Source Camera Identification using Feature Extraction

Meettu Miriam Anujan
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirapally

***Abstract*— The means and incentive to create digital image forgeries increased because of the increasing use of digital image, thus feature based source camera identification plays a crucial role in the authentication of digital images. The drawback of the conventional systems is the problem of unknown models. To rectify the disadvantage, camera model identification with unknown model was introduced but the accuracy level was found 28%. To increase the accuracy to an acceptable level, the proposed system was introduced. The new scheme consist of four stages: 1) feature extraction 2) unknown detection 3) unknown expansion 4) classification. In feature extraction, the input image is represented in 10 different formats and from each format 34 features are extracted, thus a total of 340 features are extracted from a single image. Then a KNN based unknown detection and a self training based unknown expansion is done. Finally classification is done using multi-class SVM with quadratic kernel. The experiments were carried out on Dresden image collection which confirms the effectiveness of the proposed system. The accuracy of the proposed system is found to be about 86%**

Keywords*— Unknown Models, Accuracy, Source Camera Idntification, Classification.

I. INTRODUCTION

With the popularity of digital cameras and the ease of image editing, image forgery has become a great issue and image forgery detection has become a wide area of research. Usually the goal of image forensics is either authentication or integrity validation. Authentication is to identify the source imaging device used to capture a given image and integrity validation involves determining whether the digital image has been tampered or not, and if so, what kind of tampering has been performed

There are three categories of camera model identification: 1) image metadata based 2) watermark based 3) feature based. The metadata based approach relies on investigating the image source related information such as camera brand , model, date, time etc. which are embedded in image metadata. The date and time information included in the metadata is related to the date and time of image capturing. But the drawback is the image metadata is easy to manipulate. Digital watermarking has been introduced for authenticating digital documents which embeds a watermark that carries source related information in the image. However the drawback is that watermark need to be inserted during the creation of the image which complicates the design and increase the production cost of digital cameras. This leads to the development of feature based approach which initially extracts features on intrinsic hardware artifacts or software related fingerprints left during image acquisition process. Then some multi-class classifiers such as support vector machine (SVM) are employed to classify images into predefined class of known camera model.

The conventional schemes suffer from the problem of unknown models. This is due to the fact that all cameras cannot be obtained by the identification system in advance and due to the rapid development of digital imaging technology, new camera models are continuously produced by different companies. If the testing input image is captured by a camera model which is previously unknown to the system and having some features similar to a camera model which is already trained by the system , the probability of classifying the image to the class of known camera model is more. This will adversely affects the accuracy of the system .To overcome this disadvantage a camera model identification with unknown model was developed. The disadvantage of the system is that accuracy was found to be only 28%.

In this paper, a new scheme, namely source camera identification with unknown models using feature extraction is proposed to tackle the problem of lower accuracy. The proposed scheme has 4 stages. Firstly a feature extraction where an image is represented in ten different formats based on chromatic adaptations [9] and features are extracted. Secondly a KNN based unknown detection method is developed to recognize unknown images from an unlabelled training dataset , followed by a self training based unknown expansion. Finally classification is done using multi-class SVM with quadratic model. Experimental results shows that accuracy of the proposed system is much higher compared to conventional system.

The remainder of the paper is structured as follows: related works are described in section II, section III presents a detailed description of the proposed system. Section IV reports the experimental results followed by conclusion in section V.

II. RELATED WORK

Feature based source camera identification is commonly used because of its reliability. Initially features are extracted from an image and then cast the identification as a supervised classification problem.

Intrinsic hardware artifacts and software related features are two categories of features used for camera identifications. Intrinsic hardware include features like sensor pattern noise [14], lens radial distortion, chromatic aberration, sensor dust pattern etc. Image related features include Image

Quality Matrices (IQM) features, error etc and patterns introduced by CFA are included in software related fingerprints.

Kai san choi and et al [1] believed that source camera can be identified by measuring the amount of lens aberration (ie, barrel or pincushion distortion) exhibited by each camera model. Sensor pattern noise is considered as an intrinsic fingerprint of each camera, thus SPN is commonly used for source camera identification. Lawgaly [2] proposed an efficient source camera identification based on image sharpening using an unsharp mask, The experimental results showed that identification produce higher accuracy.

