Frames Selection Strategies for PRNU-Based Source Camera Verification of Stabilized Videos

Davide Vecchi Andrea Montibeller

University of Trento, Department of Information Engineering and Computer Science, Italy

ABSTRACT

In recent years, the proliferation of Electronic Image Stabilization (EIS) technology has significantly enhanced the quality of videos by mitigating the effects of camera motion. However, this process poses challenges for forensic analysis, particularly for identification of the acquisition source of a media. The conventional approach, relying on Photo Response Non-Uniformity (PRNU), encounters difficulties due to the transformations introduced during stabilization.

Building upon the foundational work presented in the paper "GPU-accelerated SIFT-aided source identification of stabilized videos" by colleagues Montibeller et al. [1], we delve into the exploration of alternative strategies with the objective of refining the selection of a concise yet effective subset of significant frames from each video within the dataset. This pursuit is driven by the objective of alleviating the computational burden associated with extracting PRNUs from all the frames, caused by the exhaustive grid search in the space of EIS inversion parameters.

Initially, we examine the Dense Optical Flow (DOF) RAFT algorithm, to overcome the limitations of SIFT features matching, in the context of estimating camera momentum and identifying less stabilized temporal frame segments. Additionally, we introduce an adaptive frame selection mechanism driven by Peak to Correlation Energy (PCE), aiming to optimize the identification of frames crucial for PRNU estimation.

The revised version of the original code [2] is available at: `https://github.com/davidevecchi/GPU-PRNU-SIFT`.

1. INTRODUCTION

1.1. Photo Response Non-Uniformity

Photo Response Non-Uniformity (PRNU) is a fundamental concept in digital image forensics, particularly in the field of digital camera identification [3]. It stems from the inherent imperfections in the manufacturing process of image sensors within digital cameras. These imperfections lead to subtle variations in the sensitivity of individual pixels to light, resulting in a unique and fixed distinctive mark for each camera sensor.

In essence, PRNU is a statistical high-frequency noise pattern that emerges in the form of tiny pixel-wise deviations from the expected response to uniform illumination. This distinctive pattern, embedded in images and video frames akin to a sensor's "fingerprint", remains consistent for a given camera over time, serving as a signature that can be extracted from the images produced by that specific device.

Forensic analysts leverage PRNU as a powerful tool for camera identification. By comparing the extracted PRNU pattern from an unknown image or video with the known PRNU of a reference camera, investigators can establish a link between the media and its originating source. This process is particularly valuable in attributing images or videos to specific cameras, aiding criminal investigations or validating the authenticity of digital content.

In the context of electronic image stabilization and video forensics, the accurate estimation of PRNU becomes crucial. However, challenges arise due to the transformations applied during electronic stabilization, which can distort the PRNU pattern.

1.2. Electronic Image Stabilization

EIS is a widely employed in-camera processing technique designed to enhance the visual quality of videos by reducing unwanted camera movements and oscillations. This is achieved through the application of complex transformations to individual frames, compensating for motion and resulting in smoother, more visually appealing footage.

While the benefits of EIS in improving video quality are evident, its implications for forensic analysis, specifically PRNU estimation, are non-trivial. The intricate transformations applied during the stabilization process introduce distortions that can compromise the integrity of the PRNU pattern. The core challenge arises in the task of reversing the transformations to reconstruct the original pixel-wise correspondences between the stabilized frame and the reference PRNU fingerprint of the known camera. Achieving the restoration is pivotal for accurate and reliable camera attribution in the presence of EIS-induced transformations.

Compounding the task is the closed-source policy adopted by camera manufacturers, impeding a comprehensive understanding of the underlying nature of the transformations, hence their reversion process. Nonetheless, a tracking system able to investigate the differences between the existing EIS implementations was proposed by Bellavia et al. [4], aiming to revert the transformations in controlled environments.

This poses a substantial obstacle, particularly when dealing with strongly stabilized videos that necessitate complex transformation inversion settings. The transformations may display locality, and they might not be uniformly applied at the frame level, but could involve spatially varying warps [5]. Conversely, in the case of weakly stabilized videos, frame-level affine transformations can suffice for PRNU alignment and registration [6].

1.3. Related and proposed methods

This paper investigates the strengths and limitations of state-of-the-art PRNU-based source verification methodologies for EIS videos and it delves into the exploration of novel strategies to overcome the existing challenges and to contribute to improving the accuracy and reliability of camera attribution in this context.

