Machine Learning Based Detection of Digital Documents Maliciously Recaptured from Displays

Saleh Gholam-Zadeh[a], Evgeniy Upenik[a], Guy Hatarsi[b], and Touradj Ebrahimi[a]

[a]Multimedia Signal Processing Group (MMSPG),
Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
[b]Quantum Integrity SA, EPFL Innovation Park, Bâtiment C, CH-1015 Lausanne, Switzerland

ABSTRACT

We used to say "seeing is believing": this is no longer true. The digitization is changing all aspects of life and business. One of the more noticeable impacts is in how business documents are being authored, exchanged and processed. Many documents such as passports and IDs are being at first created in paper form but are immediately scanned, digitized, and further processed in electronic form. Widely available photo editing software makes image manipulation quite literally a child's play increasing the number of forged contents tremendously. With the growing concerns over authenticity and integrity of scanned and image-based documents such as passports and IDs, it is more than urgent to be able to quickly validate scanned and photographic documents. The same machine learning that is behind some of the most successful content manipulation solutions can also be used as a counter measure to detect them. In this paper, we describe an efficient recaptured digital document detection based on machine learning. The core of the system is composed of a binary classification approach based on support vector machine (SVM), properly trained with authentic and recaptured digital passports. The detector informs when it encounters a digital document that is the result of photographic capture of another digital document displayed on an LCD monitor. To assess the proposed detector, a specific dataset of authentic and recaptured passports with a number of different cameras was created. Several experiments were set up to assess the overall performance of the detector as well as its efficacy for special situations, such as when the machine learning engine is trained on a specific type of camera or when it encounters a new type of camera for which it was not trained. Results show that the performance of the detector remains above 90 percent accuracy for the large majority of cases.

Keywords: image manipulation, image forgery, forgery detection, manipulation detection, KYC, image falsification, falsification detection, machine learning

1. INTRODUCTION

Digitization has largely revolutionized our private and professional lives. Many tasks and activities that relied on paper-centred workflows have been transformed into more efficient digital counter parts that are less costly while at the same time being more efficient from many perspectives. This transformation into all digital workflows, has even transformed entire industries both in manufacturing and services, in business to business and in business to consumer, while in many cases opening doors to not only new actors but dominant players at times causing weakening if not extinction of existing dominant players. There is a large consensus that digitization has been a beneficial change with considerable advantages, but also that it has a number of drawbacks, reinforcing some if not creating new weaknesses.

One important weakness of digitization is in its growing concern regarding its security whether actual or perceived. It is a fact that in a digital world where old ways of doing things are transformed and sometimes even replaced by totally new ways, equal opportunities are created to carry out new criminal activities in particular in the areas of fraud and theft. A root cause of many of security concerns in a digital world is because of fundamental differences between information stored in digital versus physical forms. As an example, not only it

Send correspondence to Prof. Dr. Touradj Ebrahimi
E-mail: touradj.ebrahimi@epfl.ch

is trivial to duplicate a digital document, but also it is impossible to know the differences between the original and the copy of a digital document unless specific measures are put in place.

Without entering into details regarding the rather complex issue of the fundamental differences between digital and physical documents and their respective advantages and drawbacks, for the purpose of this paper, it suffices to mention two types of possible security breaches which are widespread today, namely, digital theft and digital forgery of electronic documents including physical identification documents such as passports. More precisely, in this paper, we concentrate on the former which is often the prerequisite of the latter, as it is an obvious fact that in order to digitally alter a document, an attacker needs to either posses the physical document in to produce an electronic version, or access the electronic version of the original document in order to forge it.

This type of fraud is becoming increasingly widespread because many businesses, in order to initiate a digital transaction, require assurances that the party with whom they interact is who (s)he pretends to be and is not in fact an impostor. This is referred to as KYC (Know Your Customer) and consists in a number of procedures and operations that provide sufficient and reasonable assurances in order for a provider to know that the customer is indeed who (s)he pretends to be. A corner stone of KYC is that the customer provides the service provider with a digital or physical proof of identity. Although digital identity is an active field of research and innovation, it is not yet widely adopted and therefore in most cases, the customer needs to provide an electronic version of their physical identity card or passport issued by the authorities of the country of residence. This is done in majority of cases by customers taking a picture of their physical passport or identity card and transmitting to the service provider.