Deng [3] proposed an auto white balance approximation but it has certain limitations to identify the source camera. A patch based approach for camera identification was proposed by Yue Tan [4] but it is not applicable for all kind of images. Dirik et al [6] argued that the location and shape of dust specks infornt of the imaging sensor is a useful fingerprint for source camera identification. Bayram et al [8] exploited CFA interpolation algorithm for camera identification. Khazzari et al [9] proposed a total of 34 different features which cab be extracted from an image used for source camera identification.

The next step after feature extraction is classification, usually k-class(k is the number of known models) is used. Even if an unknown image is presented to the system, it will be classified to any of the k classes and this affect the performance of the system. Inorder to rectify this defect Huang et al[p] proposed a source camera identification with unknown models which extracted 34 features [9] and then unknown detection and expansion is done, but the accuracy of the system was found lower. Inorder to tackle this problem the proposed system was introduced.

III. PROPOSED SYSTEM

This section presents a detailed description of SCI with unknown models using feature extraction. The system has four main stages namely feature extraction, unknown detection, unknown expansion, classification. The overview of the system is shown in figure1. There are three datasets: labeled training dataset which contain images from known camera model, unlabelled training dataset and testing dataset, both of which contain random images to be identified. The overview of the proposed system is shown in figure 1.

A. Feature Extraction

When a test image is presented to the system, features are extracted for source camera identification. In the conventional system 34 features, average pixel value, RGB pair correlation, neighbor distribution of center of mass, RGB pair energy ratio et were extracted [9], but the accuracy of such systems were found lower . In order to overcome this problem, in the proposed system , the given test image is represented in 10 different formats.

Chromatic adaptation is the ability of human visual system to adjust changes in the illumination in order to preserve the appearance of object colors. An object may be viewed under various conditions for eg: illuminated by sunlight, light of fire, harsh electric light etc. In all these situations an object may appear different in an image capturing device and thus features extracted will also vary which will affect the performance of the system adversely. To rectify this, in the proposed system images are represented in different chromatic adaptation methods described below:

a) Gray world : with 6 color adaptations which include von kries, diagonal, bradford, sharp, CMCCAT2000,xyz model
b) Shades of gray
c) Max RGB
d) Gray edge with differential order 1 and 2

From all these representations 34 features are extracted, thus a total of 340 features are obtained for a given test image, which will increase the accuracy to greater extend.

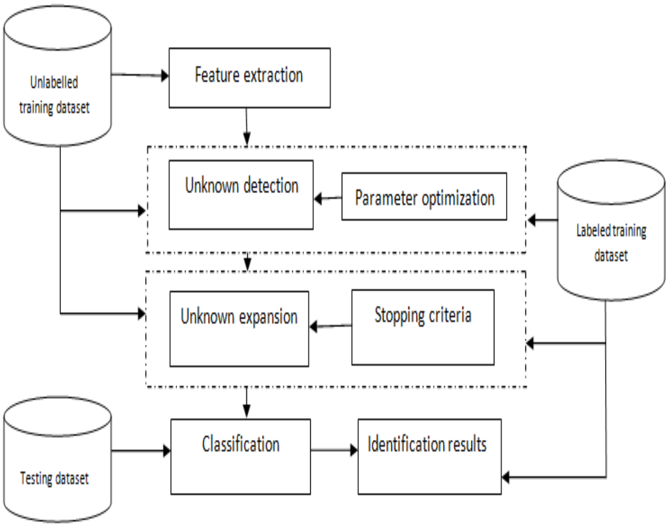

Figure 1: Overview of the proposed system

B. Unknown Detection

KNN bases unknown detection is employed to identify unknown images from the dataset. For an unlabelled image I, if it K nearest neighbor contain images of known model, the probability of I being generated by the any one of the model is more. If its K nearest neighbor does not contain any image of the unknown model, the probability of I being generated by an unknown model is more.

