E-Jigsaw

Computergestützte Rekonstruktion zerrissener Stasi-Unterlagen

Martin Marciniszyn · Angelika Steger
Andreas Weißl

Unmittelbar nach dem Fall der Berliner Mauer hat das Ministerium für Staatssicherheit der ehemaligen Deutschen Demokratischen Republik einen großen Teil seiner Akten vernichtet. Viele davon wurden zerrissen und lagern jetzt in der Behörde der Bundesbeauftragten für die Unterlagen des Staatssicherheitsdienstes (BStU, „Birthler-Behörde“) in etwa 17.000 Säcken à 20.000 Schnipsel. Seit Jahren bemüht man sich, diese Unterlagen von Hand wieder zusammenzusetzen, kommt dabei aber nur schleppend voran – etwa ein Jahr braucht eine Person für einen Sack. Aufgabe eines Studentenprojektes an der TU-München war es, durch die Entwicklung eines Prototypen die Realisierbarkeit der IT-gestützten Rekonstruktion dieser Akten nachzuweisen.

Im Herbst 2000 rief die Seniorautorin an der TU-München unter dem Namen E-Jigsaw ein Projekt ins Leben, das motivierten Studenten der Informatik nach dem Vordiplom eine attraktive Aufgabe bieten und eine echte Herausforderung mit einem spannenden Hintergrund sein sollte. Seitdem sind daraus über 20 Semester- und drei Diplomarbeiten hervorgegangen. Von den Studenten wurde ein Prototyp entwickelt, der anfangs etwa 20 Seiten rekonstruieren konnte und heute in der Lage ist, einen Sack mit 20.000 Schnipseln im Wesentlichen vollständig innerhalb von einer Woche zusammenzusetzen. Die beiden Juniorautoren promovieren heute in Angelika Stegers Gruppe am Institut für Theoretische Informatik der ETH-Zürich. Dieser Artikel stellt die technische Realisierung des Systems E-Jigsaw vor.

Lösungsansatz

Auf den ersten Blick ist die Rekonstruktion der Akten mit Rechnerunterstützung scheinbar einfach in den Griff zu bekommen. Erfahrungsgemäß kann man davon ausgehen, dass zusammengehörende Teile praktisch immer im selben Sack sind. Also beschränkt man sich auf einen davon und untersucht darin alle möglichen Paare von Schnipseln. So erhält man einen Algorithmus, dessen Laufzeit quadratisch mit der Anzahl von Schnipseln wächst, und da polynomielle Algorithmen im Allgemeinen als effizient gelten, ist diese Softwarelösung vermeintlich schnell und überdies einfach zu implementieren. Dass dieser offensichtliche Ansatz jedoch ungeeignet ist, zeigt eine Abschätzung der tatsächlich benötigten Rechenzeit unter realistischen Rahmenbedingungen. Das vollständige Ausprobieren aller Möglichkeiten nimmt selbst auf der heutigen Rechnergeneration bereits für einen einzigen Sack mehrere Jahre in Anspruch. Für eine IT-basierte Realisierung ist daher die Entwicklung einer effizienten Strategie entscheidend, mit der das Testen von möglichst vielen unnötigen Kombinationen vermieden werden kann.

Ein nahe liegender Ansatz ist die üblicherweise von Menschen angewandte Strategie beim Zusammensetzen eines Puzzles anzuwenden: als erstes versucht man alle Teile richtig zu orientieren, da-

DOI 10.1007/s00287-004-0395-8
© Springer-Verlag 2004

M. Marciniszyn · A. Steger · A. Weißl
Institut für Theoretische Informatik,
ETH Zürich,
8092 Zürich, Schweiz
E-Mail: mmarcini@inf.ethz.ch

