Automatic removal of crossed-out handwritten text and the effect on writer verification and identification

(The original paper was published in: Proc. of Document Recognition and Retrieval XV, IS&T/SPIE International Symposium on Electronic Imaging 2008.)

Axel Brink Harro van der Klauw Lambert Schomaker

Dept. of Artificial Intelligence, University of Groningen P.O. Box 407, 9700 AK Groningen, The Netherlands

ABSTRACT

A method is presented for automatically identifying and removing crossed-out text in off-line handwriting. It classifies connected components by simply comparing two scalar features with thresholds. The performance is quantified based on manually labeled connected components of 250 pages of a forensic dataset. 47% of connected components consisting of crossed-out text can be removed automatically while 99% of the normal text components are preserved. The influence of automatically removing crossed-out text on writer verification and identification is also quantified. This influence is not significant.

Keywords: Crossed-out text, off-line, writer verification, writer identification

1. INTRODUCTION

Computers can read constrained handwritten text reasonably well. While it still is a hard problem that has not yet been completely solved, fairly good recognizers exist.¹ These recognizers require that the text has been written in a strictly defined format. Several collections of such formatted handwritten text have been created in controlled conditions, for example see² or.³ These collections are very valuable for training automatic recognizers and for testing whether automatic recognition works in principle.

In practice, however, no assumptions about the input documents can be made. Therefore, automatic recognition of unrestricted handwritten text is still problematic. For example, when a student takes a page of notes written during a lecture, puts it in the scanner and instructs the computer to read it, it will probably produce garbage. One of the main problems is that the computer has no idea *where* the text is. Usually, everything that is dark, is seen as text. This can include elements like background print, stains, and physical damage. These elements hold no information about the handwriting and should be discarded. Since it is hard to define those elements explicitly, removing them is generally not straightforward.

Another such element is crossed-out text. Crossed-out text is intended not to be read and therefore it seems wise to identify them as such. Such elements impede automatic handwriting recognition.⁴ Identifying and discarding crossed-out text is not only relevant for text recognition, but also for writer verification and writer identification. It is often assumed that crossed-out text also impedes computation of writer specific features of the handwriting, because it is irregular.⁵ It is conceivable that feature extraction methods find many bogus features in the crossed-out text that may not be present in another text of the same writer, decreasing the apparent similarity. Therefore, it seems appropriate to attempt to remove crossed-out text prior to automatic writer verification.

Identifying crossed-out text has not attracted much attention yet. One approach focused on separate characters and distinguished characters from noise including crossed-out characters.⁶ In a more recent approach,

Further author information: (Send correspondence to Axel Brink) Axel Brink: E-mail: a.a.brink@ai.rug.nl, Telephone: +31 50 363 7410 Lambert Schomaker: E-mail: schomaker@ai.rug.nl

Harro van der Klauw: E-mail: lvdklauw@ai.rug.nl

Mudul ac in New York, Tokyo, Andree, Parijs , 2iaich, en Oslo waren geweest, vlogen ze urt de Usen Terrug met vlucht K.L. 650 Ze kwamen aan in Oublin om 7 ^{una} en 101 Amster Dam om grouw SAvonds. De Fiat van Bob en VW van David stonden In Vak R3 van het pankentmein. Hienvoor Moesten ze Honderd Gulden (frag.) Betalen	Hanni I. WEV YORK, TENCO, CHEBRE PAOLA, ZIANH ER COLO WARK GEDESST, HANNI ZE IN TE AM TERA DET VILKIT VE GOO'N EN-GAR. LUCAN ZE TANIMU MORTH AND ANNER EN IN NOTEDDAN ON GO AN SATURD. TE TANIMUM MORE VILLAN DANG ANNER IN VAL 8.3 INN KET (HANZER TRREIN). HURVER MAESTEN EF IN (ANDRON) FUN (RITHEN. NG PEZZETPE ANNO GEDEN EF NARR HUN VALAGEN (RISS FREC, TAN IRONE EN HENK ANDET DIN REINDRACH GETA EN HAND HISTORI (RISS FREC, TAN IRONE EN HENK ANDET DIN REINDRACH GETA EN HAND HISTORI (RISS FREC, TAN IRONE EN HENK ANDET DIN REINDRACH GETA EN HAND HISTORI (RISS FREC, TAN IRONE EN HENK ANDET DIN REINDRACH GETA EN HAND HISTORI (RISS FREC, TAN IRONE EN HENK ANDET DIN REINDRACH GETA EN HAND HISTORI (RISS FREC, TAN IRONE EN HENK ANDET DIN REINDRACH GETA EN HAND HISTORI (RISS FREC, TAN IRONE EN HENK ANDET DIN REINDRACH GETA EN HAND HISTORI (RISS FREC, TAN IRONE EN HENK ANDET DIN REINDRACH GETA EN HAND HISTORI (RISS FREC, TAN IRONE EN HENK ANDET DIN REINDRACH (RISS FREC, TAN IRONE) FRIGUNE ON FRECHENERS) FRIE UNE FRECHENER (RIKER GENERE)
Nog Dezelfole Avond Reden ze Naan Han Vaienden Chris, Emile, Jan, Inine en Henk, Nadal ze Hun Uniendinnen Greta en Maria Hoddan Opgeboeld. Samen Hadden ze vijfhondend Gestapp). (500) Pzelazame Postzegels gehadd, Bob drichondend (200) en Oavid die bondend (200). De Reis was de Mooite bached Geveert.	, h
NADAT ZE IN UGU Yaek i TobyO QueBeq, jAeiss, Zuech EU CSIO UAREN GEWERST VLOGEN ZE UIT DE USA TERUG MET VLUCHT KL 658 ON III UNC. ZE KUNEN AM IN DUDIN OM FUNC EN IN ANSTREAM ON 19:40 UNC. S'ANONS DE FAT UN BOB EN OL UN UNN DAVID STONDEN IN VAL 23. UNU HOT ARDRETES IN HIERUNC MORSTEN ZS HENDEDBOUNDAN (100) BETMEN	NADel se jn newyork Tohyo, ceuwBeh, Arejs, Zurich, lin Oslo, WarEn Dewerst. Vloben ze uit Ne USA. Terug MéTvluchT. k.L. 650 om 1200r. Ze hWamen Aanin Dublin om 7.00 UNR EN in AmsterDag og. 40 UNR avondin. De Fiat Van Bob. En De VW Van
NOC DEZELFOE AVOID REEREN ZE WARE HUU ÚRLEMEN CHRIS, EMIEC, JANJ, IRENE EN HEUR, WAAT ZE HUN UNEUDINNER GOETA EN GARIA HADREN CRECHARD, SAKEU HARDEN ZE UIJEHANBER (JSO) ZEUDIANE DISTZEGEUS GERCHT, BOB ERENDREED (JSO) EN DINID TWEFHONDERD (Goo) DE REIS WIS DE HOETE WAARD (EWENST.	Davio Stunden jn Valk R 3 VAN HET GurleEKTERE.IN. HIER VOOR MOESTEIN ZE HERE HONDERN GUlden Glood Betellen. Nog Dezelfde Guowa REDEN ZE NAAR HUN VRIENDEN CRIS, EMIL, June EN HENK, NAOSTZE HUN Vriendinn Gretz en Maria Halden Of gehaald Samen Haden ze Vij Hondern (Soo) zeldzame Pestzegels Gehocht. Bob Soo Orie Honderts (Soo) En David fuerhoeren (200). De Raja WAS De Moeite WAARD GEWEEST.

