Pattern Recognition Letters
journal homepage: www.elsevier.com

Experiencing with electronic image stabilization and PRNU through scene content image registration

Fabio Bellavia[a], Marco Fanfani[b], Carlo Colombo[b,**], Alessandro Piva[b]

[a]*Dipartimento di Matematica e Informatica, Università degli Studi di Palermo, Via Archirafi 34, 90123, Palermo, Italy*
[b]*Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Firenze, Via di S. Marta 3, 50134, Firenze, Italy*

ABSTRACT

This paper explores content-based image registration as a means of dealing with and understanding better Electronic Image Stabilization (EIS) in the context of Photo Response Non-Uniformity (PRNU) alignment. A novel and robust solution to extrapolate the transformation relating the different image output formats for a given device model is proposed. This general approach can be adapted to specifically extract the scale factor (and, when appropriate, the translation) so as to align native resolution images to video frames, with or without EIS on, and proceed to compare PRNU patterns. Comparative evaluations show that the proposed approach outperforms those based on brute-force and particle swarm optimization in terms of reliability, accuracy and speed. Furthermore, a tracking system able to revert back EIS in controlled environments is designed. This allows one to investigate the differences between the existing EIS implementations. The additional knowledge thus acquired can be exploited and integrated in order to design and implement better future PRNU pattern alignment methods, aware of EIS and suitable for video source identification in multimedia forensics applications.

1. Introduction

Photo Response Non-Uniformity (PRNU) is a unique, fixed noise pattern generated during the acquisition process by any digital sensor (Lukas et al., 2006). This makes PRNU ideal to develop robust methods for source attribution in image forensics (Chen et al., 2008). The PRNU pattern is extracted pixelwise in order to derive the fingerprint of a device, implying that PRNU patterns are best generated and compared at native camera resolution (Shullani et al., 2017). Due to their high sensitivity to pixel misalignments, PRNU patterns become particularly difficult to compare when the source images have been warped by the device internal acquisition post-processing. For this reason, the reliability of PRNU-based source attribution techniques on videos acquired with Electronic Image Stabilization (EIS) is strongly decreased. Indeed, video frames processed by EIS are typically obtained from a scaled, translated and/or rotated portion of the full sensor area in order to compensate for camera shakes and improve the final quality (Grundmann et al., 2011). Notice that the same does not apply with Optical Image Stabilization (OIS), that instead dynamically accommodates the lens, leaving the sensor response untouched. In the general case, no specifications about the EIS algorithm and the video frame transformation parameters are available from the manufacturer, making it difficult to revert back the geometrical transformation applied by EIS as a preprocessing step before perfoming PRNU-based source attribution. Current approaches attempt to find an accurate estimate of the EIS transformation by maximizing the PRNU correlation in terms of Peak-to-Correlation-Energy (PCE) either by brute-force search (Taspinar et al., 2016; Iuliani et al., 2019) or, more recently, by particle swarm (Mandelli et al., 2019) and other specific parameter search space sub-sampling and optimizations (Altinisik and Sencar, 2020). This kind of approaches can be computationally expensive, not sufficiently accurate, or demand some a priori knowledge to meet the accuracy requirements.

This paper investigates novel uses of scene content image registration to deal with EIS and PRNU. The contribute is twofold:

- A novel and robust solution, first outlined in Bellavia

**Corresponding author: Tel.: +39-329-560-3192
e-mail: carlo.colombo@unifi.it (Carlo Colombo)

et al. (2019), is designed for aligning the PRNU patterns extracted from any two output formats of a given device (i.e. photos or videos at various resolutions). Differently from previous approaches, after acquiring the same static scene in each output format, the transformation relating two different formats is found by keypoint descriptor matching (Szeliski, 2010) on the image scene content. Registration refinement by maximizing the PRNU correlation over a limited parameter search space can also be integrated to improve accuracy further. The approach does not take into account possible rotations when EIS is on. According to the experimental evaluation, the proposed solution is more reliable, accurate and faster than the state-of-the-art approaches. Furthermore, the experimental evidence has shown that PRNU pattern registration depends only on the device model, and not on the device exemplar at hand. This implies that having access to a single device for each model of interest is sufficient to estimate, with the proposed approach, the PRNU pattern transformation for any other device of the same model. Model-related transformations can then be collected into a database and employed for practical applications involving the PRNU-based analysis of videos.

- A new method to revert back (i.e., to estimate and remove from the images by map inversion) EIS frame warping in a controlled environment is devised and discussed. This is made possible by tracking points on a physical grid integral with the acquisition device, so that each tracked grid point identifies a unique pixel in the original sensor matrix. The physical support used for this aim is named "Alvaro," since the method resembles the act of peeping through a keyhole, a gag often performed in the 1970s by the Italian comic actor Alvàro Vitali. Notwithstanding some limitations, to be discussed later in the paper, this is currently the only method able to produce reliable information on the EIS frame manipulation done by the device built-in hardware, which allows one to get a deeper understanding of the specific EIS algorithms used by a particular device model. Disclosing this kind of knowledge can contribute significantly to take into account further elements in the design of more robust and efficient PRNU alignment strategies in the presence of EIS.

The rest of the paper is organized as follows. Related work is presented in Sec. 2. Scene content based PRNU registration is described and evaluated in Sec. 3. The Alvaro device and its associated estimation method is introduced together with its experimental results in Sec. 4. Conclusions and future work are discussed in Sec. 5.

