DOI: 10.2352/ISSN.2470-1173.2016.8.MWSF-068

Detecting Copy–Move Forgeries in Scanned Text Documents

Svetlana Abramova; Institute of Computer Science, Universität Innsbruck; Innsbruck, Austria
Rainer Böhme; Institute of Computer Science, Universität Innsbruck; Innsbruck, Austria

Abstract

The detection of copy–move forgeries has been studied extensively, however all known methods were designed and evaluated for digital images depicting natural scenes. In this paper, we address the problem of detecting and localizing copy–move forgeries in images of scanned text documents. The purpose of our analysis is to study how block-based detection of near-duplicates performs in this application scenario considering that even authentic scanned text contains multiple, similar-looking glyphs (letters, numbers, and punctuation marks). A series of experiments on scanned documents is carried out to examine the operation of some feature representations proposed in the literature with respect to the correct detection of copied image segments and the minimization of false positives. Our findings indicate that, subject to specific threshold and parameter values, the block-based methods show modest performance in detecting copy–move forgery from scanned documents. We explore strategies to further adapt block-based copy–move forgery detection approaches to this relevant application scenario.

Introduction

In today's media age, given the high popularity of low-cost digital imaging devices and the availability of feature-rich photo editing software, the credibility of digital image content can no longer be taken for granted. Persistent concerns about image integrity and authenticity have given rise to a rapidly growing field of digital image forensics, which relies on the statistical and structural analysis of image data and files [10]. A broad spectrum of developed techniques is mainly devoted to the three interrelated application areas: image source identification (i. e., determining a make and model of the acquisition device that captured a specific image), discrimination of computer-generated images from real-world ones, and image forgery detection (i. e., determining whether a suspicious image has undergone malicious post-processing) [22]. This paper belongs to the last direction of research, and particularly deals with *copy–move forgery detection* (CMFD) – one of the most studied topics in the image forensics literature [7].

A copy–move (CM) forgery refers to copying a portion of an image and re-inserting it (or its filtered version) elsewhere in the same image, with the intent of hiding undesirable contents or duplicating particular objects of interest. To discern this form of local image processing, a number of detectors (more specifically, feature sets) have been suggested in the literature and experimentally evaluated under different settings [3, 7, 20, 24]. Their detection performance was predominantly assessed through multiple test runs over digital or scanned images picturing *natural scenes*. However, to the best of our knowledge, none of the existing methods has yet been proven in scenarios involving *digitalized document images*. Since the strategic concept of "going paperless" and the transition to electronic document management systems is increasingly advocated among businesses as a desirable practice [11], counterfeiting scanned versions of documents of potentially high financial value (e. g., receipts, contracts, or official certificates) through a copy–move image manipulation is likely in practice. In pursuit of private benefits, even an amateur forger can undertake CM tampering operations on everyday documents lacking embedded security features and effortlessly alter their semantic content by fabricating new names, dates, or values; or by overlaying a text-free background area to conceal undesirable information. Figure 1 illustrates a realistic-looking forgery of an invoice, in which some glyphs[1] are copied to change a billing amount and a bank account number. This simple, but at the same time real-life example emphasizes the need for the design of a proper toolbox to automatically examine document images for the presence of CM modifications.

First, we must explore if available CMFD methods are applicable for the forensic inspection of scanned text images. Due to the intrinsic features of documents, certain issues are present in this application area. First, any authentic text consists of multiple, similar-looking glyphs, and even non-identical letters may have common shape structures (e. g., both symbols 'p' and 'o' contain a semicircle). Since the underlying principle behind most CMFD algorithms is to search for conspicuously similar image segments of an unknown shape located at some distance from each other, both *inter-glyph* and *intra-glyph similarity* may cause many false positives and deteriorate performance evaluation measures. In addition to duplicating certain glyphs, a forger can hide unwanted text parts by overlaying them with a white background. Most feature sets proposed for CMFD have been found to perform worse when detecting cloned regions with low entropy [7]. Printer and scanner forensics techniques examining irregularities in the background surface seem to be more suitable for this purpose. Through the execution of initial tests, we aim to reveal the pitfalls of applying the known CMFD methods to document inspection and, based on experimental findings, to formulate eligible courses of action and guidelines for the design of a CM counterfeit detector tailored to this specific domain.

With these objectives, we applied a publicly available software framework developed in [7] to carry out a series of experiments over manipulated images of black text of different font sizes on a white background. Although some block-based features have demonstrated an acceptable accuracy rate at the pixel level, the results confirmed a concern that the naive application of the current CMFD detectors to the analysis of digital documents is prone to a large number of false positive matches. Subject to the chosen block size, CMFD methods also fail to detect dupli-

[1] We use the generic term *"glyph"* in this paper to refer to any graphical text entity (e. g., letters, numbers, symbols, or other textual objects).

DOI: 10.2352/ISSN.2470-1173.2016.8.MWSF-068

cated patches of thin glyphs (e. g.,'i' or 'l'), or erroneously mark white background areas as forged. Motivated by these findings and recent progresses in related fields, we elaborate a conceptual framework for CMFD in digitalized documents. This framework specifies the need for a separate examination of text and background areas and integrates multiple aspects and forensic techniques to advance performance in terms of the detection accuracy and computational efficiency.

The remainder of this paper is organized as follows. The next section reviews related work on the current approaches for CMFD and document forensics. We then present an exploratory study of the detection performance using a limited set of the existing block-based features, the results of which serve as a base for the elaboration of the conceptual framework for CMFD in scanned text documents. Next, we make the first step towards the adaption of a general CMFD procedure in accordance with the framework's concepts, and report the detection results of our initial experiments. Finally, we conclude with a summary of the key findings and an outlook on future research.