Figure 1. Example documents in the NFI dataset, each one cut in two parts.

Markov Random Fields were used to identify crossed-out words in very challenging documents.⁷ A special property of that particular method is that it seems robust against connections between crossed-out words and normal words. The results look promising, but the performance has not been quantified.

In this paper a method for identifying crossed-out words in offline handwriting is proposed. It works on the level of connected components and classifies them based on two features of the skeleton: the *branching* feature and the *size* feature. The system is trained and tested on a part of a real forensic dataset, called the NFI dataset. This dataset was first introduced in.⁵ It consists of 3500 handwritten samples taken from suspects in criminal cases; these samples have previously been studied studied manually by the NFI, the Dutch National Forensic Institute. Apart from the textual content, the handwriting is unconstrained and contains many crossed-out words. See Figure 1 for an example. This dataset seems to be somewhat similar to the kind of data used in,⁸ which consists of spontaneous handwriting.

Training and testing was performed in three stages. In the first stage, the classification performance is assessed on the level of connected components in the first 250 pages of the NFI dataset. In the second and third stage, this classification is applied to assess the effect on writer verification and identification, respectively, on 2374 pages.

2. RECOGNIZING CROSSED-OUT WORDS

2.1 Preprocessing and segmentation

The first 250 pages of the NFI dataset were thresholded using Otsu's thresholding method.⁹ From the result, the black connected components were extracted using 8-connectivity. Two kinds of components were discarded: very

Table 1. Classes of labeled connected components.				
	normal	\mathbf{other}		
train set	43745	403	202	
test set	41640	221	326	
examples	2	E A	Ą	
	Dav	æ	po	
	se	UB	to	
	1	*	3	
	9	Ę		
	h			
	110	4		
	h	SCA		

small components with a width or height smaller than 7 pixels, and very big components with a width or height bigger than half the page. The small components can be considered to be noise or dots; the big components were caused by page border effects.

This resulted in a set of 86537 connected components. These were manually labeled into three categories: "normal", "crossed-out" and "other". The category "other" consisted of connected components that are noise or textual elements that could not be clearly categorized into one of the other categories. This categorization proved to be not straightforward during the manual labeling process, which indicates that the problem of detection of crossed-out text may actually be ill-posed. The number of components in each class and examples are shown in Table 1.

2.2 Branching feature

When handwritten words are crossed-out, one or more strokes are written over existing strokes. The result can be seen as a high number of strokes with many crossings. The branching feature takes the number of crossings into account, where each crossing is called a branching point. To find the branching points of the connected components, they were first thinned using a recent method.¹⁰ In the resulting skeleton, the branching points were identified as the black pixels that have more than two 8-connected black neighbors. This usually results in more than one branching point per actual crossing, but this is not important for the quality of the feature. The resulting number of branching points was normalized by dividing by the width of the connected component.

2.3 Size feature

The second feature exploits the fact that crossed-out text is usually a big object, because the crossing strokes add ink and usually connect individual letters or parts of a word. The size of the object is measured by counting the number of pixels in the skeleton image. Another approach would be to count the number of pixels in the original connected component; our tests indicated that that does not make much of a difference.

2.4 Training

The just mentioned features were both normalized by dividing the feature values by the standard deviation within each page. This ensures that feature values that are not common within a page can be identified as relevant extreme values. The final result is a set of labeled two-dimensional data. This data was split in two parts: a *train* set containing the connected components from page 1–125 and a *test* set involving page 126–250. The feature values of the classes "normal" and "crossed-out" in the train set are plotted in Figure 2. The feature



Figure 2. Feature values of connected components labeled as normal text ('.') and crossed-out text ('+') in page 1–125. Many pluses are inside the cloud of dots.



Figure 3. Multiple ROCs. Each ROC has a fixed θ_s ; see legend. Along each curve, θ_b varies. The curves do not reach the upper right corner because given the selections of θ_s , θ_b could not be positioned such that all data points would fall within the decision boundaries.

values of the class "other" were not plotted, since they are not relevant for determining a decision boundary between the features of normal words and crossed-out words.

