Detection and Localization of Video Object Removal by Spatio-Temporal LBP Coherence Analysis

Shanshan Bai[1,2], Haichao Yao[1,2], Rongrong Ni[1,2(✉)], and Yao Zhao[1,2]

[1] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
rrni@bjtu.edu.cn

[2] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China

Abstract. Local object removal on video can directly affect our understanding and cognition of the video content without changing the motion continuity of other moving objects in the same video frame. Forgers can use video editing tools or certain inpainting techniques to remove undesired objects easily for covering up the truth. In this paper, we present a new approach based on spatio-temporal LBP coherence analysis for detection and localization of forged regions, which are generated by removing unwanted objects from the video. The proposed method starts with frames alignment to handle camera motion. And then the coherence analysis on the spatial LBP operator between two adjacent frames is performed to find the possible forged region. Finally, the temporal LBP operator is utilized to remove the false positives so as to obtain the final abnormal area. Two common region-level inpainting methods are adopted to simulate two different types of forgery processes for performance evaluation of our scheme. The experimental results prove that our method is effective in detecting and locating the forged regions and superior to the existing two approaches.

Keywords: Video forensics · Video inpainting detection · LBP · Coherence analysis

1 Introduction

Videos are generally regarded as unbiased and reliable records of events, and they have been widely used to provide basic evidences in many different fields. However, with the rapid development of digital media editing and inpainting techniques, it becomes easier for forgers to change the facts by removing undesired target from the video. Compared with the schemes that directly delete video frames containing the target, local removal of object does not destroy the continuity of other moving objects in the same frame. Tampered videos transmitted through the Internet can disrupt people's daily lives, and even interfere with the normal social order.

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11903, pp. 244–254, 2019.
https://doi.org/10.1007/978-3-030-34113-8_21

Over the past few years, several video forensic methods for object removal have been proposed. Hsu *et al.* [1] proposed an approach for detecting and locating the forged regions using block based correlation of noise residue. This method is based on the observation that correlation between temporal noise residue in forged regions of the frame is significantly different from that in the normal regions of a frame. While the noise correlation is unstable if the test videos suffer from abrupt illumination variations and sensitive to quantization noise. And the noise-residue correlation was also used to locate forgeries in [2–4]. Singh *et al.* [5] proposed a sensor pattern noise based detection scheme, which is an improved and forensically stronger version of noise-residue based technique. Wang *et al.* [6] developed a technique to uncover copy-paste forgeries in de-interlaced and interlaced videos using correlation coefficients. For de-interlaced video, tampering will destroy the correlations introduced by de-interlacing algorithms. Bestagini *et al.* [7] proposed a similar approach to solve this problem and locate the forgeries in the spatio-temporal domain. Zhang *et al.* [8] detected video forgery based on the ghost shadow artifact which is usually introduced when objects are removed by video inpainting technology. However, this method cannot accurately locate the forged regions and is vulnerable to the effects of noise. The technique proposed by Li *et al.* [9] is to uncover object removal of surveillance videos with stationary background using motion vector correlation analysis. And it is based on the observation that the distribution of the motion vectors in the foreground area between the authentic video and the forged are quite different. Lin *et al.* [10] analyzed the abnormalities in the spatio-temporal coherence between successive frames to detect and locate forged regions. But this approach only works well on uncompressed forged videos.

Inpainting techniques are used to fill the missing holes in a visually reasonable manner when unwanted objects are removed from the video. Temporal copy-and-paste (TCP) and exemplar-based texture synthesis (ETS) are two typical inpainting methods. The TCP method replaces the forged region with the most coherent area from the nearest frame, which leads to unnaturally high temporal coherence in the forged area. The ETS inpainting method proposed in [11] individually fills in the regions from sample textures for each frame, which leads to abnormally low temporal coherence in the forged region.