In order to prevent identity theft, many reputable and reliable service providers (such as banks) do not allow their employees and operators to have direct access to the electronic version of scanned identity cards and passports sent by their customers for the purpose of KYC. In other words, it is impossible to copy, transfer to another location, or to store on an external storage drive such documents. Therefore, the only way for a malicious access to the electronic version of an identification card or a passport is that the potential attacker takes a picture of what is displayed on the screen of the computer with a mobile phone or a digital camera. In this paper we concentrate on this exact problem.

The goal of this paper is to describe an efficient detector that is proposed in order to distinguish between an electronically scanned/photographed version of a physical document and a version of an existing electronic document recaptured by means of a photographic device from a display. Doing so, we also overview the state of the art of related solutions, focus on identity cards as documents and assess the performance of the proposed detector by creating a specific database of electronic passports either directly captured or recaptured from a display, and designing a methodology to properly assess the performance of the proposed detector. We then analyse the results and highlight key points in the proposed solution.

The paper is divided into seven sections. After this introduction, in Section 2 we analyse the state of the art. Section 3 is devoted to the description of the proposed detector . Section 4 discusses the performance assessment methodology used as well as the dataset created for this specific purpose. Results are provided in Section 5 followed by a discussion in Section 6. Main conclusions of the paper as well as future directions where we intend to follow up with the reported work are provided in Section 7.

2. STATE OF THE ART

In the past decades, digital images have become increasingly popular in everyday life. At the same time, an easy accessibility of digital images has resulted in significant security challenges. Various image editing tools are available which can edit an image in different ways. By means of such tools, one can bring changes into any image without leaving visually noticeable artifacts, making the aforementioned challenges increasingly more difficult. Since images are treated as proofs in many circumstances, such as identification of a person for conditional access, such manipulations can have serious consequences. Therefore in today's world, verification of images plays an important role. Image forensics is a branch of image analysis aiming at obtaining evidence of potential manipulations of an unknown image.
Image forensic techniques can be categorized under two different approaches: Active and Passive.[1]

Watermarking and digital signature are two examples of popular active protection techniques. The former involves designing techniques that can add imperceptible marks inside an image. In the verification stage, the embedded watermark is extracted and examined in order to determine whether an image has been tampered with, and if so, the area in the image that it has occurred.[2] Although, active approaches can detect digital image tampering rather accurately, they have not been popular because it is difficult to impose that all digital images are watermarked prior to their distribution.[3] Passive approaches do not require any a priori actions, such as insertion of a watermark or digital signature in an image. Therefore, they have a broader application in image forensics. These approaches detect digital image forgeries by analyzing specific inherent clues or patterns that occur during creation or modification of a digital image.[4]

According to[5] image forensic techniques can be roughly categorized into five groups: (1) pixel-based techniques that are designed to detect alterations in the pixel domain. Figure 1 shows categories of pixel-based methods.

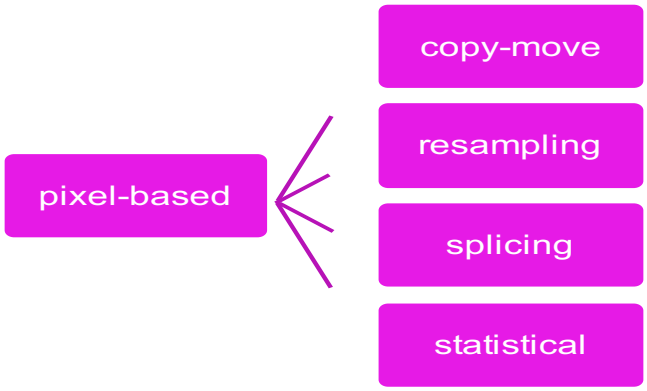

Figure 1: Categories of pixel-based methods

(2) format-based techniques that detect statistical correlations introduced by a specific lossy compression or transcoding operation. Figure 2 shows categories of format-based methods.