The figure shows that the classes "normal" and "crossed-out" mainly overlap, but not totally. It also shows that there are many more instances in the "normal" class. The classes can be separated up to a certain degree by a decision tree, which is a very simple classifier. Several other classifiers have been tried as well, including k-nearest neighbor, a linear support vector machine¹¹ and a neural network, but since their performance was not better and a decision tree is simple, the latter was used.

The decision tree was implemented by setting thresholds on each of the two normalized feature values: θ_s is the threshold on the size feature; θ_b is the threshold on the branching feature. Values above both of the thresholds were seen as positive examples, or crossed-out words. By positioning the thresholds, the ratio of true positives (TP) and true negatives (TN) can be balanced. This can be done using ROC plots which are created on the train set; see figure 3 and 4.

The optimal balance between TP and TN depends on the application, but at least it is desired that most of



Figure 4. Same zoomed figure as Figure 3(b), but now every ROC curve has a fixed θ_b (see legend); along each ROC θ_s varies.

the normal text is not seen as crossed-out and thus remains intact. For illustrative purposes, it is now assumed that TN should be at least 99%. In other words, the number of false positives (FP) should be less than 1%. Using plots 3(b) and 4, it can be derived that usable thresholds would be $\theta_s = 1$; $\theta_b = 1.5$.

2.5 Results

The thresholds $\theta_s = 1$ and $\theta_b = 1.5$ were applied to the test set, which was completely fresh: it had not been used for training or testing before. The results are: TP = 47.5% and TN = 99.1%. That means that almost half of the crossed-out words can be automatically removed while preserving 99% of the normal text. Figure 5 shows what the result would be on the images from Figure 1. Figure 6 shows the results using other thresholds. In these examples all of the crossed-out words have been successfully removed. It is clear that some of the components of normal words are removed as well, particularly bigger components, but most components of normal words remain.

To illustrate how the method scales to very big scratches, a small experiment was also performed on semiartificial data: the four pages of which the cutted versions are shown in Figure 1 have been overlaid with pages containing big scratches. For this test the condition that the crossed-out components should not be bigger than half the page was relaxed. Figure 7 shows what the result would be on such pages.

3. APPLICATION TO WRITER VERIFICATION

The proposed technique to automatically remove crossed-out words was applied in a writer verification experiment to determine whether it affects performance. Writer verification means that a decision must be made whether two documents have been written by the same person. This can be done automatically by first computing writerspecific feature vectors of the input documents and then applying a threshold on a distance measure between the feature vectors. In this experiment, the powerful $Hinge^{12}$ feature was used. This technique captures the orientation and curvature of the ink trace, encoded in a 528-dimensional feature vector. As a distance measure, the χ^2 measure¹³ was used. It is defined (after renaming) as:

$$d_{\chi^2}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{|\mathbf{a}|} \frac{(\mathbf{a}_i - \mathbf{b}_i)^2}{\mathbf{a}_i + \mathbf{b}_i}$$

where *i* is an index to the elements of feature vectors **a** and **b**. Documents with a χ^2 distance below a certain threshold were classified as belonging to the same writer.

The experiment was performed as follows. First, like in,⁵ all pages of the NFI dataset were split in an upper part and a lower part. 1127 pages had to be discarded because curved or sloped baselines made a good cut

an ze in New onk, Tok o, Rubec, Panijs Zürich, en Oslo waren geweest. vogen ze urt de Usa Terug I vlucht KL. 65 Ze waren aan in Dubin o J ^{una} en i Ams Dai on, "o una SAvon s. int va ben VW van av: ston n Hi voor o sen ze Ion n Gul n 1, c) Betalen	ADAT DE III NYEV VORK, TOKO, GUE PARIAS Z BISH E 1.96 BOUESS, UKAN ZE WIT FR UKA TIR C ET WUNT FL GAVON ROMAKA. 2 KAN E I DAWN ON JOHN R I NITTEP MON JOW UR 25 WARD. 19 THE WIN DEVININ DAND DEN IN V K 8.3 UNIKET R E T EN. HURWAR GETENGE DA (MAND) F TLEN. 66 DOZZETIPE AV ND & DEN ZEN R H N WREND N (MRIS FMEL, TAN IRENE EN N KA HT HUN IRENDINEN GETA NN N D DO OK I. 5 MADEMI (F SON CHIF D DI DEHERMIL TZEGELS GE T. 300 MA BOUND) FRIDEN OF SON CHIF D DUD HERENNIL TZEGELS GE T. 300 MAC BOUND) FRIDEN OF SON CHIF D DUD HERENNIL TZEGELS GE T. 300 MAC BOUND) FRIDEN OF SON CHIF D DUD HERENNIL TZEGELS GE T. 300 MAC BOUND) FRIDEN OF SON CHIF D DUD HERENNIL TZEGELS GE T. 300 MAC BOUND) FRIDEN OF SON CHIF D DUD HERENNIL TZEGELS GE T. 300 MAC BOUND) FRIDEN OF SON FOR D DITHERENT.
Nog Dezelfole Avond. Reden ze N n an Vaienden Chruis, Emil. Jan, I'ne en Henle, Na al ze Hun bri vi innen Greta n Varia Ha n Opg (d. Samen adden 2 vijf onderd ; (500) Zelatzami Roshzegols L Bob d'chon end(300) en Qavi Hue hon nd (200). De Reis Was Mo: Woord Gev I.	
N AT ZE IN NEW YORK, TO D QUE - 12 ISS. ZURCH EN DSID Y EEN ENERST WOG NZE UT E USA TER MET VLUCHT KL 658 OM IZUMR. ZE HEN AMN N UDIN OM FURE EN IN STER OM GIGOUNE S'ANOMS. DE FIT BOOR EN ENN STER OM GIGOUN V K B3 UAN HOT I TEGIN HIER M STEN ZS HENDEDGOM (1) B-	a jn ew ok Tohyo, cevind , P js, Zu ich, Lin Oslo, IREN Dewecst. VloGEN ZE UI DE US A. TERU MÉTVluchT. K.L. 65 OM ZUUR. ZE K IMEN 'N' USLIN OM 7.00 UUR EN IV AMSter Y Og 40 UUR 'avond'. e Fift VAN OS. EN DE V V N
Wac Dezelfae Avaro - Zeriz e Hulu úl2(cmCr C is, 6471€C, S , I we EN evik, D T 26 Hulu Unie iwn Gaera en ig Hao Cae- , S -U H ev 26 Uijarkan D (Sao)~ ; Distrzecas ∝T E (Sar EN I Tinder N (Sao) DE s W DE - W GEWE ST.	D vio Stow w Jn V & R 3 V N ET Ar EERTEREIN. IER Voor MOESTEIN ZE OUDE D Ulden 1000 I EN. Nog Dezelfde QuoU REDENZEN I HUN VRIGNDEN CRIS, EMIL, JW J e N ENK, N DRIZE UN VIE IMMN Retz EN MARI H EN O HAALD MEN GDEN VIJ ON RO (500) Zel Zame (25 Zegels GEWOCHT. BOD Rie ONDE D (300 EN Davin ec 1) (200). DE RGS. WAS MOETE WAARD GEWEEST

