Umanistica Digitale - ISSN:2532-8816 - n.3, 2018
N. Barbuti, S. Ferilli, T. Caldarola – Un innovativo Graphic Matching System per la ricerca in database di manoscritti antichi
DOI: http://doi.org/10.6092/issn.2532-8816/8144

Un innovativo Graphic Matching System per la ricerca in database di manoscritti antichi

[1]Nicola Barbuti, [1]Stefano Ferilli e [2]Tommaso Caldarola

[1]Università di Bari "Aldo Moro"

[2]D.A.BI.MUS S.r.l.

[1]nicola.barbuti@uniba.it
[1]stefano.ferilli@uniba.it
[2]t.caldarola@dabimus.com

Abstract. The paper outlines a pattern recognition process, which uses a *graphic matching* algorithm based on a *shape contour recognition* function without the need to apply segmentation. The process starts from the identification of a Region of Interest (ROI) within the image. The ROI is then managed for the creation of the *shape model* then used to perform searches for similar models in one or more target images. The system has been developed and tested with the aim of proposing an innovative approach in the research and retrieval of information in digital libraries related to ancient manuscript and printed documentation. This approach is based on the application to the data humanities of the fourth knowledge paradigm that underlies data science. Following this approach, the algorithm is used to deduce new research hypotheses through the discovery of models directly inferred from large digital libraries.

Il paper descrive un processo di pattern recognition, che utilizza un algoritmo di *graphic matching* basato sulla funzione *shape contour recognition* senza necessità di applicare funzioni di segmentazione. Il processo si avvia partendo dall'individuazione di una Region of Interest (ROI) all'interno dell'immagine. La ROI è quindi gestita per la creazione del *shape model* poi utilizzato per eseguire le ricerche di modelli simili in una o più immagini target. Il sistema è stato sviluppato e testato con l'obiettivo di proporre un approccio innovativo nella ricerca e recupero di informazioni in digital libraries relative a documentazione antica manoscritta e a stampa. Questo approccio si basa sull'applicazione ai *data humanities* del *quarto paradigma* della conoscenza che è alla base dei data science. Seguendo questo approccio, l'algoritmo si utilizza per dedurre nuove ipotesi di ricerca attraverso la scoperta di modelli direttamente derivati da digital libraries di grandi dimensioni.

Introduzione

Storicamente, lo sviluppo della conoscenza in ambito scientifico si è evoluto secondo due paradigmi: quello teorico e quello sperimentale. Negli ultimi due decenni, si sono affermati due

ulteriori paradigmi: la *simulazione computazionale* teorizzata da Ken Wilson, premio Nobel per la fisica nel 1982, è il terzo paradigma dal quale sono originate le scienze computazionali; gli studi scientifici *data-driven based* sono invece alla base del *quarto paradigma* di Gordon Bell 1. che ha dato origine alle *science informatics*. Quest'ultimo, in particolare, è oggi ampiamente utilizzato nell'analisi di dati scientifici grazie alla crescente disponibilità di enormi quantità di dati, che permettono un approccio *in silico* finalizzato alla generazione di conoscenza.

Per applicare lo stesso cambiamento di prospettiva alle *humanities*, si deve partire dall'osservare che, negli ultimi decenni, uno sforzo notevole è stato dedicato alla generazione di grandi database cui è possibile accedere on-line. Alcuni esempi sono:

- Thesaurus Linguae Graecae: raccoglie la letteratura greca da Omero (VIII sec. a.C.) alla caduta di Bisanzio (1453 d.C.), http://stephanus.tlg.uci.edu/;
- Integrated Archaeological Database (IADB): risponde alle esigenze di gestione dei dati per tutta la durata della vita dei progetti di scavo archeologico, http://www.iadb.org.uk/;
- World Digital Library (WDL): raccoglie le versioni digitalizzate dei libri rari, mappe, manoscritti, fotografie, https://www.wdl.org/en/;
- Musisque Deoque: archivio digitale di poesia latina, http://www.mqdq.it/public/;
- Trismegisto: banca dati relativa a documentazione su papiro ed epigrafica, http://www.trismegistos.org/.

Tutti questi database propongono meccanismi di interrogazione dai diversi livelli di complessità, che forniscono agli studiosi supporto per ricerche per lo più specifiche (ad esempio, il recupero di tutte le poesie scritte utilizzando una certa metrica). Vale a dire, replicano l'approccio metodologico tradizionale delle humanities in base al quale è necessario che, preliminarmente all'interrogazione, lo studioso abbia già formulato precise ipotesi di ricerca, sulle quali ci si aspetta poi di ottenere conferme grazie all'utilizzo di tecnologie digitali.

L'approccio metodologico suggerito dal *quarto paradigma* è del tutto diverso: gli algoritmi si sviluppano e applicano per trovare nuove ipotesi di lavoro tramite la scoperta di pattern dedotti direttamente da database anche di grandi dimensioni.

Per esempio: gli algoritmi possono essere applicati al fine di identificare i gruppi (cluster) di testi poetici o letterari 'simili' tra loro per argomento, o usi linguistici, o formulari, in digital libraries biblioteconomiche o archeologiche, o in corpora letteraria digitali, dai quali inferire nuove ipotesi su cui avviare ricerche utilizzando approcci tradizionali.

Nel presente intervento si descrive il modulo di *graphic matching* M-Evo (Multi-Evolution) del sistema di riconoscimento digitale ICRPad (Brevetto n. 0001407881), sviluppato tra il 2010 e il 2013 dalla spin off dell'Università degli Studi di Bari Aldo Moro D.A.BI.MUS. S.r.l. con la collaborazione di ricercatori del Dipartimento di Studi Umanistici della medesima università.

Il modulo permette di interrogare grandi database digitali di manoscritti storici applicando l'approccio metodologico definito dal *quarto paradigma*, grazie all'utilizzo di un algoritmo di *graphic matching* basato sul concetto di *shape contour recognition* che non necessita di processi di

segmentazione del contenuto dell'immagine. Il processo è stato testato con obiettivi differenti sia su medi e grandi database di varia documentazione, sia su singoli oggetti digitali riproducenti documentazione manoscritta e a stampa antica o moderna.

La prospettiva del data science negli studi sui data humanities: un'evoluzione ancora in progress

Negli ultimi quindici anni sono stati pubblicati non pochi contributi relativi a ricerche finalizzate a creare sistemi di OCR o pattern recognition per immagini riproducenti beni documentali manoscritti o a stampa antichi.