The foundational work by Montibeller et al. [1] serves as a primary reference, establishing a pipeline that aligns with previous approaches [4, 6, 7]. This general procedure encompasses four key stages:

1. Reference PRNU extraction: the fingerprint of the reference camera is computed by accumulating residual noise from a set of images captured using the known device.
2. Video frames extraction: a subset of relevant frames is selected from each video to be analyzed, employing diverse strategies tailored to each approach.
3. EIS inversion in spatial domain: consecutive frames undergo a co-registration process to align their scene content, aiming to mitigate and revert the impact of spatial distortions introduced by EIS.
4. PRNU matching: a grid search for transformation parameters, initiated from the inverted transformation computed in step 3, is conducted in the PRNU domain to achieve the optimal match with the reference fingerprint. The match is measured by the Peak to Correlation Energy (PCE) between the estimated PRNU of the current frame and the reference PRNU.

Our research mainly focuses on refining the second stage of the pipeline, specifically on improving the selection of an ideal subset of frames highly correlated with the reference PRNU. Objective of this process is to achieve a balance between high PCE values and a reduced cardinality of the selected set, in order to minimize the computational burden associated with the computation of the correlation of each frame. Our contributions introduce two distinct strategies: the first examines the advantages of leveraging a Dense Optical Flow (DOF) algorithm to identify the least stabilized Group Of Pictures (GOP) within the video. In contrast, the latter proposes a novel method based on an adaptive PCE-driven selection of frames exhibiting high correlation values.

The two approaches are explained respectively in Sections 2 and 3. Experimental analyses and results are reported and discussed in Section 4.

2. DOF MOMENTUM

2.1. Intuition

Previous strategies [1, 4, 7, 8] adopted feature detection and matching algorithms to obtain evidences about the stabilization process occurred in the scene content domain. While algorithms like SIFT are efficient and effective in generating a transformation matrix to register consecutive frames, they are subjected to limitations in finding the least stabilized GOP. Indeed, the set of keypoints these algorithms can model is tied to the local characteristics of the scene content. Notably, in regions characterized by a lack of texture or the absence of distinctive visual entities, traditional feature detection algorithms often encounter challenges in accurately generating and matching keypoints.

The underlying idea behind the selection of the least oscillating GOP, bounded by two intracoded frames (I-frames), is grounded in the assumption proposed in [1] that frames with minimal momentum –the global amount of motion between frames– experience reduced stabilization distortions.

The motivation behind embracing DOF, specifically the RAFT algorithm [9] instead of SIFT, for the GOP selection process, lies in its capability to express magnitude and direction of motion of each pixel within the frame, rather than only of a restricted set of keypoints.

Additionally, DOF allows distinguishing foreground elements from the background [10, 11, 12, 13, 14]. Background segmentation becomes particularly relevant in the context of reconstructing the high-frequency involuntary oscillations of the camera, that are compensated by the EIS. Supposedly, the estimation of camera movements can be refined by prioritizing the background over the foreground. Indeed, moving objects in the foreground of the scene should not be accounted for in the overall camera momentum, since their displacement is not tied to that of the video capture device.

To address challenges associated with varying displacement intensities induced by parallax, future investigations may also explore the integration of state-of-the-art depth estimation techniques. Notably, algorithms like [15] could complement the DOF, providing a comprehensive solution to handle scenarios where foreground moving objects and parallax effects influence the displacement patterns of the pixels within the video frames.

It's crucial to note that considerations regarding the depth of the scene are not yet integrated into the implementation and analysis proposed in this paper. The primary objective in this section is a straightforward substitution of the SIFT algorithm with RAFT in the least oscillating GOP selection process.

2.2. Formal procedure

Let $\Phi(t)$ be the dense optical flow computed between consecutive frames P_t and P_{t+1} as the pixel-wise displacement of pixels in the scene content domain. $\Phi(t)$ is a $n \times m \times 2$ matrix, where n and m are respectively the width and height of the frames and 2 represents the 2 Cartesian components $\langle \delta_x, \delta_y \rangle$ of the displacement of each pixels. The voxel $\phi(x, y, t) = \Phi(t)[x, y] = \langle \delta_x, \delta_y \rangle$ contains the optical flow components of the pixel p_{xy} at time t. Hence, magnitude of the flow can be computed for each voxel as the Euclidean norm of its components.

Moreover, the interest in reconstructing the subtle movements of the camera extends beyond the visual displacement between consecutive frames. The objective is to quantify the amount of oscillation at a specific time t. The underlying assumption is that a constant low-frequency movement of the scene content is not stabilized to the same extent as the high-frequency unintentional oscillations occurring during controlled camera movements. Therefore, pixel-wise difference between consecutive DOF volumes is computed. The resulting magnitude differences are then averaged over all the pixels of the frame to determine the overall camera momentum at time t:

$$\overline{\Delta}_t = \frac{1}{nm} \sum_{x=1}^{n} \sum_{y=1}^{m} \|\phi(x, y, t) - \phi(x, y, t-1)\|_2 \qquad (1)$$