Figure 5. Images from Figure 1; thresholds $\theta_s = 1$ and $\theta_b = 1.5$ applied. All crossed-out components have been removed at the expense of some normal text.

Table 2. Number of pages and writers in the dataset for verification and identification.

	train set	test set
original pages (writers)	250	3250
selected pages (writers)	181(87)	2193 (988)
selected parts (writers)	362 (87)	4386 (988)

impossible. After splitting, 4748 page parts remained, written by 1074 persons. The page parts were converted to monochrome using Otsu thresholding because this is required by the method to remove crossed-out words. The page parts were divided into two sets: a train set and a test set. The train set consisted of the parts of pages 1–250; the test set consisted of the parts from the other 3250 pages. See Table 2 for details. To compute the baseline performance, a verification threshold was learned from the training data by modeling the distances in the "same writer" and "different writer" classes using Parzen windowing. The threshold was selected such that the expected ratio of true positives (TP) was equal to the expected ratio of true negatives (TN); the equal-error rate (EER). This threshold was applied to the test set, yielding the experimental TP and TN.

To assess the effect of crossed-out text, the same steps were taken on the same page parts after cleaning by the crossed-out text removal method. Several values of θ_s and θ_b were tried while testing on the train set. The results are shown in Table 3. The table shows that the values for θ_s and θ_b have no big implications on writer verification, but the best result was used to determine the final values: $\theta_s = 2$ and $\theta_b = 2$.

ndal ze in New York, Tokyo, Qu'bec, Parijs 2 ürich, en Oslo waren geweest. vlogen ze urt de Usa Terrig met vlucht KL. 650 Ze kwamen aan in Dublin o Junnen II Amsteda I ori, grouwn SA vonds. D Tiat van Bob en VW van David ston n In Vak R3 van hJ par nrein. Betalen	APART 22 IN NEV KORK, TOKK, GUEBER PARIS, ZWICH PM I ATM GURESS 5 UNGAR 22 INT PC USA THR 6 MET MUNT TH GABOT RE-MAR. 2 BURNAM IN DAILY AND STUDIED IN MUTERIAN OF QU UR 3. DAILY OF THAT VAN MARE FUNN NAMO STUDIED IN VAN 8.3 VAN KET (AR DES TABRICU). HER VOR NOETON CE DOC (AND) FUN I VAN 8.3 VAN KET (AR DES TABRICU). HER VOR NOETON CE DOC (AND) FUN I VAN 8.3 VAN KET (AR DES TABRICU). HER VOR NOETON CE DOC (AND) FUN VAR BUN VAR BUN (ARST FOR JAN, MICH EN DE K.A. DT (INV VAR DODOR) GETA FINNAR BUN VAR BARDON (AND JAN) FANDA MASSIT CE STOCKNOOD CERTAFIAN KETZISTS CERCON, 728 MOR (AND MURE) FANDA MASSIT CE STOCKNOOD CERTAFIAN KETZISTS CERCON, 728 MOR (AND MURE) FANDA MASSIT CE STOCKNOOD CERTAFIAN KETZISTS CERCON, 728 MOR (AND MURE) FANDA MASSIT CE STOCKNOOD CERTAFIAN KETZISTS CERCON, 728 MOR (AND MURE) FANDA MASSIT CE STOCKNOOD CERTAFIAN KETZISTS CERCON, 728 MOR (AND MURE) FANDA MASSIT CE STOCKNOOD CERTAFIAN KETZISTS CERCON, 728 MOR (AND MURE) FENDAND STOCKNET HENDERD CERTAFIAN KETZISTS
Nog Dezelfole Avond Reden ze Naan Hun Union den Chris, Emile, Jan, Inine en Henk, Nadal ze Hun Uniondinnen Greta en Maria Ho n Opgehoeld. Samen adden 2 oijfhondend i (500) 2etatami Postzegels gehad. Rob dinichondend (300) en David Huerhon nd (200). De Reis was de Moeite Wandel Geveent.	
NADAT ZE IN NEW York, Tobyo Que - , Maiss, zuech en crib Waren geweest Wogen ze ut De Usa Teeug Met Ulucht KL 658 cm. I'L unc. Ze Kummen Ann in Dublin OH June en IN Ansteagn ch g: 40 une s'Avans. De Fitz un Bab EN e IN an Dubl Stongen In VAL B3 UAN Hot A. Tegin Hierusce Maesten zo Handesben (1) Bethen.	N Det JA NEW OKL TOLYO, CENWBOL, PARIJS, Zu ich, Lin Oslo, WAREN DEWEESL VLOGEN ZE UI DE U.S.A. TERU MÉTVLUCHT. U.L. 650 OM IZUUR. ZE LWI/MEN JANIN DUBLIN OM 7.00 UUR EN IN AMSTER IM OG.40 UUR GUORDA. DE FIAT VAN BOB. EN DE VW VAN
NOC DEZELFOR AVOND - ZE LXARE HULY ÚDLEMEN CHREIS, ETHEC, JARJ, ILENE EN HEUR, LXANT ZE HUN UME INN GAZTA EN MARIA HADDEN ORGEN ALD, SAMED HADDEN ZE UIJEHLWIBED (JW) ZELDZHAF DISTZEGAS WHT EED (JWA EN DINID TWEEHGADEED Gwo) DE DENS HUR DE - W GEWENST.	Davio Stow Dev ja VAL R 3 VAN HET PARLEEKTERE.IV. HIER Voor MOESTEW ZE KOUDERN VIELW BERLEN. NOG DEZELIDE CWOND REDEN ZE N A HUN VRIENDEW CRIS, EMIL, JAN JREME EN HENK, N DRI ZE HUN VRIENDEW GRETE EN MARIA H DEW OP HOALD SAMEN HODEN ZE VG HONDERD (SOO) Zeld Zame (Sslzegth GEWOCHT. BOD ORIEHONDERD (SOO) ZN DAVID fwee ED (200). DE RGS. WAS DE MOEITE WAARD GEWEEST.