The procedure of unknown detection is as follows: Initially we have labeled training dataset P and unlabelled training dataset Q. Initially P and Q are combined to obtain. For ach image in Q, Find K nearest images based on Euclidean distance N'. If N, does not contain any image in P, image I in Q will be labeled as unknown. In the proposed system K represent the number of nearest neighbors. If K increases the probability of labeling an image an unknown will decrease and if K decreases more images will be labeled as unknown. Therefore an optimal K should be chosen to balance the accuracy and size. For this parameter optimization is also introduced [1]. Algorithm for unknown detection is shown below:

Algorithm 1: KNN based unknown detection

```
Input   : labeled training dataset P
          Unlabelled training dataset Q
          Parameter K
Output:   Image set of unknown models U

U←Ø
T← P U Q
For every image I in Q
        N'←KNN ( I, T, K, Euclidean distance )
        If N' ∩ P == Ø
                 ← U ∪ {i}
        End
End
Return U
```

The determination of K is done using parameter optimization which is shown in algorithm 2:

Algorithm 2: Parameter optimization

```
Input   : labeled training dataset P
          Unlabelled training dataset q
          $T_{fpr}$ (default 0.5 % )

Output:   Optimal K

ΔP ←randomly selected 10 % images from P
P' ← P – ΔP
Q' ←Q + ΔP
T'← P' U Q'
For  k← 1 to $k_{maz}$ do
U'← Ø
For each image I ε ΔP do
N'← KNN (I, T', K, Euclidean distance)
If N' ∩ P' == Ø then
U' ← U' ∪ {I}
end
end
$FPR_k^{'}$ ← U/ ΔP
If  $FPR_k^{'} < T_{fpr}$  then
$K_{opt}$ ← K
Break
End
End
$K_{opt}$ ← $K_{max}$
Return $K_{opt}$
```

C. Unknown Expansion

Unknown detection can recognize some images from unknown models. In order to obtain more sample images for unknown models, a self-training based unknown expansion method is proposed. A stopping criterion of the self-training process is also used for optimal results.

Self-Training Based Unknown Expansion:

Selftraining is a bootstrapping method that aims to improve the performance of a machine learning algorithm by incorporating unlabelled data into the training procedure. In the proposed system, a self-training procedure is employed to extract more sample images of unknown models from the unlabelled training dataset. Let us consider the labelled image set P, the unknown image set through unknown detectionU0 = U, and the remained unlabelled image setQ0 = QU. The sth iteration of unknown expansion is performed as follows

- Combine P and Us-1asT s = P U Us-1, and regard Ts as training dataset to train a multi-class SVM Cs .
- Use Cs to classify images inQs-1. Suppose the unknown image set labelled by Cs is ΔUs. The labelled image set is updated asUs = Us-1 U ΔUs . The unlabelled image set is updated as Qs = Qs-1 - ΔUs.

The stopping criterion for self-training is set in consideration of the final identification accuracy and the computation cost. The stopping criteria is set default as 0.5%. Algorithm 3 depicts the algorithm for unknown expansion.

Algorithm 3: Unknown Expnsion

```
Input   : labeled training dataset P
          Unlabelled training dataset Q
          $T_{dir}$ (default 0.5 %)

Output:   Expanded unknown

$U_0$← U
 $Q_0$ ← Q -U
S ←1 to S do
$ $T_s$ ← P ∪$U_{s-1}$
//Regard $T_s$ as training dataset to train a multi-class SVM  $C_s$
$C_s$ ← Train SVM($T_s$ )$
//Use  $C_s$  to classify images in  $Q_{s-1}$
 L ($Q_{s-1}$) ← SVM Predict ($C_s$, $Q_{s-1}$)
//Us is the set of images labelled as unknown by Cs , 0
denotes the label of the class for unknown models.
$U_s$ ← $U_{s-1}$ ∪ Δ$U_s$
$Q_s$ ← $Q_{s-1}$ - Δ $U_s$
DIR ←Δ $U_s$ / $U_{s-1}$if DIR < Tdir then
Return $U_s$
end
end
return $U_s$
```

D. Classification

In traditional schemes, camera model identification is solved with the K-class classifiers, K is the number of known camera models. When the testing images are from unknown models, they will be inaccurately classified into the classes of known models. The proposed system deals with the unknown models by addressing a specific (K + 1)-class classification. In the (K + 1)-class classification, the sample images of the unknown models discovered through unknown detection and unknown expansion are regarded as 1-class, and the images of K known models are treated as K-class. When the testing images are from unknown models, (K + 1)-class classifier will classify them into the specific unknown class. Therefore, the proposed system has the capability of identifying the images of the unknown models as well as distinguishing the images of the known models.