By iterating this process for each frame in the video, the algorithm identifies the least oscillating GOP, represented by the pair of anchors $A = \langle a_0, a_1 \rangle$. The selection is based on minimizing the average of the momenta within the time range $[a_0, a_1]$, where a_0 corresponds to the index of the starting I-frame I_{a_0} of the GOP, and a_1 represents the index of the

subsequent I-frame I_{a_1} following I_{a_0}:

$$A = \underset{a_0,\, a_1}{\operatorname{argmin}} \frac{1}{a_1 - a_0 + 1} \sum_{t=a_0}^{a_1} \overline{\Delta}_t \tag{2}$$

The length of a GOP $(a_1 - a_0 + 1)$ remains almost fixed within a given video. Experimental tests indicate variations across devices, with GOP lengths typically spanning values of 24, 25, or 30 frames. The corresponding I-frame I_{A_0} represents the starting point of the successive frame-wise stabilization inversion procedures, which will be limited to the frames:

$$I_{A_0},\ P_{A_0+1},\ P_{A_0+2},\ ...,\ P_{A_1-1},\ I_{A_1}$$

where P_p, $p \in [A_0 + 1, ..., A_1 - 1]$ are predicted frames (P or B type) enclosed by the only two I-frames I_{A_0}, I_{A_1} of the GOP.

3. ADAPTIVE FRAME SELECTION

3.1. Rationale

In pursuit of an alternative approach for a faster and more effective frame selection, a novel strategy has been devised. The intuition has found theoretical support in previous researches that have explored the reliability of using different types of frames in estimating PRNU.

In [16], the authors found that intracoded frames (I type) are quantitatively more reliable than predicted frames (P and B types) for PRNU estimation and thus suggested weighting them more. This finding was essentially a consequence of the fact that I-frames typically undergo a lighter compression than other types of frames [17]. Moreover, in [6], source verification is performed by extracting the first 5-10 I-frames of each video, and then corresponding PRNU patterns are aligned with the reference PRNU pattern. Subsequently, the frames that yield a matching statistic above some predetermined PCE value are combined together to create an aligned PRNU pattern.

Our supplementary assumption posits that the likelihood of identifying a statistically significant correlation between a particular P-frame and the reference PRNU is heightened when an adjacent frame yields a notably high PCE value itself. This hypothesis is grounded in the intuition that high correlations are likely indicative of reduced stabilization distortions, thereby suggesting minimal camera oscillations which may extend also to neighboring frames. The concept of temporal locality in stabilization also seeks to address the inherent limitation associated with selecting an entire fixed-length GOP, wherein varying magnitudes of stabilization may occur. Moreover, detecting camera oscillations solely from the scene content domain, by employing algorithms such as SIFT or RAFT, is a non-trivial task in terms of accuracy and computational complexity.

The novel approach complements the original workflow, briefly described in Section 1.3, by incorporating insights about the reliability of different frame types and the concept of the temporal consistency in stabilization. The intuitions mainly translate in the frame selection process: instead of searching for the least oscillating GOP along the entire video, frames are selected in chronological order from the beginning of the video, by following an adaptive PCE-driven strategy.

Algorithm 1: Adaptive PCE-driven frame selection

Input: `hypothesis` (H1 = 1, H0 = 0)
Data: `video`
Data: `h0_pce_values` // only if `hypothesis` = 1

```
MAX_ACCEPTED ← 12

if hypothesis = 1 then
    MIN_PCE ← percentile(h0_pce_values, 95)
    TARGET_PCE ← max(h0_pce_values)
    EARLY_STOP ← 4
else
    MIN_PCE ← ∞            // only I-frames are accepted
    TARGET_PCE ← ∞
    EARLY_STOP ← ∞

I_index ← get_I_frames_index(video)
i ← 0                                   // I-frame index
p ← 0                                   // P-frame index
pce_array ← [ ]                         // pce of each frame
target_count ← 0       // counter of pce ≥ TARGET_PCE
registered ← null      // stores the co-registered frame

while (
  pce_array.length < MAX_ACCEPTED ∧
  target_count < EARLY_STOP ∧
  i < I_index.length
) do

    frame_idx ← I_index[i] + p
    frame ← get_frame(video, frame_idx)
    pce, registered ← compute_pce(frame, registered)

    if pce ≥ MIN_PCE then                // accept frame
        pce_array.push(pce)
        i, p ← next_frame_idx(i, p, I_index)
    else
        if p = 0 then            // always accept I-frames
            pce_array.push(pce)

        // skip to the next I-frame, reset registered
        i ← i + 1
        p ← 0
        registered ← null

    if pce ≥ TARGET_PCE then
        target_count ← target_count + 1

return pce_array
```

Algorithm 2: `next_frame_idx(i, p, index)`

```
if index[i] + p + 1 < index[i + 1] then
    p ← p + 1
else
    i ← i + 1
    p ← 0
return i, p
```

3.2. Algorithm

The evaluation of each model's accuracy follows a general approach wherein the analysis is conducted twice: once for hypothesis H0, comparing videos with a reference PRNU extracted from a distinct device, and once for hypothesis H1, comparing videos to the fingerprint of the same device.