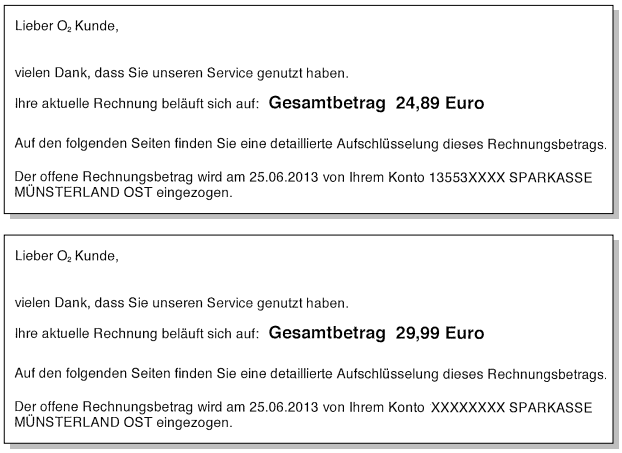

Figure 1. *Original (above) and forged (below) image snippets of an invoice*

Related Work

The image forensics literature offers a considerable number of reliable CMFD approaches, along with recent benchmarking analyses of their detection and localization performance. Most of the proposed methods can be classified as either block-based (e. g., [9, 19, 21, 26]) or keypoint-based (e. g., [2, 13, 18]), depending on whether the projection from a pixel representation of the input image into a low-dimensional feature space is performed for blocks of pixels or areas surrounding keypoints (i. e., regions with high entropy). While block-based CMFD methods partition a questionable image into (non)-overlapping blocks of pixels and extract a feature vector from each block, keypoint-based detectors compute feature descriptors only from the local area around a keypoint. The matching procedure of both outputs tuples of similar feature vectors, which undergo additional post-processing and error reduction techniques to prune outliers and finally to localize cloned image patches.

Since most, if not all, CMFD methods proceed through this common sequence of steps (with some variations), the detection performance of each approach largely depends on the selected feature representation and its robustness to different kinds of post-processing operations, such as geometrical transformations, noise, compression, or a combination of these. In the context of this study, it seems reasonable to assume that a forger may add small amounts of noise or apply JPEG compression to the falsified document image, while scaling glyphs up or down is less likely due to the use of a common font size for the main text and visible distortions. According to the results of the comprehensive CM detection evaluation of state-of-the-art feature sets in [7], the block-based methods DCT [9], PCA [19], HU [26], KPCA [4], and ZERNIKE [21] perform optimal not only in plain copy–move, but also in some post-processing scenarios relevant to our case. Since a detailed discussion of the features' properties is beyond the scope of this paper, we refer the interested reader to the original works for further information.

Other research fields closely related to the topic presented here are forensic document analysis and optical character recognition (OCR). Most prior work on passive document forensics has focused on source identification of text images, e. g., by means of texture analysis based on intrinsic signatures [17], statistical features of sensor pattern noise [16], geometric degradations [15], banding artifacts, or character morphologies caused by a printer [23]. Some researchers have extended techniques originally developed for printer or scanner identification to the detection of certain types of image tampering. For example, Khanna *et al.* [16] made use of sensor pattern noise as a scanner fingerprint to expose image contents generated from several peripheral sources. Kee and Farid in [15] established a model for printer profiling, based on distortions introduced by a printer, to detect characters initially printed on different devices (in terms of a make and model). Similarly, Shang *et al.* [23] extracted features such as noise energy, contour roughness, and average gradient from individual text characters in order to detect documents composed of parts originating from different printer types.

In general, all of these approaches exploit significant dissimilarities of selected features between various source types. However, they fail to detect traces of copy–paste and reprinting counterfeiting operations, i. e., when re-inserted document regions are generated by the same acquisition device. A meticulous analysis of the alignment of text lines and distortions in character locations is often mentioned in the literature as a potential solution for this kind of falsification [5]. Despite great efforts to paste symbols aligned to the ascender, descender, and base text-lines as accurate as possible, a fraudulent person is still unlikely to achieve an exact alignment due to technical and manual constraints. Therefore, the examination of distances between glyphs and of text-lines to respective alignment lines may provide some evidence of a digital document having been exposed to copy–paste manipulations.

Driven by strong practical demands, OCR (i. e., the conversion of a scanned text image to encoded text) has always been a popular research topic under the scope of image processing and pattern recognition. Consequently, there are a variety of techniques available for the automatic recognition of printed or handwritten text. In a simplified form, the classical process of OCR consists of a series of stages executed in a pipeline manner: format and structure analysis, character segmentation, feature extraction, and, lastly, classification [6]. Apart from required preprocessing operations, the initial step of most OCR systems is *"dissection"* – a decomposition of image content into smaller meaningful segments (e. g., text paragraphs, tables, lines, or words).

DOI: 10.2352/ISSN.2470-1173.2016.8.MWSF-068

Next, character segmentation methods are applied to split the dissected components further into individual glyphs (or words). In case of machine printing, this is commonly based on determining the location and dimensions of connected black regions and corresponding "bounding boxes" (i. e., rectangular areas around each connected component). With approximate positions and sizes of glyphs at hand, OCR schemes typically proceed with the extraction of features, followed by matching against training data sets and, finally, recognition of characters. With respect to the paper's problem, OCR offers certain potential ways to advance available CMFD methods, which will be discussed in detail later.

Exploratory Study

This section presents the procedure and the details of experiments carried out with a selected set of methods, in order to examine their performance for the practical case of authenticating scanned text documents. In this study, we have focused on block-based feature sets only which, in our view, seem to be more suitable for document forensic inspection. We neglect keypoint-based methods for two reasons. First, since text characters within a word are intrinsically located close to each other, neighbor glyphs adjacent to a specific keypoint may corrupt extracted feature descriptors and, consequently, cause many detection errors. Second, as mentioned before, keypoint-based features are less efficient when smooth areas (e. g., text background) are cloned.

Test Data Generation

To the authors' best knowledge there is no public benchmark dataset of falsified text images; in particular, of digital documents tampered through CM manipulations. Therefore, synthetic test data had to be generated first. Figure 2 presents the typical image processing chain of CM document manipulation, with optional steps represented in dashed blocks. In an attempt to defraud or deceive, the forger will most likely need to modify names, dates or values of genuine certificates and official forms. For this illicit purpose, a printed hard copy of a document is first scanned and saved in an uncompressed or compressed image format. Following this, copy–move tampering operations are undertaken on a digitalized version to change the document's semantic content. More specifically, characters, numbers or background areas from other parts of the scanned document may be copied and pasted at a required location. The forger may also apply JPEG post-compression to the forged document image as a final step. In an experimental setup, a ground truth map is additionally generated for the quantitative evaluation and comparison of the detection performance of various CMFD methods.