Figure 6. Images from Figure 1; thresholds $\theta_s = 2.5$ and $\theta_b = 2.5$ applied. With less strict thresholds, more of the normal text remains. Crossed-out text is also more likely to remain, but that does not occur in this example.

Table 3. Writer verification results on 362 thresholded half pages extracted from the train set (the first 250 pages) of the NFI dataset; 87 writers.

θ_s	$ heta_b$	TP	TN
1	1	78.1%	81.9%
1	1.5	79.3%	81.5%
1.5	1	78.7%	81.3%
1.5	1.5	79.7%	80.7%
1.5	2	80.4%	80.3%
2	1.5	80.3%	80.5%
2	2	79.0%	82.8%
2.5	2.5	79.8%	81.3%
3	3	80.5%	81.1%

Hadd an is Non York Tolaway Anchore Basir	Had	and a low of Tolard Robert Chairs
The is a fin was a generated and a second of the second and the second of the second and the second of the second and the seco	Zanita e Oslo wanten generati logen se urt de a te mel v oht kl.65 Ze m me an in Onkli om 7 e u	2001 L & Orly women growent ogen se unt de Tê to attich 65 an 1 - Tê to attich 65 Ze Mare an - Oakt o 7 a
Supplied Cash and grafian Stranfors. See First way Cash an Viel up Down for affordam.	and an insurgious Store Stores.	Hen Dars and A non Store ats.
Hidleson Morsten ze Nordend gulden (fine.) Betalen Neg Dezelfik Award Beden ze Trese to Anne de	Histories a stan ze I notered widen (frag.) Betalen Neg Deselfek Award heden se man der han de	Historyon o sen ze I nolen widen an.) Betalen Neg Dezelfek Avandae nan De her Ann
Cruis, Emile, Jan, Inine en Merte, Radaling Hun Interdision frethe an Utation Hadden Deput	Charles Emile, Jan, Jaine e en , Padal Han briendrianen ne en Karin ad - Opg (d. Samen media in Davia ad	Change Emil Jam I me en , Pa al Vin Imin in a ne en dia a opg (d. Samen add a cilitadad
Detations (Sugars and Rev Andrew (See) - Jane Liebensters (Sec) Vers was de 1906 Court (second)	D2d ame a gals of L, Bob d chord of an away heating ad (a). O Ress was de more bacand cand.	2d ani o gets L Bub d'chen o) en ange de hen est (o). O Reis was de 10, band vand.
facilit Beats on a book hour, Tona, Galler, Salan Salam at the salar blacks of security of spot Salar at darks in billing and da. In security of the salar salar spot thin Beats Salar Salar at securi	Benter an en 1973 Erre, Taria, Galler, Salar, Salare rectair - en directo/ Salare se la construcción la labor constale. In canada de la construcción de la construcción de la construcción	entran - 1974 1998, Tons, Gol - Salos i Door o the 1 Barto / 1986 - 1973 - 1976 - 1976 (1987 12 14/10 ² m-4 1. 1978 - 1989 - 1978 - 1978 - 1979 - 1979 - 1979 - 1989
fer von underskele server skie fannen in staft i verkeit fankensensen. Fansam hefelster i og andres i nær sjøler. Net stærete fanne skresser i nær sjøle (kosten disk sjøle, staftense En anså ogen sig varsefann egte medlikkringen glenskog	fer full souther derivers to the area of the souther the souther that the souther souther that the souther that the souther that the souther that the souther souther that the souther souther that the souther the souther the souther that the souther the sou	BETTER BUSHAN O BANI SARAT QLT B. HARMAN BATALT AND IT AN HISBOR BATALT AND IT AN HISBOR BATA BATA A BAN BAN AND A METAMAN A I
rener property and the second	9449 1971 1971 1970 1980 - 1984 65 2000 2003, medianangi Tali 19 (he 1940) Ti 19 (he 1960) 14 197 Ti 19 (he 1960) 14 197	эн унского го-ша 2002, метацу Тара Сери 1760 - Тара Гара 1760 - Тара Гара
$\Lambda\Lambda\Lambda$		
1000		
WHAT IS IN USA YOR, THAT COURSE, PARTS - 2016-24 BE THE LARGERINGET HOUSE IS NOT IN THE OWNER HET	WHAT TO IN YOU YOU, THE CAUSES, PARTS . 24424 BUCKE WE GUEST YOU TO S TRANSPORT	H ATTERN VEN YOUNTE D GLE - P OS 2000CH BH D N GNORT JNERNT S TRANSPORT
Vizer V Vizer 638. 34 Mary 15 and 2. 32 Azaran Ann ha bazarrega Tung Ingha Angkang an Grander 5 Mars. In Jan 1998 Baz Arha Jai an Unit Standar in Wards, and Kit Palaktares.	ULLAT 631 DH h. LUNC. Z' ar Anni lu OuiS- a Tha, in An a' Gran e S' nî Dr ⊤ Bug Gi e U I Shilo STimor in Ulik Illan (1713), Tanna	Nager book have to show a The in a give as state to be encide a state (the backet) to be
Autoria protecto, landorador de la protecta de la p	HE BETTER & HE BETTERNALDER (* 1977) He te croe fin he Zeitzich von Ubligzige CAR, (1977)C., 2 , Better all, Labor zur Unie von Terminische	ис этте 514 опорых (т. стърн. Историјана на забра на ибранија (т. стран. 3., г. наб. акулт заб. отни на трана а
Details (manuface) server (400.00 20" (1)manuface) (30) 2000 me Details of the PT . Bolt mensioners (34) an Olico Table Proceed for Details who be from while for process	Degrammer (Alberton, Sale of Baselon, 200 (Spinstandar) (Sul) Zelazawa Degrammer (Alberton, Sale terenameno (Sel in Oleno, Tallertenane) (Sul) Degrammer (Selazama) (Selazama)	мар (уль., занесі µ нозиї цункы в (за) холо разполаг на т. Кой. · е (зн. чОні Табра а чор бон) з ч. ге іноетія ч. бійде лт.
Λ/Λ		
NADd ar ju murpark Tolyro, anw Buly, Parijs, Zarid, CiriOslo, 1947 R. Dewezzel, Wolson Ze Will ar U.S.a-Fining METULOHT.	N Del sur je annykale. Tokyko, cerumPerlo, Prezis, Zuech, Cro Oslo, Maren Deureo - olasien s ur ras U.S.J. Ting Methodult.	N De the interview Toliyo, cereble, Perjs, Zoish, Enroslo, WAREN Dewes : Wesen is to the us . They Mitheleft.
22 WARNEY ARNA DURIN ON TOO YAR BU IN FRANCOC OR 40 Angla andre DURIN ON TOO YAR BU IN FRANCOC OR 40 Angla andre De FIRT VAN CAS. EXTRE VAN VAN	n n osor o izvo ize hula i davin Do Liar oz znov a Fari i tak Dari ogitar a skanda, a Fir Van ob Er Devr Van	22 Kulvi - I N'Du Zinjo Zonu R Bivi i tekany ogina nimodi, e Fi van du Si Dev Vin
lang Songhu yulu K3 VAN AETAndeeKTenzin. Hiph Van Moesik'n ze 🚥 kaylang guden (fino Belden. Nag Ozzellar fundal Recenced Vank Hyn ykensten.	Carr Dioni V je VAL 3 VAN ET value : Teac; Hi vo R noessi (v	Uni Dian no go Vilo 3 VIN ET ne En Teneri i Nor Moest-NU - a an Nober neo EN. Nog selfan a reactin ni Hiniri Enraen