2. Related work

PRNU is a unique high-frequency artifact arising during the device acquisition process, that has shown to be a valid tool for addressing the source attribution problem of digital photos (Chen et al., 2008). Flat scenes, mainly containing low-frequency signal content, are often preferred for extracting the PRNU reference fingerprint, as they reduce PRNU signal contaminations due to scene content. Extending PRNU-based methods from photos to videos is not straightforward, due to the inferior reliability of the PRNU signal on videos, whose frames have lower resolutions and stronger compression ratios than their photo counterparts (Mandelli et al., 2019). In the case of videos, it is preferable to extract the reference fingerprints from photo images obtained at native sensor resolution, as they can preserve better the original source signal (Iuliani et al., 2019). For similar reasons, only I-frames are usually employed for PRNU based verification on videos (Taspinar et al., 2016; Iuliani et al., 2019; Mandelli et al., 2019). Nevertheless, the recent H.264 and H.265 codecs can introduce intra-frame compression in I-frames that, in the case of flat content, gives rise to high compression rates but, as a side effect, also to strong PRNU degradations (Kouokam and Dirik, 2019). Moreover, an accurate alignment between the full resolution fingerprint and the video frames is needed, so as to compensate for the fixed scaling and cropping on the video frame with respect to the native image, introduced by the device to meet strict video stream computational constraints. EIS makes things even more complicated, since each frame can undergo a further distinct and unknown geometric transformation to cope with camera handshaking and rolling shutter (Grundmann et al., 2011). In the most general case, affine warping is used to model EIS frame manipulation. Nevertheless, recent experimental analyses suggest that in many circumstances only the transformation relating the different acquisition formats (i.e., the scale) is actually relevant, and one can ignore the unknown frame distortions introduced by EIS (Mandelli et al., 2019). This happens because EIS softwares usually bring back frames to their original unaltered positions as EIS becomes inactive, and larger parts of a video are composed by frame sequences obtained by smooth camera paths, i.e., by movements where the camera either stays still or moves smoothly enough as not to trigger EIS.

State-of-the-art PRNU alignment solutions for videos work by searching for the PRNU pattern transformation which maximizes the PRNU correlation between the fingerprint and the video frame under test, echoing previous works dealing with image PRNU alignment in case of digital zooming (Goljan and Fridrich, 2008) and other digital post-processing manipulations such as lens distortion correction (Goljan and Fridrich, 2012, 2013) and seam-carving rescaling (Karaküçük et al., 2015). Included transformations are translation, scale and rotation. The best translation (and implicitly the best cropping) can efficiently be found in the frequency domain when no other transformations are present, but adding scale and rotation significantly increases the search space complexity. Except for (Höglund et al., 2011), the first work dealing with PRNU on stabilized videos by compensating for frame translations only, the search in the parameter space was usually carried out by brute-force (Taspinar et al., 2016; Iuliani et al., 2019). Recently, particle swarm optimization replaced brute-force search, yielding a faster and smarter search approach (Mandelli et al., 2019). Other search space improvements have also been explored, such as hierarchical search space sampling and partial fingerprint usage (Altinisik and Sencar, 2020). Nevertheless, these last kinds of op-

timizations can be slow for some applications, and strongly depend on human skill to define a parameter setup in order for the algorithms to work properly (for an experimental assessment of this claim, see Sec. 3.2).

Keypoint image matching has a long history in computer vision (Szeliski, 2010). Although it has evolved through time, its main core remains almost unaltered and still today offers one of the best solutions for scene tracking and registration in many application contexts, such as three-dimensional reconstruction (Jin et al., 2020; Mur-Artal et al., 2015), image stitching (Zaragoza et al., 2013) and large-scale image retrieval (Zheng et al., 2018). In its essence, keypoint image matching is aimed at obtaining a set of sparse correspondences between two or more images to model the transformation between them. Three main steps can be identified for this purpose: (1) the extraction of keypoints, i.e. of distinctive yet repeatable characteristic points on the images; (2) the computation of local descriptors which encode the distinctive features of the image region surrounding each keypoints; (3) the actual matching between keypoints according to their descriptor similarity, often in conjunction with some robust outlier rejection.

3. Scene content PRNU alignment

3.1. Method description

The first step of the proposed approach is to hold the device still and acquire images of a static scene using the available photo and video formats. The native full resolution photo serves as reference for the sensor grid, which the other image formats must be mapped into. It may be argued that this acquisition step would be unpractical in many application scenarios, should one have to repeat the procedure for each specific device. However, as shown later in the experimental section, the underlying PRNU pattern transformation only depends on the device model and not on the device exemplar at hand. This implies that, once estimated for one device, the same transformation holds for any other device of the same model. Hence, source attribution inquiries where the source device model is known (e.g., for having been produced as forensic evidence) can be successfully resolved by means of the proposed approach. Our approach can be effective also to solve blind source attribution problems, where a set of videos has to be clustered according to their unknown source devices. In fact, provided that a large enough database of transformation parameters for the different device models has been built with the proposed approach, parameter search can be quickly carried out, being limited to only neighborhoods around each of the database entries. Figure 1a-b shows an example of the above acquisition step. In order to improve the registration accuracy, the scene must be on focus and include discriminative patterns distributed across the whole image area. In the case of videos, only I-frames are considered and, when available, acquired images are taken using remote or vocal controls in order to avoid any accidental misalignments due to camera shakes.

In order to register an output format to the reference, corner-like keypoints are extracted with the HarrisZ detector (Bellavia et al., 2011) and are matched with the SIFT-like sGLOH2 local image descriptor (Bellavia and Colombo, 2018). Given the initial set of correspondences, the unknown image transformation is estimated in a robust way using RANndom SAmple Consensus (Fischler and Bolles, 1981). An affine warping model is used that includes scale and translation changes, but not rotations. The scale factor is the most important parameter and it is fixed for any device output format, even in case of EIS (see later the experimental results with Alvaro). If EIS is off, then the translation is fixed. Conversely, when EIS is on, translation can be easily recovered by PCE maximization, provided that frame rotations are not involved. An example of registration is shown in Fig. 1c-d: Notice that video frames cover a smaller portion when EIS is on, in order to compensate for translations and rotations while avoiding missing image spots from areas not covered by the camera sensor.