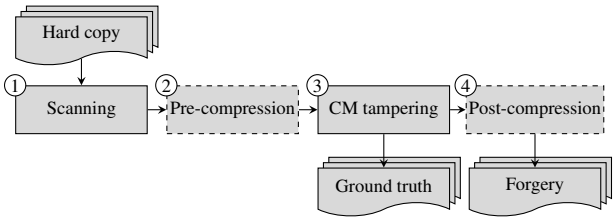

Figure 2. *Image processing chain*

Keeping this image processing chain in mind, we generated document forgeries as follows. First, we automatically produced a total of 500 random glyphs, consisting of numbers, low- and upper-case letters, and treated this set as a synthetic input document and a base for creating CM text forgeries. To study the effect of different font sizes, we typeset these glyphs (using a common word processor, in the standard Times New Roman font) with single line spacing in each of the three font sizes: 12 pt, 16 pt and 20 pt . We then printed one copy of each font size on A4 paper and scanned it at the commonly used resolution of 300 dpi. All documents were printed on a paper of the same quality to guarantee consistent settings across different test runs. Since the objects of our study are images of black text on white background (to test the simplest scenario case), scanned text images were saved in the uncompressed grayscale TIFF format. For both printing and scanning processes, we used the same device model Canon iR ADVANCE C5030i. This way, we obtained three genuine document images.

Next, we identified individual glyphs in each image by running a connected component analysis on a binarized version of a scanned document. That allowed us to identify an approximate size and a location of glyphs, which generally correspond either to single symbols or ligatures consisting of more than one character or number. We randomly chose 50 connected components out of the total sample and copied all pixels within their bounding boxes. Thus, approximately 10 % of the image's glyphs have been tampered with. As current CMFD detectors are foremost designed for the authentication of images of natural scenes, they do not account for an exact text alignment. For this reason, we decided to ignore this issue in the CM text forgery creation process. We added one or two new lines to scanned text images out of the duplicated glyphs aligned by the upper bound of their rectangular boxes. Although CM text forgeries generated this way do not look realistic and tampering traces are easily detectable by the naked eye, they still meet the primary purpose of our exploratory study by allowing us to examine the applicability of CMFD methods for the forensic analysis of digital documents.

To quantify the detection accuracy of different CMFD tools, a ground truth binary map of intact and copied pixels is needed for each test image. Following the described procedure of document forgery generation, the obvious way to create a binary mask would be to define all pixels within glyphs' bounding boxes as copied. However, the observation of detection maps produced by the block-based detectors have forced us to reconsider this approach, as the algorithms have generally marked only black pixels of glyphs as forged and ignored neighboring white pixels. Thus, labeling rectangular areas of pixels in a ground truth image as copied would significantly worsen the false negative rate. As the very act of counterfeiting a document image is still evident from the representation produced by the detectors, we decided to use an inverted binarized version of the forged document as a base for the specification of ground truth. Figure 3 demonstrates an example of a document forgery and its corresponding ground truth binary mask. Although ground truth maps generated from binarized images are not absolutely accurate, they are still more appropriate than those generated using the initial approach.

Evaluation Metrics

In literature, the performance and reliability of CMFD methods are typically assessed at image and pixel levels. The former approach focuses on the detector's ability to discern the very fact of image tampering, whereas the latter presents the detector's ca-

DOI: 10.2352/ISSN.2470-1173.2016.8.MWSF-068

pacity to accurately localize cloned image patches. Intrinsic to all document images, inter- and intra-glyph similarity calls for the evaluation at pixel level, as a clearer picture of detection results in the presence of similar image contents can be obtained for further observations and judgments. The performance on patch localization has been evaluated in our tests based on the following metrics:

- **True Positive Rate (*TPR*)**: represents the percentage of pixels correctly classified as duplicated in the detection map with respect to the real number of duplicated pixels in the ground truth map.
- **False Positive Rate (*FPR*)**: indicates the percentage of pixels erroneously classified as duplicated in the detection map with respect to the number of unmodified pixels in the reference map.
- **Accuracy (*Acc*)**: provides the quality of patch localization based on the true positive and the true negative rates. This is calculated as:

$$Acc = \frac{TPR + (1 - FPR)}{2}. \quad (1)$$

For the case of scanned document images, a high value of *FPR* can be expected for the majority of the tested detectors, whereas the true positive rate is likely to vary depending on the discriminative properties of each specific feature.

Experimental Setup

Since the principal purpose of our exploratory study is not to conduct a detailed comparative analysis of the existing methods, but rather to get an initial idea about the detection performance of block-based CMFD methods in the unexplored application domain, we have chosen a limited set of approaches for the evaluation: ZERNIKE and HU moments, PCA and KPCA. When making this decision, we have attempted to include representatives of those feature sets that, on one hand, have demonstrated reliable performance in simple, one-to-one CM forgeries; on the other hand, showed good robustness toward noise and JPEG artifacts. It is worth noting that, aside from HU moments, the selected methods have yielded decent results in the case of rotation, as well [7].

