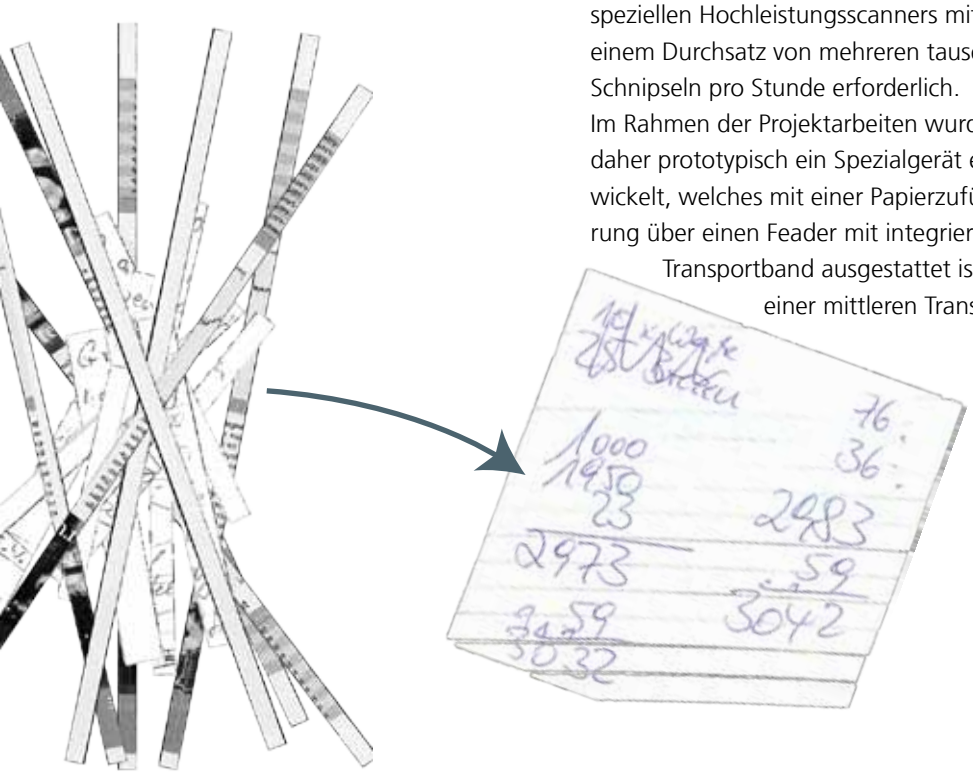
Automatisierte virtuelle Rekonstruktion vernichteter Dokumente

Das Fraunhofer IPK beschäftigt sich im Rahmen von Forschung und Entwicklung seit Mitte der 1990er Jahre mit der Digitalisierung und Rekonstruktion beschädigter und zerstörter Dokumente. Im April 2007 erhielt das Institut vom Beschaffungsamt des Bundesministeriums des Innern (BMI) den Forschungsauftrag, ein Verfahren und ein Pilotprojekt zu entwickeln, mit dem zerrissene Unterlagen des Ministeriums für Staatssicherheit (MfS) der ehemaligen DDR virtuell rekonstruiert werden können. In der auf vier Jahre angesetzten Pilotphase sollen 400 von mehr als 15 000 Säcken mit zerrissenen Dokumenten verarbeitet werden.



► Konzept des Systems

Im Rahmen der bisher erfolgten Arbeiten wurden zunächst alle Grundlagen zur schrittweisen Inbetriebnahme eines Systems zur virtuellen Rekonstruktion geschaffen. Das System umfasst folgende Komponenten: Digitalisierungs-Hardware und Serversystem für den Digitalisierungs-Workflow, Gridbasiertes Serversystem für den Rekonstruktions-Workflow, Rekonstruktionsmodule (ePuzzler), Softwareframework zur kontextabhängigen Ansteuerung der Rekonstruktionsmodule.

