TECHNIK

Traue keinem Scan, den du nicht selbst gefälscht hast

Durch einen Softwarefehler haben Scankopierer der Firma Xerox bei der Datenkompression Ziffern verändert – seit acht Jahren.

VON DAVID KRIESEL

Als mir im Sommer 2013 eine Firma ein merkwürdiges Problem vorlegte, dachte ich zunächst an einen Scherz. Man habe Baupläne mit Geräten des Marktführers Xerox gescannt und in den Scans veränderte Quadratmeterzahlen entdeckt; das Problem sei reproduzierbar. Der Fehler sei nur aufgefallen, weil ein mit 21 Quadratmetern ausgezeichneter Raum deutlich kleiner gewesen sei als der daneben liegende mit nur 14 Quadratmetern.

Es war aber kein Scherz. In den Bilddateien, die das Gerät aus den Papiervorlagen angefertigt hatte, waren einzelne Zahlen sauber – und darum völlig unauffällig – durch andere ersetzt worden. Auf einem Xerox WorkCentre 7535 lieferten mehrere aufeinanderfolgende Scans desselben Dokuments immer wieder dieselben falschen Quadratmeterzahlen. Ein leistungsfähigeres WorkCentre 7556 bot unter diesen Umständen sogar Abwechslung (Bilder unten). Ich konnte den Fehler auch mit deutlich besser lesbaren Ziffern nachvollziehen (Bild S. 21).

Als Verursacher stellte sich das Bildkompressionsverfahren JBIG2 heraus. Dieser speziell für Dokumente konzipierte Algorithmus kann ein Bild, insbesondere den Scan eines Dokuments, in viele Unterbilder zerlegen. Das ist hilfreich, weil die nachfolgende Datenkompression Texte und Graustufenbilder unterschiedlich behandelt und damit ein besseres Ergebnis erzielt.

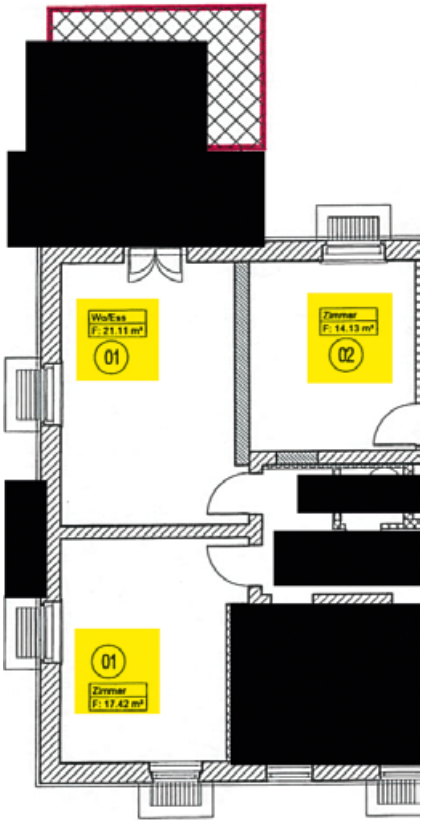
Mehr noch: Ein Unterbild kann so klein sein wie ein einzelner Buchstabe. Die Software vergleicht die Unterbilder untereinander und fasst besonders ähnliche von ihnen zu Gruppen zusammen (»pattern matching«). Eine solche Gruppe könnte beispielsweise einige hundert leicht verschiedene Versionen des Buchstabens e enthalten. Dann speichert die Software für diese Gruppe nur ein einziges e tatsächlich ab und verwendet es bei der Wiedergabe immer wieder an Stelle der anderen Gruppenmitglieder. So kann man große Datenmengen einsparen.

Gleichsetzung unterschiedlicher Zeichen

Unglücklicherweise war jedoch das Pattern Matching zu grob eingestellt. Dadurch wurden ganze Zeichenblöcke als gleich angesehen und füreinander eingesetzt, obwohl sie in Wirklichkeit verschieden waren. Das Ergebnis: perfekt aussehende Dokumente, die falsche Daten enthalten.

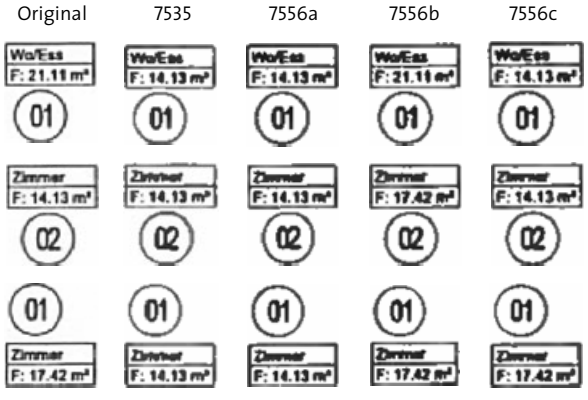
Die Scankopierer bieten verschiedene Qualitätsstufen an; je höher die gewünschte Bildqualität, desto größer der Platzbedarf des Dokuments. Allerdings heißt die Einstellung mit der niedrigsten Qualität »normal«; um zu merken, dass es – mit den Einstellungen »higher« und »highest« – auch besser geht, muss der Anwender das Benutzermenü oder das 300 Seiten starke Handbuch sorgfältig studieren.