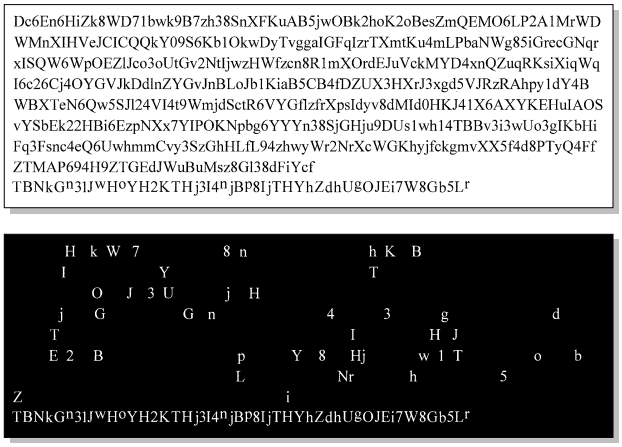

Figure 3. *Example of a generated document forgery (above) and its ground truth map (below)*

To conduct experiments in a controlled test setup, we have adopted a publicly available CMFD software framework developed for the benchmarking analysis of the known CMFD methods in [7]. Whenever possible, the relevant threshold parameters have been fixed across different test runs to allow for an unbiased comparison of the detection performance. In the case of the minimum Euclidean distance between two matched blocks, we have used the default value of 50 pixels. This limitation was deemed appropriate, since cloned glyphs were moved in the forged documents to a large distance (specifically, to the bottom of the image as described in the setup). Even if chosen from the last text line of the original document image, the distance between the original and copied patches is still greater than the defined threshold due to the space between text lines and the glyph's size.

In light of the completely different type of images being considered, we have manually adjusted several threshold values compared to those used by Christlein *et al.* [7]. One of the most crucial parameters is the block size, as it directly accounts for the detection accuracy and the number of false positive matches. If the chosen block size is too large relative to the size of a copied glyph, a CMFD method may overlook the fact of a forgery due to neighboring glyphs, being different in the original and copied patches. On the other hand, a small block size may introduce many false positives due to inter- and intra-glyph similarity. The evaluation metrics are likely to be best when the block size approximately corresponds to the glyph size, as the problem of inter-glyph similarity does not play a significant role in this situation and extracted features represent a distinct glyph. To test this hypothesis and explore the relation between block and font sizes, we have conducted test runs with varying block sizes of 10, 15, 20, and 25 pixels. We would like to note that DCT features have not been considered in our study due to the block size of 16 pixels pre-defined in the software framework. In addition, our tests have shown that although DCT features are able to detect copied glyphs, they erroneously mark large white areas as falsified, thereby yielding a high false positive rate.

In presence of inter-glyph and intra-glyph similarity, a post-processing method intended to deal with false positives should be chosen with caution. One of the most common approaches for filtering out spurious detections is to impose a minimum number of pairs of matched features sharing a similar shift vector. In addition, an area constraint can be defined to discard spuriously detected regions of small areas. Christlein *et al.* [7] have used the *Same Affine Transformation Selection (SATS)* for block-based approaches to group locations of feature vectors into larger image regions. This method yielded the most reliable results in their early experiments, and we have adopted it in our test setup as well (adjusting the relevant thresholds, i. e., the minimum number of feature pairs fulfilling the same affine transformation and the area threshold. This correction was needed due to considerably smaller image patches being copied in our case). Following the idea in [7] that these thresholds should be defined for each feature set individually, we have empirically tested the detection accuracy by running trial experiments on one image forgery with the threshold values of 50, 100, 200 and 250. By maximizing the evaluation metric *Acc*, we have obtained the best results with the threshold parameters equal to 50 and 100 for HU moments and all other features, respectively. We have used these values in the rest of our experiments.

DOI: 10.2352/ISSN.2470-1173.2016.8.MWSF-068

Performance results for plain CM at the pixel level (in %)

Font size	Size of 'e'	Block size	HU			KPCA			PCA			ZERNIKE		
			TPR	*FPR*	*Acc*	*TPR*	*FPR*	*Acc*	*TPR*	*FPR*	*Acc*	*TPR*	*FPR*	*Acc*
12 pt	20×24	10×10	58.5	2.4	78.0	67.6	3.5	82.0	63.1	2.5	80.3	53.8	2.5	75.7
		15×15	72.3	35.8	68.3	74.4	5.7	84.4	73.4	4.6	84.4	78.5	7.9	85.3
		20×20	48.3	24.9	61.8	48.0	4.9	71.5	47.6	28.3	59.6	75.0	40.5	67.2
		25×25	25.3	18.7	53.3	24.7	17.5	53.6	24.4	19.0	52.7	59.5	38.8	60.4
16 pt	26×32	10×10	53.3	1.4	75.9	63.1	2.2	80.5	59.6	1.6	79.0	42.8	2.0	70.4
		15×15	79.1	46.3	66.4	82.5	8.6	87.0	78.9	16.9	81.0	77.3	5.6	85.9
		20×20	71.5	30.9	70.3	71.4	36.1	67.7	69.8	6.2	81.8	85.5	17.9	83.8
		25×25	52.9	8.7	72.1	52.2	5.2	73.5	51.3	19.0	66.2	79.3	19.0	80.1
20 pt	32×40	10×10	35.5	0.9	67.3	46.3	58.7	43.8	43.2	1.2	71.0	28.8	0.7	64.1
		15×15	78.3	4.2	87.1	83.0	15.5	83.8	81.6	50.0	65.8	67.9	3.6	82.1
		20×20	84.1	25.0	79.6	84.2	32.6	75.8	83.4	38.5	72.5	89.3	40.7	74.3
		25×25	75.9	38.1	68.8	74.9	36.4	69.2	74.0	6.3	83.9	89.0	46.7	71.2

In order to speed up the analysis, we have ignored large image areas with no content and specified only a certain region of interest to be processed. Depending on the font size used, the regions have the dimensions of 1 900×600, 1 900×1 100 or 1 900×1 600 pixels.

Experimental Results

Our experiments have produced detection results that vary over a wide range (44 % to 87 %, see Table 1) under varying test settings. The calculated metrics indicate that the selected block-based CMFD methods have been able to detect traces of CM manipulations in scanned text images (to a certain extent and subject to specific threshold and parameter values). Figure 4 presents an example detection map that was produced from analyzing the document forgery of 12 pt font size with a block size of 15×15 pixels and ZERNIKE moments as a feature representation (shown in Figure 3).