Da una ricognizione nella letteratura specialistica sia di ricerca che tra i brevetti internazionali è emerso che lo stato dell'arte ha prodotto risultati significativi esclusivamente in ambiente di test: i diversi prototipi elaborati, infatti, hanno avuto tutti quale presupposto della ricerca la creazione di strumenti tecnologici che agevolassero le metodologie di studio tradizionali, con scarsa o nulla attenzione circa la possibilità di inferire nuovi approcci metodologici.

Inoltre, i diversi sistemi di riconoscimento analizzati sono stati testati su database di piccole dimensioni oppure su singole risorse digitali, il che non consente di ipotizzare realistici utilizzi su basi dati ampie.

Di seguito si riporta, per motivi di brevità, una selezione di contributi sul tema a nostro parere maggiormente significativi per meglio evidenziare i fattori di differenziazione del sistema descritto in questo contributo.

Shape prior model – Ben-Gurion University 16.. Una sperimentazione interessante è quella effettuata diversi anni fa da un team della Ben-Gurion University di Israele, i cui primi risultati sono stati pubblicati nel 2008.

Nel paper si descrive un metodo di segmentazione e riconoscimento di caratteri non perfettamente leggibili in manoscritti antichi danneggiati. Il processo è basato sulla costruzione manuale di *shape model*s rappresentativi della possibile variabilità dei caratteri preventivamente segmentati da immagini di manoscritti danneggiati. Su questi viene effettuato un training set che, tramite il matching con i modelli di riferimento, li riduce progressivamente a un nucleo fondamentale, generando per ciascun carattere segmentato un *shape prior* che costituisce il riferimento essenziale per la ricostruzione di caratteri danneggiati e non ben leggibili all'occhio umano.

Sebbene i risultati siano senza dubbio interessanti, primo limite del sistema è che agisce esclusivamente su immagini in scala di grigio. Limite ancora più rilevante, esso presuppone una fase preliminare di costruzione manuale di modelli di riferimento lunga e laboriosa, e necessita di più training set progressivi. Entrambi questi fattori sono decisivi nell'impedire l'utilizzo del processo su basi dati ampie..

Computer-Based Stroke Extraction in Historical Manuscripts – Universität Hamburg 12.. Il pattern grafico descritto è basato sull'estrazione e riconoscimento di sezioni o parti di grafi e sulla loro ricostruzione, finalizzata a creare modelli.

L'algoritmo di modellazione ricostruisce e riconosce la direzione in base alla quale ogni porzione di grafo è stata scritta, quindi usa le singole parti riconosciute cercando di ricostruire ogni grafo, similmente a un puzzle.

Le prove sono state condotte esclusivamente su manoscritti ideografici, non si fa menzione di test su manoscritti alfabetici: questo lascia aperte molte domande circa il funzionamento dell'algoritmo sul corsivo manoscritto, e il processo non giunge a nessuna dimostrazione dell'efficacia e efficienza della metodologia se utilizzata in un campo specifico di applicazione come la paleografia, soprattutto nel caso di basi dati ampie.

reCAPTCHA 4. Celebre sistema messo a punto nel 2008 dai ricercatori della statunitense Carnegie Mellon University, che hanno rielaborato i sistemi *CAPTCHA* rendendoli in grado di interpretare le parole dubbie individuate dai programmi OCR, secondo un sistema semplice, ma efficace.

Quando due algoritmi OCR identificano in modo diverso una parola, questa viene associata a una parola nota e inviata a un utente che deve superare un test *CAPTCHA* per accedere a un servizio. Si presuppone che, se un utente riesce ad individuare correttamente la parola nota, allora individuerà con elevata probabilità anche la parola ignota. Quando tre utenti danno la stessa risposta, il sistema archivia la parola come corretta.

Nel settembre del 2009 il progetto è stata acquisito da Google, che lo utilizza per correggere gli errori derivati dalla scansione OCR dei testi. Va però rimarcato che, relativamente a immagini di volumi stampati antecedentemente alla seconda metà dell'Ottocento, i risultati non possono essere definiti all'altezza delle aspettative createsi al momento della scoperta e diffusione del sistema. Le percentuali di restituzione sono difatti ancora oggi alquanto basse, in quanto oscillano tra il 10% e il 30% per documenti a stampa antichi, con la percentuale più elevata ottenuta esclusivamente su testi a stampa del tardo Ottocento, mentre per i manoscritti il sistema non ha mostrato alcun funzionamento degno di rilievo.

Multifont Optical Character Recognition Using a Box Connectivity Approach (EP 0649113 A2) 13.. L'approccio dell'applicativo è basato su un processo pattern recognition che si basa su un *minimal bounding rectangle* definito intorno al pattern, dividendo quindi il pattern in una griglia e confrontando un vettore partizionato derivato da questa con vettori ricavati in modo simile a partire da pattern noti.

Infine si sceglie un insieme di pattern candidati e si seleziona uno dei pattern così ricavati. Il processo, alquanto laborioso, per ammissione degli stessi creatori non è in grado di operare con efficacia su immagini di documenti antichi.

A2iA's Proprietary IWR, Intelligent Word Recognition 24.. L'applicativo, di uso commerciale ma sicuramente di rilevante interesse scientifico si basa sulla segmentazione in parole delle regioni di testo su singole immagini che riproducono documenti manoscritti. Sebbene sia stato utilizzato con successo anche in progetti di riconoscimento di documenti

antichi, il sistema è applicabile esclusivamente su grafie estremamente regolari e ricorsive, quindi facilmente inquadrabili, e per singolo documento. Se applicato su grafie leggermente difformi o su più documenti in sequenza rapida, necessita di preliminare trascrizione dei lemmi da riconoscere in specifici thesauri semantici, altrimenti risulta poco efficace. Ancora una volta, dunque, siamo in presenza di un sistema difficilmente utilizzabile per basi dati di ampie dimensioni.

Non ci soffermeremo in questa sede a discutere del recentissimo applicativo *Transkribus* (https://transkribus.eu/Transkribus/) elaborato nell'ambito del progetto europeo *READ* (https://read.transkribus.eu/), in quanto esso da solo richiede ampia e articolata discussione circa i requisiti scientifici che dovrebbero caratterizzarlo, sui quali continuiamo a nutrire qualche perplessità in virtù di personale esperienza a riguardo.