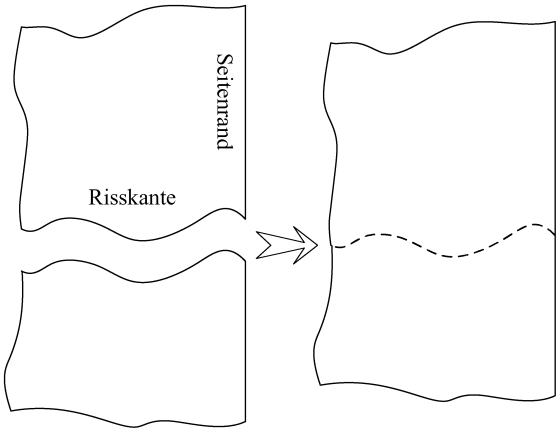

Abb. 1 Zusammenfügen von zwei Schnipseln nach dem Bottom-up-Prinzip

nach wird der Rand und am Ende die inneren Bereiche zusammengesetzt. Wieder zeigt eine Überschlagsrechnung, dass hierdurch zwar die Komplexität des Problems deutlich reduziert werden kann, aber immer noch Laufzeiten in der Größenordnung von mehreren Monaten pro Sack zu erwarten sind. Die Leistung eines solchen Systems wäre kaum höher als beim manuellen Zusammenfügen. Darüber hinaus setzt es implizit voraus, dass sich für zwei gegebene Schnipsel immer eindeutig entscheiden lässt, ob diese wirklich zusammengehören. Dies ist vor allem bei Teilen am Rand mit wenig Text und kurzen Risskanten oftmals unmöglich. Man muss die Schnipsel also in einer anderen Reihenfolge zusammensetzen.

Das von uns an der TUM entwickelte System E-Jigsaw setzt die Seiten nach dem Bottom-up-Prinzip zusammen, indem es den Ablauf des Zerreißens genau umkehrt. Diejenigen Schnipsel, die als letzte zerrissen wurden, fügt es zuerst aneinander. Dazu unterteilt es die Kontur der Schnipsel an den Eckpunkten in so genannte Segmente, d.h. eigenständige Risskanten. Zu jedem Segment, das durch den letzten Riss entstanden ist, gibt es genau ein passendes Gegenstück, das exakt dieselben geometrischen und zumindest ähnliche textliche Merkmale aufweist. Man baut ausschließlich diese Teile zusammen und fährt rekursiv fort. Die zusammengefügten Schnipsel ergeben ein neues Schnipsel, mit dem nach dem gleichen Prinzip verfahren werden kann. Abbildung 1 stellt diesen Vorgang schematisch dar: die beiden Originalteile haben jeweils drei Risskanten und einen Seitenrand. Auf dem neuen fallen zwei Segmente zusammen und die dazu benachbarten ergänzen sich gegenseitig zu längeren. Für die neuen Risskanten sucht man nach dem gleichen Prinzip das passende Gegenstück und kommt so zu immer größeren Teilen, bis man schließlich zu vollständigen Seiten gelangt.

Systemrealisierung

Das System besteht aus einem Verbund von mehreren Linux PCs, wie in Abb. 2 schematisch dargestellt ist. Der Rekonstruktionsprozess wird zentral durch den so genannten HypoManager gesteuert, dessen wichtigste Funktion das Aufstellen von Hypothesen und die Entscheidung über deren Korrektheit ist. Der HypoManager wurde wegen der hohen Leistungsanforderungen in C++ implementiert. Neue Hypothesen versendet er über CORBA in Paketen an den so genannten JBroker, damit sie parallel bewertet werden.

Der JBroker verwaltet die Rechenkapazitäten des Systems und synchronisiert parallele Abläufe. Er überwacht und steuert die so genannten Worker, deren Hauptaufgabe die Bewertung von Hypothesen ist. Die Auftragspakete des HypoManagers werden vom JBroker gleichmäßig aufgespalten und auf rechenbereite Maschinen (Worker) verteilt. Die Ergebnisse vollständig bearbeiteter Pakete reichen diese an den HypoManager zurück. Die Kommunikation zwischen dem JBroker und dem HypoManager erfolgt asynchron. Implementiert wurde der JBroker in Java, da Mechanismen zur Synchronisation nebenläufiger Prozesse dort zum Sprachschatz gehören. Auf seiner Bedienoberfläche werden dem Benutzer globale Systemparameter angezeigt, wie beispielsweise der Zustand der Worker, die Anzahl noch zu evaluierender Hypothesen und die Anzahl der fertigen Seiten. Durch Veränderung bestimmter Schwellwerte kann man zudem den Rekonstruktionsprozess steuern.