RANSAC estimation requires to set the inlier reprojection error ϵ, indirectly setting the degree of uncertainty in the final model. According to this observation, the transformation estimated so far can be refined through an exhaustive search over a small set of allowable scales, translations and rotations, operating analogously to other PRNU pattern alignment approaches. Specifically, the PRNU correlation in terms of PCE is evaluated over a limited set of transformations. Warping transformation refinement requires to extract the reference PRNU pattern and the warped PRNU pattern from flat scenes (i.e., with uniform color content), that will be used to compare PCEs. The more images are used to compute the PRNUs, the more accurate will be the refinement.

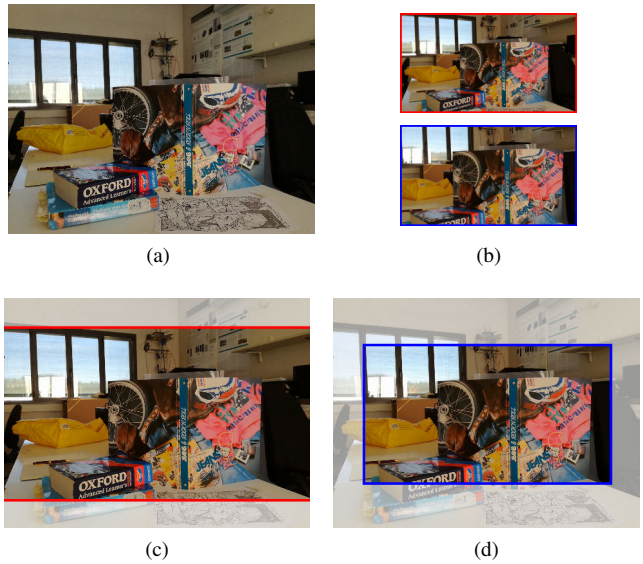

Fig. 1: Static scene image registration on a Huawei P9 Lite smartphone. The native full resolution photo (a) is used as reference to register the corresponding video frame in case EIS off (b,top) and on (b,bottom) using image keypoint matching. The final aligned video frames superimposed on the reference image are shown in (c) and (d), respectively. All images are scaled according to their resolution. The reference image on (c) and (d) is shaded for a better visual comparison.

A given length l on the reference image scales to $l' = l \times s$ on the warped image according to the initial scale factor s. The reprojection error threshold is experimentally set to $\epsilon = 4$ (i.e., the estimated average keypoint localization error) so that the

Table 1: PRNU registration evaluation results (see text for details).

