A Conditional Random Field Model for Font Forgery Detection

Romain R. Bertrand, Oriol Ramos, Petra Gomez-Krämer, Patrick Franco, Jean-Marc Ogier

A Conditional Random Field Model for Font Forgery Detection

Romain Bertrand*, Oriol Ramos Terrades†, Petra Gomez-Krämer*, Patrick Franco* and Jean-Marc Ogier*
*Laboratory L3i, University of La Rochelle, Avenue Michel Crépeau, 17042 La Rochelle, France
{romain.bertrand, petra.gomez, patrick.franco, jean-marc.ogier}@univ-lr.fr
†Computer Science Dep., Universitat Autònoma de Barcelona,
Computer Vision Center (CVC), Edifici 0, campus UAB, 08193, Bellaterra, Spain
oriolrt@cvc.uab.es

***Abstract*—Nowadays, document forgery is becoming a real issue. A large amount of documents that contain critical information as payment slips, invoices or contracts, are constantly subject to fraudster manipulation because of the lack of security regarding this kind of document. Previously, a system to detect fraudulent documents based on its intrinsic features has been presented. It was especially designed to retrieve copy-move forgery and imperfection due to fraudster manipulation. However, when a set of characters is not present in the original document, copy-move forgery is not feasible. Hence, the fraudster will use a text toolbox to add or modify information in the document by imitating the font or he will cut and paste characters from another document where the font properties are similar. This often results in font type errors. Thus, a clue to detect document forgery consists of finding characters, words or sentences in a document with font properties different from their surroundings. To this end, we present in this paper an automatic forgery detection method based on document font features. Using the Conditional Random Field a measurement of probability that a character belongs to a specific font is made by comparing the character font features to a knowledge database. Then, the character is classified as a genuine or a fake one by comparing its probability to belong to a certain font type with those of the neighboring characters.**

I. Introduction

In the context of our modern society, documents include information related to highly strategic aspects, especially in terms of industrial or intellectual property. Documents also often convey financial information concerning economic relations between different institutions and between institutions and individuals. The analysis of securing documents is a basic prerequisite for each organization, especially those who are dealing with data of high sensitivity, since a person with criminal intention might try to modify information for his own benefits or to harm others.

Military organizations, public institutions, insurance companies, and banks are among the most important organizations who need to secure the management of information contained in documents, whatever is the original format of the document: digital or paper. For the former case, i.e. digital information, there are many automated ways to secure documents from unauthorized access, e.g. by using passwords, secured connections, or local networks which are separated from the outside. Unfortunately, none of these approaches can ensure a 100% security during some particular transmission processes. Therefore, the latter case, i.e. the paper-based information exchange is often considered to be the more secure and trustworthy one. This is often enhanced through signing documents, delivering them personally, or using FAX or photocopiers instead of email transmission. However, the paper-based process also contains security leaks. If any person involved in the process tries to make some changes to the document, e.g. by changing digits in a financial document, or by modifying some terms of a contract; there is the need to develop methods that improve the way of securing paper-based documents.

Some recent studies have estimated that the cost for a company of not detecting fraudulent documents is in the range of 1.35 billion dollars, when including all the other fines, legal fees, and extra costs. These studies highlight the necessity to implement global strategies for protecting intellectual, industrial, financial information contained in documents of public and private organizations. In the context of paper documents, there is not system that enables a real protection of them. It is not possible to embed any kind of signature, or equivalent protection, that avoids illegal reproduction. Thus, the development of technologies able to detect illegal modifications of a document is becoming a very important aspect to protect original contents [1], [2]. There are several kinds of document forgeries to be detected. One of them is based on the detection of several fonts in the same document, as stated by forensics experts [3]. Indeed, in most of the "classical documents" it is quite rare to find several fonts in a same word or in a same sentence.

Consequently, we can find many works focused on font recognition [4], [5], [6]. These methods are generally based on the classical classification process. They extract a feature vector to describe font properties and train later a classifier, such as SVM or a Neural Networks, to take the final decision. In [7], the analysis of *strange* similarities between characters of the same document (too similar to be true), and the detection of outliers in a feature space (alignments, orientations, width, height of characters) in a document was also proposed for forgery detection.

In this paper, we focus on the detection of words written with different fonts. This is a first step in the ultimate goal of detecting documents where one word has been replaced by another using a different, but *similar* font. This detection has to be done on the basis of a unique document, without any

reference model to compare the analyzed one (blind approach).

We have organized the paper as follows. In section II we will briefly explain the features used in the proposed model. In section III we will introduce the CRF framework proposed in this paper. Then, in section IV we will evaluate the proposed methods and finally, in section V we will draw the conclusions and the further work.