Neben der Hypothesenevaluation bewerkstelligen die Worker die Bildanalyse und das Zusammenfügen zweier Schnipsel. Wie der HypoManager wurden sie aus Leistungsgründen in C++ realisiert, CORBA dient als Plattform für die Kommunikation. Alle benötigten Informationen beziehen sie aus einer zentralen PostgreSQL-Datenbank, die als Open Source zur Verfügung steht und mit den richtigen Optimierungen gute Dienste geleistet hat. Zugriff auf die Bilder der Schnipsel erfolgt über ein verteiltes Dateisystem basierend auf NFS, in das

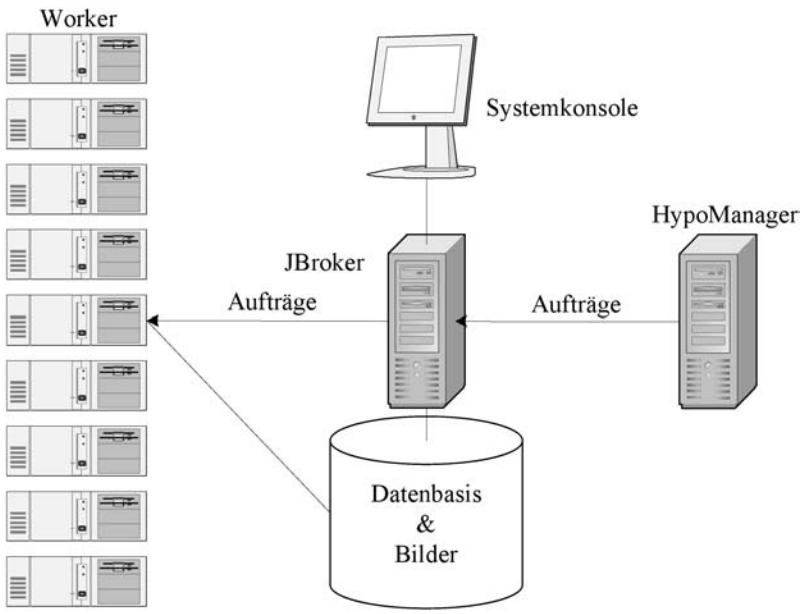

Abb. 2 Physische Verteilung des Systems

zusätzliche Mechanismen zur Replikation integriert wurden.

Scannen

Eine verlässliche Schrifterkennung setzt voraus, dass die Unterlagen mit einer Auflösung von mindestens 300 dpi digitalisiert werden. Bei der Entwicklung des Prototypen E-Jigsaw wurde zunächst ein handelsüblicher Flachbettscanner verwendet. Ein Testdatensatz von 50 Seiten ist einfach zu erzeugen und kann mit diesem Gerät bei einem ungefähren Durchsatz von einem Schnipsel pro Minute in verhältnismäßig kurzer Zeit eingescannt werden.

Analyse und Segmentklassifikation

Die Bildanalyse umfasst die Segmentierung des Schnipsels und die Extraktion verschiedener Merkmale. Abbildung 3 stellt einige von diesen dar. Man erkennt die Kontur des Schnipsels, die von vier Eckpunkten in zwei so genannte Papier- und zwei Risskanten unterteilt wird. Zur Berechnung der Kontur wird ein verhältnismäßig einfaches Schwellwertverfahren angewandt, das den Bildhintergrund von der Region des Schnipsels trennt. Eine robuste Lokalisation der Konturecken ist ausschlaggebend für die Realisierung der Bottom-up-Strategie, da sich Übereinstimmungen zwischen zwei Schnipseln immer auf einzelne Risskanten beziehen. Beispielsweise werden von zusammengehörenden Segmenten gleiche geometrische Eigenschaften wie eine übereinstimmende Länge gefordert. Eine verschobene, fehlende oder überzählige Ecke zerstört die Annahmen der Strategie. Die Implementierung der Eckenerkennung basiert auf dem bekannten Algorithmus von Ramer, der eine Kontur polygonal approximiert. Da die Wahl des Anfangspunktes für dieses iterative Verfahren kritisch ist, wird die erste Ecke bei Vorhandensein von Papierkanten mithilfe der Hough-Transformation bestimmt, mit der sich gerade Konturabschnitte extrahieren lassen. Es passiert leider häufig, dass der Ramer-Algorithmus wegen des krummen Verlaufs der Risskanten zu viele Ecken feststellt. Das System filtert die überschüssigen Ecken mit der approximierten zweiten Ableitung, d.h. der Krümmung heraus. Details zu diesen Algorithmen finden sich in Standardlehrbüchern über rechnergestütztes Bildverstehen (u.a. [1, 2, 3]), im Begleitmaterial zur Vorlesung „Methoden der industriellen Bildverarbeitung“ von Carsten Steger sowie in den Diplomarbeiten [4, 5, 6].