Device model	Recording mode		Scale					PCE					Time (sec)				
			G	G_r	G_m	P	P_r	G	G_r	G_m	P	P_r	G	$G_r - G$	G_r	P	P_r
Samsung Galaxy A3	📷 → 🎥	μ	1.6993	1.7001	1.7001	2.3125	2.2454	5826	7746	7746	1691	1358	18	44	62	524	439
		σ	–	0.0000	–	0.6059	0.5941	1426	1922	1922	2289	2191	–	0	0	100	82
		min	–	1.7001	–	1.6976	1.6967	656	820	820	42	42	–	43	61	387	341
		max	–	1.7001	–	2.9920	3.0000	6905	9176	9176	6433	8627	–	45	63	658	606
Samsung Galaxy S7 (1st device)	📷 → 🎥	μ	2.1000	2.0997	2.0997	1.7740	2.1976	1168	1233	1233	456	521	34	73	107	411	404
		σ	–	0.0000	–	0.7310	0.4968	298	313	313	473	527	–	1	1	63	72
		min	–	2.0997	–	0.5000	1.0000	434	478	478	37	37	–	73	107	329	291
		max	–	2.0997	–	3.0000	3.0000	1820	1920	1920	1453	1701	–	77	111	674	604
	📷 → 📴	μ	1.7485	1.7494	1.7499	1.3374	2.0793	212	347	336	137	117	33	81	114	384	399
		σ	–	0.0015	–	0.7728	0.6922	237	429	429	248	130	–	0	0	59	73
		min	–	1.7448	–	0.5000	1.0051	27	31	25	37	37	–	80	113	325	289
		max	–	1.7526	–	3.0000	3.0000	1255	2249	2249	1457	563	–	82	115	658	580
	📴 → 📴	μ	0.8325	0.8332	0.8333	2.5241	0.7762	1461	2700	2620	319	1648	8	14	22	552	44
		σ	–	0.0012	–	0.7787	0.1147	1533	2980	2945	851	1875	–	0	0	120	1
		min	–	0.8293	–	0.8297	0.5000	23	29	22	47	34	–	13	21	261	43
		max	–	0.8363	–	3.0000	0.9720	6749	12531	12392	4182	6004	–	14	22	669	48
Samsung Galaxy S7 (2nd device)	📷 → 📴	μ	2.0924	2.0966	2.0982	1.2509	1.8340	31	488	486	381	603	39	83	122	359	347
		σ	–	0.0029	–	0.7366	0.5014	5	360	362	522	527	–	0	0	28	32
		min	–	2.0895	–	0.5000	1.0000	26	32	26	39	38	–	82	121	326	285
		max	–	2.0982	–	2.1046	2.5687	49	925	925	1674	1559	–	84	123	399	393
	📷 → 📴	μ	1.7512	1.7515	1.7499	1.1065	1.6360	104	125	105	79	93	38	81	119	364	341
		σ	–	0.0026	–	0.5700	0.5861	93	98	92	50	78	–	0	0	17	53
		min	–	1.7471	–	0.5086	1.0000	28	32	27	45	35	–	81	119	338	281
		max	–	1.7552	–	1.7571	3.0000	409	431	361	226	355	–	82	120	399	522
	📴 → 📴	μ	0.8372	0.8356	0.8344	2.2775	0.7446	60	152	130	65	123	9	14	23	489	44
		σ	–	0.0018	–	0.8528	0.1331	34	130	129	53	96	–	0	0	132	1
		min	–	0.8335	–	0.5060	0.5000	26	32	23	38	38	–	14	23	151	44
		max	–	0.8405	–	3.0000	0.8374	129	534	534	321	414	–	15	24	705	45
Huawei P9 Lite	📴 → 📴	μ	0.7944	0.7951	0.7981	1.5721	0.7726	107	220	183	612	1019	8	14	22	343	45
		σ	–	0.0022	–	0.9340	0.1370	83	314	309	1308	2122	–	0	0	165	1
		min	–	0.7912	–	0.5000	0.5000	37	50	31	67	71	–	13	21	81	43
		max	–	0.7981	–	3.0000	1.0000	523	1535	1535	6656	8728	–	14	22	686	46
Sony Xperia XA1 G3112	📷 → 📴	μ	2.8777	2.8782	2.8759	1.5447	2.2467	95	134	129	94	86	72	182	254	834	719
		σ	–	0.0035	–	0.9944	0.6782	131	203	204	152	112	–	1	1	21	21
		min	–	2.8725	–	0.5000	1.0782	28	33	27	36	37	–	181	253	811	700
		max	–	2.8857	–	3.0000	2.9949	452	674	674	665	552	–	183	255	906	788
	📷 → 📴	μ	2.3013	2.3005	2.3003	1.2127	1.6330	38	43	38	49	48	70	200	270	824	711
		σ	–	0.0019	–	0.6753	0.7032	14	15	16	19	17	–	1	1	8	15
		min	–	2.2962	–	0.5000	1.0000	27	32	26	37	36	–	199	269	811	696
		max	–	2.3050	–	2.8759	3.0000	93	98	93	161	139	–	202	272	840	766
	📴 → 📴	μ	0.7998	0.7997	0.8001	2.5396	0.8114	784	1119	860	443	880	9	14	23	540	45
		σ	–	0.0017	–	0.6737	0.1185	839	1205	951	453	919	–	0	0	108	0
		min	–	0.7961	–	0.5104	0.5085	58	84	58	185	102	–	13	22	265	44
		max	–	0.8035	–	3.0000	0.9984	2534	3475	2632	3143	3161	–	15	24	712	45
iPhone 4S	📷 → 🎥	μ	1.3343	1.3335	1.3334	1.4217	1.7556	2974	4383	4081	1441	1852	25	57	82	359	362
		σ	–	0.0008	–	0.6236	0.6985	1614	2485	2298	1982	2425	–	0	0	96	90
		min	–	1.3327	–	0.5003	1.1685	253	453	174	47	41	–	57	82	259	227
		max	–	1.3365	–	2.9918	3.0000	5928	8361	8212	6910	7341	–	58	83	594	522
iPhone 6S	📷 → 🎥	μ	1.7754	1.7772	1.7778	1.2848	1.6941	1127	1800	1767	927	1314	36	81	117	357	315
		σ	–	0.0015	–	0.5712	0.3217	520	918	921	957	925	–	0	0	17	23
		min	–	1.7723	–	0.5000	1.0000	29	33	28	44	40	–	80	116	331	284
		max	–	1.7782	–	1.7812	2.2073	1844	3133	3105	2860	2844	–	81	117	378	365

The "recording mode" column indicates which image formats are employed for recording, the reference format being on left.
📷 photo 🎥 EIS unknown 📴 EIS off 📴 EIS on

actual scaled length l'_v ranges in the values

$$l'_v = l' + v\,, \quad v \in [-\epsilon, +\epsilon] \tag{1}$$

where v is quantized by a step of $q = 0.5$ pixels for computational efficiency. This leads to a set of $2\epsilon/q + 1 = 17$ allowable scale values $s_v = l'_v/l$ depending on v. Considering as values for l the width and height of the reference image, and repeating the process analogously on the warped image to be evaluated, the maximum number of allowable scales to be checked is $17 \times 4 = 68$, which corresponds to all the 17 possible values of v and the 4 values of l. Specifically, the PRNU pattern extracted from the warped images is rescaled with respect to the reference PRNU pattern according to each allowable scale among the 68 scales, and the one maximizing the PCE is chosen. Refined translation $\mathbf{t}_v$ is obtained from s_v as

$$\mathbf{t}_v = \sum_{k=1}^{n} \frac{\mathbf{p}_k{}' - s_v \mathbf{p}_k}{n} \tag{2}$$

where $(\mathbf{p}_k, \mathbf{p}_k{}')$ are the n RANSAC inlier keypoint pairs, being $\mathbf{p}_k$ and $\mathbf{p}_k{}'$ points in the reference image and in the video frame, respectively. While refining the scale, translation values $\mathbf{t}_v$ can be used to check the PCE peak location consistency, so as to discard solutions with relevant deviations. Note that this false alarm reduction strategy is not possible with other approaches based only on the maximization of the PRNU correlation.

