VIDEO CAMERA IDENTIFICATION USING AUDIO-VISUAL FEATURES

S. Milani[♯], *L. Cuccovillo*[♭], *M. Tagliasacchi*[♯], *S. Tubaro*[♯], *P. Aichroth*[♭]

[♯]Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy
[♭]Fraunhofer Institute for Digital Media Technology
Ehrenbergstr. 31, 98693 Ilmenau, Germany

ABSTRACT

One of the major issues in multimedia forensics is the identification of video acquisition devices. Most of the relevant state-of-the-art solutions rely on either visual or audio analysis, using feature arrays that are highly correlated with the characteristics of the respective camera or microphone.

In this work, we present a multi-modal approach that uses both video and audio information to improve the detection accuracy. For this purpose, microphone detection based on the blind estimation of the frequency response is complemented with a video camera detection based on a set of video features related to the Color Filter Array interpolation. Experimental results show that the combined approach results in an improved overall classification accuracy over the mono-modal cases.

Index Terms— microphone classification, CFA detection, device identification, image forensics, audio-visual systems

1. INTRODUCTION

One of the most-investigated issues in multimedia forensics is the identification of the acquisition devices to a given content item, which is relevant for various application scenarios, including multimedia content authentication and forgery detection [1].

State-of-the-art solutions focus on *either* audio or visual processing, looking for unique features in one of them to characterize the recording device. Hence, it is possible to classify the proposed methods in two main categories: Algorithms using visual information, and algorithms using audio material.

Approaches exploiting visual footprints include, e.g., strategies characterizing the presence of lens radial distortion [2], sensor imperfections [3], and color filter array (CFA) interpolation [4, 5]. As for audio footprints, state-of-the-art microphone classification approaches include purely statistical analysis [6, 7], processing of Fourier coefficients [8], the use of high-level model inferred from Mel-Frequency Cepstral Coefficients (MFCC) [9], and the blind estimate of the microphone frequency response [10].

In this work, we aim at improving detection accuracy by using a fusion of audio and video analysis approaches, respectively. This allows for an improved identification in cases where the poor quality of the video signal (e.g., determined by the illumination conditions or an excessive compression) leads to the estimation of a noisy set of features, or the audio spectrum does not allow for an accurate classification on its own. After splitting audio and video streams, the basic approach is as follows:

- The audio part is divided into segments, and for each segment, a set of features is computed using blind estimation of the microphone frequency response, thereby classifying the device.
- The video part is divided into multiple frames, and a set of visual features related to the adopted CFA interpolation strategy is computed.
- Then, results from the visual analysis are used to refine the device estimation provided by the audio analysis, and vice-versa.

Experimental results show that the joint strategy allows an accurate identification in many cases where a single algorithm fails.

In the following, Section 2 gives an overview of some of the works in literature targeting the device identification. Section 3 presents the audio features used for microphone identification, while Section 4 describes the visual cues applied in the camera model identification. Section 5 describes the combined audio-visual approach in more detail. Section 6 reports some experimental results, and Section 7 provides the respective conclusions.

The project REWIND acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 268478.

EUVIP 2014, Dec. 10-12, 2014, Paris, France.
978-1-4799-4572-6/14/$31.00 ©2014 IEEE.

2. RELATED WORKS

As mentioned in the previous section, the visual features employed by the proposed detector are related to the CFA interpolation strategy employed by the device. In most photo devices, it is possible to acquire a single color component per pixel, and therefore, color samples are positioned according to a specific acquisition mask (Bayer mask). The missing colors are generated via an interpolation of this mask (before compressing the image), and each camera manufacturer develops and implements its own algorithm.

Many solutions have been proposed in literature. Most of these estimate a denoised version of the image and try to characterize the type of interpolation from the residual signal between denoised and original image [5, 4]. Other solutions rely on approximating the CFA interpolation via linear filter and, after estimating the coefficient values, use them as signature [11, 12]. Other strategies perform a reinterpolation of the resampled image [13].

Although these solutions are very accurate when processing high-quality images, performances dramatically collapse whenever lossy compression is included in the processing chain.