Im nächsten Schritt bestimmt die Analyse die Textregion und richtet Zeilenmuster horizontal aus. Dann erkennt sie die Schrift und vermisst verschiedene Attribute der Zeilen wie deren Lage, Ausdehnung und Höhe der Grund- und Zentrallinie. Mit dem Verhältnis von Ober- und Unterlängen der einzelnen Buchstaben lässt sich verhindern, dass der Text nach der Orientierung auf dem Kopf steht. Da die Ausrichtung anhand der Zeilen um bis zu ±3° vom Optimum abweicht, wird sie durch Einbeziehen von Papierkanten korrigiert, die immer exakt horizontal bzw. vertikal verlaufen müssen. Damit lassen sich die meisten Schnipsel mit einer Toleranz von weniger als 1° ausrichten, sofern genügend Text darauf vorhanden ist.

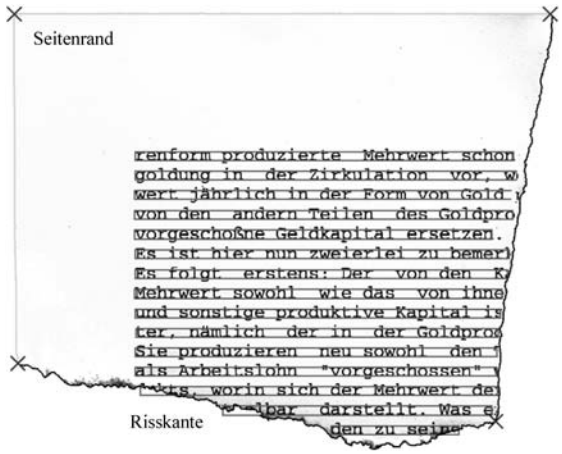

Abb. 3 Schnipsel nach der Analyse

Zuletzt werden alle Risskanten durch mehrere Merkmale charakterisiert, die zur Auswahl potentieller Gegenstücke nach der Bottom-up-Strategie dienen. Dazu zählen unter anderem

- die Länge eines Segmentes,
- der Typ dazu benachbarter Segmente,
- eine robuste Ausgleichsgerade nach Tukey,
- der Winkel dieser Geraden zum Seitenrand,
- die Streuung der Segmentpixel um die Ausgleichsgerade,
- die Anzahl an dem Segment durchtrennter Zeilen,
- die Farbe des Papiers.

Bei der Implementierung der Bildanalyse konnten wir auf den reichen Vorrat an Bildverarbeitungsoperatoren der Software-Bibliothek HALCON zurückgreifen, die uns die Firma MVTec freundlicherweise kostenlos zur Verfügung gestellt hat.

Aufstellen von Hypothesen

Der HypoManager stellt Hypothesen für jedes verfügbare Schnipsel einzeln auf. Lose Schnipsel werden als aktiv bezeichnet und in einer Ringliste verwaltet, die vom HypoManager zyklisch durchlaufen wird. Für jedes Schnipsel betrachtet dieser nacheinander die Risskanten und sucht passende Gegenstücke nach den oben aufgezählten Kriterien. Er vergleicht jeweils die Merkmalsvektoren der zwei Segmente und entscheidet regelbasiert, ob eine Paarung sinnvoll und daher eingehender zu bewerten ist. Die Regeln besagen z.B., dass die Längen der Segmente übereinstimmen und benachbarte Papierkanten wiederum durch Papierkanten fortgesetzt werden müssen. Die Winkel zwischen der Ausgleichsgeraden und der Horizontalen sollten etwa gleich sein, die zwischen der Ausgleichsgeraden und dem Seitenrand sollten sich wechselseitig zu 180° ergänzen. Die Streuung der Segmentpixel um die Ausgleichgerade quantifiziert die Krümmung der Risskante und muss jeweils gleich ausfallen. Kritisch ist beispielsweise die Anzahl der an einem bestimmten Segment durchtrennten Zeilen, da hier viel Spielraum notwendig ist. Man erwartet zwar, dass sich die meisten der Zeilen auf der anderen Seite fortsetzen, gewisse Schwankungen sind jedoch wegen Absatzumbrüchen erlaubt. Im Extremfall hören sogar alle Zeilen genau am Riss auf. Die Papierfarbe hingegen ist ein Kriterium mit wenig Toleranz.