CRUS LEHIL, Yege Yeen, Kin HERU, NAOKHERUN Kinadaan Goeta an Maaca, HEREN OP gathalio Santini Ukoevi ze Valtheudaro (2002 zuiliziane Palzanda	CNUS, EMIL, J., Jeene, Tor H. K. N. OZ Hon Universitient Gente to Marka H. Berr of Gelsallo . Sentin Ukozo ze. 14. Handren (son zeldzenne Polizerio).	CRUS, LHIL, Y., JOON, W. K., N. OR, UN LHIE HAV Garber CU Made: H. EV D. gerhaald Sarribin allen ze vig aa 190 (jero zel zonne Paszarko,
Gebookt. Boh 3000 Decembert God Eu Ouger There have Gara) De ban 3485 Decharite WARRS Decare CT	(Gelocht Bob) Riemonie So Eulour reholding (Geo) Qelaix AS DE TO : E WAA - WEEST	Gebookt. Boh Rie ander so Europer arkbakeg (2007) De kon AS De ta '6 WA - were CT
Radal as in New York, Tolyon, Anches, Ange 2 Strick on Orlo waven generated vlogen 20	Midrit ve in New York T. andree, Barro 2007 - an Osto ware over vlagen se	2 in the la New role T. as bee, Busie 2 in . en Orlo worke see. Jogen se
Ze lewanen aan in Ochin om 7000 en m Ansterlanden aan in Ochin om 7000 en m	ann un leang met o 64.6580 Ze Gwamen aan in Dublin am Zum en is Anslas Dam an is Dublin am Zum en is	and in leaving wet a let 650
White also Conto an Vill your David of the day	in Vale 23 von " when to von den	n Vale R3 un 7 andream and at and an
Bellen Morental de Stondand Gallen (2001) Bellen Dissifie (2001) geten sel Apol Hen Velable, C Califie d	Max ondered it in (frag.) Nog Deselfthe Award, in 20 - Wahlbert	Noe ondered in man.) Nog Dezelfek Award - 22 - Walkthel
Margarite, Salar Viene en Heads, Radal Se Han miteraturan Gata en Martin Hadden Gardweiter Samen Haden X. Sprinderd Martinger (1920)	briendrienen Geta en Merrige He - gebeurte Sanger Hingden & Officialent - (200)	brinn innen Greta en 1 in H - gebra 5 men den de aj o dand - gebra
en Owiel (tese hereford General), (tele dechenderd(tere) Marson Merekowski (General) Marson Merekowski (General) Marson Merekowski (General)	en Canife lives legiting general, the decondered (200) Mars with our product (200) Mars with de 1989 whether whethe	an Daniel Krache and Low Variation (m) Opperson de la basin variation X
have been as a liter base from first they also a literature person to be the second	Break and ATA (B)	ener un a 1970 10
F. Saran Rabins et al. (2007) The West in Function and the Construction of the Construction of the Function of the Construction of the Construction of the Construction of the Construction of the Construction of the Construction of the Construction of the Construction of the Construction of the Construction of the Construction of the Construc- tion of the Construction of the Construction of the Construction of the Construction of the Construction of the Construction of the Construc- tion of the Construction of the Cons	HARRING MUTTERS, INC. M. T. T. M. T. M.	HARDER MARTING C. M.C. 144 M.S.DELLING, F.W.O. J. KARTER, J. & A. BARTAN, ODB. THEI, JAN, HON, DA. J. L. S. CHARTER, GRAV, AND J. B. A. BARTAN, S. M. MARTER, MARTING, GRAV, N. W. J. B. 200, J. S. M. MARTER, M. S. MARTING, MARTIN
	reall al bell The bolt and a so so	nan fill a is s.
MAT NOT MAL MA	MAN NOV MAN	MAN NAN MAN MA
* * * * * * *	* * * * * * *	* * * * * * *
x <u> x x X X X X</u>	$X_{-} \times X_{-} \times X_{-$	$X \times X \times X \times X \times X$
WANT ZE HE HERE YOT TO BE COULDED, MADE - ZWEET BIT OF LARD STRUCT HOTE FOR HIT OF USA TONIG HAT	10/047 · · · · · · · · · · · · · · · · · · ·	த் AT - பல் கொட்டிப் அதை பலை பலை பல காழ முத்தா
Twe count account of the source of the sourc	The part of the pa	There is a standard and the standard and
BOG VIETNO DE MARTINELLA LAN ARTINA BOG VIETNO SOM OLDER DE VIEL HAR VIETNA (ANT, EMILE, Martin BOF BU VIEN, ADAT DE VIEL VIETNA (ANT, EMILE,	the brittle finds there is whe has all child of the print. Has there as the Has an are the shift	CONTRACT - TO ENDEDDEDIC S. PHILE. MALLINE DI DI MALUM 4
Sugar constant state barrow of the state of	Details 3 26 (Vpticologi) (So) 20120006 Detrocour gelocity . Bob transported Sol to Theme Jahresia Viewst	Detrocase a 7 Bot threades (1) 223- (1) 000 10 (1) (1) (1) (1) (1) (1) (1) (1) (1) (1)
* * * * * * *	* * * * * * * *	$\times \times \times \times \times \times \times$
x <u>x x x</u>	x <u>x x x</u> x x x	x <u> x x x</u> x x x
NA Del no ja margade Belge councility, Rogis, Zurch in Oslo, MARENT Deweiselt, Massin zu ant the USA Tenung Mittalacht. 11. 1520-2011 Town	NADd the for one refer to a start time period on Oslo, WA period and then a usin time period and the WA period and then a usin time period and the	N De ser je new o k. Tok (* ² , Rejs, Zo ok GN Oslo, 1921 - Dewestle Wolsen (* 6 US i Tom, Mittacht) 161 - 163 - Soure
26 LUMAMEN AANIN PORTIN OM 700 UVR EN IN ANGLARDAN OG 40 DIE WARDAN DE FAT VAN BYLE DIE VAN VAN Die on Einender Van Byle En Die Van Van	20 Loulanese Jani in on 700 ont En in Ansterber on to one woods. De Fiat van Bo, En De V. V.N. Dern	28 LIGHMEN NI W OM 700 LOR EN LY AMSTERDAY ONLY UNE NEEDED. DE FLAT VAN BE . SN DE V V N D con
HIT WORK MOST PUT 25 HER BUDGED GUDGE DESCRIPTION	Hire o ce E 6 Noveen uses from EV. Nog Dezeller word w 26 N A i-	R 0 OE 6 6 OUDED WHEN NO EN. Nog Oczaller curry 1126 V 1 i-
and an and the of several states and the second	Grade en Markia Altheris p gallisato Sanggy ze 0 C. Me is g	Gara and Maria and Alling Maria and Alling A
yennett. 1904 teo D iceihandari (soe) zii Olevin hueededa 2007 Octofi Westbe rentifi WAARD Geweest	y yerroc ob Orierbondian (soo) Eu Orierbondian (2001) (2011) (2013) WAS Die Hurther WAARD Geweesst	yernor of Onic and to Good Su Davin ectioning (200) De g. Wits the norther WAARD GENEEST
\times \times \times \times \times \times \times \times	$X \times X \times X \times X$	$\times \times \times \times \times \times \times \times$