This paper aims to address the problem of detecting and locating forged regions based on the coherence analysis of spatio-temporal local binary patterns (LBP). LBP is a popular operator for describing the spatial structure of image texture, and it is not affected by illumination variations because of its invariance to monotonic gray level changes. It is robust to video compression since LBP describes the distribution of regional gray space, and compression does not change this relationship significantly. In view of its simplicity and effectiveness in image representation and classification, LBP and its variants have been applied in many research fields, such as facial image analysis and digital image/video forensics [12,13]. The major procedures of the proposed algorithm are as follows: (i) the motion vector (MV) of the background for each frame is computed to align video frames so as to realize the preprocessing of video captured by mobile camera; (ii) the coherence analysis on the spatial LBP operator between

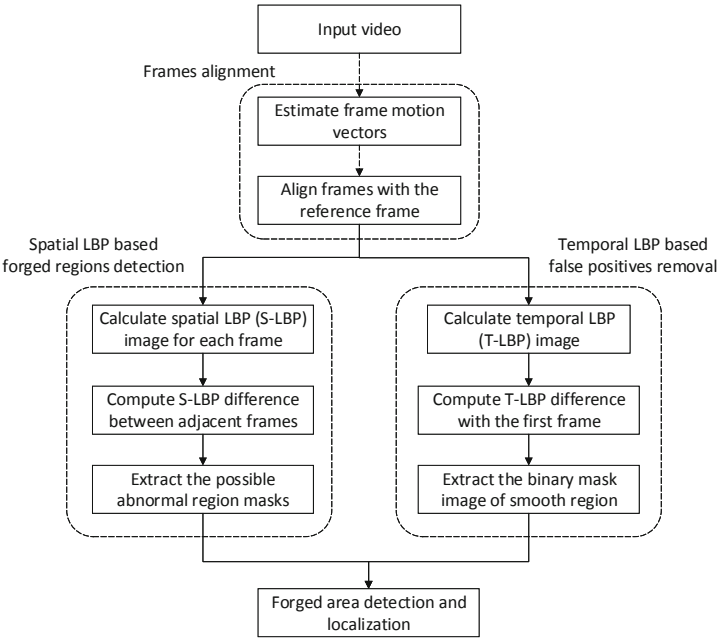

Fig. 1. Flowchart of our proposed method.

two adjacent frames is performed to find the possible forged region; (iii) the temporal LBP is utilized to remove false positives and the final abnormal region is located. Our method can be applied to videos taken by moving cameras. And the experimental results prove that it is effective and relatively robust to detect the forged region manipulated by well known inpainting methods such as TCP and ETS in video sequences.

The remainder of this paper is organized as follows. Section 2 gives the details of the proposed video forgery detection scheme based on spatio-temporal LBP coherence analysis. The experimental results are presented in Sect. 3. Finally, Sect. 4 summarizes the highlights and discusses the future work.

2 Proposed Method

The proposed method aims to expose the traces of object removal forgery in static and dynamic scene videos. Our idea is to detect the forged regions manipulated by TCP and ETS by means of finding the region with abnormal temporal correlation, because video subjected to such forgery exhibits unnaturally high or low correlation between region of successive video frames. As shown in Fig. 1, the proposed detection scheme consists of three major steps: (i) frames alignment, (ii) spatial LBP (S-LBP) based forged regions detection, and (iii) temporal LBP (T-LBP) based false positives removal. The details of the proposed method are given in following subsections.

2.1 Frames Alignment

In order to handle video motion caused by camera movement or shaking of the mobile phone, we adopt a simple block matching motion estimation algorithm to obtain the motion vector of each video frame to achieve frames alignment. In this paper, we use $\{F_1, F_2, \cdots, F_L\}$ to represent a video sequence V of length L, $L \in \mathbb{Z}^+$. And F_t represents the t^{th} frame, $\mathbb{Z}^+$ is the set of positive integers. It is obvious that the background motion vector (Vx_t, Vy_t) of t^{th} frame can be utilized as the t^{th} frame motion vector.