Die vorangegangene Diskussion über die Anzahl durchtrennter Zeilen lässt erkennen, dass die Festlegung der Übereinstimmungskriterien nicht starr sein darf. Anfangs ist eine möglichst gezielte Suche nach dem Gegenstück zur Einschränkung der Rechenzeit zweifellos sinnvoll, doch müssen die Kriterien mit der Zeit schrittweise abgeschwächt oder sogar völlig deaktiviert werden, damit die richtige Paarung zustande kommt. In E-Jigsaw wurde dafür ein Rundenkonzept entworfen, dass die Toleranz der Merkmale eines Schnipsels mit jedem vollständigen Durchlauf über die Ringliste sukzessive erhöht. Damit wird sichergestellt, dass meist wenige Hypothesen pro Teil zu evaluieren sind und lediglich in Problemfällen unspezifischere und somit mehr Paarungen zugelassen werden.

Die Auswertung der Regeln für die charakteristischen Merkmale wird durch einen zweistufigen Trie unterstützt, der alle Segmente nach ihrer Länge und nach dem Winkel ihrer Ausgleichsgeraden zur Horizontalen vorklassifiziert. Die übrigen Kriterien werden innerhalb einer solchen Klasse für alle möglichen Tupel überprüft.

Bewerten von Hypothesen

Für jede aufgestellte Hypothese berechnet das System einen so genannten Score zwischen 0 und 1, der die Übereinstimmung von zwei Schnipseln an vorgegebenen Risskanten quantifiziert. Anhand des Rissverlaufs und der Eckpunkte wird das eine Schnipsel möglichst nahtlos mit dem anderen verbunden, d.h. ohne die Entstehung starker Überlappungen oder großer Löcher. Dies erfordert zwei Schritte: erst richtet man die Eckpunkte und Papierkanten aufeinander aus und optimiert dann die

Position durch Verschieben und Drehen innerhalb kleiner Bereiche.

Es werden mehrere Scores berechnet, die unterschiedliche Übereinstimmungskriterien widerspiegeln. Die geometrische Bewertung stützt sich auf den Verlauf beider Risse, indem die Entfernung zwischen diesen jeweils punktweise gemessen wird. Die Abstände sollten immer möglichst klein sein, da sonst Löcher oder überlappende Bereiche vorhanden sind. Der textbasierte Score quantifiziert die Übereinstimmung der Zeilenfragmente, denn man erwartet, dass sie sich über den Riss hinweg gegenseitig ergänzen. Da diese Annahme nicht in allen Fällen zutrifft, ist das Aufstellen einer geeigneten robusten Bewertungsfunktion für dieses Merkmal nicht einfach. Grund dafür sind Absatzumbrüche oder eine unregelmäßige Absatzstruktur.

Der Engpass der Hypothesenbewertung ist die zentrale Datenbank: sie konnte die parallelen Anfragen der bis zu 25 Worker trotz eines schnellen Serverrechners mit 2 GB Hauptspeicher nicht in angemessener Zeit beantworten. Auf lange Sicht muss man die Möglichkeiten der Datenreplikation in Erwägung ziehen.

Nach der Hypothesenbewertung steht zum einen die Transformation fest, um zwei Schnipsel möglichst optimal zu verbinden, und zum anderen die Scores der geometrischen und zeilenbasierten Übereinstimmung.

Auswahl zusammengehörender Paare

Ist die Evaluation aller Hypothesen zu einem Schnipsel abgeschlossen, wählt der HypoManager die richtige aus diesen aus. Die korrekte Paarung sollte sich durch einen besonders hohen Score von den anderen abheben. Man muss die Option vorsehen, dass eine Entscheidung im Zweifelsfall interaktiv durch den Systemoperator getroffen wird, falls mehrere aussichtsreiche Hypothesen mit ähnlicher Bewertung entstanden sind. Erfahrungsgemäß hängt die Auswahl von einigen Schwellwertparametern ab, die sich mit jeder Runde ändern, die das Schnipsel durch die Hypothesengenerierung absolviert. Auch die Formel zur Berechnung des Gesamtscores aus dem geometrischen und dem zeilenbasierten wird sukzessive angepasst, damit ungewöhnliche Absatzstrukturen nicht irrtümlich zur Ablehnung einer Anordnung führen.