Sintetizzando, dunque, quanto rilevato da una più ampia ricognizione della letteratura scientifica a oggi prodotta sul tema, risulta agevole rilevare come quasi tutte ricerche di settore si siano basate principalmente su due diversi processi di sviluppo:

a. a. segmentale: è il più utilizzato in un ampio numero di prototipi, per lo più strutturati secondo lo Hidden Markov Model (HMM) (3., 4., 5., 6., 7.);
b. b. olistico: è preferito in alcune più recenti sperimentazioni che hanno prodotto risultati di rilievo per percentuale di contenuto riconosciuto, ma i test sono stati eseguiti su campioni di immagini del tutto esigui e perciò non assumibili come significativi 8..

La maggior parte delle ricerche si è basata sui seguenti approcci metodologici:

- segmentazione: gli *shape model*s sono creati da regioni segmentate (porzioni di grafemi, grafemi, parole, ecc.), quindi i modelli sono classificati in thesauri di riferimento per essere utilizzati nell'esecuzione del matching con i testi di interesse: approccio tanto laborioso quanto potenzialmente infinito;
- riuso e adattamento di processi esistenti (word spotting, HMM, ecc.), basati sull'estrazione dei dati e su processi di matching utilizzati soprattutto per fini statistici; in particolare, la maggior parte di questi prototipi si basa principalmente su immagini digitali di documenti stampati, mentre non si sa nulla della loro efficacia usabilità su manoscritti;
- matching manuale integrale del contenuto digitale con testo elettronico corrispondente precedentemente trascritto manualmente da un operatore, o con thesauri di parole selezionate, sempre strutturate in modo manuale; questo approccio limita seriamente la possibilità di processare una grande quantità di dati, in quanto richiede un lavoro manuale preliminare troppo lungo e complesso (9., 10., 11., 12., 13., 14., 15., 16., 17., 18., 19., 20., 21., 22.).

Inoltre, le ricerche su entrambi i processi presentano a nostro parere i seguenti limiti, che condizionano notevolmente la possibilità di ulteriori sviluppi e l'utilizzo pratico dei risultati:

- i risultati consistono per lo più in prototipi dei quali non si ha alcuna certezza se siano efficaci su ampi database digitali, in quanto sono stati testati esclusivamente su risorse

digitali singole o comunque quantitativamente e qualitativamente molto limitate;

- quasi sempre i set di immagini utilizzati sono immagini singole con grafia manoscritta o a stampa assolutamente omografa e uniforme, quanto più possibile prive di rumorosità e ulteriormente ripulite con accurati interventi di post processing; non vi sono indicatori che attestino altrettanta qualità di risultati qualora i prototipi siano utilizzati su database di medie o grandi dimensioni contenenti immagini eterogenee;
- in molti casi, il processo necessita di attività manuali lunghe e complesse che inevitabilmente incidono su risultati pur rilevanti: l'eccessivo lavoro manuale, infatti, limita notevolmente la possibilità di utilizzare questi sistemi su database di grandi dimensioni, e ne restringe l'uso a singole risorse digitali, o a database contenenti poche risorse possibilmente omografe o quantomeno omogenee.

Il modulo M-Evo del sistema ICRPad

Il modulo M-Evo del sistema ICRPad è stato sviluppato con l'obiettivo di facilitare a studiosi e utenti la fruizione di database digitali contenenti documentazione manoscritta e a stampa antica, interrogandoli, oltre che secondo le metodologie tradizionali, anche utilizzando l'approccio definito dal *quarto paradigma*. Un approccio del tutto nuovo nei domini della cultura e del cultural heritage, che consente di inferire nuove o inattese ipotesi di ricerca direttamente dai database interrogati.

Il modulo, infatti, utilizza un algoritmo di *graphic matching* basato sul concetto di *shape contour recognition* che, senza necessità di preliminari laboriose attività manuali o complessi training di segmentazione del layout e riconoscimento intelligente dei grafi, permette di individuare e selezionare in immagini regioni grafiche o porzioni di esse.

Caratteristiche tecniche del modulo

L'idea che sta alla base dell'algoritmo di matching è di utilizzare una regione grafica di partenza (un'immagine o parte di essa), crearne un modello e recuperare regioni omografe o graficamente simili in una o più immagini destinazione. Una volta creato il modello da cercare, tramite una gamma di operatori a sua disposizione l'utente può impostare i seguenti parametri e soglie secondo cui eseguire la ricerca, che gli consentono di ottimizzarla in termini di prestazioni e di qualità dei risultati:

- la posizione localizzata all'interno del documento oggetto della ricerca;
- l'angolazione;
- la scala della porzione di immagine trovata rispetto al modello fornito in input;
- un indicatore di score della similarità del risultato rispetto al modello di partenza.

Il sistema consente di gestire le impostazioni utilizzando come parametri di base per la ricerca sia un set completo di *primitive*, che funzioni grafiche. Alcune delle principali funzioni grafiche

gestite sono: angolazione, rotazione, scala, sovrapposizione, contrasto, luminosità, colore, trasparenza, messa a fuoco, distorsione, occlusione, deformazione.

Il modulo è provvisto di un'interfaccia user-friendly che l'utente può utilizzare intuitivamente per personalizzare le sue ricerche, utilizzando le seguenti funzionalità:

1. connessione in tempo reale a n database esistenti sul web, utilizzando la funzionalità "repository selection" dell'interfaccia "system setting";
2. esplorazione delle immagini memorizzate nei database connessi per la scelta degli elementi con cui creare i modelli da utilizzare come chiavi di ricerca;
3. parametri di impostazione della ricerca modificabili in tempo reale per personalizzare le ricerche e migliorare quantità e qualità dei risultati;
4. creazione in tempo reale di *shape model*s: dopo aver scelto una o più immagini, l'utente seleziona le Regioni di Interesse (ROI) direttamente su esse (singolo grafo, più grafi, un'intera parola, una linea, un capolettera miniato, ecc.), quindi le estrae tramite una funzione automatica inclusiva di zoom che gli consente di modellarle in base alle sue esigenze; un tool di rilevazione delle rumorosità dell'immagine consente di evidenziare fattori che possono condizionare l'affidabilità della ricerca;
5. personalizzazione delle ricerche tramite il salvataggio in apposita repository di sistema dei *shape model*s creati e utilizzati per la ricerca.
6. ricerca nell'intero database connesso.

Funzionamento dell'algoritmo

L'algoritmo di pattern matching, essendo *shape based*, riconosce le ROI in base alla loro forma, senza tenere conto di dimensioni e valori di scala di grigi dei pixel, e non ha bisogno di una segmentazione preliminare delle pagine del documento.