Figure 7. Result of removing big scratched-out elements. Column 1: original overlaid with scratches, column 2: result with $\theta_s = 2.5$, $\theta_b = 2.5$, column 3: result with $\theta_s = 1.0$, $\theta_b = 1.5$.

Table 4. Verification results on 4386 thresholded half pages; 988 writers.

	TP	TN
Baseline	76.6%	84.0%
$\theta_s = 2, \theta_b = 2$	77.1%	83.6%

Table 5. Identification results on 362 thresholded half pages extracted from the first 250 pages of NFI dataset; 87 writers. $\theta_s = \theta_b \mid \text{Top-1} \quad \text{Top-10} \quad \text{Top-100}$

θ_s	θ_b	Top-1	Top-10	Top-10
1	1	85.4%	95.6%	98.9%
1	1.5	87.9%	96.1%	99.2%
1.5	1	86.7%	95.0%	99.2%
1.5	1.5	87.3%	96.1%	99.2%
1.5	2	88.4%	95.0%	99.2%
2	1.5	87.0%	96.1%	99.2%
2	2	88.1%	95.6%	99.2%
2.5	2.5	88.1%	95.0%	98.9%
3	3	87.9%	95.0%	99.2%

3.1 Results

The values $\theta_s = 2$ and $\theta_b = 2$ were used to remove crossed-out text in the test set (4386 half pages) of the NFI dataset. Table 4 shows the result together with the baseline performance. It is clear that automatically removing crossed-out text using the proposed method has no substantial influence on writer verification performance.