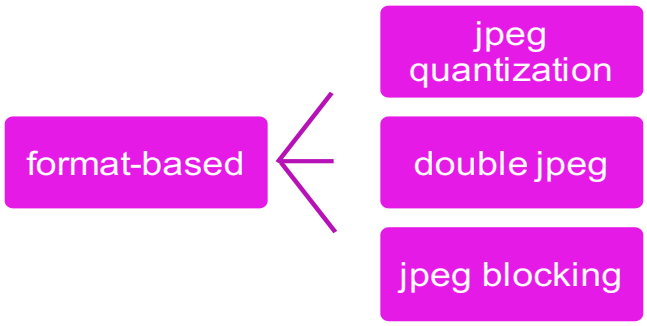

Figure 2: Categories of format-based methods

(3) camera-based techniques that leverage on artifacts generated by the camera lens, sensor or on-chip post-processing. Figure 3 shows categories of camera-based methods.

(4) physical environment-based techniques explicitly model and detect anomalies in the three dimensional interaction between physical objects, light, and the camera. These technique are divided into three categories as shown in the figure 4.

and (5) geometry-based techniques make measurements of objects in the world and their positions relative to the camera. Geometry-based image forgery techniques are divided into two categories as shown in the figure 5.

For passive image forensics, various kinds of traces can be exploited to distinguish tampered from authentic images.[6] In the paper[3] these traces have been categorized into three groups: traces left in image acquisition, traces left in image storage, and traces left in image editing. Here we briefly describe traces left during image

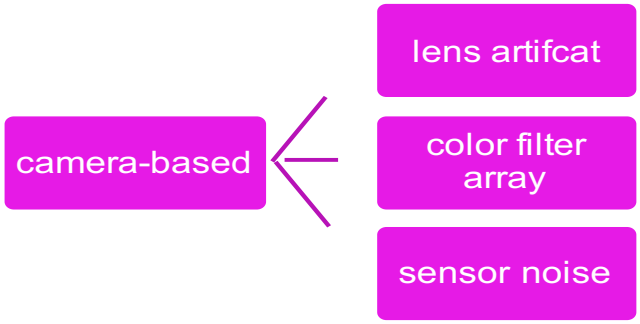

Figure 3: Categories of camera-based methods

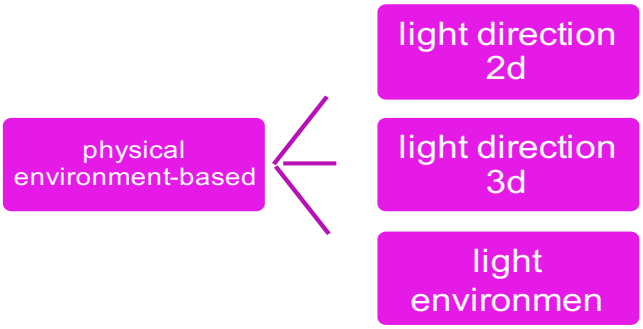

Figure 4: Categories of physical environment-based methods

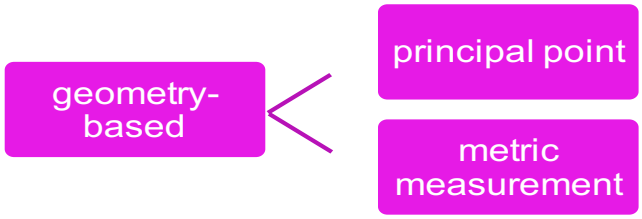

Figure 5: Categories of geometry-based methods

acquisition process since it is most relevant to our work. Image acquisition pipeline is shown in the figure 6. When a digital image is captured, several processing steps are performed prior to storage. Before hitting the sensor, the light from the scene goes through a series of lenses. The sensor then captures the incident rays of light through a color filter array (CFA), configured according to a specific color mosaic, which only allows a specific color component of the light to pass through a specific area of the sensor. In most digital cameras, only one color component (e.g. R, G and B) is allowed to reach each pixel. After CFA, the light reaches the sensor which is the essential part of any digital camera. The large majority of cameras use one of the following two sensor technologies to convert light into electric signal: charge-coupled device (CCD) or complementary metal-oxide semiconductor (CMOS). A number of photo detectors are contained within the sensor and each of their output is related to a pixel of the image. In each detector, the filtered light is transformed into a corresponding voltage; So, the output of the sensor is a mosaic of three color components with various intensity values. In order to obtain the integrated color information for every pixel in the image, a demosaicing procedure must be performed. Demosaicing is an interpolation process applied to all color channels in order to produce the missing pixels for every color component.