Bei diesen höheren Qualitätsstufen finde kein Pattern Matching statt, so die erste öffentliche Reaktion von Xerox auf meinen Bericht. Tatsächlich konnte ich aber nachweisen, dass die interne Software der Scankopierer ebenfalls einen Fehler enthält, so dass – entgegen



ALLE ABBILDUNGEN: DAVID KRIESEL

In diesem Bauplan (links, teilweise geschwärzt) wurden die drei gelb markierten Stellen während des Scans verändert: Die Tabelle (rechts) zeigt sie im Original, vom Xerox WorkCentre 7535 gescannt sowie in drei verschiedenen Scans vom WorkCentre 7556.



den Absichten von Xerox – auch auf den höheren Qualitätsstufen Pattern Matching nicht abgeschaltet wird. Selbst ein ungewöhnlich sorgfältiger und qualitätsbewusster Anwender konnte also diese Art Verfälschungen nicht verhindern.

Kopierer, die auf derart unmerkliche Weise Zeichen vertauschen, sind eine potenziell lebensbedrohliche Gefahr. Man stelle sich nur vor, dass jemand auf Grundlage der falschen Zahlen eine Autobahnbrücke baut, ein Gerichtsverfahren führt oder gar die Dosierung eines Medikaments für ein ganzes Pflegeheim bestimmt.

Nachdem ich Xerox angesprochen und zunächst keine hilfreiche Antwort erhalten hatte, dokumentierte ich den Fehler auf Deutsch und Englisch in meinem Blog unter www.dkriesel.com/xerox und fand damit nicht nur eine halbe Million Leser, sondern auch Erwähnung in Hunderten von Massenmedien weltweit. Von Xerox gab es zunächst Dementis – aber dann Neufassungen der Software für die betroffenen Kopierer. Wie sich herausstellte, steckte der Fehler in Hunderttausenden von Geräten verschiedenster Baureihen (Tabelle rechts), und das seit mehr als acht Jahren.

Auf ein kleines Pixelbild wurden drei verbreitete Kompressionsverfahren angewandt. Während JPEG die Vorlage in 8 · 8 Pixel große Blöcke zerlegt und jeden Block mit Artefakten wiedergibt, neigt GIF dazu, ein Pixel mit zweifelhaftem Grauwert eher weiß als schwarz abzuspeichern. Demgegenüber sieht die JBIG2-Wiedergabe sehr sauber aus. Die gezeigten Fehler entstehen bereits bei Schriftgrößen von 7 oder 8 Punkt und Abtastraten von 200 oder 300 dpi.

Der Großteil der Scankopierer wird professionell von einer Vielzahl von Anwendern genutzt. Krankenhäuser, Finanzämter, Ingenieursbüros, Regierungs- und Forschungseinrichtungen scannen mittlerweile sämtliche eingehenden Schriftstücke an einer zentralen Poststelle und arbeiten nur noch mit den Scans weiter; das Original wird archiviert oder gleich entsorgt.

Risiken und irreparable Schäden

Es ist kaum abzuschätzen, welche Gefahren für Menschen oder Vermögenswerte durch acht Jahresproduktionen an möglicherweise subtil falschen Dokumenten entstanden sind – und welche Schäden irreparabel sind. Betroffen sind ja nicht nur die betreibenden Institutionen, sondern auch alle, die diesen ein Schriftstück geschickt haben und nun fürchten müssen, falsch verstanden worden zu sein.

Leiter von großen Archiven, die vor Jahren ihre Bestände mit Xeroxgeräten auf PDF-Dateien umgestellt hatten, haben mich gefragt, was sie tun können, um Fehler zu finden. Die Antwort ist wenig befriedigend: Man kann zwar feststellen, ob ein Scan mit einem betroffenen Gerät ausgeführt, nicht aber, was verändert wurde. Wenn die Archivare, wie oft üblich, die Originale vernichtet haben, bleibt ihnen nur, Jahresproduktionen an Dokumenten auf Plausibilität zu überprüfen. »Xerox kann die zahlendrehenden Scanner reparieren, aber nicht die veränderten Dokumente«, titelte später treffend die Wirtschaftszeitschrift »Bloomberg Businessweek«.

WorkCentre	232, 238, 245, 255, 265, 275, 5030, 5050, 51xx, 56xx, 57xx, 58xx, 6400, 7220, 7225, 75xx, 76xx, 77xx, 78xx
WorkCentre Bookmark	40, 55
WorkCentre Pro	232, 238, 245, 255, 265, 275
ColorQube	8700, 8900, 92xx, 93xx

Die von dem Softwarefehler betroffenen Geräte der Firma Xerox. Der Buchstabe x steht für eine beliebige Ziffer; hier geht es um ganze Gerätefamilien.

Xerox hat eine Korrekturdatei (einen »Patch«) für seine Software veröffentlicht; aber bis der Anwender sie installiert hat, enthält das Gerät den Fehler immer noch. Wer eines der in der Tabelle oben aufgeführten Geräte betreibt, sollte sich vergewissern, dass das Problem zumindest für die Zukunft behoben ist. Bei den Scans der letzten acht Jahre hilft nichts als eine gesunde Skepsis – wie gegenüber unserer Technikabhängigkeit insgesamt.

David Kriesel ist Diplominformatiker und arbeitet als Systemingenieur bei der IVU Traffic Technologies in Aachen. Unter dkriesel.com findet man seinen Blog zu technischen und satirischen Themen sowie sein zweisprachiges E-Book über neuronale Netze.

Eine längere Version dieses Artikels ist in den »Mitteilungen der Deutschen Mathematiker-Vereinigung« (Band 22, S. 30–34, 2014) erschienen.