Document Recapture Detection: an overview

Jaime Corsetti, Samuel Kostadinov, Michele Presutto, Kaleem Ullah

University of Trento

jaime.corsetti@studenti.unitn.it, kaleemullah.ullah@studenti.unitn.it,
michele.presutto@studenti.unitn.it, samuel.kostadinov@studenti.unitn.it

1. Introduction

1.1. Motivation

The problem of image recapturing detection aims at distinguish an image acquired in a real scenario from a re-acquired one, usually from a physical or a digital medium (i.e. a photo of the image shown on a digital screen). The need for a robust solution to this problem is particularly strong in the more specific case of documents. A common anti forensic tool is the recapturing attack: recapturing an image (possibly of a document) can help in hiding traces of forgery. In this document, we present some recent development in the field of image recapturing detection (IRD) and in the more specific setting of document recapturing detection (DRD).

We have noticed that methods specifically designed for DRD are not very common, and often are designed for specific types or documents, or only work under certain assumptions. For example, [6] has been used for detecting traces of recapturing on Euro bills and in Spanish ID, while [7] uses Siamese networks [2] to process pairs of original and recaptured images of the same document type.

Since most of these works rely on a machine learning / deep learning approaches, the availability of large datasets is crucial. In this respect, we have found that datasets for IRD are common and used ([15], [5], [1] among others), while datasets for DRD are not. This is due to the legal difficulties of building a dataset using real documents, which under some conditions is often restricted by law. A promising recent work in this direction is DLC2021 [13], a dataset which adds recaptured images of different IDs to the original ones in a previous dataset, MIDV2020 [3].

2. Related Works

2.1. Common datasets

Most of the related works use the datasets ICL [15] and NTU-ROSE [5]. The ICL database was created from the Imperial College of London [15], and contains 1035 single captured images and 2520 recaptured images, of which 900 single captured images and 1440 recaptured images are freely available.
The NTU-ROSE database [5] is composed of 2710 single captured images and 2776 recaptured images.
Both these databases are acquired with multiple cameras and different LCD screens were used for recapturing.

2.2. Document Liveness Challenge (DLC-2021) [13]

2.2.1. Dataset description

The dataset follows the structure of a previous work , MIDV-2020 [3]. The documents used are of 10 different types, evenly divided between ID cards and passports. The dataset presents the original versions of the documents, gray and color copies without lamination, and video recaptures of the documents.
The original documents are captured in a variety of different lighting and perspective conditions. Moreover, the videos are also shot with some occlusions like the ones due to fingers holding the document. Regarding the recapture, two desktop and two notebook LCD monitors were used.
In order to be GDPR compliant all documents are synthetic mock-ups.

2.2.2. Experimental baseline

The authors also present some baselines based on CNNs and sklearn dummy classifier for all the purposes of the datase, which are screen recapture, color copy detection and gray copy detection. In particular, for screen recapture the baseline method uses a ResNet-50 [9], replacing the last layer with 2 outputs and training only these. The achieved accuracy is 89.67%.

2.3. A Simple and Effective Image-Statistics-Based Approach to Detecting Recaptured Images from LCD Screens [16]

2.3.1. Methodology

The authors observed that recapturing an image introduces alterations mostly on the spatial domain, so the proposed approach focuses on using features that can be extracted directly from the pixel domain. The extracted features should be easy to compute and discriminative enough to be used to detect whether an image is recaptured or not. Moreover, the features shouldn't be content-dependent.
To achieve this result, the method relies on residual images. The residual is an image in which every pixel has, as value, the original value minus the mean of the two neighbouring pixel in the horizontal axis. The same procedure is then applied with the vertical axis.