As for microphone identification, in this work we focus on the blind estimate of the microphone frequency response which permits identifying the acquiring device and environment. More precisely, the implemented approach is derived from the work [10]. A related approach can be found in [14], where Gaubitch *et al.* provide an algorithm for the blind estimation of a channel response, where the channel was representing the surrounding acoustic environment. The method in [14] was meant to work on noiseless speech-only recording in case of uncompressed audio content, and has the advantage of estimating the channel starting from observations acquired by a single microphone.

3. MICROPHONE DETECTION STRATEGY

As mentioned in Section 2, the microphone detection strategy relies on the method proposed in [10].

The adopted method is largely based on [14], and is designed to work on real-life recordings with environmental noise and background music, thus overcoming the speech-only content limitation of [14]. Moreover, [10] has been tested with encoded audio files, and proved to be robust against lossy compression, i.e. AMR, MP3 and AAC audio files.

3.1. Feature vector

Let us denote with $s(t)$ an audio signal in the time domain, and with $S_l(k)$ the short-term Fourier Transform of its l-th frame, with L_S the number of analysis frames, and with N_{fft} the number of frequency bins per frame. The log power spectrum of each l-th frame is normalized according to

$$Z_{S,l} = log\left(|S_l|\right) - \frac{1}{N_{\text{fft}}} \sum_{k=1}^{N_{\text{fft}}} log\left(|S_l(k)|\right), \; \forall l \in L_S$$

and all the normalized log power spectra are used to build the matrix Z_S, where $Z_S \in \mathbb{R}^{L_S \times N_{\text{fft}}}$.

From Z_S an average normalized log power spectrum can be computed as follows:

$$\hat{p} = \frac{(Z_S)^t \cdot \mathbf{1}}{\bar{L}_S},$$

where $\mathbf{1}$ is a $\bar{L}_S \times 1$ vector with all elements equal to one, the superscript t denotes the matrix transpose, and $\bar{L}_S$ denotes the number of frames in a selected subset of the audio file.

For an audio signal $s(t)$ we thus compute two main features from a subset of frames $\bar{L}_S$, i.e.

1. The blind estimate $\hat{h}$ of the amplitude of the microphone frequency response.

2. The average normalized log power spectrum $\hat{p}$ of the audio content.

Following the algorithm in [10], these initial features are then manipulated and combined in a final high-level feature vector

$$f = [f_1\,,\; f_2\,,\; f_3] \tag{1}$$

where f_1 keeps all the information about the channel estimate $\hat{h}$, f_2 is a descriptor of the correlation between $\hat{h}$ and $\hat{p}$, and f_3 defines the properties of the average normalized log power spectrum $\hat{p}$.

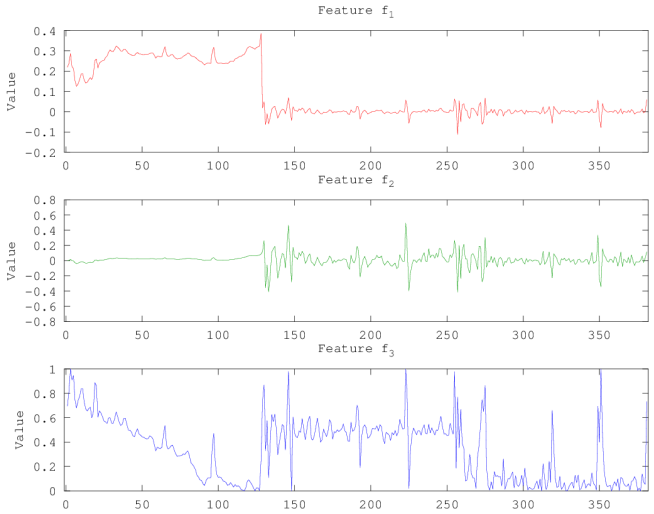


Fig. 1. Graphs of the features f_1, f_2, and f_3 for a test audio file.

A visual comparison of f_1, f_2, and f_3 is shown in Figure 1, the x-axis representing the dimension index. Features are computed only for the frequency bins from 1 to $N_{\text{fft}}/2$, in order to avoid redundancies.

The complete, detailed description of the features involved in eq.(1) is reported in [10].

3.2. Training and classification

The classification task relies on a multi-class Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel.

The training audio signals are downsampled to 8 kHz, if necessary, and split in training signals of 5 seconds length. For each training signal a training label is stored according to the original recording device, and a feature vector is computed according to eq. (1). Prior to the SVM training, each dimension of the full training set of feature vectors undergoes a normalization between -1 and +1, and a feature selection based on the F-score as defined in [15].