Esistono diversi modi per determinare o descrivere la forma di un oggetto. In M-Evo l'estrazione del *shape model* avviene selezionando tutti quei pixel il cui contrasto rispetto ai pixel vicini supera una determinata soglia: tipicamente, tali pixel appartengono al contorno dell'oggetto. Quindi, dato un modello di forma, lo scopo principale del processo di matching è trovare nelle immagini di destinazione le sue occorrenze (tutte, o un numero massimo di esse, se inizialmente specificato).

In particolare, il sistema consente di specificare quali pixel fanno parte del modello per velocizzare la ricerca utilizzando un sotto-campionamento, specificare un intervallo di orientamento, specificare un intervallo di scala e così via.

Infine, l'algoritmo di matching consente di ricercare contemporaneamente n modelli all'interno di di n immagini e di parallelizzare tutte le elaborazioni per uno o più modelli, con l'obiettivo di ottimizzare i tempi di ricerca.

A seconda delle impostazioni dei parametri date dall'utente, l'algoritmo può fornire le seguenti informazioni per ciascun modello recuperato:

- posizione, angolo di inclinazione e score;
- posizione, angolo di inclinazione, un fattore di risoluzione uniforme e score;
- posizione, angolo di inclinazione, differenti fattori di risoluzione per x e y, score.

Se la ricerca riguarda diversi modelli, vengono fornite anche informazioni sul modello al quale ciascuna istanza trovata si riferisce.

Creazione del shape model

Il *shape model* caratterizza e definisce una rappresentazione interna della parte dell'immagine che sarà utilizzata come elemento di ricerca. Questa immagine dovrebbe essere mostrata nella sua forma ideale, vale a dire la più nitida possibile, senza occlusioni e possibilmente ben allineata con l'asse orizzontale.

Il formato di immagine sorgente per definire il modello può essere uno dei formati elettronici comuni, come TIFF, BMP, GIF, JPEG, PPM, PGM, PNG, PBM e così via. L'immagine può essere di qualsiasi forma (ellittica, circolare, poligonale o anche delineata a mano libera) e avere un angolo arbitrario.

La Figura 1 mostra la creazione di modelli. La cornice che circonda il modello definisce la ROI. L'ottimizzazione del processo di ricerca inizia dalla definizione di un buon modello. Una volta selezionata la parte dell'immagine da utilizzare come modello, alcuni dei suoi parametri possono essere modificati per il processo di ricerca; infatti, un modello può essere memorizzato per recuperarlo e modificarlo in un secondo momento per uso futuro.

Illustrazione 1: Creazione del modello

Per ottenere un modello valido, il parametro di contrasto deve essere scelto includendo nel modello i pixel significativi per l'identificazione dell'oggetto. Per pixel significativi intendiamo quei pixel che caratterizzano l'oggetto e consentono di differenziare chiaramente la forma da cercare da altri oggetti e dallo sfondo. Il modello deve avere una rumorosità minima e deve escludere eventuali regioni di non interesse, cioè regioni non pertinenti all'oggetto da cercare.

Parametri di ricerca

I parametri più importanti per la ricerca di un modello sono:

- contrasto: attraverso la definizione di una soglia (low-high) consente di differenziare i pixel

appartenenti a porzioni dell'immagine sporche o irrilevanti;

- numero dei livelli di piramide che costituiscono il modello: l'insieme di immagini a risoluzioni diverse che rappresentano la stessa immagine sorgente, ordinate per risoluzione decrescente, è la piramide, e le immagini in una piramide costituiscono i livelli di piramide, dove il vertice della piramide è l'immagine alla risoluzione più bassa (Figura 2). Se l'immagine originale ha una risoluzione di 600x400 dpi, la piramide sarà costituita dall'immagine di primo livello 600x400 dpi, immagine di secondo livello 300x200 dpi, immagine di terzo livello 150x100 dpi e così via. Questo è un fattore cruciale per le prestazioni e la precisione del risultato: come regola generale, si ha un buon risultato se una regione di interesse con una larghezza di 2LevelNumber – 1 pixel, cioè 8 pixel di larghezza, consente di utilizzare quattro livelli di piramide. Dopo aver impostato la regione, l'immagine può essere utilizzata come modello per la creazione della forma di riferimento;

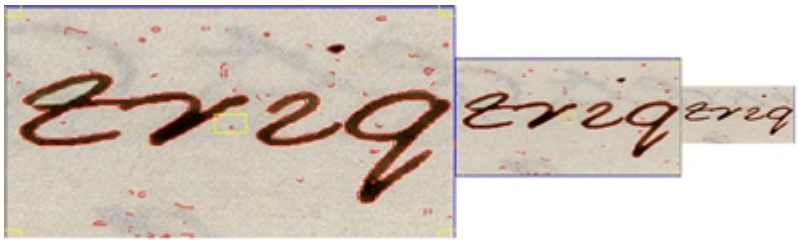

Illustrazione 2: Numero dei livelli di piramide

- angolo di rotazione ed estensione del modello: impostando l'angolo a 5° e l'estensione a 10°, la ricerca può essere eseguita utilizzando immagini con una tolleranza di rotazione di circa +/- 5°;
- scala per i fattori x, y: consente di definire la coppia min/max per ciascun asse necessaria a definire un tratto del modello;
- timeout: consente di velocizzare il processo di ricerca fino al 10%; questo è utile quando si desidera soddisfare un determinato lasso di tempo per ogni immagine di destinazione.

Parametri di configurazione

I parametri di configurazione sono stati calibrati su test case composti da immagini diverse, variando gradualmente la percentuale di riconoscimento al fine di migliorare le diverse fasi applicative dell'algoritmo.

I parametri utilizzati durante questa prova e nella successiva definizione finale delle diverse fasi sono:

- di base:
  - minimum score: la misura di somiglianza tra il modello da cercare e le occorrenze candidate nelle immagini di destinazione; più alto è lo score, più veloce è la ricerca, perché le occorrenze non candidate vengono scartate prima.