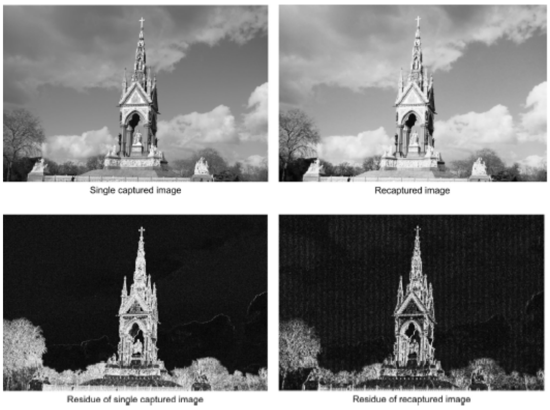

Figure 1: *An example of image residuals*

In Figure 1 an example of residuals is shown. It is possible to see that in the case of the recaptured image (on the right) the corresponding residual image shows some artifacts that in the original are not present.
After obtaining the residual images, every overlapping 5x5 patch in the image is extracted, then the results are recombined into 25 feature vectors. After that, the correlation coefficients between the feature vector of the central pixel and all the vectors are computed. The correlation coefficients (CC) are rearranged in a matrix, of which only the upper triangle is considered since the matrix is symmetric. Moreover, the central value is always 1 so it is discarded. This procedure is performed for both the residuals images, and the obtained features are concatenated.
The features extracted in this way are second order statistics, and are expected to be sensitive to recapturing, while being independent of the content of the image.

2.3.2. *Experimental setting*

The extracted features are fed to an SVM to perform classification. The experiments involved different patches dimensions and different images dimensions, in particular 2048, 3072 and 4096 pixels as width.

2.4. Image Recapture Detection with Convolutional and Recurrent NN [10]

2.4.1. *Methodology*

The intuition behind the work is that when an image is recaptured it introduces loss-of-the-detail artifact, color distortion, and anti-aliasing texture. Such features can be detected by CNNs. Secondly, the authors states that images have high dependency between the blocks, and using this fact they use a DAG-RNN [14]to learn those dependencies.

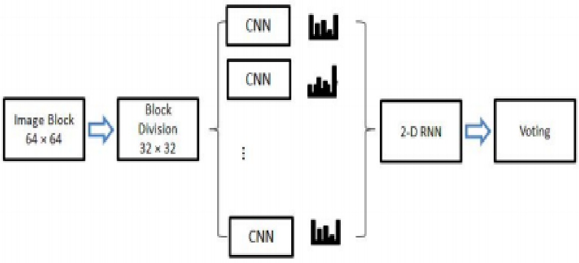

Figure 2: *The proposed feature extraction framework for recapturing detection*

2.4.2. *Experimental setting*

They have illustrated the performance of their approach with state-of-the-art handcrafted features and also with deep learning features.

The results are evaluated by accuracy and on different datasets. The methods presented as baseline include.

- Multi-scale Local Binary Pattern
- Multi-scale wavelet statistics:
- CNN model with and without Laplacian filter
- CNN model with learned filter:

Evaluation is performed on public available datasets ASTAR [8], NTU-ROSE [5], ICL cite[15] and the results obtained are 93.29% 98.67% 99.54%, respectively.

2.5. Domain agnostic document authentication against Recapturing Attacks [7]

2.5.1. *Methodology*

The method presents a unique approach to generalize Recapture Detection. They proposed the usage of Siamese Networks with CNN model and a metric learning approach that involves the use of Forensic Learning [12]. The overall structure of the model is displayed in 3.

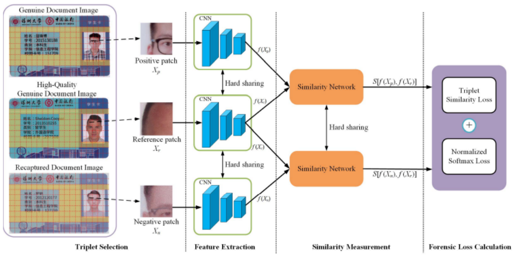

Figure 3: *The network architecture of the proposed recaptured document authentication scheme*

Three inputs are provided to the model: genuine, recaptured, and reference (high-quality genuine) documents. One patch from each input is forwarded into the network. The authors employ a specific Triplet Sampling Strategy to sample effective triplets. Once the model is trained the output of the network described as f (xpi), f (xni) and f (xri), respectively, are passed through subnet-CNN to extract the Learn-able similarity [12] between {f (xpi), f (xri)} and {f(xpi),f (xri)} in the embedding space. Once the similarity measure is obtained, it is embedded in the Triplet loss.

2.5.2. *Experimental setting*

The dataset used for experiments involves the Identity Documents of five different universities, but it is not publicy available. The resulting performance expressed for the proposed approach are AUC 1 and EER 0.01% respectively.