After the training of the SVM, an audio signal with sampling frequency 8 kHz can be classified independently from its length, provided that each dimension of the feature vector undergoes the same feature selection of the training vectors, and is normalized according to the range of the SVM original training set.

Each test audio signal is classified according to the label assigned by the SVM. The estimate of the microphone frequency response becomes unreliable for short time intervals: Input signals should thus be in general longer than 1 second.

4. CFA RELATED VISUAL FEATURES

Several works in literature have proved that it is possible to identify the acquiring device by detecting the CFA interpolation strategy that was implemented on the camera. A set of visual features is computed from each frame, organized in an array, and classified via a multiclass approach.

The proposed camera identification strategy falls among the set of strategies based on reinterpolation and has been derived considering the idempotence property of many operators. Given a set of data Y which has been obtained processing data X via operator o (i.e. $Y = o(X)$), it is possible to say that o satisfies the idempotence property if the correlation between Y and $o(Y) = o(o(X))$ is maximal. Therefore, given a set $\mathcal{O}$ of operators o_t, it is possible to state that $o \equiv o_{t^*}$ if the correlation between Y and $o_{t^*}(Y)$ is maximal. In the literature, the idempotent property has been successfully exploited for the identification of the quantizer [16], the traces left by JPEG compression antiforensics [17], and the adopted video coding architecture [18, 19, 20].

These approaches imply that $o \in \mathcal{O}$; in case $o \notin \mathcal{O}$, the correlation values between Y and the different $o_t(Y)$ provide useful information to identify o. Similar approaches have been exploited in [21, 22], but the strategy of this work is focused on detecting the CFA interpolation strategy from video frames.

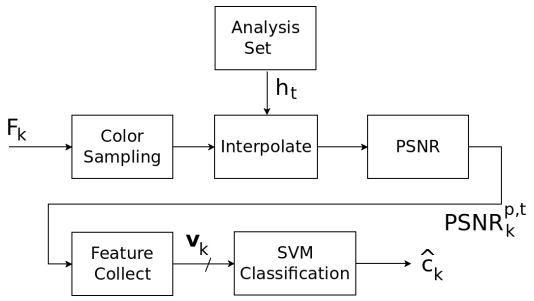
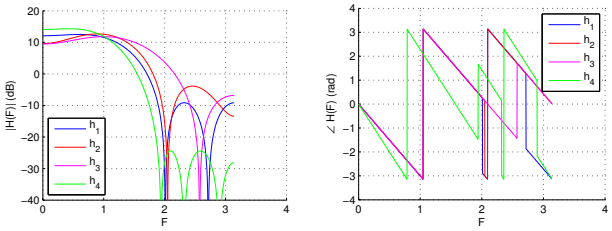


Fig. 2. Block diagram for the extraction of video features.



(a) (b)

Fig. 3. Frequency response for different filters h_t; a) $|H_t(F)|$; b) $\angle H_t(F)$.

Let us denote with F_k the k-th frame acquired using the c-th camera (camera models involved in the testing are detailed in Table 1). F_k is resampled according to a specific Bayer pattern p, generating the sampled frame F_k^p. The image F_k^p is then re-interpolated in the image $F_k^{p,t}$ with algorithm t, and the algorithm computes the $\text{PSNR}_k^{p,t}$ between the frame F_k and the frame $F_k^{p,t}$. The obtained PSNR value represents one of the features that characterized the algorithm. The whole chain of operations (which is schematized in Fig. 2) is repeated for different analysis algorithm t and different CFA sampling patterns p generating the array $\mathbf{v}_k = [\text{PSNR}_k^{p,t}]_{p,t}$. The rationale of the approach relies on the fact that the behavior of most of the CFA interpolation algorithm is quite similar at low frequencies, the value of the PSNR will be mostly characterized by the difference of signals at medium frequencies. Changing the algorithm t and the Bayer pattern p employed in the resampling, it is possible to generate a set of different PSNR values and characterize the unknown strategy t.