Das Fehlen der richtigen Hypothese in einem Paket kann viele Gründe haben. Untypische Zeilenmuster sowie ausgefranste Risskanten beispielsweise können die Bewertung negativ beeinflussen. Es kann auch vorkommen, dass zu dem Zeitpunkt der Hypothesengenerierung das passende Gegenstück noch nicht wiederhergestellt war, was durch parallele Abläufe nochmals komplizierter wird. Dann gibt es viele Möglichkeiten für den nächsten Rekonstruktionsschritt, angefangen bei der Anpassung der Toleranzen für die Hypothesengenerierung bis hin zum Absenken der Schwellwerte für die Auswahl richtiger Paare. Hier gibt es viel Spielraum für Optimierungen.

Digitales Zusammenfügen

Als korrekt eingestufte Hypothesen werden vom HypoManager in einem so genannten Merge-Auftrag an den JBroker geschickt und von dort an freie Worker zur Ausführung weitergeleitet. Aus zwei Originalschnipseln entsteht ein virtuelles neues, für das jeweils ein Bild von der Vorder- und der Rückseite konstruiert wird. Danach bestimmt das System die Merkmale des neuen Objektes wie beispielsweise dessen Kontur und legt sie in der Datenbank ab.

Das Vorgehen beim Zusammenfügen weist eine große Ähnlichkeit zur Analyse von Schnipseln auf, doch extrahieren wir die Objektattribute diesmal nicht aus Bilddaten, sondern berechnen sie aus den Attributen der ursprünglichen Teile. Zum Beispiel lokalisieren wir Ecken nicht erneut, sondern übernehmen deren Position aus den alten Koordinaten. Die Ecken an der zusammengefügten Riss-Stelle verschwinden. Von da ab stellt sich das zusammengesetzte Objekt für das System wie ein eingescanntes, analysiertes Originalteil dar und wird wie ein solches behandelt.

Bei der Auswahl von richtigen Paaren sind Fehlentscheidungen bis zu einem gewissen Grad unvermeidbar. In einer Menge von 20.000 Schnipseln finden sich viele Paare mit ähnlichen Risskonturen, insbesondere treten annähernd gerade Risse im rechten Winkel zum Seitenrand häufig auf. Da sich dort meist wenig Schrift befindet, können manchmal selbst Menschen nicht mit eindeutiger Sicherheit sagen, ob zwei kleine Stücke zusammengehören. Deshalb entsteht aus zwei falschen Einzelteilen selten ein verkehrtes neues Schnipsel. Dieses fällt später im Rekonstruktionsprozess auf, da kein geeigneter Partner dazu gefunden wird. Solche Problemfälle werden nach erfolg-

loser Suche dem Systemoperator präsentiert, der dann den Befehl zur Wiederauftrennung geben kann.

Der Härtetest

Als vorläufigen Abschluss des Projektes sollte E-Jigsaw seine Leistungsfähigkeit an einem Sack selbst erzeugter, zerrissener Unterlagen unter Beweis stellen. 1700 weiße DIN-A4-Seiten wurden beidseitig mit verschiedenen englischen und deutschen Texten bedruckt und jeweils in 10 bis 20 Teile gerissen. Insgesamt entstanden 20.709 Schnipsel, d.h. durchschnittlich 12,3 Teile pro Seite, die mit freundlicher Unterstützung des Münchener Scandienstleisters MFM innerhalb einer Woche digital erfasst werden konnten. Für das Einscannen wurde eine Kombination aus einem Gerät von Kodak mit Bandtransport und einem schnellen A3-Flachbettscanner der Firma Lanier gewählt, mit dem etwa 20% der Teile, die für den automatischen Einzug zu klein waren, digitalisiert wurden. Insgesamt waren fünf Tage zum Einscannen notwendig, und einen Tag etwa dauerte der Import der Bilder in die Datenbank von E-Jigsaw. Einen weiteren Tag benötigte deren Analyse. Vor der eigentlichen Rekonstruktion galt es dann noch einige Anlaufschwierigkeiten zu bewältigen.