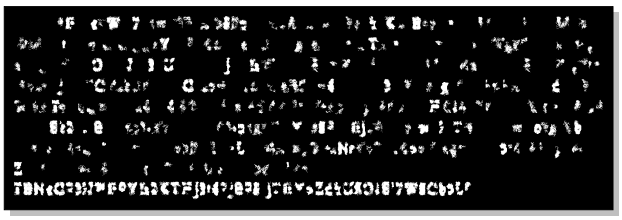

Figure 4. *Detection map (*ZERNIKE*, 12 pt font size, 15× 15 block size)*

The test experiments have also provided some evidence for a relation between block and font sizes with respect to the *TPR*. To give the reader an idea about an average width and height of symbols in the forged images of various font sizes, we have specified, as a reference, the dimensions of the rectangular bounding box of the character 'e' – the most frequent letter in English text. In most cases the highest *TPR* was obtained with the block size of 15×15 pixels for font sizes of 12 pt and 16 pt, and with 20×20 for 20 pt. With respect to large values of the *FPR* demonstrated in some test runs, a visual observation of the generated detection maps has revealed the primary cause to be the erroneous determination of blank spaces between text lines as copied.

However, the presented results cannot be generalized to all possible experimental settings, and therefore should be taken with some caution. The final decision of whether a pixel is either authentic or forged is influenced by a combination of several parameters. In order to gain a complete picture of the detection performance of CMFD methods in the case of scanned document images, it is necessary to carry out a large number of experimental runs that cover all combinations of parameter values. At its core, this is a multidimensional optimization problem. The applied decision logic is however "hidden" in the adopted software framework, making the construction of a complete ROC-curve impractical in our setup.

Nevertheless, our initial experiments support our conjecture that a block-based CMFD approach, in its current design and in the considered application scenario, is prone to many false positive matches due to inter- and intra-glyph similarity and the presence of smooth background areas. To obtain reliable results, one should first inspect document content and carefully choose threshold parameters. In order to alleviate this need for human interaction, in the next section we propose some modifications of a prototypical CMFD approach based on OCR technology.

Analysis Framework for CMFD in Text Images

Unlike well-crafted forgeries depicting natural scenes, by their nature falsified scanned documents convey some a priori information about potential locations and sizes of copied glyphs. Backed up by a OCR system, this intrinsic property opens up a promising approach for the development of a CMFD toolbox tailored exclusively for the analysis of text images. When all crucial aspects are carefully considered at the design stage, such a toolbox may significantly improve the detection performance of the known feature representations in the studied application domain. Based on the detection results and shortcomings demonstrated by the block-based methods, we have made a first attempt at designing a holistic framework for the automatic forensic analysis of document images for the presence of CM manipulations (henceforth referred to as the conceptual analysis framework).

DOI: 10.2352/ISSN.2470-1173.2016.8.MWSF-068

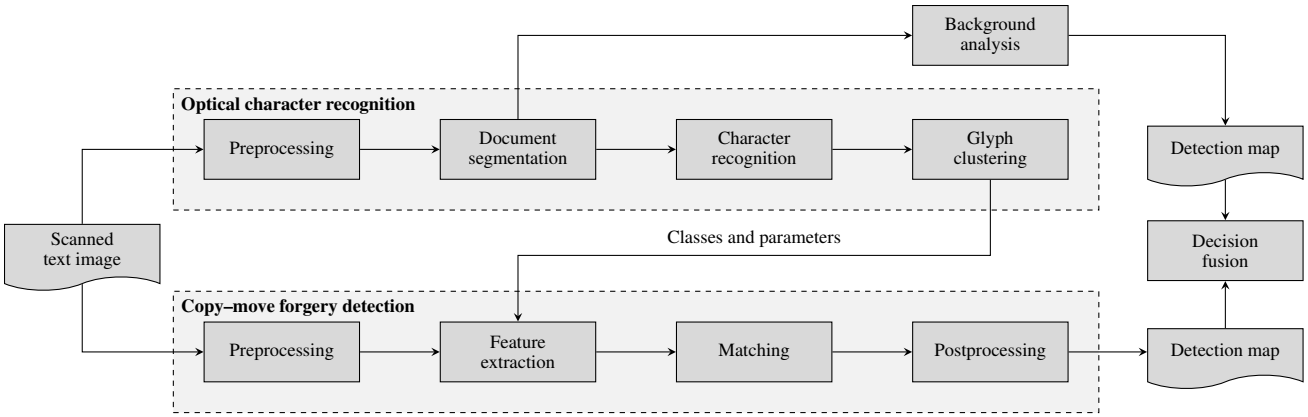

Figure 5. *Analysis framework for copy–move forgery detection in scanned text images*

To build a fine-tuned approach for CMFD in digital document images, some combination of the mentioned techniques must be integrated together into a coherent processing chain. Figure 5 illustrates our analysis framework for CMFD in scanned text images. It is presented in a pipeline manner, meaning that each stage passes its results on to the next one and depends on the success of the preceding steps. The proposed framework combines the leading streams of research in digital image forensics and pattern recognition: image source identification, forgery detection, and OCR technology. OCR serves as a preceding step to the common workflow of CMFD, as character recognition results are employed as input parameters in the typical processing steps of original CMFD methods. This is intended to solve the problem of intra-glyph similarity as well as to speed up the detection process by constraining the search space. We now discuss particular steps of the framework in greater detail.

Preprocessing. A scanned text image has to first be preprocessed and transformed into a form that can be more effectively analyzed by the corresponding OCR and CMFD methods. In general, OCR schemes apply a broad set of different preprocessing techniques (e. g., skew detection/correction, image enhancement techniques, noise removal, edge detection, binarization, morphological operations etc.) to improve the image quality of a digitally converted document and thus to increase OCR recognition performance [1]. Some of these operations may cause a loss or a change of image pixel information. Due to other required preprocessing operations, a CMFD method needs to operate on a separate image that is pre-processed in line with a selected feature representation (e. g., converted to a grayscale image or represented as the integral image). This must still preserve the original image data in order to guarantee the reliable detection of copied image snippets. The fact that OCR schemes usually run on binarized versions of images further justifies the separation of the preprocessing steps between OCR and CMFD.