In the presented approach, we approximated the CFA interpolation strategy with a set of 135 linear interpolation filters h_t whose frequency responses differ in the transient bandwidth, as Fig. 3 shows. Filters were obtained via the Parks-McClellan optimization imposing 15 different spectral behaviors and 3 different filter supports (4, 6, 8 taps, respectively). Moreover, we considered 4 different CFA pattern sampling

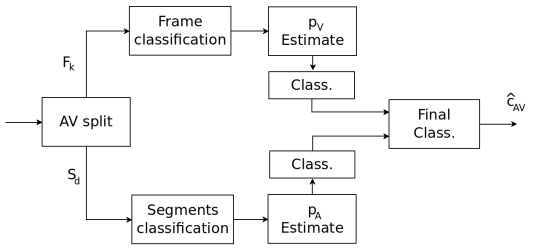


Fig. 4. Block diagram for the joint audio-visual detector.

(RGGB,BGGR,GRBG,GBRG) leading to 520 video features.

The dimensionality of this array is the reduced to 40 via a PCA decomposition of the features. The obtained descriptors are then fed to a set of multi-class Support Vector Machine classifiers using an RBF kernel.

More precisely, we computed a set of SVM classifier D_c, where the index c is referred to one of the cameras to be detected, such that the output value of $D_c(\mathbf{v}_k) \in \{+1, -1\}$ indicates whether the frame F_k has been acquired by camera c or not.

The different classifiers are composed into the variable Δ_c such that

$$\Delta_c = \sum_k \left(2\, I(D_k > 0) - 1\right) \tag{2}$$

where $I(\cdot)$ is the indicating function. Final decision for the current frame is then computed as $c_V = \arg\max_k \Delta_k$ such that $D_k > 0$.

5. THE PROPOSED MULTIMODAL DETECTOR

Previous sections have described audio and video features computed by the proposed approach and how their values are employed in a disjoint classification. The results of these operations are then merged together in order to increase the robustness of the final classification.

The block diagram of the whole scheme is reported in Fig. 4. At the beginning, the detector extracts a set of frames and the audio track from the input video file. For every frame F_k, a set of video features is computed and classified as described in Section 4. A probability mass function (pmf), named $p_V(c)$ is estimated according to the outcomes of the classification: each probability value provides a likelihood for that estimation results. The video detector chooses the vale $\hat{c}_V$ such that

$$\hat{c}_V = \arg\max_c p_V(c). \tag{3}$$

As for audio, the signal is divided into segments S_d of 2 sec. For every segment S_d, the detector estimates the acquiring microphone following the approach of Section 3. A second pmf $p_A(c)$ is computed from the classification outcomes, and the final detected camera $\hat{c}_A$ is found via eq. (3).

Table 1. Experimental setting.

Cameras		
Canon (C1)	HTC Hero (C2)	Panasonic (C3)
Samsung Galaxy (C4)	Sony HD (C5)	
Environments for training		
anechoic	lobby	lake
street with cars	mensa hall empty	park
Environments for testing		
meeting room	public square	

Experimental results have shown that video features are more accurate in characterizing some devices, while audio features prove to be more efficient for others. As for the camera set reported in Table 1, video features accurately identify the first three cameras and prove to be inaccurate in distinguishing Samsung Galaxy from Sony HD. Similarly, audio features mistake Sony HD for Canon cameras. From these premises, we designed the classification strategy reported in the following.

The input video sequence is partitioned into wider time segments S'_d such that they include multiple frames and audio segments. For every segment S'_d, audio and video estimates $\hat{c}_A$ and $\hat{c}_V$ are computed considering the most-frequently detected device within S'_d. Then, estimates are merged according to the following strategy.

if $\hat{c}_V \equiv \hat{c}_A$ **then**
  $\hat{c}_{AV} = \hat{c}_A = \hat{c}_V$
else
  if $\hat{c}_V \neq \hat{c}_A \wedge \hat{c}_V \in \{C1, C2, C3\}$ **then**
    $\hat{c}_{AV} = \hat{c}_V$
  else
    $\hat{c}_{AV} = \hat{c}_A$
  end if
end if

6. EXPERIMENTAL RESULTS

Tests were run considering a set of 5 camcorders, which acquired 32 sequences in 8 different environments. Camera models and environments are reported in Table 1, where a label is assigned to each camera for the sake of conciseness.