La sperimentazione ci ha permesso di concludere che, dato un campione positivo, vale a dire documenti che certamente contengono il *shape model* da trovare, la percentuale di riconoscimento necessaria a definire lo score ottimale per tipi di immagine uniformi è stata di circa il 90%; tale condizione significa che i parametri di definizione del modello e quelli per l'esecuzione dell'algoritmo assicurano il risultato atteso;

- maximum number of items found per image: è il valore per recuperare da ogni immagine tutti i potenziali modelli;

- avanzati (sono essenziali per la ricerca):
  - completeness: determina il trade-off tra efficienza ed efficacia dei risultati della ricerca. Un valore basso determina una ricerca completa, ma piuttosto lenta; più alto è il valore, più veloce è la ricerca, ma la completezza ne risente (cioè, un'occorrenza del modello potrebbe non essere trovata anche se è visibile nell'immagine);
  - overlap: specifica in che modo due aree grafiche possono sovrapporsi a un'immagine; in caso di simmetria, la sovrapposizione consentita dovrebbe essere ridotta per evitare corrispondenze multiple sullo stesso oggetto;
  - sub-pixel: definisce l'accuratezza selezionando il fattore di precisione nel calcolo della posizione, dell'orientamento e della scala; queste caratteristiche possono essere definite attraverso alcune impostazioni predefinite, come il calcolo della posizione, che può essere determinato solo con 'pixel' di precisione, mentre l'accuratezza dell'orientamento e della scala è rispettivamente uguale ai valori dell'angolo e delle dimensioni della scala specificati durante la costruzione del modello. In questo modo, la posizione è stimata con accuratezza fino al livello di pixel, e dalla dimensione dell'oggetto dipende l'accuratezza della scala e dell'orientamento stimati: maggiore è la dimensione, più precisi saranno l'orientamento e la scala;
  - deformation: a volte gli oggetti nelle immagini destinazione non vengono trovati, o sono trovati solo con un grado di precisione basso perché sono leggermente deformati rispetto al modello. In queste occorrenze, è possibile utilizzare un parametro di deformazione che esprime quanti pixel di deviazione devono essere tollerati tra i contorni trovati nell'immagine destinazione e quelli del modello. Questo parametro deve essere impostato al minimo valore possibile; valori elevati possono essere utilizzati solo per ricerche mirate. Infatti, maggiore è questo valore, maggiore è il rischio di recuperare 'falsi positivi' (v. infra) e, nel contempo, aumenta il tempo di elaborazione. Entrambi i problemi si hanno principalmente nella ricerca di oggetti piccoli/raffinati/sottili, poiché questo genere di oggetti, se soggetti a deformazioni, perdono la loro forma caratteristica, che è importante per una ricerca efficiente.

Ricerca del shape model

In risposta al lancio della ricerca nel dataset di immagini destinazione scelto, la posizione e la

rotazione delle istanze trovate ritornano come valori di riga, colonna e angolo. Inoltre, ogni istanza trovata è contrassegnata da uno score. Sono restituite informazioni aggiuntive, come la scala: se il *shape model* è stato creato, il rapporto di risoluzione tra il modello e l'immagine trovata viene parametrizzato.

Il sotto-campionamento può essere abilitato per accelerare il processo di matching, vale a dire che è possibile utilizzare immagini a bassa risoluzione. Come detto sopra, nella piramide composta da immagini a risoluzione decrescente la parte superiore è l'immagine a risoluzione più bassa. Quando si definisce un modello, viene creata una serie di immagini con diversa risoluzione. Quindi, il modello viene creato ed è ricercabile su più livelli della piramide. È possibile specificare il numero di livelli della piramide da utilizzare: è buona norma scegliere il livello più alto della piramide con un modello contenente almeno da 10 a 15 pixel, in modo che il *shape model* assomigli alla forma dell'oggetto.

In ogni caso, il sistema fornisce le *primitive* che consentono di impostare automaticamente questi parametri mediante un'analisi interna della regione coperta dal modello.

Risultati della sperimentazione

Sono stati eseguiti numerosi test per verificare le funzionalità del sistema e la sua validità. I test hanno riguardato sia oggetti digitali riproducenti diverse tipologie documentali manoscritte e a stampa antiche e moderne, sia database di diversa dimensione e complessità.

Come detto all'inizio, in questa sede si presentano i risultati delle sperimentazioni eseguite in ambiti della ricerca paleografica, privilegiando l'approccio metodologico definito dal *quarto paradigma*.

Il modulo M-Evo, infatti, permette agli studiosi di paleografia di eseguire ricerche in diversi tipi di database interrogandoli secondo un approccio assumption-free: vale a dire che il sistema non è destinato a interagire con un singolo, specifico, pre-definito database, ma può essere collegato in tempo reale con molteplici database disponibili on-line che l'utente può selezionare al momento e a sua discrezione in quanto potenzialmente rilevanti per i suoi obiettivi di ricerca.

Nella cornice classica della metodologia di ricerca umanistica, l'interrogazione dovrebbe essere effettuata presupponendo una determinata ipotesi di ricerca con l'obiettivo di verificare se i risultati la confermino.

Nell'impostazione metodologica utilizzata nella sperimentazione, la ricerca è stata casuale, cioè senza alcuna aspettativa sui risultati, e le suggestioni per la costruzione di nuove ipotesi di ricerca sono state dedotte dall'analisi dei dati inferiti dall'interrogazione.

Tuttavia, anche nel caso di modalità di interrogazione dei database partendo da precise ipotesi di ricerca, possono essere dedotte ulteriori ipotesi di studio inizialmente impreviste anche dall'analisi attenta dei risultati 'falsi positivi': questi ultimi, infatti, possono rivelarsi di grande

interesse, perché alcuni, anche se formalmente diversi rispetto al modello utilizzato, possono tuttavia rivelare somiglianze non altrimenti rilevabili a occhio nudo, ma nondimeno reali, che aprono la via a ipotesi diverse e interessanti da indagare.

Sono, questi ultimi, due degli approcci del data science che aprono la via al rinnovamento metodologico nello studio dei data humanities tramite la consultazione di banche dati, aprendo nuovi scenari di ricerca.

1. Test case e sperimentazione

Per la sperimentazione si è simulata la metodologia di studio di un paleografo impegnato in studi su codici manoscritti tramite l'interrogazione di database on line, con l'obiettivo di esplorare nuove ipotesi ricerca da dedurre direttamente dai risultati dell'interrogazione che non fossero assumibili tramite approcci metodologici tradizionali. A tal fine, sono stati utilizzati due diversi campioni di manoscritti:

- un set di 3500 immagini estratte da sette codici latini conservati nella Biblioteca Apostolica Vaticana e datati tra tra l'XI e il XIII secolo, tutti di autori e scriptoria sconosciuti ma considerati diversi tra loro per data e provenienza (numerati da 1 a 7);
- un set di 30 immagini relative a due dei tre manoscritti greci che compongono il codice Sinaitico conservato presso la British Library (denominati A e B).