Schwierigkeiten und Lösungen

Die Umstellung von dem langsamen Flachbettscanner auf die Hochleistungsgeräte bereitete anfangs einige Sorgen. Zuvor hatten wir beispielsweise beim Einscannen einen farbigen Hintergrund verwendet, um die Segmentierung des Schnipsels zu vereinfachen. Dieser Farbwert war entweder bekannt oder wurde durch Stichproben aus dem Randbereich des Bildes bestimmt. Er ließ sich zur Selektion der Hintergrundregion ausnutzen, die wir lediglich invertieren mussten, um die Region des Schnipsels zu erhalten.

Der Hochleistungsscanner mit automatischem Bandeinzug erlaubte jedoch – zumindest ohne gravierende Veränderungen – ausschließlich einen schwarzen Hintergrund. Da sich dieser nicht von der Druckerschwärze der Schrift unterscheiden lässt, werden Buchstabenfragmente bei der Selektion der Farbe Schwarz dem Hintergrund zugeordnet, und die Region des Schnipsels zerfällt unter Umständen in mehrere Komponenten. Üblicherweise erhält man eine große Hauptkomponente, um deren Rand herum ein paar kleinere liegen, die

Abb. 4 Probleme der Konturerkennung mit schwarzem Hintergrund: das linke Bild zeigt die erste Version der Kontur, das rechte ihren gefilterten Verlauf

ebenso wichtig sind, da die Hauptkomponente allein den wirklichen Umriss des Schnipsels unzureichend wiedergibt. Da verfälschte geometrische Eigenschaften des Rissverlaufs sehr negative Auswirkungen auf die Bewertung der Hypothesen hätten, müssen die kleineren Zusammenhangskomponenten geeignet mit der großen verbunden werden.

Abbildung 4 illustriert zwei Probleme bei der Segmentierung des Schnipsels mit schwarzem Hintergrund. An dem Bruchstück des Buchstabens „u“ tritt die beschriebene Abspaltung kleiner Bereiche auf. Die Kontur macht hier einen Bogen nach innen, obwohl sie eigentlich geradeaus verlaufen sollte. Bei dem „m“ entstehen schmale Einbuchtungen, die jedoch keine weißen Bereiche vollständig abtrennen. Da die Geometrie der Risskanten bei der Rekonstruktion eine wesentliche Rolle spielt, wurden einige Filter zur Verbesserung der Konturen implementiert. Die rechte Abbildung deutet den approximierten Verlauf der Kontur an.

Ein Flachbettscanner hat weiterhin den Vorteil, dass sich die Schnipsel während des Abtastens in Ruhe befinden. Bei Hochleistungsscannern mit Bandtransport werden sie hingegen bewegt, und es kann leicht passieren, dass sich die kleinen Teile währenddessen wegen unzureichender Führung verdrehen. Dies hat geometrische Verzerrungen auf den Bildern zur Folge. Diese Maschinen ermöglichen jedoch einen hohen Durchsatz von mehr als einem Teil pro Sekunde und sind in der Lage, beide Seiten in einem Durchgang im Duplexbetrieb zu erfassen. Bei den Flachbettscannern müssen die Schnipsel zwischendurch manuell umgedreht werden.

Es gab auch einige prinzipielle Schwierigkeiten, die sich nicht beheben ließen. Sie sind für einen gewissen Anteil der Seiten verantwortlich, die das System nicht rekonstruieren konnte. Auf manchen

Bildern wurden zum Beispiel Ecken des Schnipsels abgeschnitten, weil der Scanner mit Bandtransport den Papieranfang nicht richtig registriert hat. Außerdem sind einige Teile zwischen den Transportbändern während des Ablichtens verrutscht, sodass sie stark verzerrt wurden. Nicht unerwähnt bleiben soll, dass wir beim Erzeugen der Testdaten das System ebenfalls in Bedrängnis gebracht haben, indem wir manche Seiten in mehr als 20 Teile zerrissen haben.

Ergebnisse

Das System konnte 1502 der 1700 Seiten vollständig rekonstruieren und erzielte somit eine Erfolgsquote von 88,4%. Dafür benötigte es insgesamt etwa eine Woche unterbrechungsfreie Rechenzeit. Wegen der prototypischen Implementierung wurde der Prozess allerdings mehrfach angehalten, um beispielsweise die Strategie zum Aufstellen neuer Hypothesen zu verfeinern. Dieser umfangreiche Testfall brachte auch eine Reihe von Leistungsengpässen und Fehlern der Software zutage, die dann behoben worden sind. Stark betroffen war davon beispielsweise das Datenbankzugriffs-Modul, denn viele SQL-Anfragen, die bei wenigen Datensätzen keine nennenswerten Leistungseinbußen bedeutet haben, mussten unter den gegebenen Umständen neu formuliert werden.