II. Features

We have used the typographical features introduced in [5], [6]. These features are computed from segmented characters after applying an OCR system. In order to compute these features, four structuring lines have to be extracted before, as show in figure 1. Then, twelve features are computed regarding the dimension of a connected component (ie. high, short, large, squared...). The following six features at a character level are computed: density of the horizontal projection, density of the squared values of the derivative of horizontal projection, average height of vertical blacks-run, average width of horizontal blacks-run, average normalized height and average normalized width [5]. Moreover and also at character level, the following five features are computed: unnormalized height and width, xheight (height of the character central zone), xratio (ratio of xheight to height) and finally, ratio of width to height [6]. Additionally, the space between pairs of characters is also included in the model.

Since we aim to detect the forgery at the word level, our system differs from the above-mentioned work in how we compute the normalization. In fact, instead of considering the character features of all the words within the complete line, we normalize it by only taking into account the character features of the regarding word.

Fig. 1. Illustration of four structuring lines obtained by typographical classification at a word level.

We have embedded these features in a CRF model. Thus, for each letter, c; each font, f; and each size s; we have defined the unary term of a feature $x_{i,l}$:

$$\phi_{i,l,c,f,s}(x_i) = \exp\left\{-\frac{(x_{i,l} - \mu_{l,c,f,s})^2}{2\sigma^2_{l,c,f,s}}\right\} \quad (1)$$

Similarly, we have defined the pairwise terms corresponding to features $x_{i,i+1}$ of two consecutive characters, c and d by the factor:

$$\phi_{i,c,d,f,s}(x) = \exp\left\{-\frac{(x_{i,i+1} - \mu_{c,d,f,s})^2}{2\sigma^2_{c,d,f,s}}\right\} \quad (2)$$

III. Proposed method

We propose to use of conditional random fields (CRF) to represent the knowledge related to each font contained in a document [8]. CRF are powerful tools allowing to precisely describe correlations between fonts, styles and sizes of characters with features that are generally used to describe them. A CRF models the probability of a target random variable given a set of observations. In our case, target variables are the type of font f while the observations will correspond to the feature vectors $x = (x_1, x_n)$ computed from the document and the word w recognized by an OCR system. Thus, at word level the problem of font recognition is modeled by the conditional probability:

$$p(f|w,x) = \frac{1}{Z(x,w)} \prod_n \phi_n(x_n) \quad (3)$$

where $Z(x,w)$ is the partition function given the observations w and x and $\phi_n(x_n)$ are the so-called *factor* functions that describes each font f depending on the font, the word and the features. Then, the font is recognized by solving the *Maximum a posteriori* (MAP) problem:

$$\hat{f} = \underset{f}{\operatorname{argmax}}\, p(f|w,x) = \underset{f}{\operatorname{argmax}} \prod_n \phi_n(x_n) \quad (4)$$

The CRF models which used here are applied for both font recognition and forgery detection and are a bit more sophisticated than the CRF model briefly described in (3). The main difference is that most of the features described in section II are not scale invariant and hence depend on the font size. Thus, we have to deal with a CRF model with latent variables, like the font size. Thus, in the following we will describe a CRF model with latent variables that will perform font recognition at word level regardless font size. Then, in subsection III-B we will extend this model to forgery detection.

A. Font recognition

To apply suspicious word detection based on the presence of a wrong typeface, our system should be able to recognize font first. This results in the ability to assign for each word w its font among an a priori known set of typefaces.

We define the CRF as shown in figure 2. We denote a *word* by a sequence of characters $w = c_1 \cdots c_N$ and by $f = \{\text{Arial, Times, Courier}...\}$ a random variable to model all the possible existing fonts. Then, a random variable $s \in S = \{8\text{pt, } 9\text{pt, } 10\text{pt,}...\}$ is used to denote word's size in points. Finally, we indicate by $x_{i,l}$ the l-th feature as defined in section II, computed from the character c at position i.

Font recognition consists in looking for the probability to obtain a specific font f regarding a word w and a set a features x. Therefore, the font recognition system is modeled by the conditional probability $p(f|w,x)$ as defined in (5). Here, the font size s does not appear in this probability whereas, as denoted in sections II and III, somes features x strongly depend on s. Thus, we include the latent variable s in the model by marginalizing the conditional probability. At this point, we can express this probability by using the Bayes' theorem. Therefore, the problem is reduced to the estimation of the probability $p(w,x|f,s)$ which can be easily computed in regard to the training set.