Codici BAV

Si è scelta casualmente un'immagine digitale di una pagina del codice ms 1 e tramite il tool create model è stato estratto automaticamente da essa una regione in cui è rappresentato il grafo &, in quanto certamente comune a tutti i manoscritti di interesse.

Prima di procedere con il matching, sono stati impostati i parametri deformation, different resize e minimum score per ottimizzare la ricerca dei modelli all'interno del set di immagini.

Nell'ipotesi di utilizzare il *shape model* creato per altre interrogazioni in dataset diversi, lo si è salvato nel repository di sistema.

Il *shape model* è stato quindi lanciato sull'intero dataset di immagini per verificare l'eventuale presenza delle stesse caratteristiche grafiche in immagini destinazione riferentesi a manoscritti diversi da quello dell'immagine scelta. L'algoritmo ha scansionato l'intero set restituendo tutte le occorrenze coerenti con i parametri utilizzati, e perciò da ritenere uguali, o comunque molto simili al modello.

La tabella seguente illustra i risultati del test case. Creato il *shape model* (col. 2), variando il rapporto tra i tre parametri deformation (col. 1), different resize (col. 4) e minimum score (col. 5) sono stati ottenuti risultati diversi per ciascuna immagine campione (col. 3), il che ci ha

permesso di valutare l'efficacia di ciascun modello creato dal medesimo grafo.

I risultati migliori li ha dati il modello creato con deformation 3 (sul massimo di 5 possibile), different resize del 40% e minimum score compreso tra 60% e 80%. Impostando questi parametri abbiamo avuto un elevato riscontro di 'positivi' nei quattro ms. numerati 1, 2, 4 e 5: circa 80%. Alcuni falsi positivi sono stati rilevati con un'impostazione del minimum score al 60% (parametri weak/low).

Inoltre, eseguendo la ricerca con altri grafi (C ed S), a un'analisi più approfondita sulle immagini abbiamo notato che alcuni 'falsi positivi' avuti in risposta costituiscono in realtà un'ulteriore prova dell'omografia dei quattro ms, poiché, sovrapponendoli con alcuni positivi dal tratto simile (ad esempio, la lettera C con la O, la S con la F secondo lo stile antico), alcuni tratti dei diversi grafi sono risultati del tutto identici tra loro.

Deformation	Search model	Document Image	Resize	Minimum score (threshold)	Notes	Ms. 1	Ms. 2	Ms. 3	Ms. 4	Ms. 5	Ms. 6	Ms. 7
Def 4	et	Ms.1004v	40%	20%	Good	yes	yes, few false	only false	only false	yes, few false	only false	only false
Def 4	et	Ms.1004v	30%	50%	Good	excellent	yes, few false	only false	only false	yes, few false	only false	only false
Def 4	et	Ms.1004v	40%	50%	Good	excellent	yes, few false	only false	only false	yes, few false	only false	only false
Def 4	et	Ms.1004v	40%	60%	Good	excellent	excellent	only false	only false	yes, few false	only false	only false
Def 4	et	Ms.1004v	40%	70%	Good for the Ms. 1 and 2	excellent	excellent	no	no	no	no	no
Def 4	et	Ms.1004v	40%	100%	No matches	no	no	no	no	no	no	no
Def 4	et	Ms.5003r	40%	50%	Sufficient	yes, but with false	yes, but with false	only false	yes, but with false	excellent	only false	only false
Def 4	et	Ms.2	40%	50%	Very	excelle	excelle	only	yes,	excelle	only	only

		002r			interesti ng	nt	nt	false	but with false	nt	false	false
Def 3	et	Ms.2 002r	40%	40%	Better than deforma tion 4, but it should not give the last two manuscr ipts	all positiv e	all positiv e	only false	sufficie nt, with some false	good, with some false	only false	only false
Def 3	et	Ms.2 002r	40%	50%	Better than deforma tion 3, but it should not give the last two manuscr ipts	all positiv e	all positiv e	only false	sufficie nt, with some false	good, with some false	only false	only false
Def 3	et	Ms.2 002r	40%	70%	Better than deforma tion 3 at 50%, but it doesn't find some positive	all positiv e	all positiv e	no	no	no	no	only false
Def 3	et	Ms.2 002r	40%	60%	Better than deforma tion 3 at 70%, it finds some other	all positiv e	all positiv e	only false	2 positiv e	some positiv e and few false	1 false	some false

					positive, but some false on Ms. 7 too.							
Def 4	proxim	Ms.1004v	40%	50%	Almost all the false words start/contain the letter p	yes, but with false	only false	only false	only false	only false	only false	only false
Def 4	proxim	Ms.1004v	40%	70%	Not all of the false are considered with this doc %	yes, but with false	only false	only false	only false	only false	only false	only false
HC (def 4)	proxim	Ms.1004v	40%	80%	Document of extraction only	yes, one doc only	no	no	no	no	no	no
HC (def 3)	proxim	Ms.1004v	40%	60%	Excellent, but with some false	excellent	three false	no	no	no	no	no
HC (def 3)	proxim	Ms.1004v	40%	70%	Excellent	excellent	one false only	no	no	no	no	no
HC (def 3)	terra	Ms.5002v	40%	60%	Excellent if we consider the similarity with Ms. 1 2 and 4, very	only false	only false	no (correct)	only false	excellent	only false	only false

					good on the 14, there is a problem on last two ms.							
HC (def 3)	terra	Ms.5 002v	40%	70%	Excellent	2 false	no	no (correct)	no	good, with some false	no	no
HC (def 4)	singolo (e)	Ms.7 f.003r	30%	80%	Excellent, it find the one match in the sample image	no	no	no	no	no	no	1 match
HC (def 4)	sillaba (ta)	Ms.1 004r	30%	80%	Excellent	excellent	good	1 match	excellent	no	no	no
HC (def 4)	sillaba (ta)	Ms.1 004r	30%	90%	It only considers the manuscript extraction	ottimo	no	no	no	no	no	no
HC (def 3)	sillaba (ta)	Ms.1 004r	40%	60%	Good, it matches some heterogeneous positive and some false	some positive	excellent	some positive with different calligraphy	excellent	good, with some false	false	some positive and some false
HC (def 3)	sillaba (ta)	Ms.1 004r	40%	70%	Better than threshold set at 60%	it find positive more than with the thresho	excellent	some positive and some false	excellent	good, with some false	false	some positive and some false