2.6. Near-duplicate detection for LCD screen acquired images using edge histogram descriptor

2.6.1. *Methodology*

Near-duplicated image are high-quality recaptured images, difficult for the human naked eye to recognize the difference from the original. It has been discovered in [11] that near-duplicated images contain edges of different direction compared to the original one. The re-acquired images lose information as edge counts and edge sharpness, thus to differentiate between the two classes the authors use the count of edges and the orientation.
The image is first converted to gray-scale than the edges are extracted using a Canny edge detector [4]. The next step is to split the image in 16 (4 X 4) parts and compute the Edge Histogram Descriptor [17] for each. This descriptors combines both textures and shape knowledge taking in consideration the frequency and the orientation of the brightness change in the

image. Five different directional edges are considered (horizontal,vertical, diagonal 45°, diagonal 135°, non directional edges) and each sub-image is represented as an histogram of five bin, resulting with a total of 80 (16 X 5) bins for the entire image. Further more to extract the local edge orientation, a convolution operation It's applied on each sub-image using 5 different directional filters (2x2). Than the maximum value amongst the 5 convolutions is compared with a threshold T.

The step is repeated for each 2x2 pixels of the sub-block in one $(\frac{M}{4}, \frac{N}{4})$ sub-image, so for each of this block we will get the complete 5 bins and in total for the entire image the overall count of bins will be of 80.

$$Bin_{all} = \begin{bmatrix} b_{00} & b_{01} & b_{02} & b_{03} & b_{04} \\ b_{ij} & b_{ij} & b_{ij} & b_{ij} & b_{ij} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{15,0} & b_{15,1} & b_{15,2} & b_{15,3} & b_{15,4} \end{bmatrix} \quad (1)$$

The parameters Bin_{all} will give the Local Edge Histogram LEH. It is been computed also the Global Edge Histogram along with the local and global edge count for the whole image.

$$Bin_{Global_j} = \frac{1}{16} \sum_{i=0}^{15} b_{ij}$$

$$Bin_{LC_j} = \sum_{i=0}^{15} b_{ij}$$

$$Bin_{GC} = \sum_{j=0}^{4} Bin_{LC_j}$$

The final Edge Histogram Descriptor vector is obtained by concatenating all parameters Bin_{all}, Bin_{Global}, Bin_{LC}, Bin_{GC}. The final length of the vector is 91 (80 + 5 + 5 +1). The SVM was used to carry out the classification using a non-linear Radial Basis Function Kernel. The authors discovered that introducing an anti-aliasing filter(Median filter of 3x3) before this stage will strengthen the resulting descriptor features.

2.6.2. Experimental setting

The experiments were performed using three publicly available datasets NTU-ROSE[5], Mturk [1], and ICL[15]. The evaluated results shows that this approach works well with different datasets and reach state of the art performances, for the ICL dataset the accuracy is up to 100%, For the NTU-ROSE dataset,the classification accuracy of 99.57% and for the Mturk dataset, the accuracy achieved is 99.77%.

2.7. Exposing Recaptured Images with Constrained Convolutional Neural Network [18]

2.7.1. Methodology

The main idea in [18] is to extract features from the prediction error maps of an image and using the data to learn the artifacts introduced by the recapturing action. To suppress the image content and jointly learn features for the recaptured image detection it's been used a CNN and in particular the architectures include a constrained convolution layer. The whole architecture is composed by two parts: 1) deep residual module 2) the convolution modules and decision layer. The Residual Module is been used to extract the prediction error maps, where the image pass through two convolutional layer, the first one used to learn the channel correlation of a color image and the second is a constraints convolutional layer that acts as a predictor by using a set of learnable constraints prediction error filters. in particular, $w^{(k)}$ is the k-th filter in constraints convolution layer and the spatial index of the center of a filter is (0,0), the prediction error maps can be obtained by using the following constraints

$$\left\{ \sum_{(m,n)\neq(0,0)} w^{(k)}(m,n) = 1. \right. \quad (2)$$

To calculate different residual maps a diverse set of prediction filters can be used. To update the filter weight it's used the SGD(stochastic gradient descend) during backpropagation steps. The Residual module consist of three identical branches functioning as described previously, to obtain the output of this deep residual module, a concatenation of the output of the three branches is performed. In total 24 different prediction error maps are obtained. The concatenated output than is given to the five consecutive convolution modules to learn higher-level representative features and the associations between the prediction error maps. Ultimately, a 2 neuron fully-connected layer is used and a SoftMax activation function to make the final decision.

2.7.2. Experimental setting

The training was done using three different datasets: NTU-ROSE [5]., Dartmounth [1], ICL [15]. The proposed method can achieve an average accuracy of 99% on the three different datasets, demonstrating the dominance of the proposed constrained convolution network.