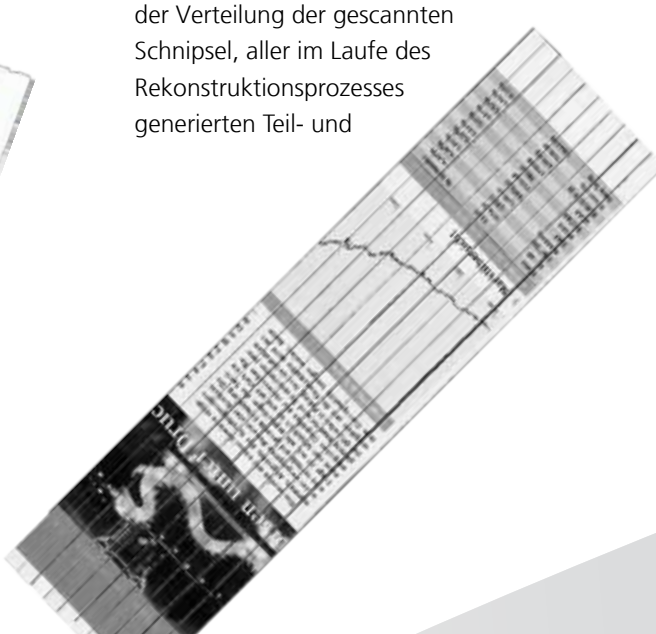
► Digitalisierung

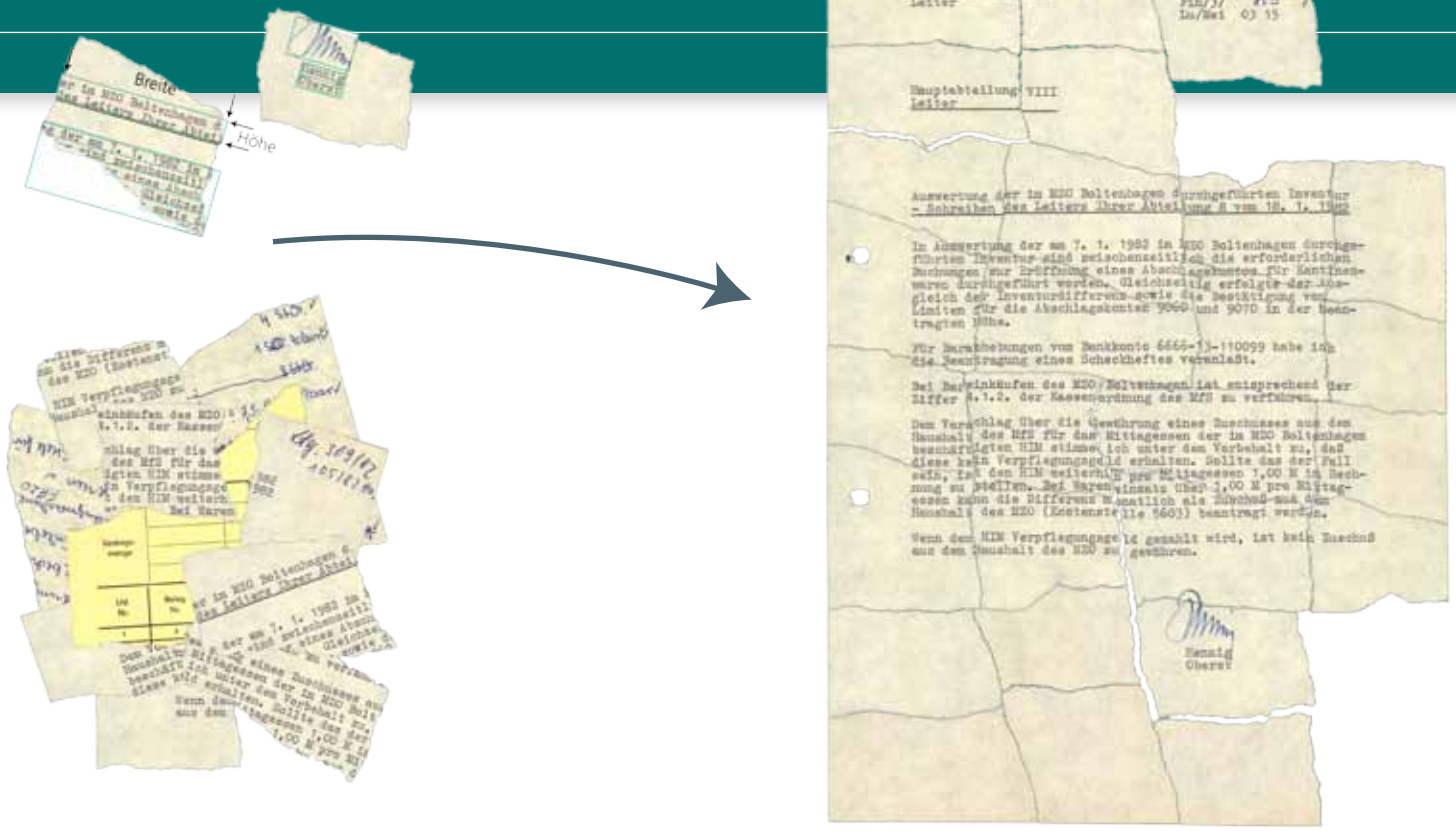
Vor der virtuellen Rekonstruktion müssen die Schnipsel digitalisiert werden. Für die Digitalisierung von mehreren Millionen Schnipseln ist die Verwendung eines speziellen Hochleistungsscanners mit einem Durchsatz von mehreren tausend Schnipseln pro Stunde erforderlich.
Im Rahmen der Projektarbeiten wurde daher prototypisch ein Spezialgerät entwickelt, welches mit einer Papierzuführung über einen Feader mit integriertem Transportband ausgestattet ist. Mit einer mittleren Transportgeschwindigkeit von 30 cm pro Sekunde werden dem Gerät die Papierschnipsel zugeführt, welches diese dann beidseitig mit 300 dpi digitalisiert. Ein farbiger Scanhintergrund ermöglicht analog zu Blue-Screen-Verfahren ein präzises und pixelgenaues Ausschneiden der Schnipselbilder.