For computational efficiency, we first convert the video sequence from three-dimensional color space to two-dimensional grayscale space. Then each frame is divided into non-overlapping b blocks with $M \times N$ pixels, and the motion vector of the i^{th} block of t^{th} frame can be denoted by (vx_t^i, vy_t^i). We use the exhaustive search (ES) algorithm and mean absolute deviation (MAD) matching criterion for each block between successive frames F_{t-1} and F_t to find the most similar block, and then obtain the motion vector of each block. In typical applications, the area of foreground regions is usually much smaller than that of the background region. Based on this assumption, choose the most frequent (vx_t^i, vy_t^i) as the background motion of F_t as follows:

$$Vx_t = mode\{vx_t^1, vx_t^2, \cdots, vx_t^b\} \tag{1}$$

$$Vy_t = mode\{vy_t^1, vy_t^2, \cdots, vy_t^b\} \tag{2}$$

where $mode(\cdot, \cdot, \cdots, \cdot)$ denotes that the value with the highest frequency in parentheses will be selected as the result. After the motion vector of each frame is obtained, the pixels in frame F_t are shifted by the cumulative vector (Cx_t, Cy_t) of the motion vectors of all frames before F_t. Cx_t and Cy_t can be calculated as follows:

$$Cx_t = \sum_{j=1}^{t} Vx_j \tag{3}$$

$$Cy_t = \sum_{j=1}^{t} Vy_j \tag{4}$$

2.2 Spatial LBP Based Forged Regions Detection

The spatial LBP (S-LBP) based forged regions detection is performed on the aligned frames. In this section, S-LBP is defined in 3×3 window as shown in Fig. 2. We take the center pixel of the window as the threshold and compare the gray values of 8 adjacent pixels with it: if the surrounding pixel value is greater than the center, then the binary code of the corresponding position is 1, otherwise 0. Finally, the S-LBP coded frame SL of each original video frame is obtained. The definition of SL is given by Eqs. (5) and (6).

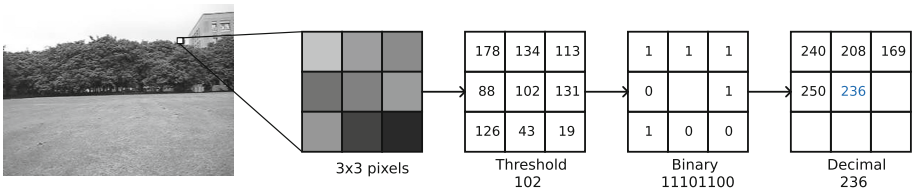

Fig. 2. The computation process of LBP.

$$SL(x_c, y_c) = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{5}$$

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & otherwise \end{cases} \tag{6}$$

where (x_c, y_c) represent the coordinates of center pixel, p denotes the serial number of the sampling point around (x_c, y_c), and g_c, g_p represent the gray value of (x_c, y_c) and its adjacent pixel respectively. P is the number of pixels around the center pixel, which is set to 8 here.

For analyzing the correlation between the previous frame F_{t-1} and the current frame F_t, we first calculate the frame difference S_d of two adjacent LBP frames. Then S_d is divided into non-overlapping blocks, and the number of zeros in the histogram vector for each block is counted. If a block is forged, the number of zeros in the block varies (increased or decreased) substantially depending on the forgery scheme (TCP or ETS). Figure 3 shows the average distribution of histograms of block-level (8×8 block) differences between every two consecutive LBP frames in three different cases. Note that the ordinates of the three figures are different. Obviously, the numbers of zeros and the distributions of histograms in forged region are significantly different from those of the original area. As a result, the forged region and non-forged one can be distinguished by analyzing the number of zeros Q in the histogram vector in each block of S_d. The preliminary classification is defined as follows:

$$Class_i = \begin{cases} 0, & T_1 < Q < T_2 \\ 1, & otherwise \end{cases} \tag{7}$$

where $Class_i$ denotes the binary classification mask of the i^{th} block, and a value of 1 indicates that the block has been forged. T_1 and T_2 are thresholds for dividing the forged region and the normal. Finally, the pre-classification mask image of every original video frame is obtained by combining these block-level binary mask. Since the large smooth areas like sky can also lead to abnormally high correlation between two adjacent frames and interfere with the detection result, we elaborate the scheme for removing the false positive areas in the next section.