3. DETECTOR OVERVIEW

In this section, we describe the proposed machine learning-based detector that is designed to distinguish between authentic and recaptured images, through a binary classification. As in any learning-based method, the approach is divided into a training and a testing phase. The block diagram of the proposed detector in the training phase

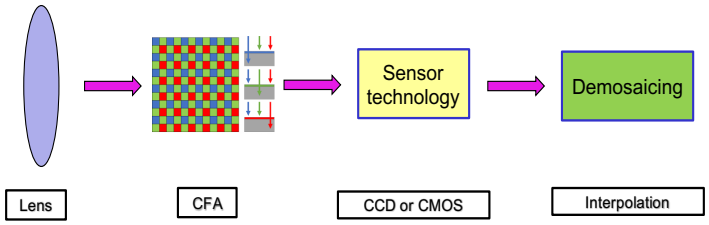

Figure 6: Image acquisition pipeline

is shown in the figure 7. Recaptured images can be of any size and captured from any distance and any angle. They can also be resized and even rotated prior to their deployment, e.g. before sending to a service provider as a proof of identity in a KYC use case. Therefore, the detector should be robust to any such transforms. To cope with the above-mentioned situation, as it will be descried further, the collected dataset contains pictures of various sizes captured from different distances. Furthermore, in order to make the detector more robust, data augmentation is performed by rotating every picture in the dataset during training. In this paper, we only considered rotations of 90, 180 and 270 degrees. Prior to augmentation, the first step of the training consists in resizing every image to 256*256 pixels. This operation is performed to normalize the input to the classification algorithm and can be seen as a pre-processing operation. In our approach, actual pixels of the resized and augmented pictures are used directly as features that will consequently will be subjected to Principal Component Analysis (PCA). The most relevant principal components are then retained. In our implementation, a threshold of 500 principal components was selected based on trials.
The last stage of the training phase consists in feeding the most relevant principal components and their corresponding labels into a support vector machine classifier with radial basis function (RBF) kernel to to train the parameters of the model.

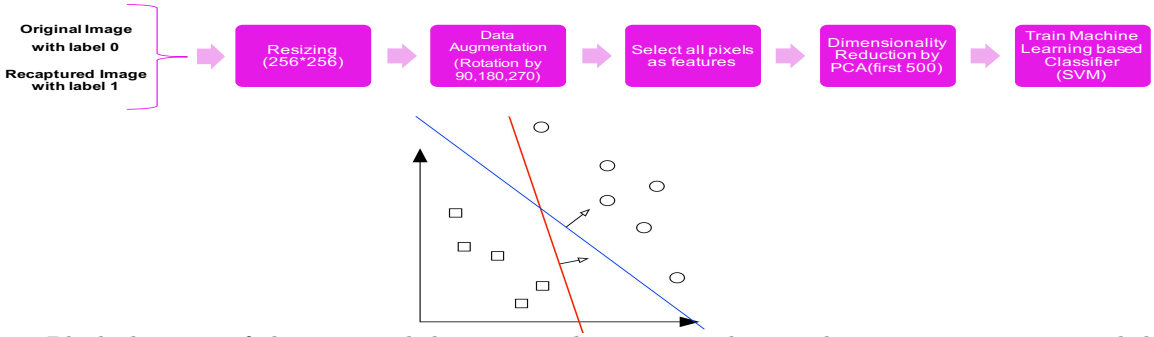

Figure 7: Block diagram of the proposed detector in the training phase. The inputs are images with known labels.

During the testing phase, as shown in the figure 8, the input is an unknown image. Similar to the training phase, the first step in the testing phase is to resize the input unknown image to 256*256 pixels. Here again, all pixel values of the resized image are considered as features. The next step consists in PCA after which the first 500 principal components are retained. Finally the trained SVM classifier is applied to predict the label and consequently predict if the unknown image is authentic or recaptured.