In the novel approach, slightly different procedures are employed for H0 and H1. Under hypothesis H0, a predetermined and fixed number of I-frames (e.g. `MAX_ACCEPTED` $\leftarrow 12$) are selected in chronological order from the beginning of the video, and all the 12 corresponding PCE values are stored. No P-frames are involved in H0, and since there aren't consecutive frames in the selection, the EIS inversion process in the PRNU domain always commences from the original non-transformed position of the frame, rather than from the co-registered transformation parameters, as proposed in [1].

Conversely, hypothesis H1 adopts an adaptive PCE-driven frame selection procedure. If the current frame (I or P type) attains a PCE value higher than a predefined threshold, then the subsequent P-frame is selected for PCE computation. Should the PCE value of a P-frame exceed the threshold, its PCE is stored, and the analysis proceeds to the next P-frame. Otherwise, if the PCE value of a P-frame falls below the threshold, then it is discarded, and the algorithm advances to the subsequent I-frame. All PCE values of the I-frames are stored, regardless of their value, as in H0.

In the case of consecutive frames, the latter should be co-registered to the former in the scene content domain. The process aims to facilitate the initialization of an effective set of the transformation parameters, from which the EIS inversion parameters search in the PRNU domain starts. The procedure is entirely performed by the `compute_pce`(`frame`, `registered`) function, described and implemented in [1, 2]. The function accepts in input the current non-transformed frame and the immediately preceding frame, co-registered in turn with the frame preceding it. It returns the maximum PCE value detected during the grid search of the inversion parameters and the transformed current frame, co-registered with the `registered` frame. If `registered` is `null`, no co-registration is performed, as in H0.

The PCE threshold value necessitates to be accurately chosen to ensure robust generalization properties. Thereby, an optimal approach is to define it as a statistic of the correlation values computed for each frame under hypothesis H0. For instance, the `MIN_PCE` threshold is defined as the 95th percentile of these values. Another threshold, denoted as `TARGET_PCE`, assumes the maximum correlation value found under hypothesis H0. This value represents the target PCE used for early stopping. If a specified number of frames (e.g. `EARLY_STOP` $\leftarrow 4$) in a given video under hypothesis H1 exhibit PCE values exceeding this threshold, then the iteration for that video terminates in advance. This strategy serves a dual purpose: it reduces computational time and implicitly prioritizes higher correlation values by preventing the acceptance of additional low-confidence I-frames, which could diminish the average PCE value of the video. Additionally, basing the thresholds on statistics of the PCE values observed under hypothesis H0 promotes generalization across diverse datasets. Specifically, in the VISION dataset [18] employed for evaluation, these values are respectively `MIN_PCE` $= 32$ and `TARGET_PCE` $= 45$.

3.3. Discussion

The fact that in H0 no P-frames are accepted is motivated both by practical and theoretical motives. Indeed, H0 is computed as the first, so it doesn't have PCE values already computed for that specific dataset from which deriving the thresholds. Thus, thresholds are assigned to infinity in the code, to force the algorithm to accept all and only the first 12 I-frames. This practical limitation is supported by the assumption that under hypothesis H0 the PCE values are independent and identically distributed random variables, as they shouldn't really match with the reference PRNU of a different device. Therefore the assumption of temporal consistency of the correlations should not hold under H0. Additionally, since the threshold is defined as the 95th percentile, the number of possible P-frames to be selected and eventually accepted is negligible.

Conversely, under hypothesis H1, the fact of accepting all the I-frames is consistent with H0, mitigating the little differences in the two procedures. Moreover, experiments show that H1 procedure tends to accept more I-frames than P-frames. Since both types of frames are equally weighted, the unbalance in their number implicitly weights more I-frames.

The proposed procedure is faster for manifold reasons. Firstly, it doesn't require the prior GOP selection at all. This procedure is indeed highly computational demanding, as it has to scan the entire video heavy algorithms such as SIFT or RAFT. In fact, the novel approach can rely on a reduced number of frames (e.g. 12 instead of 24-30) to compute co-registration and correlation operations on. Lastly, the early stopping criterion reduces the computational time even further.