IV. EXPERIMENTAL RESULTS

The following section describes the experimental setup and results obtained from the implementation of the system. A large number of real world images were used to evaluate the performance of the system

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Dataset

Dresden image collection is used in the proposed system for empirical study. The open image collection was specifically built for the purpose of development and bench-marking of camera-based digital forensic techniques. It was created using different scenes of natural and urban environments as well as indoor and outdoor environments.

In the proposed system the test image for which the source camera is identified is given as an input to the system. In the existing system 34 features were extracted and based on these features the source camera is identified. The identification accuracy of such systems were found lower. In the proposed system each input image is represented in 10 different formats and from these 10 representations 34 features are extracted. Thus in total 340 features are extracted from a given input image.

B. Performance Evaluation

The overall accuracy (OACC), precision and recall are used to measure the performance of camera model identification.

_ OACC is used to measure the overall identification accuracy, which is the ratio of the number of all correctly identified images to the number of all identified images.

$$\text{OACC} = \frac{\text{correctly identified images}}{\text{Total no:of images identified}}$$

_ precision is defined as the ratio of number of correctly identified images to the number of images identified.

$$\text{precision} = \frac{\text{correctly identified images}}{\text{total number of images identified}}$$

_ recall is defined as the ratio of number of images identified from a camera model to the total number of imaged identified from the model .

$$\text{recall} = \frac{\text{correctly identified images from a camera model}}{\text{total number of images identified from the model}}$$

In this experiment the influence of unknown models is evaluated to conventional identification schemes. The experiment was carried out to compare our proposed SCI scheme to the state-of-art camera model identification methods including binary SVMs method (BSVM), combined classification framework method (CCF) and decision boundary carving method (DBC).

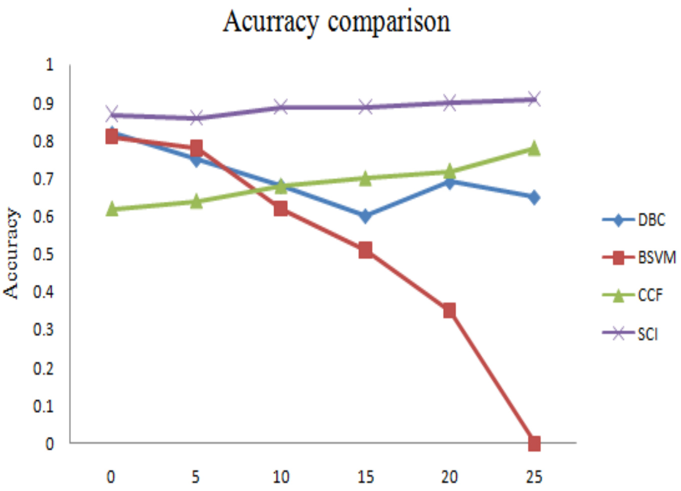

Figure 2: Accuracy comparison of various systems

Fig 2 shows the accuracy comparison of these methods with the proposed system. From the experimental results, it is found that the proposed SCI scheme is significantly superior to the other four methods. The second best is the CCF method. The accuracy difference between SCIU and CCF is about 18on average. The BSVM method has the worst performance and DBC is superior to BSVM. The cause of the low accuracy exhibited by BSVM is the inaccurate classification of the images from unknown models into known models. BSVM and DBC both take advantage of the images of other known models to approximate the unknown class. However, DBC adjusts the decision boundary of the SVM to alleviate the negative effect of unknown models. Therefore, DBC achieves better performance than BSVM. The CCF, DBC, and BSVM methods dont make use of the information of unknown models, thus the performance improvement is limited. The proposed SCI is able to utilize the information of unknown models discovered by unknown detection and expansion, so as to improve the identification accuracy significantly.