Method	Accuracy	StD
Wang [10](center crop)	84.58	3.01
Wang (5 crops)	82.69	1.78
Wang (whole image)	87.19	2.64
Mehta et al [11] (whole image)	84.12	3.07
Mehta et al (whole image) w. filter	80.68	3.50
ResNet18 [9](center crop)	87.89	7.62
ResNet50 (center crop)	89.68	5.26

Table 1: *"center crop" refers to cropping a 224x224 portion at the center of the image. "whole image" uses all the document part of the image using the ground truth detection, and in "5 crops" we extract 5 224x224 random crops from the document part of the image.*

3. Results

In order to evaluate the quality of the features from these methods in the DRD task, we re-implemented [16] and [11] and evaluate them on DLC2021. We compare the two by evaluating the performance of an SVM trained on their features. As an additional baseline, we also evaluate an SVM trained on the features extracted from two pretrained ResNets [9] from the PyTorch library. The results are reported in 10.

A surprising result is the one obtained by the ResNet models: despite not being trained for the specific tasks, they achieve comparable or higher results with respect to the other evaluated methods. This suggests that, given enough data, CNN models are capable of handling the complex task of DRD.

For reproducibility and additional details on the experiments, we released the code at `https://github.com/jcorsetti/TACV-DocumentRecapturing`

4. Extra tables

Method	A	P	R
Wang	50.87	51.09	98.46
Mehta w. filter	47.89	47.71	16.58
Mehta	48.19	46.84	7.34
Resnet18	63.57	72.50	46.70
Resnet50	61.15	70.60	41.59

Table 2: *Results of training on DLC2021 and testing on our dataset.*

Method	Transfer Mean			Same document			Our dataset		
	A	P	R	A	P	R	A	P	R
Wang	82.08	83.82	82.95	87.14	92.87	82.48	50.87	51.09	98.46
Mehta et al	77.65	82.62	80.14	87.20	91.99	83.53	48.19	46.84	7.34
Mehta et al w. filter	72.14	77.54	74.84	84.78	88.28	82.74	47.89	47.71	16.58
ResNet18 (frozen)	83.37	82.68	83.02	97.94	98.17	98.01	63.57	72.50	46.70
ResNet50 (frozen)	81.58	80.40	80.99	98.25	98.08	98.69	61.15	70.60	41.59
MobileNetV2	91.15	93.46	90.69	99.92	99.95	99.90	64.28	79.65	40.79
EfficientNetV3	91.92	98.83	86.23	99.94	100.0	99.90	71.21	93.33	47.25
ResNet50	93.71	96.13	92.26	99.94	99.95	99.95	68.00	85.03	45.66

Table 3: *Results on transfer from each of the presented methods*

Method	Mean			Same doc			Our dataset		
	A	P	R	A	P	R	A	P	R
MobileNetV2	91.15	93.46	90.69	99.92	99.95	99.90	64.28	79.65	40.79
EfficientNetV3	91.92	98.83	86.23	99.94	100.0	99.90	71.21	93.33	47.25
ResNet50	93.71	96.13	92.26	99.94	99.95	99.95	68.00	85.03	45.66

Table 4: *Results to transfer of different documents* .

FDA interval	Accuracy	Precision	Recall
None	64.28	79.65	40.79
0, 0.01	57.31	70.31	29.03
0, 0.05	60.46	78.03	31.91
0, 0.1	58.20	69.99	32.41
0.01, 0.02	57.13	70.41	28.34
0.02, 0.05	57.99	64.92	39.40
0.05, 0.1	54.68	82.19	14.89

Table 5: *Results of training on DLC2021 and testing on our dataset.*

Edge type	Horizontal	Vertical	Diagonal 45	Diagonal 134	Non-directional
Original	135	18	2593	1	3823
Recaptured	26566	29179	74367	16855	145612

Table 6: *Results of training on DLC2021 and testing on our dataset.*

	Kaleem	Jaime	Michele	Samuel
Phone	iPhone 13	OPPO A16s	iPhone X	LG G8s Thinq
Resolution	3840×2160*	720x1280	1920x1080	720x1280

Driving license	1505
UniTN card	1990
PoliTO card	476
Tessera sanitaria	1316
Supermarket card	481
Passport	552
Id card	1045
Residency permit	492