4. EXPERIMENTAL ANALYSES

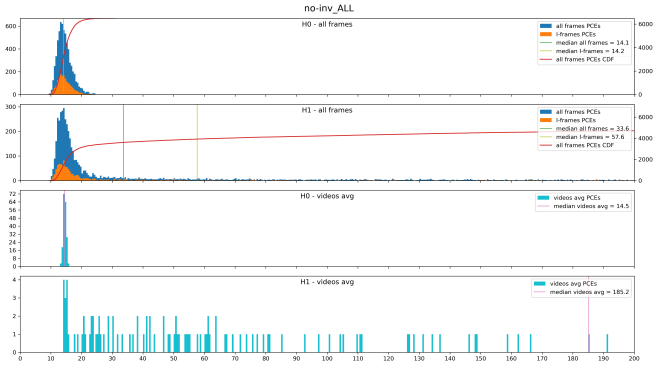
4.1. PCE histograms

To gain deeper insights and validate our hypotheses, analysis of the distribution of PCE values was conducted. Three distinct cases for both hypotheses H0 and H1 were investigated:

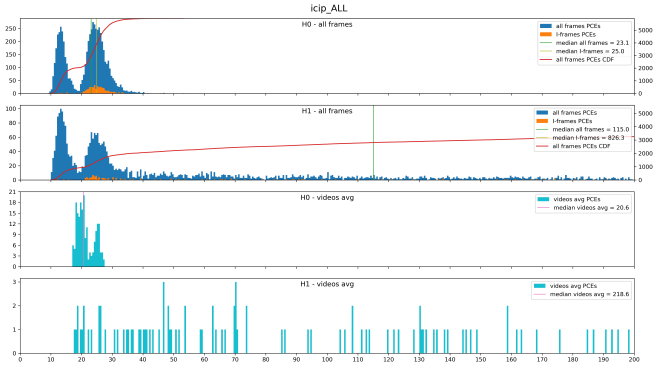
- NO-INV: a naive method that excludes the EIS inversion procedure, hence correlation is computed without performing the transformations;
- ICIP: the original implementation of the code [2];
- NEW: the novel adaptive PCE-driven approach (Algorithm 1).

The histograms presented in Figure 1 illustrate the frequency distribution of PCE values observed in a specific test. The first two graphs in each figure refer to single-frame counts. In these graphs, blue histograms depict the sum of the counts of both I and P frames, while orange histograms, where present, represent I-frames exclusively. Notably, in the NEW method, the distinction between I-frames and P-frames is not preserved due to the absence of saved frame-type information. Conversely, the two bottom graphs (in cyan), delineate the average PCE value across each video. Additionally, Cumulative Distribution Function (CDF) plots and median values are provided for comprehensive analysis.

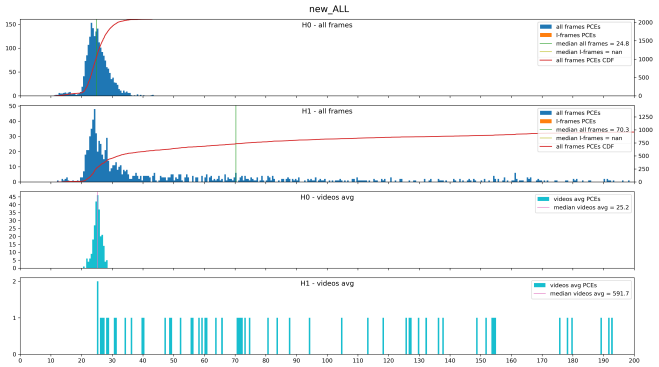
In NO-INV test, illustrated in Figure 1a, under hypothesis H0 all PCE values predominantly cluster within the range of [9,24]. Conversely, under H1, a notable extension of PCE values beyond 24 is evident. The CDF allows indeed a better visualization of the actual distribution of the frames whose correlation represents a notable match with the reference PRNU.



(a) *NO-INV: no EIS inversion applied.*



(b) *ICIP: original implementation of the GOP selection.*



(c) *NEW: adaptive PCE-driven frame selection.*

Figure 1: *PCE value counts for all individual frames (blue) and for average PCE of each video (cyan).*

In both the ICIP and NEW tests, general considerations can be advanced concerning the approximate ranges of the PCE values computed for individual frames.

- [0, 10): almost no frame obtained PCEs less than 10.
- [10, 20): frames within this range do not achieve higher PCE values following the EIS inversion process, suggesting the inability to attain a valid inversion. Notably, in the NEW test, P-frames are absent from this range, and only a minimal number of I-frames aren't promoted to the next range.
- [20, 40): frames within this range are typically promoted by the EIS inversion process. In the NEW test, P-frames are only observed under hypothesis H1 and when PCE values exceed approximately 32.
- [40, ∞): frames within this range demonstrate high correlation with the reference PRNU. This range is virtually empty under hypotheses H0, but vastly populated under H1.

Regarding the range [10,20), analysis of Figure 1b reveals that the majority of I-frames (orange) are promoted by the EIS inversion process to the subsequent range. This observation also extends to the NEW case, as illustrated in Figure 1c. Indeed, despite the absence of explicit frame type information in NEW histograms, the algorithm's design ensures that under hypothesis H0, all frames are of type I. Conversely, under hypothesis H1, P-frames are accepted only if their PCE value exceeds the threshold of approximately 32.

Finally, across all H1 scenarios, the median PCE values of exclusively the I-frames notably surpass those of all frames, predominantly of P type. This evidence confirms the higher reliability of intracoded frames compared to predicted ones in the context of source camera verification. It also substantiates the prevailing preference of the novel algorithm to select I-frames over P-frames. Quantitative results corroborating these observations are presented in Table 1.