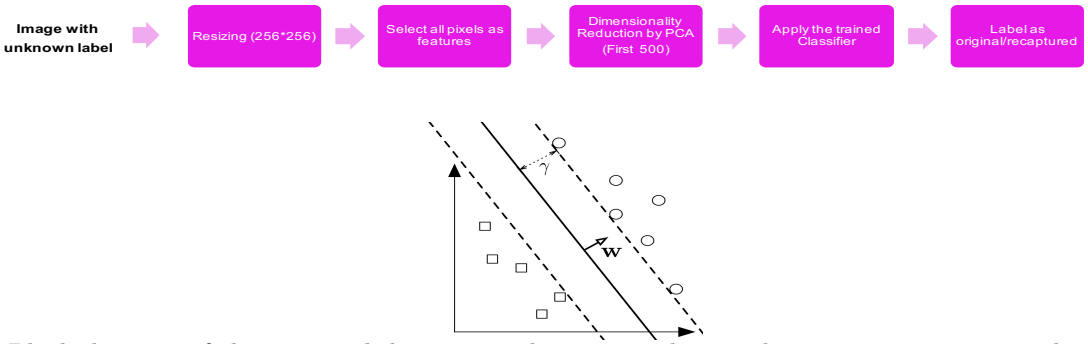

Figure 8: Block diagram of the proposed detector in the testing phase. The input is an image with unknown label.

4. EXPERIMENTS

4.1 Dataset

In order to asses the performance of the proposed detector, a labelled dataset consisting of 1499 scanned original passports and 1499 recaptured passports was collected. The original data were stored in JPEG format with quality factors varying between 0.65 and 0.96 and the recaptured passports are JPEG compressed with a quality factor between 0.95 and 0.98. As mentioned before, pictures in the dataset were selected to be of different sizes and taken from different distances and different angles. The recaptured passports were taken with 9 different types of cameras as shown in figure 9: NEX5R, RICOHCX2, IPHONE-6S, IPHONE6S-HDR, NIKOND100, NIKOND3400, IPHONE-SE, IPHONE-SE-HDR,SONYNEX5R. A typical example of an original and a recaptured passport is shown in figure 10.

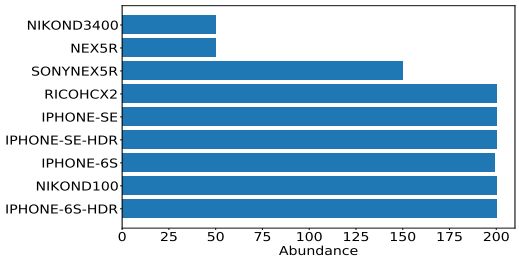

Figure 9: Distribution of recaptured data.

Since a rotated passport is considered as a valid passport, the dataset was further augmented with rotated versions of all images in the initial dataset. For augmentation only 90, 180 and 270 degree rotations were applied.

(a) original passport (b) recaptured passport

Figure 10: Example of an original and a recaptured passport

4.2 Evaluation criteria

In the experiments, accuracy, precision and recall are used as evaluation criteria. When denoting the number of true positives, true negatives, false positives and false negatives by TP, TN, FP, FN respectively, one can define Accuracy, Precision and Recall as follows.

$$Accuracy = \frac{\text{Number of correctly classified samples}}{\text{Total number of tested samples}} = \frac{\text{TP + TN}}{\text{TP + TN + FP + FN}} \tag{1}$$

$$Precision = \frac{\text{TP}}{\text{TP + FP}} \tag{2}$$

$$Recall = \frac{\text{TP}}{\text{TP + FN}} \tag{3}$$

In the above a recaptured picture of a passport is considered as being positive and an authentic picture of a passport (original) is considered as negative. Therefore, a true positive denotes a recaptured passport that has been correctly detected as a recaptured passport, a false positive is an authentic passport that is wrongly detected as recaptured, a true negative is a an authentic passport that is correctly detected as authentic and a false negative represents a recaptured passport that is falsely detected as authentic .

4.3 Assessment methodologies

This section describes in more details the assessment approach adopted in the paper in order to evaluate the performance of the proposed detector. Three types of experiments were designed and carried out.
To assess the overall performance of the detector a first experiment randomly selected 80% of the augmented dataset for training and the remaining for testing. This process was repeated 5 times.