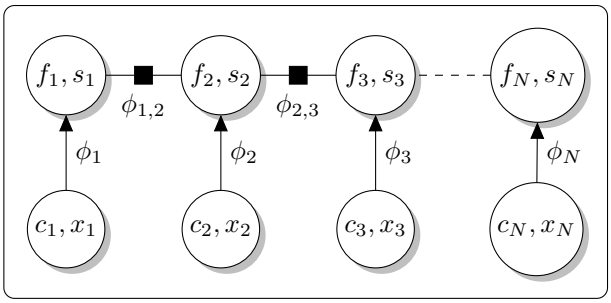

Fig. 2. Graphical representation of random variables involved in the font recognition model. x_i denotes a vector with l features and factor ϕ_i stands for $\prod_l \phi_{\{i,l,c_i,f_i,s_i\}}(x_{i,l})$. Similarly, factor $\phi_{i,i+1}$ stands for $\prod_i \phi_{\{i,c_i,c_{i+1},f_i,s_{i+1}\}}(x_{i,i+1})$.

$$p(f|w,x) = \sum_{s\in S} p(f,s|w,x) = \sum_{s\in S} \frac{p(w,x|f,s)p(f,s)}{p(f,s)} = \\ = \frac{1}{Z(x,w)} \sum_{s\in S} \prod_{i,l,f} \phi_{i,l,c,f,s}(x_{i,l}) \prod_{i,f} \phi_{i,c,d,f,s}(x_{i,i+1}) \tag{5}$$

In (5), $Z(x,w)$ is the partition function, given the features x and the recognized word w. We have defined factors $\phi_{i,l,c,f,s}$ and $\phi_{i,c,d,f,s}$ in (1) and (2), respectively. Moreover, we assume that the *a priori* probabilities $p(f,s)$ are equally distributed. We finally find the font f that maximizes (5):

$$\hat{f} = \underset{f}{\operatorname{argmax}}\, p(f|w,x) = \\ = \underset{f}{\operatorname{argmax}} \sum_{s\in S} \prod_{i,l,f} \phi_{i,l,c,f,s}(x_{i,l}) \prod_{i,f} \phi_{i,c,d,f,s}(x_{i,i+1}) \tag{6}$$

B. Suspicious word detection

We have done the suspicious word detection on the basis of the result of a previous font recognition. Thus, we consider the categorical random variable y denoting if a word has been forged, or not, together with the same categorical random variables defined in the above subsection. Thus, the probability of a *non suspicious* word given the font f, the word w and the feature vector x is given by (5) as follows:

$$p(y = \text{non suspicious}|f,w,x) = \frac{p(\text{non suspicious},w,x,f)}{p(w,x,f)} \\ = \frac{p(f|w,x)p(\text{non suspicious})}{p(f,w,x)} \tag{7}$$

Similarly, we define the probability of a fraudulent word as:

$$p(y = \text{suspicious}|f,w,x) = \frac{p(\text{suspicious},w,x,f)}{p(w,x,f)} \\ = \frac{p(f|\text{suspicious},w,x)p(\text{suspicious})}{p(w,x,f)} \tag{8}$$

We estimate the *a priori* probabilities $p(\text{suspicious})$ and $p(\text{non suspicious})$ from the training dataset and we model the conditional probability $p(f|\text{suspicious},w,x)$ using the unary and pairwise factors defined in (1) and (2), respectively. We assume that font forgery is done by copying and paste similar fonts from other document to the target one. Thus, for each letter, c; each font, f; and each size s; we define the unary term of a feature $x_{i,l}$ as in (1). On the contrary, the space between two consecutive characters written with different fonts and/or size, can not been modeled by (2) and we define the pairwise term corresponding to features $x_{i,i+1}$ of two consecutive characters, c and d by the factor:

$$\psi_{i,c,d,f,s}(x) = 1 - \exp\left\{ -\frac{(x_{i,i+1} - \mu_{c,d,f,s})^2}{2\sigma^2_{c,d,f,s}} \right\} \tag{9}$$

Note that with the above definition we are modeling the conditional probability that two consecutive characters to be written with different fonts and/or size. The probability $p(f|\text{suspicious},w,x)$ is given by:

$$p(f|\text{suspicious},w,x) = \frac{1}{Z(\text{suspicious},w,x)} \cdot \\ \cdot \sum_{s\in S} \prod_{i,l,f} \phi_{i,l,c,f,s}(x_{i,l}) \prod_{i,f} \psi_{i,c,d,f,s}(x_{i,i+1}) \tag{10}$$

The final decision is taken by comparing the probabilities obtained in (7) and (10).