						ld set at 60%						
HC (def 3)	sillaba (ta)	Ms.1004r	40%	80%	Good, it find heterogeneous positive except Ms. 6	it find positive more than with the threshold set at 60%	excellent	1 positive	excellent	good with some false	no	some positive and some false
HC (def 4)	frase	Ms.1004r	40%	90%	Extraction document only	1 positive	no	no	no	no	no	no
HC (def 4)	frase	Ms.1004r	40%	80%	Extraction document only	1 positive	no	no	no	no	no	no
HC (def 4)	frase	Ms.1004r	40%	70%	Extraction document only	1 positive	no	no	no	no	no	no
HC (def 4)	frase	Ms.1004r	40%	60%	Good, with Ms. 7 exception	yes, with some false	only false, because the ms does not contain any positive	no (correct, ms has another calligraphy)	only false, because the ms does not contain any positive	false in each doc	1 false only	many false (it's no good)
HC (def 3)	frase	Ms.1004r	40%	60%	Extraction document only	1 positive	no	no	no	no	no	no
HC (def 3)	frase	Ms.1004r	40%	50%	Extraction document only	1 positive	no	no	no	no	no	no

HC (def 3)	frase	Ms.1 004r	40%	40%	It's interesting, but with this false deformation parameter and the low threshold it excludes a lot of false	yes, with	no positive	no	no	no	no

Tabella 1: Risultati del test case

Nel caso i risultati siano insoddisfacenti, i tre parametri di cui sopra possono essere modificati e reimpostati anche durante la ricerca.

Sorprendentemente rispetto a quanto stabilito dalla storia degli studi tradizionali, è risultato che i ms 1, 2, 4 e 5 possono essere stati vergati dalla stessa mano, o in un medesimo scriptorium, e in ogni caso utilizzando un canone rigorosissimo e molto ben definito.

I risultati 'positivi', sia pure incerti, potevano essere attesi dall'utente al momento del lancio della ricerca, e tuttavia potevano non essere sufficienti, da soli, a sostenere l'ipotesi che i codici abbiano la medesima provenienza. In tale direzione, i 'falsi positivi', pur attesi, assumono però un rilievo tanto inatteso quanto decisivo, in quanto sono risultati contenere tratti omografi rispetto al modello, il che è possibile solo ipotizzando un canone scrittorio identico per i quattro manoscritti.

I dati analizzati sollecitano, quindi, ulteriori approfondite indagini sia sui dati digitali che sugli artefatti originali, diventando premessa per formulare nuove ipotesi su paternità, provenienza e datazione dei quattro manoscritti, tra le quali, a esempio:

- che siano opera di uno stesso amanuense operativo in un singolo scriptorium in un determinato arco di tempo, o in scriptoria differenti in archi temporali non distanti tra loro;
- che alcuni dei manoscritti siano opera di due diversi amanuensi attivi in tempi diversi (anche nel corso di secoli diversi) nel medesimo scriptorium, che perciò hanno utilizzato il medesimo canone rigoroso riuscendo a perfezionare la loro abilità scrittoria al punto da rendere le grafie quasi omografe;
- che alcuni dei manoscritti siano stati prodotti nel medesimo periodo storico ma in due scriptoria diversi, nei quali era utilizzato il medesimo canone;
- che alcuni dei manoscritti siano stati prodotti in secoli diversi e in scriptoria differenti, ma collocati nella stessa area geografica, che quindi avrebbero utilizzato

e perfezionato un canone scrittorio comune mantenutosi nel corso del tempo,
e così via ipotizzando.

Quanto rileva senza dubbio è che, approfondendo l'analisi dei dati risultanti dall'interrogazione, risposte normalmente definibili come "rumorosità" che pregiudicano l'efficacia di un algoritmo diventano invece in questo caso ulteriori fonti attendibili da cui dedurre nuove ipotesi di ricerca altrimenti non formulabili, in quanto non rilevabili con gli approcci metodologici tradizionali.

Codice Sinaitico

I manoscritti che chiamiamo A e B si trovano nel codice Sinaitico rispettivamente in prima e terza posizione, e sono correntemente considerati nella letteratura di riferimento come scritti da amanuensi diversi.

Anche in questo caso la sperimentazione è stata mirata a verificare eventuali possibilità di utilizzare l'algoritmo di matching per inferire dalle immagini ipotesi differenti da quelle comunemente accettate.

Sono state scelte casualmente 30 immagini da entrambi i manoscritti. Quindi, secondo il procedimento sopra illustrato, da uno dei due, che chiameremo ms A, è stato selezionato il grafo y come ROI da cui creare il modello utilizzando lo strumento create model. Salvato il *shape model* nel repository di sistema per ulteriori necessità, è stata eseguita la ricerca. L'algoritmo ha scansionato l'intero dataset restituendo tutte le occorrenze coerenti con i parametri deformation, different resize e minimum score preliminarmente impostati, e perciò da ritenere uguali o comunque molto simili al modello.

Anche in questo caso, pur variando il rapporto tra i tre parametri, i risultati migliori sono stati ottenuti con i medesimi parametri dei ms BAV.

Nello specifico, sono stati ottenuti i seguenti risultati:

- ms A, grafo y:
  - positivi (omografi): 75% (50% in ms A, 25% in ms B)
  - falsi positivi: 25% (10% in ms A, 15% in ms B); tra i quali:
    - grafi quasi omografi: 20% (5% in ms A, 15% in ms B), nello specifico i grafi f e j, i cui tratti si sovrappongono perfettamente ad alcuni tratti di y;
    - grafi approssimatamente omografi: 5% (tutto in ms B), tutti grafi u, alcuni tratti delle cui linee curve sono sovrapponibili ai tratti corrispondenti di y.

Ancora una volta, differentemente dalle tesi comunemente accettate, risultati che normalmente dovrebbero essere considerati solo come rumorosità del sistema sono interpretabili quali ulteriori conferme che entrambi i manoscritti sono opera di un medesimo amanuense, o comunque hanno medesima provenienza. Ulteriori indagini diventano perciò irrinunciabili.

Conclusioni

In questo paper descriviamo le caratteristiche funzionali dell'algoritmo di *graphic matching* del modulo ICRPad M-Evo per il recupero di informazioni in database digitali di manoscritti antichi, con l'obiettivo di proporre un nuovo approccio metodologico agli studi delle humanities basato sull'applicazione del *quarto paradigma* del data science che è alla base dell'informatica scientifica.