A set of 6 environments was used to train both audio and video detectors, while testing was performed on a different set of sequences. Each video sequence was divided into segments of 2 seconds. For every segment, video frames and audio signals were extracted and classified as described in the previous sections. Since every camera acquires video frames with different resolutions (1080×1920, 288×352, 1080×1920, 240×320, 1080×1440), we selected a central area sized 240×320 from every frame in order to compute the CFA related features.

In first set of experiments, we tested both detectors sep-

Table 2. Confusion matrix for audio detector.

	C1	C2	C3	C4	C5
C1	0.5	0.0	0.0	0.0	0.5
C2	0.0	1.0	0.0	0.0	0.0
C3	0.0	0.0	1.0	0.0	0.0
C4	0.0	0.0	0.0	1.0	0.0
C5	0.05	0.0	0.0	0.0	0.95

Table 3. Confusion matrix for video detector.

	C1	C2	C3	C4	C5
C1	1.0	0.0	0.0	0.0	0.0
C2	0.0	0.96	0.0	0.04	0.0
C3	0.04	0.0	0.80	0.16	0.0
C4	0.0	0.0	0.0	0.63	0.37
C5	0.0	0.0	0.0	0.4	0.6

arately. Table 2 reports the confusion matrices for detector using audio features only. Every row is related to a different device, while columns are referred to the classification outcomes of the detector. Cell (i, j) report the percentage of segments acquired by device i and classified as j. It is noticed that the audio detector finds very difficult to discriminate the C1 camera from the C5 camera: half of the times frames belonging to C1 are labeled as belonging to C5. The average accuracy of the audio-only device identification approach is approximately 89 %,

As for the video feature, Table 3 reports the confusion matrix obtained from the detector using visual features only. The overall performance is affected by the different resolutions and the strong compression level which has been applied to the video signal. It is noticed that video detector finds difficult to discriminate camera C4 from camera C5. The average accuracy of the video-only device classification is around 80 %.

Lastly, results related to the joint detector results are reported in Table 4. It is noticed that the accuracy of the classification is enforced by the adoption of the most accurate classifier according to the possible camera model: the average accuracy obtained by the joint approach is thus increased up to approximately 98 %.

7. CONCLUSIONS

The paper presents a multi-modal video camera detection approach combining audio and visual features to maximize the accuracy of the estimation. Experimental results show that such fusion allows to compensate for limitations of the respective uni-modal audio and visual approaches, resulting in an overall average accuracy beyond 98 %. Future work will be devoted to further increasing the accuracy by merging together additional features, e.g., video-codec-related footprints, audio-codec-related footprints and PRNU patterns.

Table 4. Confusion matrix for joint audio-video detector.

	C1	C2	C3	C4	C5
C1	1.0	0.0	0.0	0.0	0.0
C2	0.0	1.0	0.0	0.0	0.0
C3	0.0	0.04	0.96	0.0	0.0
C4	0.0	0.0	0.0	1.0	0.0
C5	0.04	0.0	0.0	0.0	0.96

Moreover, the accuracy of the detector could be improved by adopting more sophisticated compositions of the results from the different detectors, like the fusion strategies based on Dempster-Shafer theory.

8. REFERENCES

[1] S. Milani, M. Fontani, P. Bestagini, M. Barni, A. Piva, M. Tagliasacchi, and S. Tubaro, "An overview on video forensics," *APSIPA Transactions on Signal and Information Processing*, vol. 1, pp. e2, 2012.

[2] Kai San Choi, Edmund Y. Lam, and Kenneth K. Y. Wong, "Source camera identification using footprints from lens aberration," in *Proc. SPIE 6069, Digital Photography II*, Feb. 2006, vol. 6069J, pp. 60690J–60690J–8.

[3] J. Lukas, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," *Information Forensics and Security, IEEE Transactions on*, vol. 1, no. 2, pp. 205–214, June 2006.

[4] Shang Gao, Guanshuo Xu, and Rui-Min Hu, "Camera model identification based on the characteristic of CFA and interpolation," in *Proceedings of IWDW 2011*, Oct. 2011, pp. 268–280.

[5] Sevinc Bayram, Husrev T. Sencar, and Nasir Memon, "Improvements on source camera model identification based on CFA interpolation," in *International Conference on Digital Forensics*, Jan. 29 - Feb. 1, 2006.