Die 198 nicht vollständig rekonstruierten Seiten ließen sich auf 641 Einzelteile reduzieren, d.h. im Durchschnitt 3,2 Schnipsel pro Seite im Vergleich zu 12,2 zu Beginn. Diese Seiten waren nahezu fertig, doch verhinderte oft beispielsweise einverzerrtes oder fehlendes Bild deren vollständige Zusammensetzung. Einige der Fehlerquellen liegen also noch beim Einscannen, doch auch bei der Software gibt es Verbesserungspotential. Probleme bereiten beispielsweise Handschrift und Graphiken, da die Ausrichtung solcher Schnipsel sowie die Erkennung von Absatzstrukturen darin schwieriger ist. Ausbaufähig ist auch die Spezifität des Klassifaktors, der richtige von falschen Hypothesen trennt.

Ausblick

Im Sommer 2001 haben wir die BStU über den ersten funktionierenden Prototypen unterrichtet. Im Herbst desselben Jahres schrieb die Behörde eine Machbarkeitsstudie über die computergestützte Rekonstruktion der Unterlagen europaweit aus. Den Zuschlag zur Durchführung dieser Studie bekam leider das Berliner Fraunhofer-Institut für Produktionsanlagen und Konstruktionstechnik IPK gemeinsam mit der Gesellschaft für beleglose Dokumentenbearbeitung (GbD). Einem studentischen Team wurde die notwendige Kompetenz anscheinend nicht zugetraut. Am 20. Mai 2003 haben wir das Projekt E-Jigsaw am Institut für Informatik der TU-München der Öffentlichkeit vorgestellt. Es wurden der Hintergrund des Projektes, dessen technische Umsetzung sowie die Ergebnisse präsentiert. Dies löste ein breites Echo in der Presse aus.

Die Machbarkeitsstudie wurde im Herbst 2003 abgeschlossen, und als Ergebnis gab die BStU in ihren Pressemitteilungen bekannt, die Rekonstruktion und Grobsichtung der Unterlagen aus 600 Millionen Schnipseln sei innerhalb von fünf Jahren möglich. Dazu benötige die Behörde pro Haushaltsjahr eine einstellige Millionensumme zusätzlich. In anderen Presseberichten kursierte auch die Summe von 60 Millionen EUR. Konkrete Zahlen über die Dauer und die Erfolgsquote einer etwaigen Rekonstruktion von einem Sack im Rahmen dieser Studie nennen weder die BStU noch das Fraunhofer-Institut, das eigens zu diesem Projekt einen Film produziert hat. Nun liegt es in der Hand des Deutschen Bundestages, über den Fortgang des Vorhabens zu entscheiden.

Weitere Informationen unter http://www.e-jigsaw.de und in den Diplomarbeiten der Juniorautoren.

Literatur

1. da Fontoura Costa, L.; Marcondes Cesar, R.: Shape analysis and classification. Theory and Practice. CRC, 2001
2. Haralick, R.M.; Shapiro, L.G.: Computer and robot vision. Addison-Wesley, 1992
3. Jähne, B.: Digital image processing. 5th revised and extended edn. Berlin Heidelberg New York Tokio, Springer 2002
4. Marciniszyn, M.: Algorithmen zur Rekonstruktion zerrissener Seiten: Konstruktion und Bewertung von Hypothesen. Diplomarbeit, Institut für Informatik, Technische Universität München, 2002
5. Micklitz, S.: Algorithmen zur Rekonstruktion zerrissener Seiten: Anwendungen optischer Zeichenerkennung auf qualitativ schlechten und unvollständigen Eingabedaten – Segmentierung und Klassifikation. Diplomarbeit, Institut für Informatik, Technische Universität München, 2002
6. Weißl, A.: Algorithmen zur Rekonstruktion zerrissener Seiten: Anwendungen optischer Zeichenerkennung auf qualitativ schlechten und unvollständigen Eingabedaten – Ausrichtung und Vervollständigung. Diplomarbeit, Institut für Informatik, Technische Universität München, 2002