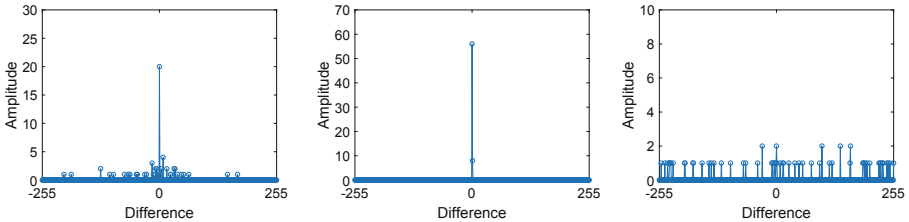

Fig. 3. The comparison of the histograms of block-level differences between every two consecutive LBP frames in three different cases: (a) normal region, (b) the block inpainted by TCP, and (c) the block inpainted by ETS. The abscissa represents the difference of pixels between adjacent LBP frames (ranging from −255 to 255), and the ordinate represents the number of zeros in the 8×8 block of S_d.

2.3 Temporal LBP Based False Positives Removal

In this section, temporal LBP (T-LBP) operator extended from the spatial domain is utilized to remove the false positives. The value of each pixel in the aligned frame obtained by Sect. 2.1 is computed by weighting the symmetric pixels within the range of 8 adjacent frames in temporal domain. That is, each T-LBP coded frame TL_t carries the information of video frames within 8 neighborhoods (16 frames in total). The mathematical definition of TL_t is as follows:

$$TL_t(x, y) = \sum_{r=1}^{R} s(G_{t-r}(x, y) - G_{t+r}(x, y))2^{r-1} \tag{8}$$

where R is the neighborhood radius in temporal domain, which is set to 8 here. $G_t(x, y)$ represents the gray value of pixel point whose coordinates are (x, y) in the t^{th} aligned frame. Figure 4 shows the pixel pairs and their weights in the process of calculating $TL_t(x, y)$. Thus, the TL sequence of length $L - 16$ consisting of LBP-coded frames with the same size as the original video frames are obtained.

Large smooth areas causing false alarms are found by means of extracting the regions that remain stable for a period of time in TL sequence. The specific method is described as follows: similar to the previous section, we first calculate the frame difference between each current frame TL_t and the first LBP-coded frame to obtain the difference sequence of length $L - 17$. Then each difference frame is divided into non-overlapping blocks, and the number of zeros in the histogram vector of each block is counted. Finally, we convert each difference frame to a binary image based on a proper threshold as follows:

$$Class_i' = \begin{cases} 1, & Q' > T_3 \\ 0, & otherwise \end{cases} \tag{9}$$

where $Class_i'$ denotes the binary classification mask of the i^{th} block, and a value of 1 indicates that the block belongs to the large smooth area. The binary mask

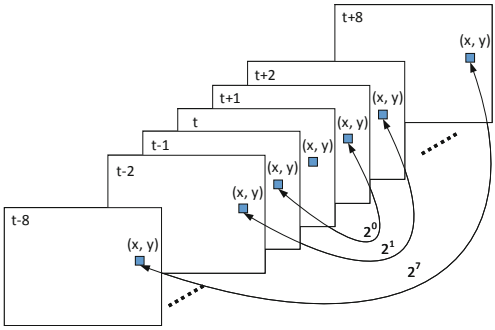

Fig. 4. Pixel pairs and their weights in the process of calculating $TL_t(x, y)$.