IV. Experimental results

In this section, we present how we process the evaluation of our system. Firstly, font recognition and classification is evaluated followed by forgery detection. The dataset corresponding to each experimentation is also described.

A. Font recognition and classification

In order to train our system for font recognition and classification, ninety nine documents were generated. Each one corresponds to a 300 dpi noise-free document image containing two hundred and fifty words from *Lorem Ipsum* generated by a computer in eleven typeface and nine different font sizes. These documents are considered as the training set. Afterwards, eleven documents of each typeface comprising fifty words were generated as test evaluation files. Their font size contents vary randomly from 9 to 15 points. Once these two sets of documents were generated, feature extraction is performed as described in section II and it is used as input in our system (see section III).

As a result, we measure the capability of the system to correctly label (eg. typeface class *Arial*) all the words within the eleven test documents as illustrated in table I.

The numbers displayed in the confusion matrix (table I) show that the use of the CRF model at the word level lead to a reasonable classification process and, thus, to an adequate forgery detection. In fact, nine typefaces out of eleven, present a considerably high score when detecting the right label. However, if we look in detail, the difference of some classification results can be explained based on the typeface properties as pixel distribution. For instance, *Courier New* differs considerably from *Liberation Sans* while this latter is very similar to *Arial* as shown in figure 3. It is important to

Typeface	Arial	Calibri	Cambria	Cantarell	Courier New	DejaVu Sans	Franklin Gothic	Garamond	Liberation Sans	Tahoma	Times New Roman
Arial	44	1	0	1	0	0	0	0	2	0	0
Calibri	0	38	0	1	0	2	0	0	0	2	0
Cambria	0	0	46	0	0	5	0	0	0	0	0
Cantarell	2	0	0	34	0	2	0	0	0	0	0
Courier New	0	0	0	0	48	0	0	0	0	0	0
DejaVu Sans	0	0	0	0	0	45	0	0	1	0	0
Franklin Gothic	0	0	0	0	0	0	47	0	0	0	0
Garamond	0	1	3	0	0	0	0	25	0	0	5
Liberation Sans	19	0	0	8	0	0	0	0	15	0	0
Tahoma	0	0	0	5	0	2	1	0	1	37	0
Times New Roman	0	1	3	0	0	0	0	1	0	0	35

TABLE I. TYPEFACE CONFUSION MATRIX AT A WORD LEVEL. ROWS DENOTE GROUND TRUTH WHILE COLUMNS SHOW RECOGNIZED TYPEFACES. EACH UNIT IN THE MATRIX REPRESENTS A WORD.

note that some of the test words do not appear in the confusion matrix. In fact, due to OCR errors, some of them have been removed from the testing data.

Fig. 3. Illustration of three different typefaces. From left to right: *Arial*, *Liberation Sans*, *Courier New*. As we can see, *Arial* and *Liberation Sans* are visually similar.

B. Forgery detection

Forgery detection process was undertaken by performing the same training process explained in the previous step followed by the evaluation of a series of tests. When modifying information in a document, the chances of making font types errors are often high. Therefore, the process allowing the evaluation of forgery detection was divided in a set of three tests. Each test considered a potential committed error when falsifying a document based on the hypothesis of the proposed method described in section III. At the word level, it is possible to distinguished three type of errors when the word has been modified (figure 4):

- Copy/paste: the space between a pair of characters differs from the ones between the rest of the pair of characters. It can be either smaller or bigger ;
- Imitation: a single word appears to be written in two different typefaces ;
- Copy/paste and Imitation: both errors, above described, might be found in the same word.

In order to evaluate the system, we have conceived a software allowing to generate some documents that we considered as *forged* based on the three types of errors. Each error was generated randomly on three words out of three-hundred (1% of forgery) over sixty four documents. Only six typefaces are considered here (*Arial*, *Calibri*, *Cantarell*, *Courier New*, *Liberation Sans* and *Tahoma*) in eleven sizes (from 8 to 16 points).

Fig. 4. Illustration of three different type of errors due to forgery. From the top to the bottom: Copy/paste forgery (space between *o* and *n*), Imitation forgery (two typefaces used), Copy/paste and Imitation forgery (a space separating two typefaces)

As a result, 192 documents were generated and used to evaluate the performance of the system as illustrated in table II.

	Recall	Precision
Copy/paste	0.10	0.25
Imitation	0.31	0.22
Copy/paste and Imitation	0.48	0.25

TABLE II. RECALL AND PRECISION OF THE THREE EXPERIMENTS.

