reported in the literature.

References

[1] A. Rocha, W. Scheirer, T. Boult, and S. Goldenstein, "Vision of the unseen: current trends and challenges in digital image and video forensics," *ACM Computing Surveys*, vol. 43, no. 4, pp. 26–40, 2011.

[2] D. Liao, R. Yang, H. Liu, et al., "Double H.264/AVC compression detection using quantized nonzero AC coefficients," in *Proc. SPIE, Media Watermarking, Security, and Forensics III* , 78800Q, 2011.

[3] M.C. Stamm, W.S. Lin, and K.J.R. Liu, "Temporal forensics and anti-forensics for motion compensated video," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 4, pp. 1315–1329, 2012.

[4] J. Zhang, Y. Su, and M. Zhang, "Exposing digital video forgery by ghost shadow artifact," in *Proc. 1st ACM Workshop on Multimedia in Forensics, (MiFor 09)*, pp. 49–54, 2009.

[5] V. Conotter, J. OBrien, and H. Farid, "Exposing digital forgeries in ballistic motion," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 1, pp. 283–296, 2012.

[6] R. Chen, G. Yang, and N. Zhu, "Detection of object-based manipulation by the statistical features of object contour," *Forensic Science International*, vol. 236, pp. 164–169, 2014.

[7] S. Chen, S. Tan, B. Li, and J. Huang, "Automatic Detection of Object-based Forgery in Advanced Video," *IEEE Trans. Circuits Syst. Video Technol.*, DOI: 10.1109/TCSVT.2015.2473436, 2015.

[8] X. Qiu, H. Li, W. Luo, and J. Huang, "A universal image forensic strategy based on steganalytic model," in *Proc. 2nd ACM Information Hiding and Multimedia Security Workshop (IH&MMSec' 14)*, pp. 165–170, 2014.

[9] J. Kodovsky, and J. Fridrich, "Calibration revisited," in *Proc. 11th ACM Workshop Multimedia Security (MMSec' 09)*, pp. 63–74, 2009.

[10] J. Kodovsky, and J. Fridrich, "Ensemble classifiers for steganalysis of digital media," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 2, pp. 432–444, 2012.

[1] http://sulfa.cs.surrey.ac.uk/forged_1.php