image of each difference frame is obtained by combining these block-level binary mask. And the mask image of the smooth region is obtained after OR operation and mathematical morphological processing as follows:

$$smooth = ((B_1 \bigcup B_2 \bigcup \cdots \bigcup B_{L-17}) \oplus E) \ominus E \tag{10}$$

where B_t is the binary image filtered by the threshold of each difference frame. $\bigcup$ is the logical OR operator. E is a structuring element. $\oplus$ and $\ominus$ represent the morphological close and open respectively. Finally, the false positives removal operation is performed on each pre-classification mask image given in Sect. 2.2 according to binary mask image of the smooth region, and the final binary classification image and the localization result are obtained, as shown in Fig. 5.

3 Experimental Results

To evaluate the performance of our method, twenty test video sequences were prepared for the experiments. We classify these videos into three groups according to their sources and the states of the video background: group I contains 7 test videos with still background which were obtained from SULFA data set [14], and the resolution of each frame is 320 × 240 pixels. Group II contains 8 test videos that we have taken with static camera and group III contains the

Fig. 5. The process of locating the forged region in a frame: (a) pre-classification mask image, (b) mask image of the smooth area in video, (c) final binary classification image, and (d) localization result.

remaining 5 videos with dynamic background taken by ourselves. The resolution of each frame in group II and III is 352×288 and the frame rate for all test videos is 30 fps. All the videos were forged by TCP and ETS inpainting methods respectively, and then re-encoded to H.264/AVC (with bitrates in the range of 1 Mbps to 5 Mbps) after the forgery. Through a large number of experiments, T_1, T_2 and T_3 are empirically set to 18, 63 and 30 for 8×8 blocks.

As shown in Table 1, the detection performance is measured by precision rate P, recall rate R, and F1-score $F1$, which are calculated as below:

$$P = TP/(TP + FP) \tag{11}$$

$$R = TP/(TP + FN) \tag{12}$$

$$F1 = (2 \times P \times R)/(P + R) \tag{13}$$

where TP denotes the number of correct detections, FP represents the number of false positives, and FN is the number of misses. Table 1 shows the average values of the experimental results for all videos in each group at 5 different bitrates. And it can be seen that the proposed method achieves high precision for both two video inpainting attacks, especially for TCP scheme. The performance of ETS tampered videos with a large amount of dynamic background is degraded because the errors of frames alignment operation make the authentic regions to be falsely classified as forged. Figure 6 shows the screenshots of the original frames, their inpainted frames forged by two inpainting schemes, and the corresponding localization results using the proposed method. The red blocks indicate the forged regions detected.

Table 1. Average performance of the proposed method for videos forged by TCP and ETS.

Group	TCP inpainting			ETS inpainting		
	Precision	Recall	F1	Precision	Recall	F1
I	0.9677	0.8937	0.9274	0.9483	0.8513	0.8969
II	0.9441	0.8640	0.9021	0.9243	0.8925	0.9077
III	0.9724	0.8623	0.9134	0.8271	0.7767	0.8002
Average	0.9614	0.8733	0.9143	0.8999	0.8402	0.8683

In addition, we make a comparison between the proposed approach and the existing methods presented by Hsu *et al.* [1] and Lin *et al.* [10], and the comparison results are shown in Table 2. It can be seen that our approach outperforms the other two algorithms and it achieves higher performance especially for ETS inpainting attack. There is no mechanism to remove false positive regions in the noise residual based method [1], so the performance of data set in group II with large smooth areas is obviously decreased. And since it is not available to dynamic background, we have not shown the relevant experimental results