3.2. Experimental results

The proposed PRNU pattern registration approach is compared on seven different devices against particle swarm optimization, which provides better accuracy and computational efficiency than brute-force approaches. For each device, the PRNU pattern of video I-frames from a flat homogeneous scene content is warped according to the transformation parameters found by the corresponding method into the reference PRNU pattern. The PCE between the warped and reference PRNUs is then evaluated. The reference PRNU pattern was extracted from photos at native resolution but also from video I-frames acquired with EIS off when the source video to check was acquired with EIS on. Smooth video paths are assumed, so no rotations were taken into account. For each device, tested format and compared method, Table 1 reports the estimated scale, the accuracy in terms of PCE, and the running time. Results are presented in terms of mean μ, standard deviation σ, and minimum and maximum statistics (detailed results, dataset and code are available as additional material for further analysis and reproducibility[1]). The proposed scene content PRNU alignment before and after refinement are indicated as G and G_r, respectively. Additionally, G_m represents the results obtained by averaging G_r scales while discarding video I-frames with low PCE values (i.e., less than 50) on G, as a fast way to skip unreliable frames. For particle swarm, implemented using the Matlab built-in function, two different setups were evaluated. In detail, setup P uses 35 particles and a scale search range in $[0.5, 3]$, while setup P_r uses 30 particles and a scale search range in $[1, 3]$ and $[0.5, 1]$, respectively when the reference PRNU pattern is extracted from native full resolution photos or video frames captured with EIS off.

The mean PCE value obtained with the scene content registration method G only is in most cases quite accurate, even without scale refinement (method G_r). The average registration G_m gives values very close to those given by G_r. The almost identical scale values obtained with the two different Samsung S7 devices witness that *the warping transformation between the different image formats does not change among devices of the same model*. This is quite reasonable since, differently from acquisition, the warping process is not analog but digital and depends on the device firmware (actually, firmware updates may exist for the same device model, yet it is very unlikely that these contain changes in their low-level camera acquisition code). This implies that G_m warping information can be used with other devices of the same model, thus avoiding one to acquire each time ad-hoc static scene images or videos. Moreover, as reported in the additional material, transformations across the different image acquisition formats can be concatenated without any accuracy degradation. Concerning particle swarm optimization, P_r results are usually more accurate and reliable than those obtained with P, confirming that the particle swarm approach can lead to unstable or even wrong solutions if no clues about the allowable transformation parameter range are available.

[1] https://drive.google.com/open?id=1hfqqWDBZRxTErDNAjQg-GtTYUj267gdl

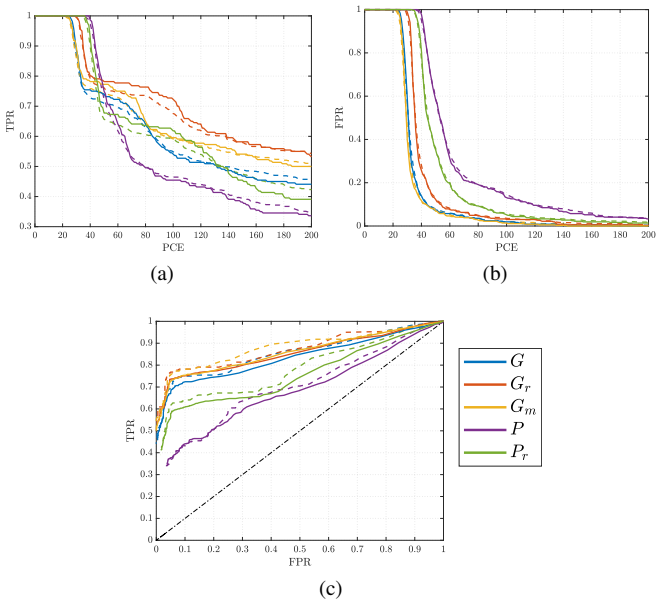

Fig. 2: (a) True and (b) false positive attribution rates for increasing PCE threshold, solid and dashed curves refer respectively to rely on a single frame and a majority voting over 3 frames in order to get the final decision. Corresponding ROC curves are shown in (c). Best viewed in color and zoomed in.

The proposed scene-based PRNU pattern registration is in general more accurate and reliable than that obtained by particle swarm optimization, also considering the lower excursion range in the scale and PCE values by inspecting the σ, minimum and maximum related values. The only relevant exception is the Huawei P9 Lite, for which the proposed approach obtains lower PCE values even if the PCE variation σ among all the tested frames is lower than particle swarm. As detailed in the next section, this device model has some peculiarities that could have affected the comparison. Notice also that in this case P obtains the highest average PCE value after P_r, but it is more distant in terms of the retrieved scale from P_r than the proposed scene content based methods, underlining accidental inconsistencies that may occur due to the stochastic nature of PRNU. Analogous considerations about consistency and stability of the scales and PCE values hold for the Samsung Galaxy S7 (2^{nd} device, mapping from photos to I-frames with EIS off) and the Sony Xperia XA1 G3112 (mapping from photos to I-frames with EIS on), whose average PCE values can be slightly better for the P_r particle swarm than the proposed scene content approaches.

The previous analysis is corroborated by the source attribution test whose results are reported in Fig. 2 in terms of True Positive Rate (TPR), False Positive Rate (FPR) and Receiver Operating Characteristic (ROC) curves. Two alternative approaches were investigated to establish whether an unknown video was recorded by a specific source device. In the first approach, the decision was taken according to a preset threshold on the PCE value between the PRNUs of a single frame of the unknown input video and the reference device to be queried. In the second approach, three frames of the unknown video were considered, and the final decision was taken according to a majority voting scheme on the same basis of the single

frame check. To define a query, the same dataset of the previous experiment was used, with 17 device/acquisition format source pairs (thirteen pairs as reported in Table 1, plus other four obtained by concatenating transformations as detailed in the additional material) and flat unknown videos, considering only I-frames. The total number of true positive queries evaluated is 678 for the single frame test, corresponding by construction to $678/3 = 226$ queries in the case of the majority voting test, i.e., in the latter case, three queries from different I-frames of the same unknown video were merged into a single query. The respective numbers of true negative queries are 1098 and $1098/3 = 366$.