Optical Character Recognition. This overarching process is composed of a number of distinct steps, which together resemble a typical workflow of the majority of known OCR systems. The main rationale for the incorporation of OCR techniques in the analysis framework is threefold. First, OCR schemes are purposefully designed to automatically analyze the structure of a digital version of the document and to segment it into regions free of text (i. e., background) and connected components (e. g., at the level of text lines, words, or individual glyphs). As we have seen in our initial experiments the existing block-based methods, which were originally proposed to deal with other kinds of images (and therefore do not differentiate between text and background areas), have incorrectly identified blank space between text lines as forged in certain instances. We believe that these two kinds of image areas should be considered in the detection analysis separately, in order to decrease both false detections and computational overload. Secondly, CMFD methods may significantly benefit from character recognition techniques in a way that font sizes, in which symbols are typeset in the text image, can be estimated and used for the determination of optimal block sizes. Lastly, OCR is certainly one of the most promising ways to handle the critical issues of intra-glyph similarity and false positives, as it recognizes specific text characters and permits adjustment in the feature matching process of CMFD methods, such that feature vectors extracted only from similar characters are matched against each other.

With preprocessing as the first step, OCR proceeds with *document structure segmentation* which is intended first to delineate text lines and background areas in the preprocessed document image and then to isolate glyphs from each other [25]. Thus, sizes and positions of glyphs, as well as of identified text-free areas, are obtained at this stage and passed to the next processing steps. Under the generic term *"character recognition"*, we have encompassed several prototypical subprocesses of core OCR algorithms: feature extraction, matching, classification, and error reduction [14, 25]. As follows from its name, this stage is primarily responsible for the conversion of scanned document image into encoded text. In this regard, research and practice in the field of OCR offers a range of established approaches and advanced methods; the detailed discussion of which is not a concern here. Character recognition is followed by a *glyph clustering* routine that discriminates various font sizes used in the scanned document and assigns recognized characters to symbolic classes of same font size and case. To accomplish this task in document analysis, OCR algorithms usually extract features from global properties of the text image (e. g., character size, text density etc.) [27]. Following this sequence of steps, the OCR method thus delivers clusters of recognized characters of the same case and font size, along with their locations (i. e., pixel coordinates) in the image and dimensions.

DOI: 10.2352/ISSN.2470-1173.2016.8.MWSF-068

Copy-Move Forgery Detection. In essence, the CMFD workflow presented in our analysis framework is similar to those originally proposed for the forensic inspection of digital images of natural scenes. However, we have reconsidered certain aspects in the original CMFD pipeline in order to tailor it for the purpose of forensic document analysis. In the context of our framework, block-based CMFD methods rely on output results of the preceding OCR process. They operate on the input image of scanned text, preprocessed in line with a chosen feature representation, and continue with *image localization* (i. e., block tiling) and *feature extraction*. Document font sizes, identified in the OCR stage, play a leading role in specifying respective block sizes. However, the best range of block sizes for a specified font size remains an open question. We assume that this critical decision is subject to a trade-off between multiple factors, and therefore encourage further research in this direction. Thus, the framework analysis envisions the determination of the best block sizes for image fragmentation and feature extraction, which are obtained with respect to base font sizes used for typesetting document content.

After the bounding boxes around recognized glyphs are fragmented into blocks of pixels, and feature statistics are computed for each of these blocks, *matching of feature vectors* takes place. From a pragmatic perspective, this process usually requires a substantial amount of time and computational resources, depending on the image size, number of features, and the chosen matching method. Motivated by the need to speed up matching and resolve the issue of false positives caused by intra-glyph similarity, we propose to compare and find similar feature vectors within classes defined by same character (of same font size and case). Due to visual distortions, applying scaling operations to copied glyphs is rather unlikely in real life, justifying the suggested matching routine. One can argue that matching feature descriptors across the same recognized characters may lead to missed forgery detections because of potential errors in OCR. If either a genuine or a duplicated character is wrongly recognized due to its visual resemblance to other characters and is consequently clustered into another class than its counterpart, the proposed matching procedure will fail to detect these forged glyphs. An alternative method is to match extracted feature statistics across all glyphs of same font and size (i. e., without paying attention to syntactical content), and check in the next post-processing stage whether matched feature vectors correspond to similar-looking characters. Even though we do not consider this approach as completely inappropriate, we adhere to the less time- and resource-consuming one, as modern OCR systems have advanced to a level of high recognition accuracy.

The key goal of *post-processing* step is to handle outliers and prune false positive matches. In this respect, the translation vector analysis traditionally used in many of the existing CMFD methods can be applied to pairs of matched feature vectors, achieving greater detection accuracy. In the case of document images being analysed, post-processing can also be augmented with syntactical models and statistical techniques from natural language processing, allowing the examination of syntactic and semantic grammars and linguistic structures.

Background Analysis. With regard to the forgery detection in text-free image areas, the analysis framework assumes the use of *scanner and paper forensics techniques* instead of the conventional CMFD approaches. As evident from the empirical findings, feature sets are generally sensitive to white background areas; thereby making CMFD techniques unsuitable for the application in this context. Recent research in scanner and paper forensics, on the other hand, offers many potential ways to analyze the background of printed and scanned documents for the presence of inconsistencies (e. g., in pattern noise, color degradation, etc). Additionally, if a questionable image has been saved in a compressed format *JPEG compatibility analysis* [8, 12] may reveal the fact that the background area was copied and used as an overlay to conceal some undesirable image contents.

Decision Fusion. Both the CM forgery detection process and the background analysis produce a binary detection map as an output, depicting identified image duplicates. As the framework provides for the separate analysis of content and text-free image areas, these binary masks have to be combined in the *decision fusion* step in order to construct a complete detection map. This map is employed in the following comparison against a corresponding ground truth map, and the ultimate evaluation of detection accuracy. However, it should be noted that the generation of detection maps and their fusion are associated with partial information loss in the sense that detection maps do not allow us to control for the correct relation between a copied glyph and its original counterpart. Detection results may be misleading when one glyph has been correctly identified as cloned, while its counterpart, being marked in the detection map as forged, is, in fact, not the one originally used by the forger but just a similar glyph.