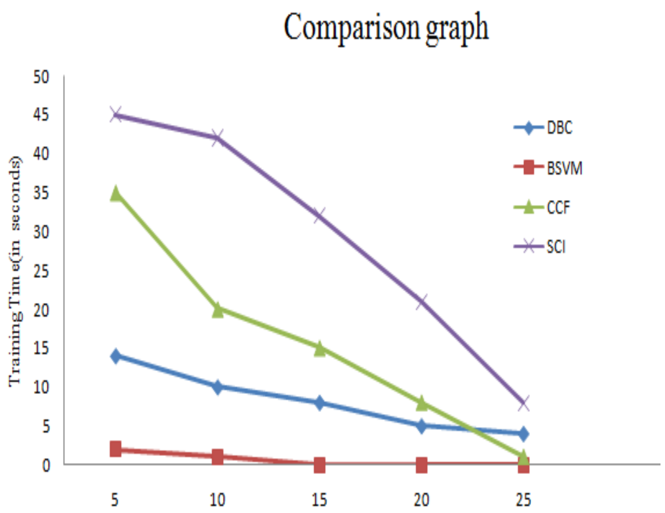

Figure 3: Training time comparison graph

Fig 3 shows the training time comparison of different source camera identification system. From the graph it is clear that the training time required for the SCI scheme is higher compared to other conventional source camera identification scheme but it will not affect the accuracy or performance because the training process is done offline

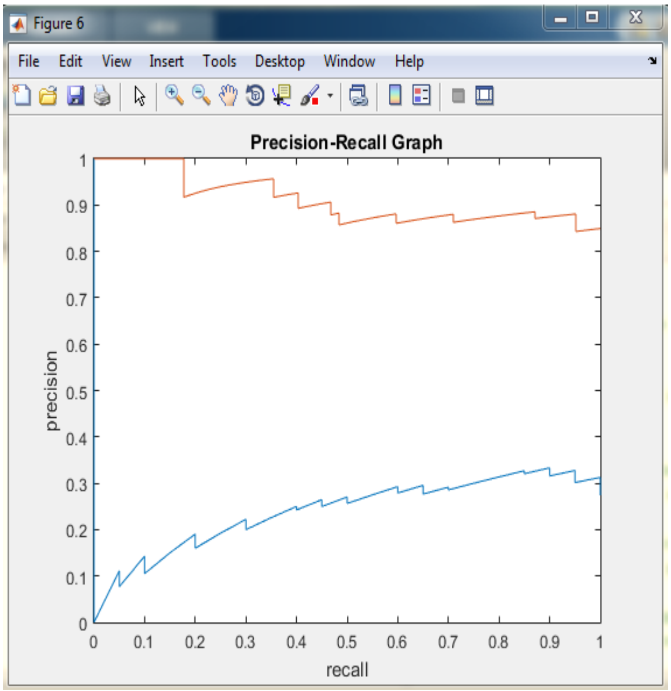

Figure 4: Precision recall graph

Fig 4 depicts the precision-recall graph of the proposed system

Receiver Operating Characteristics (ROC) curve is a plot of the true positive rate against the false positive rate for the different possible cut points of a diagnostic test. It shows the trade-off between sensitivity and specificity. Closer the curve follows the left hand border and the top border of the ROC space, more accurate the test. Figure 7 shows the ROC curve for the proposed system which depicts that the accuracy of the system is found higher.