		All frames	I-frames	#I/#P frames
H0	NO-INV	14.1	14.2	0.374
H1	NO-INV	33.6	57.6	0.371
H0	ICIP	23.1	25.0	0.070
H1	ICIP	115.0	826.3	0.076

Table 1: *Comparison of median PCE values of I-frames only versus all the frames, under hypothesis H0 and H1. Ratio of the number of I and P frames is shown in the last column.*

4.2. First I-frame and GOP

In the original (ICIP) implementation, the less oscillating GOP found using the SIFT algorithm could be discarded if its average momentum calculated between the two anchors frames of the GOP (i.e. the first and last I-frames of the GOP) exceeds a predetermined threshold, fixed for instance to 10 pixels. In such cases, the very first GOP of the video is selected instead. Additionally, the failure of the frame selection process could also prompt the selection of the initial GOP.

This heuristic wittingly exploit the characteristic of the EIS to be weaker at the beginning of the video. In fact, the first frame of a video is often less affected from stabilization as most motion smoothing methods correct motion with respect to a reference frame, which is typically the first frame [19]. In [20], it is reported that the first frame of videos in the VISION dataset [18] are mostly non-stabilized. In [6], authors verified that the first frame in 80% of the videos in that dataset yields a PCE value higher than 60. However, they also discovered that, for some camera models, it seems stabilization gets activated when the camera is set to video mode, even before recording starts. Consequently, these findings underscore the necessity of formulating more comprehensive strategies for reliable source verification, which do not primarily depend on the first frame. Such strategies must also accommodate potential post-processing cropping of the beginning of the video.

As a result, each analyzed approach has undergone rigorous testing using varied modalities:

- ALL: considers all frames within the video, thus initiating analysis with the very first frame, which is invariably of type I;
- I0: discards the very first I-frame, thus initiating analysis from the second frame, of type P;
- GOP0: excludes the initial GOP, thus initiating analysis from the second I-frame.

In the case of the ICIP-GOP0 method, the heuristic involving the momentum threshold is disregarded, ensuring consistent selection of the least oscillating GOP.

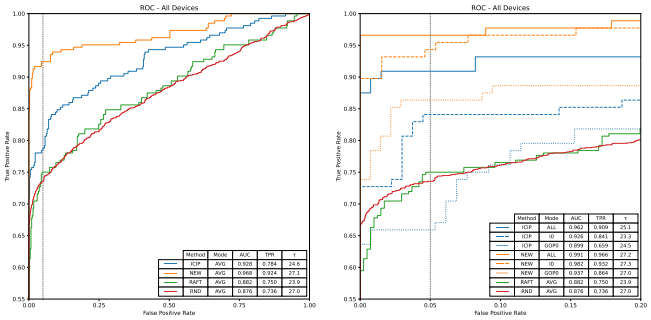
4.3. Evaluation procedure

The two proposed methods are compared with the state-of-the-art work presented in [1], along with a test based on random selection of frames, which serves as the base reference. The methods are denoted respectively as ICIP ([1]), RAFT (Section 2), NEW (Section 3) and RND (random selection). Random test is performed replacing the frame selection process by choosing 12 or 24 random frames from the current video. It is repeated for all the modalities ALL, I0 and GOP0, eventually discarding the first I-frame or the first GOP. Multiple iterations of the random tests are conducted to ensure robust statistical analysis.

In addition to the modalities ALL, I0 and GOP0 described in Section 4.2, another mode denoted as AVG is defined as the union of the results of the three tests, hence providing a form of composite or averaged result. Notably, due to the stochastic nature of frame selection in the RND method, the impact of discarding the first I-frame or GOP is less pronounced compared to ICIP and NEW methods. Hence, the RND method is solely presented in AVG mode. Similarly, the RAFT method, which consistently selects the least oscillating GOP without resorting to momentum thresholding heuristics, is also exclusively displayed in the AVG mode. Indeed, it is statistically less probable that the first GOP is selected, thus the difference between ALL and the other two modalities is negligible.

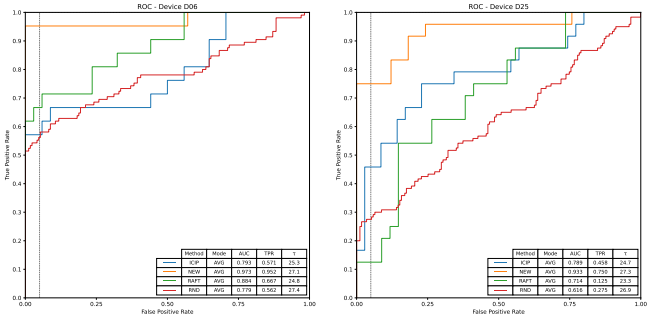
The experiments are conducted using the VISION [18], comprising 88 videos captured by 12 distinct devices, primarily iPhones, with the exception of one device (D25) manufactured by OnePlus. Only horizontal videos taken from “indoor” and “outdoor” environments, featuring camera movements labeled as “move” and “panrot” are considered. Videos categorized with “flat” or “still” tags, or featuring vertical rotation, are discarded. Under hypothesis H0, for each of the 12 reference PRNUs, one for each device, 10 videos from devices distinct from the reference PRNU are randomly selected for comparison. Conversely, under hypothesis H1, each reference fingerprint is evaluated against all videos from the corresponding device.