A second experiment was designed to assess the robustness of the detector to a new type of camera, i.e. to evaluate the performance of the detector in presence of data captured by a new type of camera. To do so, the data in the dataset that was captured from one pre-selected type of camera was excluded during the training phase in order to prevent the detector encounters any data from that specific type of camera. The same amount of samples from original data was considered for testing in order to have a balanced test set. Finally the detector was trained on the rest of the dataset prior of its testing on a balanced test set. This process was repeated separately for each specific type of camera in the dataset.

To assess the testing performance of the detector when both tested and trained on a specific type of camera, in a third experiment a separate model was produced for each camera type and tested on the data corresponding to the same camera carefully avoiding that the same data is used for both training and testing.

In all experiments, during the test phase, the number of authentic and recaptured images were the same.

To better visualize the distribution of the collected data within the collected dataset and to better interpret its potential classification feasibility, t-SNE[7] algorithm was applied to project the data from each camera into a two-dimensional feature space, as shown in figure 11. One can observe that IPHONE camera types have a closer distribution to the original data, hence, it could be expected that the classifier shows lower performance for these types of cameras.

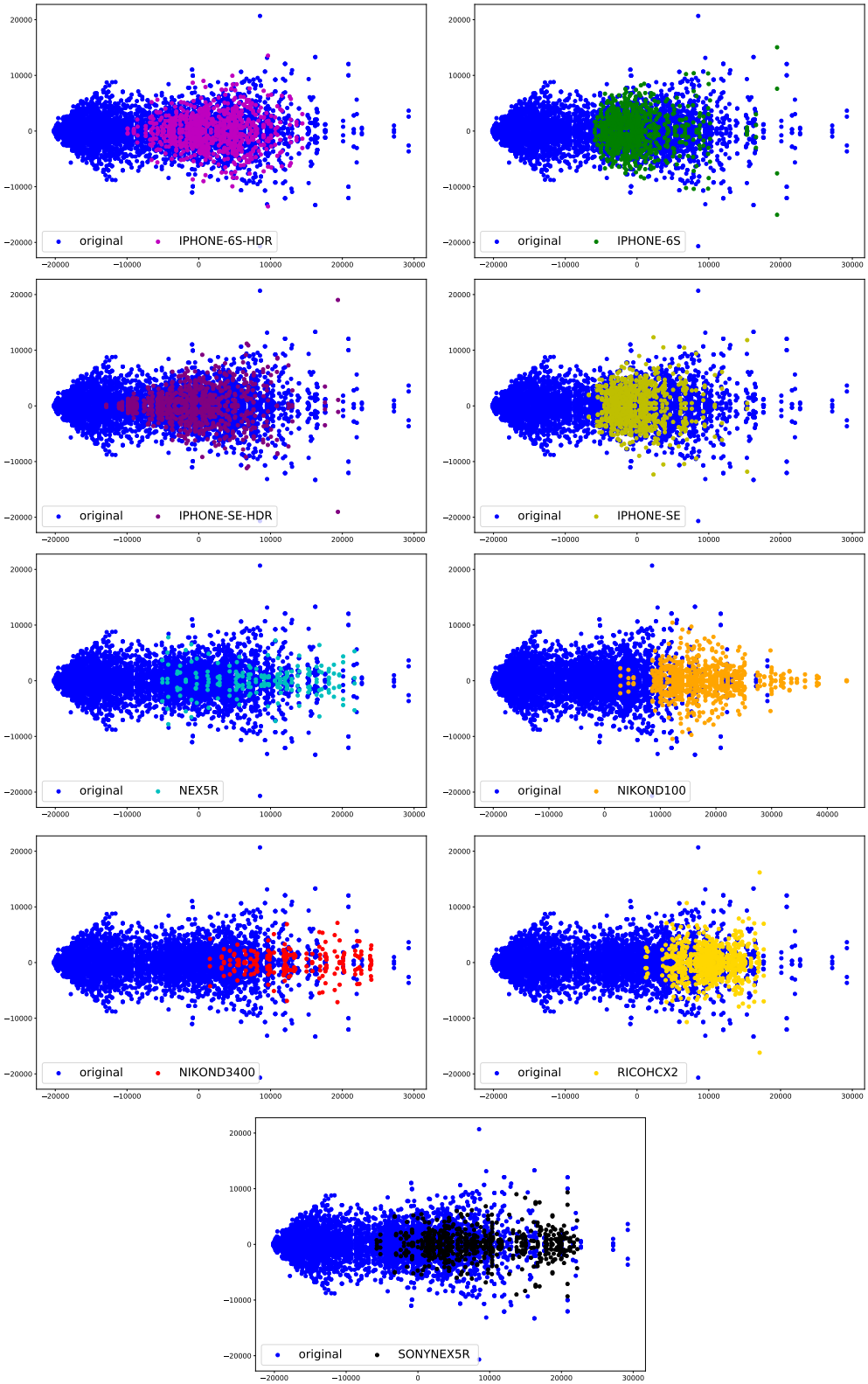

Figure 11: Visualization of our data using t-SNE algorithm

5. RESULTS

5.1 Experiment 1

To assess the overall performance the detector was trained using 80% of the entire augmented dataset randomly selected for training and the remaining 20% for the testing and this process was repeated 5 times. The results are shown in the table 1

Table 1: Results of the first experiment in terms of accuracy, precision and recall. Recaptured samples are considered as positive samples. All accuracy values are above 90%

	trial 1	**trial 2**	**trial 3**	**trial 4**	**trial 5**	**Avg**
Accuracy	0.9120	0.9086	0.9025	0.9107	0.9064	0.908
Precision	0.8919	0.8927	0.8763	0.9018	0.8941	0.891
Recall	0.9329	0.9242	0.9332	0.9170	0.9226	0.926

5.2 Experiment 2

In this experiment data corresponding to a specific type of camera was left out during the training phase and that data together with the same amount of data from authentic data was used for the testing phase. This process was repeated for all types of cameras. Results are shown in table 2 in which each column corresponds to a specific type of camera that was left out during the training phase. Results of this experiment show that there is a small and insignificant drop of around 1% in average accuracy with respect to the previous experiment. This shows that the detector is robust to new types of cameras.