To sum up, the presented analysis framework for CMFD in text images can serve as a design starting point for tackling the problem of detecting CM counterfeiting operations in digitalized documents. In addition to integrating OCR techniques into the common CMFD pipeline, it promotes several other ideas with respect to different processing steps in order to optimize the detection results. The framework envisions a separate examination of content and background image areas, block size identification in relation to font sizes and feature matching within classes of same recognized characters, font size and case. It should be mentioned that the analysis framework presented here should not be considered comprehensive, as there are other relevant methods available that may enhance the detection performance in this application scenario.

Results of a First Tailored Approach

To get an initial idea about potential advantages of applying OCR principles to the problem of CMFD in scanned text images, we have conducted a series of preliminary tests with a document forgery of 12 pt font size. Instead of starting by partitioning the input image into overlapping blocks of pixels, we have first performed the connected component analysis in order to identify positions and dimensions of the rectangular boxes that bound glyphs or ligatures. Next, we have calculated logarithms of HU moments of the pixels inside each bounding box as a glyph's feature representation. In spite of the fact that the recognition of individual characters has not been implemented in our practical setup, we have obtained decent detection results in the basic CM and compression scenarios. As a threshold for determining whether an individual glyph has been duplicated or not, we have measured the Euclidean distance between the computed feature vectors of

DOI: 10.2352/ISSN.2470-1173.2016.8.MWSF-068

each identified glyph and of its nearest neighbor. In the simplest case of a plain CM forgery, the Euclidean distance between feature vectors of two copied glyphs was always zero. This resulted in a correct detection of all forged glyphs without any false positive matches.

The second batch of experiments was aimed at testing a more realistic CM document forgery. For that, we have applied JPEG compression with varying quality factors to the whole falsified document. This operation corresponds to the last (post-compression) step of the image processing chain presented in Figure 2.[2] The quality factors of the post-compression varied between 100 and 80 in steps of 5. As JPEG compression introduces a common global disturbance [7], we have adjusted the threshold for the Euclidean distance to 0.1. This value was determined through a set of trials conducted with varying thresholds, in the search of a good trade-off between the number of true and false positives. Computing feature vectors per bounding box has enabled us to analyze the detection performance at the level of individual glyphs instead of the original pixel-based approach. Table 2 presents the detection results for JPEG post-compression with different quality factors. We have obtained the highest detection accuracy of 42 glyphs out of 50 originally copied (i. e., 84 %) at the quality factor of 100. As expected, the detection accuracy decreases with a lower quality factor, whereas the false positive rate remains relatively stable.

Performance results at the glyph level (in %)

JPEG quality factor	*TPR*	*FPR*	*Acc*
100	84.0	4.7	89.7
95	76.0	5.3	85.4
90	54.0	5.1	74.5
85	42.0	4.7	68.7
80	36.0	5.1	65.5
No compression (for comparison)	100.0	0.0	100.0

These initial results have demonstrated the general feasibility of enhancing the performance of existing CMFD approaches to document forensic analysis by incorporating techniques and principles of OCR. Clearly, there is much room for improvement and further research that addresses all aspects of the conceptual analysis framework.

Concluding Remarks

The main contribution of this work is fourfold. First, we have raised and sought to draw academic attention to the practical problem of copy–move forgery detection in the so-far unexplored domain of scanned document images. To date, most CMFD methods proposed in the literature have been validated only on digital or scanned images of natural scenes, whereas digitalized versions of documents have been left out of researchers' consideration. Secondly, we have empirically studied the detection performance of block-based methods when applied to scanned images of black text on a white background. Our results have demonstrated that the existing CMFD methods, originally designed for the forensic analysis of natural images, have significant shortcomings in this case and yield many false positive matches. Third, we have adapted the common processing pipeline for CMFD and elaborated an integrated analysis framework for detecting CM tampering in text images. This theoretical framework offers a systematic way to address the critical issue of intra-glyph similarity, intrinsic to any digital document image. In essence, it integrates the two leading streams of research in digital image forensics – image source identification and forgery detection – with optical character recognition techniques, making use of all available information about the document structure, positions and sizes of individual characters. Forth, we have presented first promising evidence for the effectiveness of CM forgery detection in text images using a preliminary instantiation of the proposed analysis framework.

More generally, we consider this study as the first step toward addressing the problem of copy–move tampering in digitalized documents. The presented analysis framework is intended to provide a starting point and stimulate further thoughts regarding the design of a tailored approach for CM forgery detection in text images. The next crucial task is to implement a more comprehensive CMFD-T ('T' for text) toolbox and evaluate its detection performance in practical settings, with different feature representations and post-processing operations on the text image. This will serve to further validate the overall feasibility of the proposed framework. In addition, we see a need for future research examining the relation between font and block sizes as factors jointly influencing detection performance. We also encourage further research on the development of new (or adjustment of the existing) scanner and paper forensics techniques in order to enable a reliable detection of copied background areas.

[2] The optional pre-compression at step 2 has been omitted in our experiments. Scanned images of the genuine documents had been initially saved in lossless TIFF format.

Acknowledgments

Part of this research and the associated conference presentation have been supported by Archimedes Privatstiftung, Innsbruck.

References

[1] Yasser Alginahi. "Preprocessing Techniques in Character Recognition". *Character Recognition*. Ed. by Minoru Mori. InTech, 2010, pp. 1–20.

[2] Irene Amerini, Lamberto Ballan, Roberto Caldelli, Alberto Del Bimbo, and Giuseppe Serra, A SIFT-Based Forensic Method for Copy–Move Attack Detection and Transformation Recovery, IEEE Transactions on Information Forensics and Security, 6(3), pp. 1099–1110. (2011).

[3] Irene Amerini, Roberto Caldelli, Alberto Del Bimbo, Andrea Di Fuccia, Luigi Saravo, and Anna Paola Rizzo, Copy–move forgery detection from printed images, Proc. SPIE 9028, Media Watermarking, Security, and Forensics, pp. 1–10. (2014).

[4] M.K. Bashar, K. Noda, N. Ohnishi, and K. Mori, Exploring Duplicated Regions in Natural Images, IEEE Transactions on Image Processing, accepted for publication. (2010).