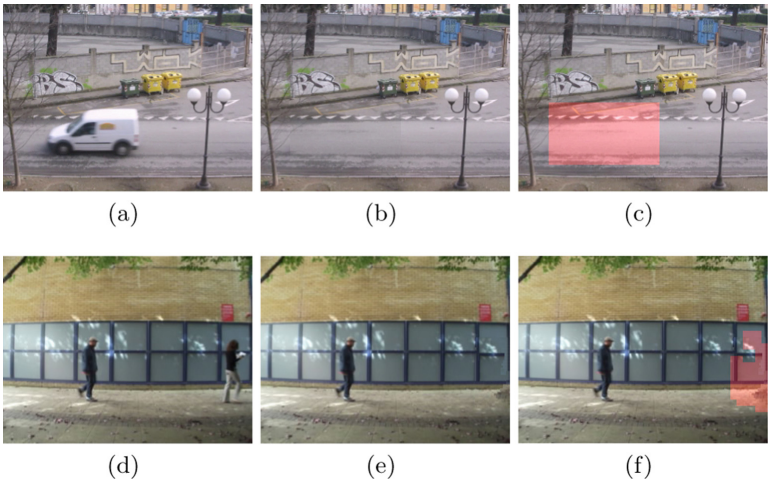

Fig. 6. Screenshots of the test video sequences: (a), (d) original frames, (b) the inpainted frame forged by TCP, (e) the inpainted frame forged by ETS, and (c), (f) the corresponding detection result. (Color figure online)

of group III. The performance of [10] drops significantly for videos forged by ETS inpainting compared with TCP since this method relies heavily on the edge detection of forgery region, and it is difficult to accurately extract the region boundary forged by ETS. The LBP operator and its variants used in our method are not affected by the change of illumination due to their invariance to monotonic gray level changes. In addition, the frames alignment operation enables video captured by the mobile camera to be detected.

Table 2. Comparison results between our method and two existing schemes presented by Hsu *et al.* [1] and Lin *et al.* [10].

Group	TCP inpainting			ETS inpainting		
	Hsu *et al.*	Lin *et al.*	Ours	Hsu *et al.*	Lin *et al.*	Ours
I	0.8672	0.9278	0.9677	0.8860	0.8521	0.9483
II	0.5448	0.9076	0.9441	0.8947	0.8148	0.9243
III	–	0.9311	0.9724	–	0.7238	0.8271
Average	0.7060	0.9222	0.9614	0.8904	0.7969	0.8999

In-depth analysis of the literatures revealed that the primary factors affecting the performance of inpainting detection techniques are the bitrates and compression quality of the test videos. Therefore, we present the forgery detection capabilities of these three forensic schemes for video sequences with bitrates in

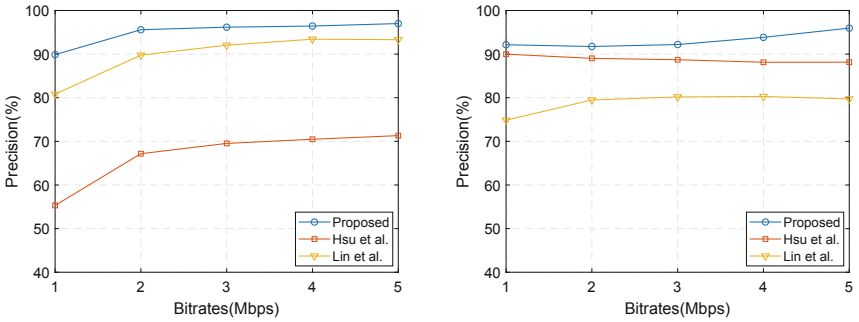

Fig. 7. Comparison of detection precision under different bitrates settings: (a) forged by TCP inpainting scheme, and (b) forged by ETS inpainting scheme.

the range of 1 Mbps to 5 Mbps as shown in Fig. 7. It can be seen that the spatio-temporal LBP based approach still has high precision rate in the case of decreasing the bitrate. This is because the LBP operator and its variants describe the distribution of regional gray space, which does not change significantly during the compression process.