Nel processo di riconoscimento delle immagini digitali, l'algoritmo utilizza la funzione di *shape contour recognition* senza necessità di attivare fasi di segmentazione e di utilizzare i valori di scala di grigi dell'immagine. Esso, infatti, utilizza i pixel per la copertura della forma e il parametro del numero di livelli della piramide che costituiscono il modello: vale a dire, l'insieme di immagini composto da modelli di diversa risoluzione grafica. Questo fattore è cruciale per le prestazioni e la precisione del risultato perché, dopo aver impostato una regione, l'immagine ridotta può essere utilizzata come modello per la creazione del *shape model*.

L'utilizzo del modulo in test eseguiti su dataset di differenti codici manoscritti latini e greci ha confermato l'ipotesi di partenza, in quanto i risultati hanno portato in rilievo ipotesi di ricerca inferite direttamente dalle immagini ricercate che altrimenti, con le metodologie tradizionali, difficilmente sarebbero emersi.

L'ultima versione del sistema è in grado di processare circa 240.000 immagini/h con il 60-90% di corrispondenze positive, avviando la ricerca dai parametri di minimum score del 60% (weak), 70% (low), 80% (medium), 85-90% (high).

References

1. Hey, T., Tansley, S. and Tolle, K. (ed. By, 2009), The Fourth Paradigm. Data Intensive Scientific Discovery, Microsoft Research, Redmond, Washington.
2. Barbuti, N., & Caldarola, T. (2012). An innovative character recognition for ancient book and archival materials: A segmentation and self-learning based approach. In M. Agosti, F. Esposito, S. Ferilli, N. Ferro (Ed.), Communications in Computer and Information Science. Vol. 354: Digital Libraries and Archives, IRCDL 2012, Heidelberg: Springer, (pp. 261-270).
3. Fischer, A., & Bunke, H. (2011). Character prototype selection for handwriting recognition in historical documents. In Proceedings of 19th European Signal Processing Conference, EUSIPCO (pp. 1435–1439).
4. von Ahn, L., Maurer, B., McMillen, C., Abraham, D. and Blum, M (2008), reCAPTCHA: Human-Based Character Recognition via Web Security Measures, «Science», 321 (5895), p. 1465–1468, doi: 10.1126/science.1160379, PMID 18703711

5. Indermühle, E., Eichemberger-Liwicki, M., Bunke, H. (2008). Recognition of Handwritten Historical Documents: HMM-Adaptation vs. Writer Specific Training. In Proceedings of 11th International Conference on Frontiers in Handwriting Recognition, Montreal, Quebec, Canada (pp. 186-191).

6. Bulacu M., & Schomaker L. (2007). Automatic Handwriting Identification on Medieval Documents. In ICIAP 2007: 14th International Conference on Image Analysis and Processing (pp. 279-284).

7. Rath, M. T., Manmatha, R.A., & Lavrenko, V. (2004). Search Engine for Historical Manuscript Images. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. (369-376).

8. Srihari, S., Huang, C., & Srinavasan, H. (2005). A Search Engine for Handwritten Documents. In Document Recognition and Retrieval XII, vol. 154, no. 3. (pp. 66-75).

9. Adamek, T., O' Connor, E. N., & Smeaton, A. F. (2007). Word matching using single closed contours for indexing handwritten historical documents. In International Journal of Document Analysis and Recognition (IJDAR), Volume 9, Issue 2-4, (pp. 153-165).

10. Le Bourgeois, F., & Emptoz, H. (2007). DEBORA: Digital AccEss to BOoks of the RenaissAnce. IJDAR, vol. 9(2-4), 193-221.

11. Stokes, P. A. (2009). Computer-aided Palaeography, Present and Future, in M. Rehbein [et al.] (Eds.), Codicology and Palaeography in the Digital Age, Schriften des Instituts fur Dokumentologie und Editorik, Band 2, Norderstedt: Book on Demand GmbH.

12. Fischer, A., Wüthrich, M., Liwicki, M., Frinken, L., Bunke, H., Viehhauser, G., & Stolz, M. (2009). Automatic Transcription of Handwritten Medieval Documents. In Proceedings of 15th International Conference on Virtual Systems and Multimedia (pp. 137-142).

13. Herzog R., Neumann B., & Solth A. (2011). Computer-based Stroke Extraction in Historical Manuscripts, Manuscript Cultures. Newsletter No. 3, (pp. 14-24).

14. Krtolica, R. V., & Malitsky, S. (2012). Multifont Optical Character Recognition Using a Box Connectivity Approach (EP0649113A2). Retrieved May, 20, 2012 from http://worldwide.espacenet.com/publicationDetails/biblio?CC=EP&NR=0649113&KC=&FT=E&locale=en_EP

15. Leydier, Y., Le Bourgeois, F., & Emptoz, H. (2005). Textual Indexation of Ancient Documents. In Proceedings of the 2005 ACM Symposium on Document Engineering (pp. 111-117).

16. Dalton, J., Davis, T., & van Schaik, S. (2007). Beyond Anonymity: Paleographic Analyses of the Dunhuang Manuscripts. Journal of the International Association of Tibetan Studies, No. 3, 1–23.

17. Bar-Yosef, I., Mokeichev, A., Kedem, K., & Dinstein, I. (2008). Adaptive shape prior for recognition and variational segmentation of degraded historical characters. Pattern Recognition, vol. 42(12), 3348-3354.

18. Gordo, A., Llorenz, D., Marzal, A., Prat, F., & Vilar, J. M. (2008). State: A Multimodal Assisted Text-Transcription System for Ancient Documents. In DAS '08. Proceedings of 8th IAPR International Workshop On Document Analysis Systems (pp. 135-142).

19. Cheriet, M. [et al.] (2009). Handwriting recognition research: Twenty years of achievement... and beyond, Pattern Recognition, vol. 42, 3131–3135.

20. Le Bourgeois, F., & Emptoz, H. (2009). Towards an Omnilingual Word Retrieval System for Ancient Manuscripts. Pattern Recognition, vol. 42(9), 2089-2105.

21. Nel, E.-M., Preez, J. A., & Herbst, B. M. (2009). A Pseudo-skeletonization Algorithm for Static Handwritten Scripts. International Journal on Document Analysis and Recognition (IJDAR) 12, 47–62.

22. Toselli, A. H., Romero, V., Pastor, M., & Vidal, E. (2010). Multimodal Interactive Transcription of Text Images. Pattern Recognition, vol. 43(5), 1814-1825.

23. Fischer, A., M. Wüthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehhauser, and M. Stolz (2009). Automatic Transcription of Handwritten Medieval Documents. Proceedings 15th International Conference on Virtual Systems and Multimedia, pp. 137-142.

24. http://www.a2ia.com