TPR and FPR plots show that the proposed scene content PRNU alignment works with lower PCE thresholds (about 10 units less) with respect to particle swarm, which implies a more accurate PRNU registration. Likewise, the ROC curves evidence a better behavior of the proposed approach in the source attribution task. Furthermore, it can be noted that aggregate decisions on the majority voting scheme can improve the final decision. Globally, G_m works slightly better than G_r, which are both better than the unrefined strategy G.

Finally, concerning running times, according to Table 1 scene-based PRNU registration G is very fast and even by summing up the further refinement step G_r, the approach is faster than particle swarm optimization. Notice that in the table the total running time for G_r is obtained by adding the corresponding columns G and $G_r - G$ (the refinement step alone). In particular, the full approach G_r is about four times faster than particle swarm optimization. The only exception is when G_r is compared against P_r and the transformation involves mapping from video frames with EIS off to frames with EIS on, for which the proposed approach is only twice faster due to the lower resolutions involved. As matter of fact, running times depend on the image resolution and the scale search range. Clearly, particle swarm accuracy can be improved by employing more particles at the expense of higher computational time.

4. Reversing EIS

4.1. Alvaro description

Alvaro, the support employed to extract EIS data, is shown in Fig. 3. It is composed of a cubic brass frame of side 1 meter, with plywood panels on each face except one to strengthen the structure and decrease nonrigid oscillations when it is moved or shaken. On the empty face there is a thin grid made by a stretched nylon thread, whose intersections are evidenced by markers (see Fig. 3a). On the opposite face to the grid, in the center, there is a slot on which the device acquisition sensor under test is firmly fastened by strings (see Fig. 3b). Once put in place, the device becomes integral with Alvaro and the grid. Each grid marker is virtually anchored to a location inside the device sensor matrix grid, so that markers visible on a warped EIS frame can individually be mapped back onto the sensor matrix grid. Knowing the marker correspondences with respect to the reference frame enables one to find the EIS warping transformation for the current frame. Notice that the Alvaro approach (1) requires a controlled environment for studying the effect of EIS and (2) it cannot work if OIS is simultaneously enabled on the device, since in that case grid markers would not be anchored to the device matrix grid.

(a) (b)

Fig. 3: (a) Front side and (b) back side of Alvaro, notice the device placed in (best viewed in color and zoomed in).

In order to obtain a frame by frame correspondence between the markers, an automatic tracking system was developed. Due to the really small size of the markers, chosen for avoiding interferences with the EIS system, state-of-the-art tracking solutions would not work. To solve this issue, an ad-hoc tracking system was developed based on keypoint matching. Being K_i the set of HarrisZ keypoints on frame i-th, the set of keypoints $M_i \subset K_i$ associated to the markers are matched to the keypoints K_{i+1} on the next i+1-th frame, through sGLOH2 descriptor matching. Notice that using all the keypoints K_i is not a good choice, since this would lead to erroneously estimate the dominant transformation of the scene undergone by non-marker keypoints. Hence, to improve the matching accuracy, the putative corresponding keypoints of the i+1-th frame (K_{i+1}) are constrained to stay inside a circular window of 50 px radius from the marker keypoints of the i-th frame (M_i). The process is further refined using RANSAC, finding the transformation $H_{i,i+1}$ from the i-th frame to the i+1-th frame, modeled as a planar homography in order to be the most general, and the putative marker keypoints $\overline{M}_{i+1}$. In order to avoid accumulating errors when concatenating successive transformations with respect to the reference 0-ed frame, i.e. $\overline{H}_{0,i+1} = H_{i,i+1}H_{0,i}$, the homography is re-estimated as $H_{0,i+1}$ using RANSAC between keypoint markers $M_{0\to i}$ and $\overline{M}_{i+1}$, where $M_{0\to i}$ are the marker keypoints of the reference frame that have a corresponding marker on the i-th frame, notice that markers can provisionally go out of sight in a frame. The next frame marker keypoints M_{i+1} are then found by re-projecting through $H_{0,i+1}$ the whole set of marker keypoints M_0 of the reference frame (provided manually as input to the tracker) and associating to them the closest keypoints in K_{i+1}, if they fall inside the frame canvas. This expedient allows the recovery of lost markers that went out of sight at some frame. Figure 4 shows the results of reversing EIS warping according to the transformation estimated with Alvaro for two video frames obtained with the devices analyzed hereafter. Notice that in Fig. 4b the bottom marker row went out of sight.

4.2. Case study

Two mid-range smartphones, the Huawei P9 Lite and the Xiaomi M2 A1, were considered for analyzing EIS. The testing video sequences were obtained by moving and shaking the devices on Alvaro in front of a fixed background, and additionally introducing moving foreground objects of different sizes

(i.e. walking or jumping people, and fluttering flyers). For each frame, the homography obtained by tracking the grid markers was employed to revert back the frame EIS transformation as described previously (see again Fig. 4). The resulting videos, together with the input sequences and the tracking code, are freely available[2]: The reader is strongly invited to examine these videos for a better understanding of the process. The tracking is quite stable, except in some cases due to motion blur effects that decrease keypoint localization accuracy and for some unabsorbed, non rigid oscillation of the Alvaro structure with respect to the camera.