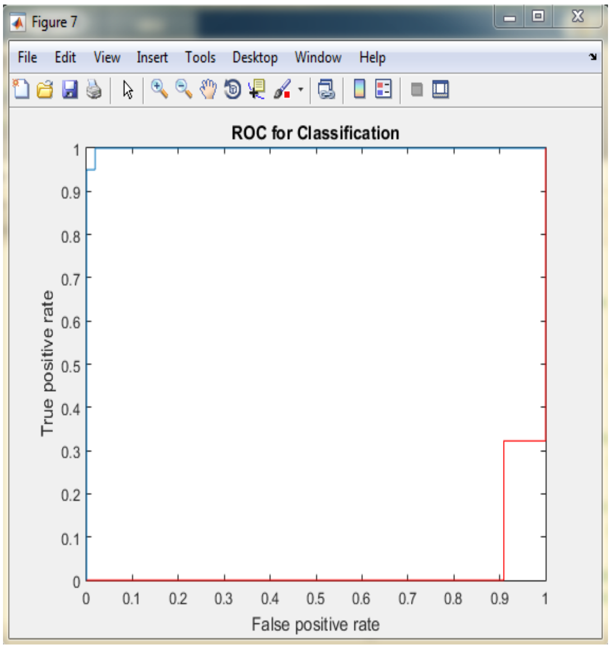

Figure 5: ROC curve

V.CONCLUSION

The proposed work addressed the critical problem of unknown models in camera model identification. Most of existing camera model identification methods suffer from poor performance when the unknown models present. This is due to misclassification of images of the unknown models into the known models. A new scheme is proposed, SCI which can improve the identification accuracy by identifying the images of unknown models. The proposed SCIU consists of three stage: unknown detection, unknown expansion and (K + 1)-class classification. In order to optimize the performance and efficiency of the proposed solution, a parameter optimization method is developed for unknown detection and investigated the stopping criterion for unknown expansion. To evaluate the new scheme, a large number of experiments were carried out on a real-world image collection. The results demonstrate that the proposed SCI scheme significantly outperforms four state-of-the-art methods. The future work will focus on applying the proposed scheme on large-scale image collections, following the way of Goljan et al. [17]

REFERENCES

[1] Yonggang Huang, Member, IEEE, Jun Zhang, Member, IEEE, and Heyan Huang,"Camera Model Identification With Unknown Models", IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 10, NO. 12, DECEMBER 2015

[2] Lawgaly, Fouad Kheli, Ahmed Bouridane Image Sharpening for Effcient Source Camera Identification Based on Sensor Pattern Noise Estimation, 2013 Fourth International Conference on Emerging Security Technologies.

[3] Z. Deng, A. Gijsenij, and J. Zhang, Source camera identification using auto-white balance approximation, in Proc. 13th IEEE Int. Conf. Comput. Vis., Barcelona, Spain, Nov. 2011, pp. 5764.

[4] Yue Tan, Bo Wang, Meijuan Zhao, Xiangwei Kong, and Ming Li, Patch-Based "Sensor Pattern Noise for Camera Source Identification", in Proc. IEEE 12th ICCV, Sep./Oct. 2009

[5] Yongjian Hu1, Chang-Tsun Li, Changhui Zhou Selecting Forensic Features for Robust Source Camera Identification.

[6] A. E. Dirik, H. T. Sencar, and N. Memon, Source camera identification based on sensor dust characteristics", in Proc. IEEE Workshop Signal Process. Appl. Public Secur. Forensics, Washington, DC, USA,Apr. 2007, pp. 16.

[7] Kai San Choi, Edmund Y. Lam, Kenneth K.Y. Wong Source Camera Identification by JPEG Compression Statistics for Image Forensics.

[8] S. Bayram, H. Sencar, N. Memon, and I. Avcibas, Source camera identification based on CFA interpolation, in Proc. 12th IEEE Int. Conf. Image Process., Genoa, Italy, Sep. 2005, pp. 6972.

[9] Kharrazi, H. T. Sencar, and N. Memon, Blind source camera identification, in Proc. 11th IEEE Int. Conf. Image Process., Singapore, Oct. 2004, pp. 709712.

[10] F. de O. Costa, E. Silva, M. Eckmann, W. J. Scheirer, and A. Rocha, Open set source camera attribution and device linking, Pattern Recognit. Lett., vol. 39, pp. 92101, Apr. 2014.

[11] H. Farid and S. Lyu, Detecting hidden messages using higher-order statistics and support vector machines 5th International Workshop on Information Hiding.,2002

[12] J. Luk, J. Fridrich, M. Goljan, Digital camera identification from sensor pattern noise IEEE Trans. Inf. Forensics Secur. 1(2) (2006) 205214.

[13] C.-T. Li, Source camera identification using enhanced sensor pattern noise, IEEE T.IFS, vol. 5, no. 2, pp. 280287, June 2010.

[14] T. Cover and P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 2127, Jan. 1967

[15] T. Gloe, Feature-based forensic camera model identification, in Transactions on Data Hiding and Multimedia Security VIII. Berlin, Germany: Springer-Verlag, 2012, pp. 4262.

[16] B. Wang, X. Kong, and X. You, Source camera identification using support vector machines, in Advances in Digital Forensics V. Berlin, Germany: Springer-Verlag, 2009, pp. 107118.

[17] M. Goljan, J. Fridrich, and T. Filler, Large scale test of sensor fingerprint camera identification, Proc. SPIE, vol. 7254, p. 72540I, Feb. 2009.

[18] K. S. Choi, E. Y. Lam, and K. K. Y. Wong, Source camera identification using footprints from lens aberration, Proc. SPIE, vol. 6069, pp. 60690J-160690J-8, Feb. 2006.