Performance evaluation encompasses computational cost analysis and Receiver Operator Characteristics (ROC) assessment, with a focus on Area Under the Curve (AUC) and True Positive Rates (TPR) at a fixed False Positive Rate (FPR $\approx$ 0.05). Computational experiments are conducted on a desktop computer equipped with 32GB RAM, Intel(R) Core(TM) i7-13700K 13th Gen 3.40GHz CPU, and NVIDIA GeForce RTX 4090 24GB GPU. A detailed test procedure is outlined in [1], providing comprehensive insights into the experimental setup and methodology.



(a) *Average TPRs.* **(b)** *ALL, I0, GOP0.*

Figure 2: *ROC curves of all the devices. TPR and PCE threshold τ refer to the fixed FPR $\approx$ 0.05.*



(a) *Device D06.* **(b)** *Device D25.*

Figure 3: *ROC curves of the two critical devices.*

4.4. Experimental results

The Receiver Operating Characteristic (ROC) curves for the methods under investigation are presented in Figure 2. In particular, Figure 2a depicts the ROC curves for the AVG mode, which combines the PCE values from the ALL, I0, and GOP0 modalities. Notably, the novel approach based on adaptive PCE-driven frame selection, denoted as NEW, demonstrates superior performance in terms of both AUC and TPR. This evidence underscores the efficacy of the proposed algorithm in source camera verification tasks.

Figure 2b further elucidates the impact of discarding the first I-frame and the first GOP, focusing on the FPR range [0, 0.2]. Notably, at low FPR values, ICIP-GOP0 and RAFT methods exhibit significant overlap with the RND test, indicating comparable performance to random selection. This evidence underscores the limited reliability of selecting an entire GOP distinct from the initial one. Moreover, each mode on which the NEW method is tested largely outperforms the corresponding mode of the ICIP approach.

Particular emphasis is placed on devices D06 (Apple iPhone 6) and D25 (OnePlus A3000), which in [1] exhibited lower TPR as compared to other state-of-the-art approaches [20, 21]. Figure 3 reveals that NEW method consistently demonstrates high reliability for these devices. A comprehensive comparison of TPR across each device is presented in Figure 4, further corroborating its effectiveness across diverse device types.

Finally, Figure 5 provides insights into the computational cost of the three algorithms, quantified as the total execution hours for running both hypothesis H0 and H1. The time advantage afforded by the NEW method is primarily stems from

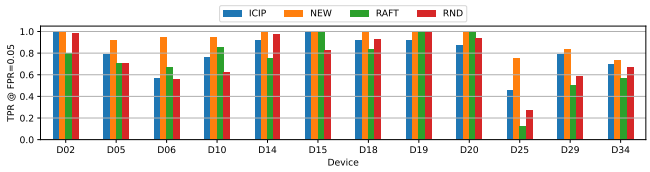


Figure 4: *True Positive Rates of each device, at a fixed False Positive Rate FPR ≈ 0.05.*

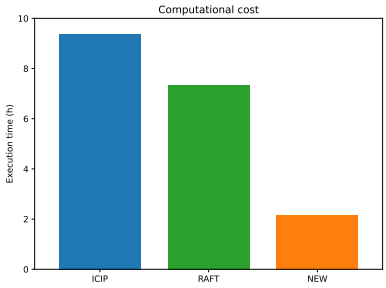


Figure 5: *Sum of H0 and H1 execution times.*

the exclusion of the prior selection of the least oscillating GOP, whose computational overhead is dominant. Contribution to the efficiency gain is also provided by the early stopping criterion.

5. DISCUSSION AND CONCLUSIONS

This study introduces two novel frame selection strategies aimed at improving the source camera verification methodology established by Montibeller et al. [1]. The first strategy involves a straightforward substitution of the feature matching SIFT algorithm with the Dense Optical Flow RAFT algorithm for selecting the least oscillating Group Of Pictures. Conversely, the second strategy proposes a radical paradigm shift by replacing the conventional GOP selection with an adaptive PCE-driven approach for optimal frame selection.