Figure 5 depicts for each analyzed sequence the decomposition of the EIS frame transformation into a similarity. In addition to each component of the similarity transform, it is also indicated the reprojection error obtained as the more general homography estimated by tracking with Alvaro is fitted into a similarity, having less degrees of freedom. This reprojection error is fully compatible with the keypoint localization accuracy of the scene content of the current frame. Moreover, the slight variation of the scale component, that is associated to the temporary loss of keypoint accuracy discussed above, suggests that frame transformations inside a EIS video are only metric, i.e. they involve only rotation and translation. In addition, no rotation of more than 5 degrees was observed. Both scale constraints and rotation limits can be exploited for designing new PRNU registration methods. For the sake of completeness, affine frame decomposition, that confirms the absence of shearing, i.e., of different scaling factors in the horizontal and vertical directions, is also reported in the additional material.

Furthermore, from the analysis carried out emerges that different device models use distinct EIS implementations. In particular, observing the second part of the sequence where Alvaro is fixed while a person is moving from one side to the other, it comes out that Huawei P9 Lite triggers EIS according to scene visual flow, opposing to the Xiaomi M2 A1, for which EIS is based on physical movement sensors (i.e., gyroscopes or accelerometers). Moreover, unlike the Xiaomi M2 A1, it can be observed that EIS frame rotation steps seem quantized for the Huawei P9 Lite, maybe due to the usage internally of Look-up Tables (LUT) for an efficient computation of the frame warp. Finally, the Huawei P9 Lite tends to maintain the last frame transformation over the next frames even if the condition that has triggered EIS disappears, while the Xiaomi M1 A2 in this case tends to smoothly come back to the original reference status. The latter seems to be the most common behavior across many devices, and explains why PRNU alignment can often work, disregarding of EIS, on sufficient long sequences, as most of the frames will result aligned to the reference first frame.

[2] https://drive.google.com/drive/folders/1vdRpYe9pC_kSPknWQj3Vu9pzd27GxCCX

(a) (b)

Fig. 4: Alvaro tracking results on EIS videos obtained from (a) Huawei P9 Lite and (b) Xiaomi M2 A1 devices. EIS warped frame (red border) is superimposed on the reference frame (blue border, best viewed in color and zoomed in).

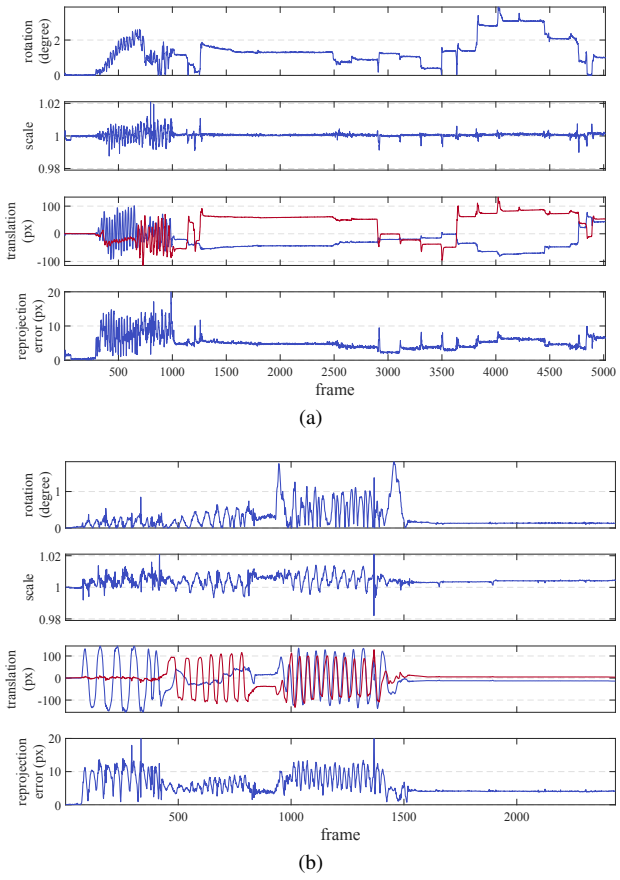

Fig. 5: Decomposition of EIS frame transformation according to a similarity for (a) Huawei P9 Lite and (b) Xiaomi M2 A1 videos. x and y translation components are indicated respectively in blue and red (best viewed in color and zoomed in).

Concerning the Huawei P9 Lite, it was also experimentally verified that its EIS aversion to get back to the reference frame status holds until exiting from the software interface, i.e., it holds even if one stops and starts again the video recording in the same session. Moreover, for this device the PRNU signal appears to be very weak. This is especially true for flat I-frames that seem not to be the best choice for extracting PRNU, as the intra-frame compression in case of highly compressible images, such as for flat scenes, may strongly disrupt the PRNU component. Structured scenes characterized by homogenous content, defined borders and low frequency textures seem more appropriate. The experimental validation of this claim is reported in Fig. 6.

5. Conclusions and future work

This paper explored and discussed content-based image registration with the purpose of dealing with (and better understanding) EIS in the context of PRNU alignment. In particular, a novel and robust solution to extrapolate the transformation relating the different image output formats for a given device model was proposed. This general approach can be adapted to