► Gridbasiertes Serversystem für den Rekonstruktionsworkflow

Das Serversystem bildet die Basis der verteilten Anwendung des Rekonstruktionssystems. Die Hardware besteht aus einer hochparallelen, clusterbasierten Architektur verschiedener Prozessknoten, die einen jeweiligen Softwareteil des Gesamtworkflows ausführen. Die Serverhardware ist modular ausgelegt, um eine optimale Skalierbarkeit des Gesamtsystems zu gewährleisten, welche nach Abschluss der Pilotphase die Voraussetzung für eine möglichst weitgehende Beschleunigung des Gesamtprozesses durch weiteren Hardwareeinsatz ist.

Ein »Storage Area Network«, kurz SAN, ist der zentrale Ort der Speicherung und der Verteilung der gescannten Schnipsel, aller im Laufe des Rekonstruktionsprozesses generierten Teil- und





Vollrekonstruktionen sowie aller für die Auswertung der rekonstruierten Seiten benötigten Metainformationen. Das SAN ist in einen sehr schnell zugreifbaren Cachespeicher für flüchtige Rekonstruktionsdaten (derzeit 10TByte) und in einen langsameren Archivspeicher zur Ablage persistenter Daten aufgeteilt (derzeit 26TByte).

► Rekonstruktionsmodule (ePuzzler) und Softwareframework

Der ePuzzler ist ein vom Fraunhofer IPK entworfenes und stetig weiter entwickeltes System zur virtuellen Rekonstruktion von zerrissenen, geschredderten oder anderweitig beschädigten Dokumenten. Der ePuzzler gliedert sich in die drei Hauptmodule Merkmalsextraktion, Suchraumreduktion und Matcher, die auf komplexen Methoden der Mustererkennung und digitalen Bildverarbeitung basieren. Die Hauptmodule sind durch ein Softwareframework in einen nicht deterministischen und von der Beschaffenheit des zu verarbeitenden Materials abhängigen, adaptiven Workflow eingebettet.

Die Methodik der virtuellen Rekonstruktion ist vergleichbar mit der eines Menschen bei der Lösung eines Puzzles. Anhand einer Vielzahl von Merkmalen entscheidet dieser, ob zwei Teile zusammen passen oder nicht. Analog zur menschlichen Vorgehensweise werden daher zunächst vom Rekonstruktionssystem verschiedene Merkmale wie beispielsweise Kontur, Papierfarbe, Schrift oder Linierung aus den Schnipselbildern extrahiert. Aufgrund der sehr großen Datenmenge, die in diesem Projekt zu beherrschen ist, werden diese Merkmale genutzt, um den kombinatorischen Aufwand beim eigentlichen Puzzeln so weit wie möglich zu reduzieren. Dafür werden jeweils ähnliche Schnipsel mittels intelligenter Suchraumreduktion gruppiert, das heißt in einer Untermenge zusammengefasst. Innerhalb dieser reduzierten Mengen findet dann die eigentliche Rekonstruktion, das »Matchen«, statt. Dazu werden Schnipsel entlang ihrer Konturen auf Merkmalsübereinstimmung hin verglichen. Werden passende Schnipsel gefunden, so werden diese zu einem größeren Teil des Dokumentes zusammengefasst, es werden erneut die Merkmale des zusammengesetzten Stücks berechnet und dieses wird als neuer Schnipsel in der weiteren Rekonstruktion berücksichtigt. Das Bild oben zeigt exemplarisch einen auf Papierfarbe und Schriftmerkmalen basierenden Suchraum sowie eine entsprechende Teilrekonstruktion.

Ausblick: Rechnerbasierte Aktenformierung und -erschließung

Die stetig wachsende Zahl von Anfragen aus aller Welt zeigt, dass Bedarf besteht, die Potenziale der digitalen Bildverarbeitung und Mustererkennung auch für die Formierung von Akten und deren Erschließung zu nutzen. Da bei der Rekonstruktion der Seiten Informationen über Papierart und -farbe sowie Schriftbild und Ähnliches anfallen, können diese als Meta-Daten für die Formierung von Seiten zu Dokumenten und jene von Dokumenten zu Akten genutzt werden. Das Fraunhofer IPK erweitert zurzeit das Pilotsystem mit entsprechenden Systemkomponenten, um Interessenten ein integratives Gesamtsystem anbieten zu können.

Ihr Ansprechpartner
Dr. Bertram Nickolay
Tel.: ++49 (0) 30/390 06-2 01
bertram.nickolay@ipk.fraunhofer.de