4 Conclusion

In this paper, we have presented a detection and localization method for video object removal forgery based on spatio-temporal LBP coherence analysis. We first perform frames alignment to handle camera motion. Then we use spatial LBP operator making coherence analysis to find the possible abnormal areas. Finally, the temporal LBP operator is utilized to remove the authentic regions that are falsely classified as forged to locate the final forged areas. In our experiments, two video inpainting schemes (TCP and ETS) are used to simulate two different types of tampering processes for performance evaluation. The experimental results prove that our method can detect and locate the forged regions effectively and keep stability with respect to decreased bitrates. It can also be applied to videos taken by mobile cameras or handheld phones. However, great shaking and even slightly rotating of the forged video will cause unsatisfactory experimental results. The main reason is that the coherence analysis does not work well under the above conditions because the difference between two normal frames can be very large. In the future, we will explore ways to solve the problems above and improve the scope of the applicability.

Acknowledgements. This work was supported in part by the National Key Research and Development of China (2016YFB0800404), National NSF of China (61672090, 61532005), and Fundamental Research Funds for the Central Universities (2018JBZ001).

References

1. Hsu, C.-C., Hung, T.-Y., Lin, C.-W., Hsu, C.-T.: Video forgery detection using correlation of noise residue. In: 2008 IEEE 10th Workshop on Multimedia Signal Processing, pp. 170–174. IEEE (2008)
2. Chetty, G.: Blind and passive digital video tamper detection based on multimodal fusion. In: Proceedings of 14th WSEAS International Conference on Communications, Corfu, Greece, pp. 109–117 (2010)
3. Goodwin, J., Chetty, G.: Blind video tamper detection based on fusion of source features. In: 2011 International Conference on Digital Image Computing: Techniques and Applications, pp. 608–613. IEEE (2011)
4. Pandey, R.C., Singh, S.K., Shukla, K.K.: Passive copy-move forgery detection in videos. In: 2014 International Conference on Computer and Communication Technology (ICCCT), pp. 301–306. IEEE (2014)
5. Singh, R.D., Aggarwal, N.: Detection and localization of copy-paste forgeries in digital videos. Forensic Sci. Int. **281**, 75–91 (2017)
6. Wang, W., Farid, H.: Exposing digital forgeries in video by detecting duplication. In: Proceedings of the 9th Workshop on Multimedia & security, pp. 35–42. ACM (2007)
7. Bestagini, P., Milani, S., Tagliasacchi, M., Tubaro, S.: Local tampering detection in video sequences. In: 2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP), pp. 488–493. IEEE (2013)
8. Zhang, J., Su, Y., Zhang, M.: Exposing digital video forgery by ghost shadow artifact. In: Proceedings of the First ACM Workshop on Multimedia in Forensics, pp. 49–54. ACM (2009)
9. Li, L., Wang, X., Zhang, W., Yang, G., Hu, G.: Detecting removed object from video with stationary background. In: Shi, Y.Q., Kim, H.-J., Pérez-González, F. (eds.) IWDW 2012. LNCS, vol. 7809, pp. 242–252. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40099-5_20
10. Lin, C.-S., Tsay, J.-J.: A passive approach for effective detection and localization of region-level video forgery with spatio-temporal coherence analysis. Digit. Invest. **11**(2), 120–140 (2014)
11. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. **13**(9), 1200–1212 (2004)
12. Li, L., Li, S., Zhu, H., Chu, S.-C., Roddick, J.F., Pan, J.-S.: An efficient scheme for detecting copy-move forged images by local binary patterns. J. Inf. Hiding Multimedia Signal Process. **4**(1), 46–56 (2013)
13. Zhang, Z., Hou, J., Ma, Q., Li, Z.: Efficient video frame insertion and deletion detection based on inconsistency of correlations between local binary pattern coded frames. Secur. Commun. Netw. **8**(2), 311–320 (2015)
14. Qadir, G., Yahaya, S., Ho, A.T.S.: Surrey university library for forensic analysis (SULFA) of video content (2012)