Table 7

iPhone 13	2494
OPPO A16s	1218
iPhone X	1908
LG G8s Thinq	2237

Table 8

Document	Accuracy	Precision	Recall
Driving license	59.89	69.81	31.27
UniTN card	59.22	97.25	20.78
PoliTO card	83.00	83.80	86.23
Tessera sanitaria	50.23	73.08	5.62
Supermarket card	47.26	43.27	40.54
Passport	90.11	100.00	81.82
Id card	65.83	70.83	61.59
Residency permit	97.36	100.00	94.74

Table 9

Person	Accuracy	Precision	Recall
iPhone X	56.06	56.06	15.01
iPhone 13	83.21	86.15	81.28
OPPO A16s	59.70	100.00	14.04
LG G8s Thinq	57.34	70.48	27.01

Table 10

References

[1] Shruti Agarwal, Wei Fan, and Hany Farid. “A Diverse Large-Scale Dataset for Evaluating Rebroadcast Attacks”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 1997–2001.

[2] Jane Bromley et al. “Signature Verification Using A ”Siamese” Time Delay Neural Network”. In: *International Journal of Pattern Recognition and Artificial Intelligence*. Vol. 07. 04. 1993, pp. 669–688.

[3] K.B. Bulatov et al. “MIDV-2020: a comprehensive benchmark dataset for identity document analysis”. In: *Computer Optics*. Vol. 46. 2. Samara National Research University, Apr. 2022, pp. 252–270.

[4] John Canny. “A Computational Approach to Edge Detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. PAMI-8. 6. 1986, pp. 679–698.

[5] Hong Cao and Alex C Kot. “Identification of recaptured photographs on LCD screens”. In: *2010 IEEE International conference on acoustics, speech and signal processing*. 2010, pp. 1790–1793.

[6] A. Berenguel Centeno et al. “Recurrent Comparator with Attention Models to Detect Counterfeit Documents”. In: *International Conference on Document Analysis and Recognition (ICDAR)*. IEEE Computer Society, Sept. 2019, pp. 1332–1337.

[7] Changsheng Chen et al. “Domain-Agnostic Document Authentication Against Practical Recapturing Attacks”. In: *IEEE Transactions on Information Forensics and Security*. Vol. 17. 2022, pp. 2890–2905.

[8] Xinting Gao et al. “A Smart Phone Image Database for Single Image Recapture Detection”. In: *Digital Watermarking*. 2011, pp. 90–104.

[9] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[10] Haoliang Li, Shiqi Wang, and Alex C. Kot. “Image Recapture Detection with Convolutional and Recurrent Neural Networks”. In: *Electronic Imaging*. Vol. 2017. Jan. 2017, pp. 87–91.

[11] Preeti Mehta and Rajiv Kumar Tripathi. “Near-duplicate detection for LCD screen acquired images using edge histogram descriptor”. In: *Multimedia Tools and Applications*. Vol. 81. Sept. 2022, pp. 1–19.

[12] Matthew C. Stamm Owen Mayer. “Forensic Similarity for Digital Images”. In: *IEEE Transactions on Information Forensics and Security (2019)*. Vol. 15. 2019, pp. 1331–1346.

[13] Dmitry V Polevoy et al. “Document Liveness Challenge Dataset (DLC-2021)”. In: *Journal of Imaging*. Vol. 8. 7. 2022, p. 181.

[14] Bing Shuai et al. “DAG-Recurrent Neural Networks for Scene Labeling”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 3620–3629.

[15] Thirapiroon Thongkamwitoon, Hani Muammar, and Pier-Luigi Dragotti. “An image recapture detection algorithm based on learning dictionaries of edge profiles”. In: *IEEE Transactions on Information Forensics and Security*. Vol. 10. 5. IEEE, 2015, pp. 953–968.

[16] Kai Wang. “A simple and effective image-statistics-based approach to detecting recaptured images from LCD screens”. In: *Digital Investigation*. Vol. 23. Elsevier, 2017, pp. 75–87.

[17] Chee Sun Won. “Feature extraction and evaluation using edge histogram descriptor in MPEG-7”. In: *Pacific-Rim Conference on Multimedia*. 2004, pp. 583–590.

[18] Nan Zhu, Hanchen Xiang, and Zhiqin Liu. “Exposing Recaptured Images with Constrained Convolutional Neural Network”. In: *2022 7th International Conference on Signal and Image Processing (ICSIP)*. 2022, pp. 463–467.