The DOF approach leverages the full pixel displacement information of each frame of the video, compared to the limited subset of keypoints for pairs of frames utilized by SIFT. Although this theoretical advantage, empirical results indicate minimal improvements. This outcome underscored the necessity for alternative frame selection methodologies beyond those based on choosing full-length GOPs. Consequently, other hypotheses involving depth estimation were set aside in favor of exploring innovative approaches.

The intuitions that laid the foundation for the second approach stem from two key insights: the heightened reliability of intra-coded frames compared to predicted frames [16] and the temporal locality and consistency inherent in EIS. Indeed a full-length GOP typically comprise a small fraction of I-frames relative to P-frames, ranging from 2 I-frames for every 23 to 29 P-frames. Moreover, diverse degrees of stabilization could occur within it, compounding the EIS inversion process. Building upon these insights, the adaptive PCE-driven frame selection algorithm emerges as a robust solution, showcasing superior performance in both identification accuracy and computational efficiency. Empirical findings from experiments conducted on stabilized videos sourced from the VISION dataset [18] surpass the capabilities of previous methodologies, underscoring the efficacy of the proposed approach.

References

[1] Andrea Montibeller et al. “GPU-accelerated SIFT-aided source identification of stabilized videos”. In: *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2022, pp. 2616–2620.

[2] Andrea Montibeller and Stefano Dell’Anna. *GPU-accelerated SIFT-aided source identification of stabilized videos*. `https://github.com/AMontiB/GPU-PRNU-SIFT`.

[3] Jan Lukas, Jessica Fridrich, and Miroslav Goljan. “Digital camera identification from sensor pattern noise”. In: *IEEE Transactions on Information Forensics and Security* 1.2 (2006), pp. 205–214.

[4] Fabio Bellavia et al. “Experiencing with electronic image stabilization and PRNU through scene content image registration”. In: *Pattern Recognition Letters* 145 (2021), pp. 8–15.

[5] Feng Liu et al. “Content-preserving warps for 3D video stabilization”. In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 2023, pp. 631–639.

[6] Enes Altinisik and Hüsrev Taha Sencar. “Source camera verification for strongly stabilized videos”. In: *IEEE Transactions on Information Forensics and Security* 16 (2020), pp. 643–657.

[7] Fabio Bellavia et al. “PRNU pattern alignment for images and videos based on scene content”. In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pp. 91–95.

[8] Sebastiano Battiato et al. “SIFT features tracking for video stabilization”. In: *14th international conference on image analysis and processing (ICIAP 2007)*. IEEE. 2007, pp. 825–830.

[9] Zachary Teed and Jia Deng. “Raft: Recurrent all-pairs field transforms for optical flow”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer. 2020, pp. 402–419.

[10] Junjie Huang et al. “Optical flow based real-time moving object detection in unconstrained scenes”. In: *arXiv preprint arXiv:1807.04890* (2018).

[11] Jing Ding et al. “A Novel Moving Object Detection Algorithm Based on Robust Image Feature Threshold Segmentation with Improved Optical Flow Estimation”. In: *Applied Sciences* 13.8 (2023), p. 4854.

[12] Wenlong Zhang, Xiaoliang Sun, and Qifeng Yu. “Moving object detection under a moving camera via background orientation reconstruction”. In: *Sensors* 20.11 (2020), p. 3103.

[13] Ibrahim Delibasoglu et al. “Motion detection in moving camera videos using background modeling and FlowNet”. In: *Journal of Visual Communication and Image Representation* 88 (2022), p. 103616.

[14] Arati Kushwaha et al. “Dense optical flow based background subtraction technique for object segmentation in moving camera environment”. In: *IET Image Processing* 14.14 (2020), pp. 3393–3404.

[15] Lihe Yang et al. “Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data”. In: *arXiv:2401.10891* (2024).

[16] Wei-Hong Chuang, Hui Su, and Min Wu. “Exploring compression effects for improved source camera identification using strongly compressed video”. In: *2011 18th IEEE International Conference on Image Processing*. IEEE. 2011, pp. 1953–1956.

[17] Enes Altinisik, Kasim Tasdemir, and Husrev Taha Sencar. “Mitigation of H. 264 and H. 265 video compression for reliable PRNU estimation”. In: *IEEE Transactions on information forensics and security* 15 (2019), pp. 1557–1571.

[18] Dasara Shullani et al. “Vision: a video and image dataset for source identification”. In: *EURASIP Journal on Information Security* 2017.1 (2017), pp. 1–16.

[19] Matthias Grundmann, Vivek Kwatra, and Irfan Essa. *Cascaded camera motion estimation, rolling shutter detection, and camera shake detection for video stabilization*. US Patent 9,635,261. Apr. 2017.

[20] Sara Mandelli et al. “Facing device attribution problem for stabilized video sequences”. In: *IEEE Transactions on Information Forensics and Security* 15 (2019), pp. 14–27.

[21] Sara Mandelli et al. “A modified Fourier-Mellin approach for source device identification on stabilized videos”. In: *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2020, pp. 1266–1270.