4. APPLICATION TO WRITER IDENTIFICATION

The same kind of experiment was performed to determine the effect of automatically removing crossed-out text on writer identification. Writer identification means returning a *hit list*, a sorted list of documents of which the handwriting is similar to that of a questioned document. In this experiment, similarity was again determined by the hinge feature and χ^2 -distance. The pages of the NFI dataset were split in parts, thresholded, and divided into a train set and a test set as described in section 3. The baseline performance was computed by treating every document in the test set as a questioned document, then yielding the hit list and finally counting how often a matching document appeared in the top-1, top-10 or top-100.

This was also done with pages of which crossed-out text was automatically removed using several values of θ_s and θ_b in the train set. The performance using these thresholds on the train set is shown in Table 5. Although the differences are again very small, the best selection of thresholds could be identified: $\theta_s = 1.5$ and $\theta_b = 2$.

4.1 Results

The thresholds $\theta_s = 1.5$ and $\theta_b = 2$ were applied to remove crossed-out text in the test set. On the resulting documents, writer identification was performed. The results are shown in Table 6, together with the baseline performance. The table shows that automatically removing crossed-out text does not improve writer identification performance.

5. CONCLUSION

In this paper a simple method to identify and remove crossed-out text was presented. It can remove 47% of crossedout text while 99% of the normal text is preserved. There is no important effect on writer verification or identification based on the hinge feature.¹² This is an indication that the effect of crossed-out text on writer verification and identification may be overestimated.

Table 6. Identification results on 4386 thresholded half pages; 988 writers.

	Top-1	Top-10	Top-100
Baseline	76.5%	88.1%	95.0%
$\theta_s = 1.5, \theta_b = 2$	75.5%	87.7%	94.8%

6. DISCUSSION

Although our result suggests that removing moderate crossed-out text may not be worth the effort, there are options to make make this statement more firm. It is conceivable that the Hinge feature that was used for the writer verification and identification experiment, is just quite robust for crossed-out text. Therefore, other features should be tried for this too, for example the Fraglets feature (also called fCO3).¹⁴ It is also possible that the automatic method to remove crossed-out words does improve verification or identification performance, but at the same time reduces performance because also some good text is removed. Therefore the next step should be to improve the method to detect crossed-out words. 3 One way to improve the method could be to use textural features such as Hinge on the level of connected components. Alternate thinning methods could be tried for the branching feature because artefacts in the skeleton have a big influence on the performance. The method could also be adapted to work with grayscale images, which would make the method more versatile. That would slightly improve writer verification and identification performance, since the hinge feature was designed for grayscale and performs a bit worse on black and white images. The best values for θ_s and θ_b could be determined in a more thorough way by using steepest descent or genetic algorithms. Furthermore, other classifiers can be tried, and line segmentation should be applied to disconnect connected components that are big because they consist of intersecting text from multiple text lines.

A final question that remains unanswered for now is how much text can be crossed-out without disturbing automatic verification or identification. It is hard to imagine that everything could be crossed-out for free, so we wonder how far one can go.

REFERENCES

- R. Plamondon, S. Srihari, E. Polytech, and Q. Montreal, "Online and off-line handwriting recognition: a comprehensive survey," *Pattern Analysis and Machine Intelligence*, *IEEE Transactions on* 22(1), pp. 63–84, 2000.
- U. Marti and H. Bunke, "A full english sentence database for off-line handwriting recognition," in Proc. of the 5th ICDAR, pp. 705–708, 1999.
- 3. L. Schomaker and L. Vuurpijl, "Forensic writer identification: A benchmark data set and a comparison of two systems," tech. rep., NICI, Nijmegen, 2000.
- 4. E. Lecolinet, L. Likforman-Sulem, L. Robert, F. Role, and J. Lebrave, "An integrated reading and editing environment for scholarly research on literary works and their handwritten sources," *Proceedings of the third* ACM conference on Digital libraries, pp. 144–151, 1998.
- A. Brink, L. Schomaker, and M. Bulacu, "Towards explainable writer verification and identification using vantage writers," in *ICDAR*, pp. 824–828, 2007.
- 6. J. Arlandis, J. C. Perez-Cortes, and J. Cano, "Rejection strategies and confidence measures for a k-nn classifier in an ocr task," in *ICPR*, 2002.
- 7. S. Nicolas, T. Paquet, and L. Heutte, "Markov random field models to extract the layout of complex handwritten documents," in *Proc. of the 10th IWFHR*, 2006.
- A. H. Toselli, A. Juan, and E. Vidal, "Spontaneous handwriting recognition and classification," in *ICPR*, pp. 433–436, 2004.
- N. Otsu, "A threshold selection method from gray-level histograms," IEEE trans. on Systems, Man and Cybernetics 9(1), pp. 62–66, 1979.
- L. Huang, G. Wan, and C. Liu, "An improved parallel thinning algorithm," in *ICDAR '03: Proceedings* of the Seventh International Conference on Document Analysis and Recognition, p. 780, IEEE Computer Society, (Washington, DC, USA), 2003.
- T. Joachims, "Making large-scale svm learning practical," in Advances in Kernel Methods Support Vector Learning, B. Schölkopf, C. Burges, and A. Smola, eds., MIT-Press, 1999.
- M. Bulacu and L. Schomaker, "Text-independent writer identification and verification using textural and allographic features," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)* 29(4), pp. 701– 717, 2007.

- L. Schomaker and M. Bulacu, "Automatic writer identification using connected-component contours and edge-based features of uppercase western script," *IEEE Trans. Pattern Anal. Mach. Intell.* 26(6), pp. 787– 798, 2004.
- L. Schomaker, M. Bulacu, and K. Franke, "Automatic writer identification using fragmented connectedcomponent contours," in *9th IWFHR*, F. Kimura and H. Fujisawa, eds., pp. 185–190, (Tokyo, Japan), October 2004.