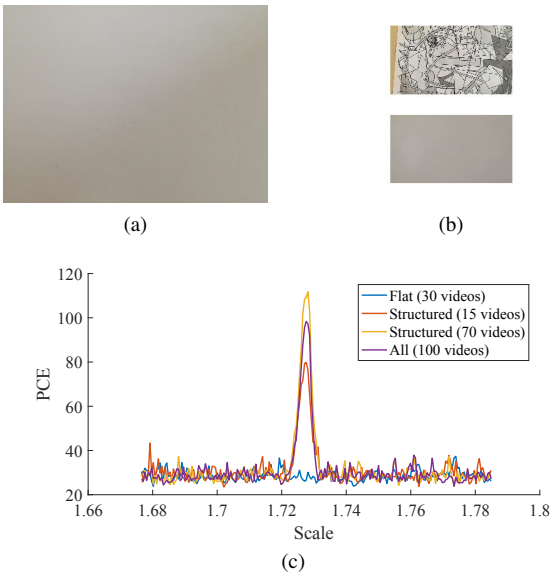

Fig. 6: Huawei P9 Lite PRNU signal evaluation. The reference PRNU fingerprint extracted from 15 flat photos at native resolution is compared against the fingerprint extracted from the first I-frames of EIS videos. (a) Sample flat scene employed for the acquisition of the native resolution photos. (b) Sample structured (top) and flat (bottom) scenes employed for the acquisition of the first I-frames of EIS videos. (c) PCE between the native resolution photo and EIS video fingerprints according to different scale factors, evidencing how flat I-frames contain weak PRNU signal (best viewed in color and zoomed in).

extract the scale factor, and when appropriate the translation, so as to align native resolution images to video frames, with or without EIS on, and eventually to compare PRNU patterns. The proposed approach has shown to be more reliable, more accurate and faster than state-of-the-art approaches based on brute-force and particle swarm optimization. Furthermore, a tracking system able to revert back EIS in a controlled environment was designed. This allows one to investigate the differences between the existing EIS implementations. The additional knowledge thus acquired can be exploited and integrated in order to design and implement better future EIS-aware PRNU pattern alignment methods.

Future work will move towards this direction, aimed at devising approaches able to better deal with EIS frame rotation. Additionally, further device models will be investigated, so as to extend the knowledge provided by this data into a shared database.

Acknowledgments

This material is based on research partially sponsored by the Air Force Research Laboratory (AFRL) and the Defense Advanced Research Projects Agency (DARPA) under agreement number FA8750-16-2-0188. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL and DARPA or the U.S. Government.

We also thank Massimo Iuliani for his help in highlighting the main PRNU-related problems, and Giuseppe Colombo for building the Alvaro device.

References

Altinisik, E., Sencar, H.T., 2020. Source camera verification for strongly stabilized videos. IEEE Transactions on Information Forensics and Security 16, 643–657.

Bellavia, F., Colombo, C., 2018. Rethinking the sGLOH descriptor. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 931–944.

Bellavia, F., Fanfani, M., Iuliani, M., Piva, A., Colombo, C., 2019. PRNU pattern alignment for images and videos based on scene content, in: Proceeding of the IEEE International Conference on Image Processing (ICIP).

Bellavia, F., Tegolo, D., Valenti, C., 2011. Improving Harris corner selection strategy. IET Computer Vision 5, 86–96.

Chen, M., Fridrich, J., Goljan, M., Lukas, J., 2008. Determining image origin and integrity using sensor noise. IEEE Transactions on Information Forensics and Security 3, 74–90.

Fischler, M., Bolles, R., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24, 381–395.

Goljan, M., Fridrich, J., 2012. Sensor-fingerprint based identification of images corrected for lens distortion, in: Media Watermarking, Security, and Forensics 2012, SPIE. pp. 132–144.

Goljan, M., Fridrich, J., 2013. Sensor fingerprint digests for fast camera identification from geometrically distorted images, in: Media Watermarking, Security, and Forensics 2013, SPIE. pp. 85–94.

Goljan, M., Fridrich, J.J., 2008. Camera identification from cropped and scaled images, in: Proceeding of Security, Forensics, Steganography, and Watermarking of Multimedia Contents X.

Grundmann, M., Kwatra, V., Essa, I., 2011. Auto-directed video stabilization with robust l1 optimal camera paths, in: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 225–232.

Höglund, T., Brolund, P., Norell, K., 2011. Identifying camcorders using noise patterns from video clips recorded with image stabilization, in: Proceedings of the International Symposium on Image and Signal Processing and Analysis (ISPA), pp. 668–671.

Iuliani, M., Fontani, M., Shullani, D., Piva, A., 2019. Hybrid reference-based video source identification. Sensors 19.

Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M., Trulls, E., 2020. Image matching across wide baselines: From paper to practice, in: arXiv.

Karaküçük, A., Dirik, A.E., Sencar, H.T., Memon, N.D., 2015. Recent advances in counter prnu based source attribution and beyond, in: Media Watermarking, Security, and Forensics 2015, pp. 201 – 211.

Kouokam, E.K., Dirik, A.E., 2019. PRNU-based source device attribution for YouTube videos. Digital Investigation 29, 91–100.

Lukas, J., Fridrich, J., Goljan, M., 2006. Digital camera identification from sensor pattern noise. IEEE Transactions on Information Forensics and Security 1, 205–214.

Mandelli, S., Bestagini, P., Verdoliva, L., Tubaro, S., 2019. Facing device attribution problem for stabilized video sequences. IEEE Transactions on Information Forensics and Security 15, 14–27.

Mur-Artal, R., Montiel, J., Tardos, J., 2015. ORB-SLAM: a versatile and accurate monocular slam system. IEEE Transactions on Robotics 31, 1147–1163.

Shullani, D., Fontani, M., Iuliani, M., Shaya, O.A., Piva, A., 2017. VISION: a video and image dataset for source identification. EURASIP Journal on Information Security 2017, 15.

Szeliski, R., 2010. Computer Vision: Algorithms and Applications. 1st ed., Springer-Verlag.

Taspinar, S., Mohanty, M., Memon, N., 2016. Source camera attribution using stabilized video, in: Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6.

Zaragoza, J., Chin, T.J., Brown, M.S., Suter, D., 2013. As-Projective-As-Possible image stitching with moving DLT, in: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zheng, L., Yang, Y., Tian, Q., 2018. SIFT meets CNN: A decade survey of instance retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 1224–1244.