Table 2: Results of the second experiment in terms of accuracy, precision and recall. Recaptured samples are considered as positive samples. Average accuracy is close to 90% which is only 1% less than the previous experiment, indicating robustness of the detector to new types of cameras.

	NEX5R	RICOHCX2	IPHONE 6S-HDR	NIKON D100	IPHONE SE	IPHONE 6S	IPHONE SE-HDR	SONY NEX5R	NIKON D3400	AVG
Accuracy	0.9475	0.9444	0.8869	0.8775	0.8881	0.8989	0.8088	0.9408	0.88	0.897
Precision	0.9091	0.8939	0.8736	0.8792	0.8838	0.8863	0.8709	0.8982	0.8780	0.886
Recall	1.0000	1.0000	0.8912	0.8550	0.8938	0.9108	0.7250	1.0000	0.9000	0.908

5.3 Experiment 3

In a third experiment a specific SVM model was trained for each type of camera, to assess performance of the detector if tailored for a specific type of camera. In the other words, in the training phase only data from a specific camera as well as well as their authentic data was considered and the same procedure was repeated for the testing phase. Results are shown in table 3 where each column corresponds to a specific type of camera that was used for training and for testing carefully avoiding that the same data is used in both. As expected, the performance figures are higher in average than the two previous experiments.

Experiments 2 and 3 further indicate that the performance of the proposed detector is camera dependent.

Another important observation is that the performance of the proposed detector on IPHONE cameras are lesser than for other cameras which is in line with figure 11, as distribution of IPHONE cameras visually are closer to those of the authentic data.

Table 3: Results of the third experiment. Recaptured samples are considered as positive samples. Here a single camera type was used for both training and testing. As expected, the performance figures are higher in average than the two previous experiments.

	NEX5R	RICOHCX2	IPHONE 6S-HDR	NIKON D100	IPHONE SE	IPHONE 6S	IPHONE SE-HDR	SONY NEX5R	NIKON D3400	AVG
Accuracy	0.9750	0.9625	0.8781	0.9438	0.9031	0.8836	0.8594	0.9458	0.9375	0.921
Precision	0.9524	0.9302	0.8954	0.9128	0.9057	0.8675	0.8958	0.9023	0.9487	0.912
Recall	1.0000	1.0000	0.8562	0.9812	0.9000	0.9057	0.8063	1.0000	0.9250	0.930

6. DISCUSSION

Let us interpret in this section the results of the experiments performed using the proposed machine-learning based system for detection of maliciously recaptured digital documents and discuss possible causes of the outcomes of the performed evaluations.