[5] Joost van Beusekom, Faisal Shafait, and Thomas M. Breuel, Document Inspection Using Text-Line Alignment, Proceedings of the 9th IAPR International Workshop on Document Analysis System, pp. 263–270. (2010).

[6] Richard G. Casey and Eric Lecolinet, A Survey of Methods and

DOI: 10.2352/ISSN.2470-1173.2016.8.MWSF-068

Strategies in Character Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(7), pp. 690–706. (1996).

[7] Vincent Christlein, Christian Riess, Johannes Jordan, Corinna Riess, and Elli Angelopoulou, An Evaluation of Popular Copy–move Forgery Detection Approaches, IEEE Transactions on Information Forensics and Security, 7(6), pp. 1841–1854. (2012).

[8] Jessica Fridrich, Miroslav Goljan, and Rui Du, Steganalysis based on JPEG compatibility, Proc. SPIE 4518, Multimedia Systems and Applications IV, pp. 275–280. (2001).

[9] Jessica Fridrich, David Soukal, and Jan Lukas, Detection of Copy Move Forgery in Digital Images, Proceedings of Digital Forensic Research Workshop. (2003).

[10] Thomas Gloe, Matthias Kirchner, Antje Winkler, and Rainer Böhme, Can we trust digital image forensics?, Proceedings of the 15th ACM International Conference on Multimedia, pp. 78–86. (2007).

[11] Andrew D. Gross, Daniel G. Neely, and Juergen Sidgman, When Paper Meets the Paperless World, The CPA Journal, 85. (2015).

[12] Nicholas Zhong-Yang Ho and Ee-Chien Chang, Residual Information of Redacted Images Hidden in the Compression Artifacts. Information Hiding (10th International Workshop). Ed. by Kaushal Solanki, Kenneth Sullivan, and Upamanyu Madhow, 5284, Lecture Notes in Computer Science (LNCS), Berlin Heidelberg: Springer-Verlag, pp. 87–101. (2008).

[13] Hailing Huang, Weiqiang Guo, and Yu Zhang, Detection of Copy-Move Forgery in Digital Images Using SIFT Algorithm, Proceedings of the Pacific-Asia Workshop on Computational Intelligence and Industrial Application, 2, pp. 272–276. (2008).

[14] S. Impedovo, L.Ottaviano, and S.Occhinegro, Optical Character Recognition – A Survey, Int. J. Pattern Recog. Artif. Intell., 5(1-2), pp. 853–862. (1991).

[15] Eric Kee and Hany Farid, Printer Profiling for Forensics and Ballistics, Proceedings of the 10th ACM Workshop on Multimedia and Security, pp. 3–10. (2008).

[16] Nitin Khanna, George T. C. Chiu, Jan P. Allebach, and Edward J. Delp, Scanner Identification with Extension to Forgery Detection, Proc. SPIE Electronic Imaging, 6819. (2008).

[17] Nitin Khanna and Edward J. Delp, Intrinsic Signatures for Scanned Documents Forensics: Effect of Font Shape and Size, IEEE International Symposium on Circuits and Systems (ISCA), pp. 3060–3063. (2010).

[18] Xunyu Pan and Siwei Lyu, Region Duplication Detection Using Image Feature Matching, IEEE Transactions on Information Forensics and Security, 5(4), pp. 857–867. (2010).

[19] Alin C. Popescu and Hany Farid, Exposing Digital Forgeries by Detecting Duplicated Image Regions, Department of Computer Science, Dartmouth College, Tech. Rep. TR2004–515. (2004).

[20] Seung-Jin Ryu, Matthias Kirchner, Min-Jeong Lee, and Heung-Kyu Lee, Rotation Invariant Localization of Duplicated Image Regions Based on Zernike Moments, IEEE Transactions on Information Forensics and Security, 8(8), pp. 1355–1370. (2013).

[21] Seung-Jin Ryu, Min-Jeong Lee, and Heung-Kyu Lee, Detection of Copy-Rotate-Move Forgery using Zernike Moments, Proceedings of Information Hiding Conference, pp. 51–65. (2010).

[22] Husrev Taha Sencar and Nasir Memon, Overview of State-of-the-Art in Digital Image Forensics, Algorithms, Architectures and Information Systems Security, 3, pp. 325–348. (2008).

[23] Shize Shang, Nasir Memon, and Xiangwei Kong, Detecting documents forged by printing and copying, EURASIP Journal on Advances in Signal Processing, 2014(1), pp. 1–13. (2014).

[24] Ewerton Silva, Tiago Carvalho, Anselmo Ferreira, and Anderson Rocha, Going deeper into copy–move forgery detection: Exploring image telltales via multi-scale analysis and voting processes, J. Vis. Commun. Image R., 29, pp. 16–32. (2015).

[25] Øivind Due Trier, Anil K. Jain, and Torfinn Taxt, Feature extraction methods for character recognition – A survey, Pattern Recognition, 29(4), pp. 641–662. (1996).

[26] Junwen Wang, Guangjie Liu, Zhan Zhang, Yuewei Dai, and Zhiquan Wang, Fast and Robust Forensics for Image Region–Duplication Forgery, Acta Automatica Sinica, 35(12), pp. 1488–1495. (2009).

[27] Abdelwahab Zramdini and Rolf Ingold, Optical Font Recognition Using Typographical Features, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), pp. 877–882. (1998).

Author Biography

Svetlana Abramova *is a PhD student in the Security and Privacy Lab at the Institute of Computer Science, Universität Innsbruck, Austria. She received her BSc ('10) and MSc ('12) degrees in Information Systems from the National Research University – Higher School of Economics, Russia, and the University of Münster, Germany. Her research interests focus on multimedia security and digital image forensics in particular as well as on virtual currencies.*

Rainer Böhme *is Professor of Security and Privacy at the Institute of Computer Science, Universität Innsbruck, Austria. A common thread in his scientific work is the interdisciplinary approach to solving exigent problems in information security and privacy, specifically concerning cyber risk, digital forensics, cyber crime, and crypto finance. Prior affiliations in his academic career include TU Dresden and Westfälische Wilhelms-Universität Münster (both in Germany) as well as the International Computer Science Institute in Berkeley, California.*