The *Copy/paste* experimentation reveals the difficulty for our system to retrieve fraudulent words when the only clue about the manipulation of the fraudster is a spacing error between a pair of characters. Even if one over ten forgeries is highlighted, character spacing error does not appear to be discriminative enough to discover forgeries regarding the recall and precision values. If the score appears to be really low in this experiment, it is important to note that a word forgery will certainly lead to more than one imperfection caused by the fraudster. The next two experiments are designed on this assumption.

In the case of *Imitation* experimentation, we can see that one over three forgeries is retrieved by our system while this score is close to one over two concerning *Copy/paste and Imitation* experiment. Some forgeries aren't retrieved because of the difficulty for the system to accurately classify each character of a word. Thus, the system incorrectly labels some fraudulent words composed of two typefaces promoting one typeface over the other, resulting by an incapacity to highlight font forgeries as demonstrate in figure 5.

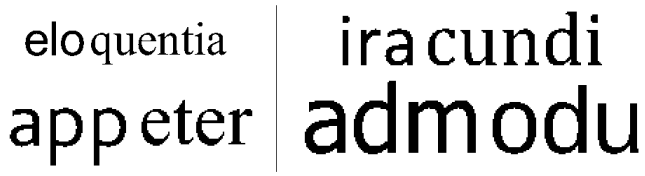

Fig. 5. Illustration of four forged words. Two are correctly classified as forgeries by our system (left column) while the other two are classified as genuine (right column) resulting in a classification error. Here, the classification error is surly due to the shape proximity of the font used to forge the word regarding the original one.

Concerning the precision of these two experiments, the low values obtained here are mainly caused by *Liberation Sans* and *Tahoma* typeface as shown in table III. This drop in the

precision value is surely due to the inability of the method to correctly classify this two typefaces as shown in table I.

Finally, we have to point out that for all these experiments, we have to deal with the error propagation issue caused by the OCR step, based of the whole features extraction process.

	Recall	Precision
Arial	0.44	0.35
Calibri	0.33	0.30
Cantarell	0.48	0.39
Courier New	0.63	0.34
Liberation Sans	0.55	0.1
Tahoma	0.48	0.13

TABLE III. RECALL AND PRECISION FOR EACH TYPEFACE REGARDING THE THIRD EXPERIMENTATION (COPY/PASTE AND IMITATION FORGERIES).

V. CONCLUSIONS AND FURTHER WORK

We present in this paper an automatic forgery detection method based on document font features. The method is based on a Conditional Random Field model which first allows to recognize and classify typefaces and then to highlight font forgeries. This work is a first approach to address the growing issue of document manipulation by fraudsters. Even if the actual state of the system suffers from a lack of precision, we obtained encouraging results which could be improved by easily adding new font features to the model. Moreover, adding information relative to the neighboring words could increased the font recognition rate and maximized the font forgery detection.

REFERENCES

[1] S. Elkasrawi and F. Shafait, "Printer identification using supervised learning for document forgery detection," in *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, April 2014, pp. 146–150.

[2] A. Piva, "An overview on image forensics," in *ISRN Signal Processing*, 2013, pp. 146–150.

[3] M. I. Malik, M. Liwicki, L. Alewijnse, M. Blumenstein, C. Berger, R. Stoel, and B. Found, Eds., *Proceedings of the 2nd ICDAR International Workshop on Automated Forensic Handwriting Analysis, AFHA 2013, Washington DC, USA, 22-23 August 2013*, ser. CEUR Workshop Proceedings, vol. 1022. CEUR-WS.org, 2013. [Online]. Available: http://ceur-ws.org/Vol-1022

[4] B. Bataineh, S. Abdullah, and K. Omar, "A statistical global feature extraction method for optical font recognition," in *Intelligent Information and Database Systems*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, vol. 6591, pp. 257–267.

[5] A. Zramdini and R. Ingold, "Optical font recognition using typographical features," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 8, pp. 877–882, Aug 1998.

[6] A. Satkhozhina, I. Ahmadullin, and J. P. Allebach, "Optical font recognition using conditional random field," in *Proceedings of the 2013 ACM Symposium on Document Engineering*, ser. DocEng '13. New York, NY, USA: ACM, 2013, pp. 119–122. [Online]. Available: http://doi.acm.org/10.1145/2494266.2494307

[7] R. Bertrand, P. Gomez-Kramer, O. Ramos Terrades, P. Franco, and J.-M. Ogier, "A system based on intrinsic features for fraudulent document detection," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, Aug 2013, pp. 106–110.

[8] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Foundations and Trends in Machine Learning*, vol. 4, no. 4, pp. 267–373, 2011. [Online]. Available: http://dx.doi.org/10.1561/2200000013