[6] Christian Krätzer, Andrea Oermann, Jana Dittmann, and Andreas Lang, "Digital audio forensics: a first practical evaluation on microphone and environment classification," in *Proceedings of the 9th ACM workshop on Multimedia & security (MM&Sec '07)*, Dallas, TX, USA, Sep. 2007, pp. 63–74, ACM.

[7] Christian Krätzer, Kun Qian, Maik Schott, and Jana Dittmann, "A context model for microphone forensics and its application in evaluations," in *Proc. of SPIE 7880, Media Watermarking, Security, and Forensics III, 78800P*, Feb. 2011.

[8] Robert Buchholz, Christian Krätzer, and Jana Dittmann, "Microphone classification using Fourier coefficients," in *International Workshop on Information Hiding*, Stefan Katzenbeisser and Ahmad-Reza Sadeghi, Eds., 2009, pp. 235–246.

[9] D. Garcia-Romero and C.Y. Espy-Wilson, "Automatic acquisition device identification from speech recordings," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, Dallas, TX, USA, Mar. 2010, pp. 1806–1809.

[10] L. Cuccovillo, S. Mann, M. Tagliasacchi, and P. Aichroth, "Audio tampering detection via microphone classification," in *Proceedings of the IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*, Pula, Italy, Sep. 2013, pp. 177–182.

[11] Hong Cao and A.C. Kot, "Accurate detection of demosaicing regularity for digital image forensics," *Information Forensics and Security, IEEE Transactions on*, vol. 4, no. 4, Dec. 2009, pp. 899–910.

[12] Matthias Kirchner, "Efficient estimation of CFA pattern configuration in digital camera images," in *Proc. SPIE 7541, Media Forensics and Security II, 754111*, Jan. 27 2010, vol. 7541, pp. 754111–754111–12.

[13] A. Swaminathan, Min Wu, and K.J.R. Liu, "Digital image forensics via intrinsic fingerprints," *Information Forensics and Security, IEEE Transactions on*, vol. 3, no. 1, Mar. 2008, pp. 101 – 117.

[14] N. D. Gaubitch, M. Brookes, P. A. Naylor, and D. Sharma, "Single-microphone blind channel identification in speech using spectrum classification," in *Proc. EUSIPCO 2011*, Barcelona, Spain, Aug. 2011, pp. 1748 – 1751.

[15] Yi-wei Chen and Chih-jen Lin, "Combining SVMs with various feature selection strategies," *Feature Extraction Studies in Fuzziness and Soft Computing*, vol. 207, no. II, 2006, pp. 315 – 324.

[16] Z. Zhu and T. Lin, "Idempotent H.264 intraframe multigeneration coding," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, Taipei, Taiwan, Apr. 2009, pp. 1033 – 1036.

[17] G. Valenzise, V. Nobile, M. Tagliasacchi, and S. Tubaro, "Countering JPEG anti-forensics," in *Proceedings of the IEEE International Conference on Image Processing (ICIP 2011)*, Brussels, Belgium, Sep. 2011, pp. 1949–1952.

[18] P. Bestagini, A. Allam, S. Milani, M. Tagliasacchi, and S. Tubaro, "Video codec identification," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, Mar. 2012, pp. 2257–2260.

[19] M. Sorell, "Video provenance by motion vector analysis: A feasibility study," in *Proceedings of International Symposium on Communications Control and Signal Processing (ISCCSP 2012)*, Rome, Italy, May 2012, pp. 35–42.

[20] P. Bestagini, S. Milani, M. Tagliasacchi, and S. Tubaro, "Video codec identification extending the idempotency property," in *Proc. of 2013 European Workshop on Visual Information Processing (EUVIP 2013)*, Paris, France, June 10–12, 2013, pp. 220–225.

[21] S. Milani, M. Tagliasacchi, and S. Tubaro, "Identification of the motion estimation strategy using eigenalgorithms," in *Proceedings of IEEE International Conference on Image Processing (ICIP 2013)*, Vancouver, BC, Canada, Sep. 15 – 18, 2013, pp. 4477 – 4481.

[22] S. Milani, P. Bestagini, M. Tagliasacchi, and S. Tubaro, "Demosaicing strategy identification via eigen algorithms," in *Proc. of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, Florence, Italy, May 4–9, 2014, pp. 2678 – 2682.