As we have seen in Section 5, the performance of our system may depend on the camera model that was used to maliciously recapture authentic images of passports from a display. One can argue that the more data we have from a particular type of camera, the higher accuracy will be achieved. It is, however, very often the case that a new type of camera becomes available. In Experiment 2, the issue of detection in presence of an unknown camera is investigated. Moreover, the results in Table 2 show that, even though, in some cases the accuracy drops when the data for a particular camera model have been left out from the training, the average accuracy changes only around 1% comparing to the first experiment. That is to say that in the long term and, if the unknown cameras only appear occasionally in the analysed data, the statistical ability of the proposed system to detect fraud attempts stays virtually the same and does not result in any financial loss when deployed in practical operations.

The single cases of the drop in performance need however, to be investigated. From the values in Tables 1 and 3 one can notice that for the major part of camera models, namely, NEX5R, RICHOCX2, NIKON D100, SONY NEX5R, and NIKON D3400, the accuracy increases, when the system is trained on the data containing images from only a single camera model, whilst for a few other camera models, namely, IPHONE 6S-HDR, IPHONE 6S, and IPHONE SE-HDR it slightly drops. In addition, according to Figure 9, the relative amount of training data per camera model does not correlate with this drop of accuracy. The above observations allow one to hypothesize that the selection of features is not optimal for the latter camera models and the performance can be possibly improved without a need of a larger training dataset.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we presented an overview and classification of malicious image manipulation methods. We described the main state-of-the-art counter measure techniques and methods designed to address these types of digital forgery. Then we addressed a rather specific case of KYC fraud attempts, which consists in malicious recapturing of digital identification documents displayed on a computer screen.

We proposed an end-to-end system to detect inauthentic digital documents that were recaptured without authorisation in order to be further used in, for example, deception activities that intend to steal a person's identity. The proposed system is based on SVM that requires fewer training data when compared to the now more popular deep-learning based approaches. We then performed a thorough evaluation of the performance of the proposed end-to-end system. The results of the evaluation show that the accuracy of the detection of inauthentic content is higher than 90% when averaged among five trials.

The future work should address the limitations of the proposed system when it is fed with the data that were produced using camera models that were absent in the training data. Nonetheless, the average drop in accuracy for unknown camera models did not exceed 1%, the single cases for particular cameras show a more significant accuracy reduction. The issue of unknown cameras may be addressed either by increasing the number of types of cameras used in the training which would possibly lead to a better generalisation ability of the model or by improving the feature selection stage with an attempt to find representation vectors that are invariant to the camera differences.

ACKNOWLEDGMENTS

Authors would like to acknowledge contribution from Innosuisse under grant number 34270.1 lP-lCT for the project entitled "Deep Fake Factory".

REFERENCES

[1] Kumar, S. and Das, P., "Copy-move forgery detection in digital images: progress and challenges," *International Journal on Computer Science and Engineering* **3**(2), 652–663 (2011).

[2] Zhou, G. and Lv, D., "An overview of digital watermarking in image forensics," in [*2011 Fourth International Joint Conference on Computational Sciences and Optimization*], 332–335, IEEE (2011).

[3] Lin, X., Li, J.-H., Wang, S.-L., Cheng, F., Huang, X.-S., et al., "Recent advances in passive digital image security forensics: A brief review," *Engineering* **4**(1), 29–39 (2018).

[4] Farid, H., "How to detect faked photos: Techniques that analyze the consistency of elements within an image can help to determine whether it is real or manipulated," *American Scientist* **105**(2), 77–82 (2017).

[5] Farid, H., "A survey of image forgery detection," *IEEE Signal Processing Magazine* **2**(26), 16–25 (2009).

[6] Birajdar, G. K. and Mankar, V. H., "Digital image forgery detection using passive techniques: A survey," *Digital investigation* **10**(3), 226–245 (2013).

[7] Maaten, L. v. d. and Hinton, G., "Visualizing data using t-sne," *Journal of machine learning research* **9**(Nov), 2579–2605 (2008).