DOI: 10.26314/GARR-Conf17-proceedings-32

Un innovativo graphic matching system per il recupero di informazioni di contenuto in database digitali di manoscritti antichi

Nicola Barbuti[1], Stefano Ferilli[2], Tommaso Caldarola[3]

[1]Università degli Studi di Bari Aldo Moro, Dipartimento di Studi Umanistici,
[2]Università degli Studi di Bari Aldo Moro, Dipartimento di Informatica,
[3]D.A.BI.MUS. S.r.l.

Abstract. Il paper descrive il sistema di graphic matching ICRPad M-Evo, sviluppato con l'obiettivo di consentire agli studiosi di humanities di effettuare ricerche su grandi database di manoscritti storici applicando ai data humanities l'approccio metodologico definito dal "quarto paradigma" del data science (data intensive scientific discovery – Gordon Bell, 2012). Secondo tale approccio, gli algoritmi si sviluppano e applicano per trovare nuove ipotesi di lavoro tramite la scoperta di pattern estratti direttamente da database di grandi dimensioni.

Keywords. Graphic Matching, Data Humanities, Digital Recognition

Introduzione

Nel presente intervento si descrive l'innovativo sistema di graphic matching ICRPad, che utilizza un algoritmo sviluppato con l'obiettivo di consentire agli studiosi di humanities di effettuare ricerche su grandi database di manoscritti storici applicando ai data humanities l'approccio metodologico definito dal "quarto paradigma" del data science (data intensive scientific discovery – Gordon Bell, 2012). Secondo tale approccio, gli algoritmi si sviluppano e applicano per trovare nuove ipotesi di lavoro tramite la scoperta di pattern estratti direttamente da database di grandi dimensioni.

A oggi, infatti, i database digitali a disposizione degli studiosi del CH utilizzano processi di interrogazione che replicano il medesimo approccio metodologico di tipo tradizionale, il cui presupposto indispensabile è l'elaborazione preliminare di ipotesi precise sulle quali si vanno poi a formulare le query. Un approccio che, con lo sviluppo di database sempre più ampi e complessi, risulta ormai inadeguato a soddisfare pienamente i bisogni di chi li interroga.

1. ICRPad M-Evo

L'algoritmo utilizzato nel modulo M-Evo di ICRPad è stato sviluppato avendo quale obiettivo la costruzione di uno strumento tecnologico che consentisse agli studiosi di paleografia di avvalersi nelle proprie ricerche dei database digitali esistenti, interrogandoli sia secondo metodi di approcci tradizionali (primo e secondo paradigma), sia utilizzando

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

l'approccio definito dal quarto paradigma, del tutto nuovo nel dominio di riferimento, di modo da poter inferire nuove o inattese ipotesi di ricerca dall'analisi dei dati risultati dall'interrogazione dei database.

L'algoritmo si basa sul concetto di shape contour recognition, che consente di evitare laboriose attività manuali o complessi training preliminari per la segmentazione del layout e il riconoscimento delle regioni grafiche. L'utente seleziona direttamente sul layout di un'immagine da lui preliminarmente scelta una regione grafica, che l'algoritmo codifica come lo shape model da utilizzare quale chiave di ricerca per recuperare regioni omografe o graficamente simili in una o più immagini di destinazione.

Per eseguire il matching con le immagini di destinazione, l'algoritmo utilizza non i valori in scala di grigio dell'immagine, ma i pixel della forma che costituisce il modello scelto dall'utente e il parametro del numero di livelli della piramide che ne strutturano la rappresentazione iconica.

In tal modo, il processo di interrogazione del modulo M-Evo consente la massima efficacia nella ricerca e, contestualmente, le più ampie potenzialità di effettuarla sia secondo metodi tradizionali che secondo il quarto paradigma, in quanto:

- permette di collegarsi real time come client a n database esistenti on line le cui immagini sono fruibili liberamente, grazie alla funzione di selezione e scelta di "repository" prevista nel sistema;
- consente di visualizzare ed esplorare le immagini contenute nei diversi database per valutare eventuali elementi di interesse, anche secondo scelta casuale, da selezionare per creare shape models da utilizzare quali chiavi di ricerca;
- consente di variare, modulare e personalizzare in qualsiasi momento i parametri di setting per la ricerca, la quantità e la qualità delle risposte, in relazione alle attese di maggiore o minore quantità di dati da rilevare (soglie di deformazione, etc.);
- consente di creare gli shape models in tempo reale secondo le esigenze dell'utente: visualizzate una o più immagini, egli può selezionare le regioni di interesse direttamente sulle immagini e modellarle secondo le sue necessità (fermarsi a un singolo grafo, comprendere più grafi, un'intera parola, etc.); un tool di rilevazione delle rumorosità dell'immagine gli consente di verificare i livelli di "sporcizia" che potranno in qualche modo compromettere l'affidabilità della ricerca (Figura 1);
- consente di personalizzare le ricerche salvando le regioni selezionate e utilizzate come modelli per la ricerca in apposita repository di sistema.

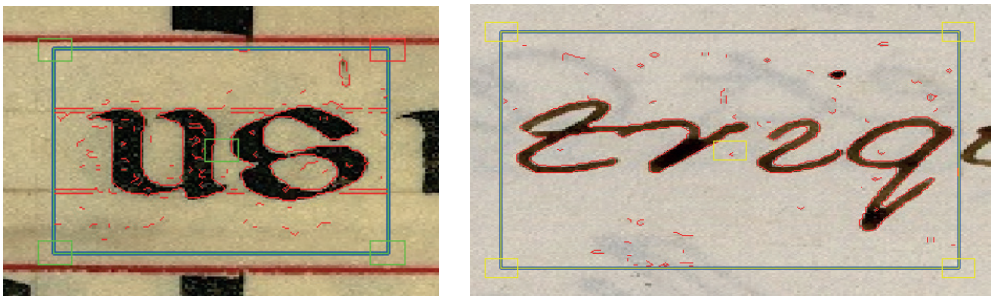

Fig 1
Creazione dei modelli

2. Risultati della sperimentazione

Sono stati eseguiti numerosi test per verificare le funzionalità del sistema e la sua validità. In particolare, sono state effettuate simulazioni prendendo in considerazione gli ambiti di ricerca paleografica. Si è simulato un approccio metodologico di ricerca tramite l'utilizzo di database digitali basato sul quarto paradigma, secondo il quale il paleografo sceglie di analizzare oggetti digitali contenuti in alcuni tra i più importanti (Biblioteca Apostolica Vaticana, British Library, Trinity College, BNF, John Rylands Library of Manchester, etc.), senza formulare preliminarmente una precisa ipotesi da cui partire, ma volendo valutare le possibili ipotesi deducibili dalle risposte alle query che andrà a fare.

Tra i vari test validi, si descrive in questa sede quello eseguito su due manoscritti greci (A e B) contenuti nel codice Sinaitico conservato presso la British Library, considerati opera di due diversi amanuensi, in quanto ha prodotto risultati a nostro parere di particolare interesse.

Il test è stato effettuato allo scopo di verificare se, lanciando query su un campione significativo di immagini scelte casualmente da entrambi i codici, i risultati consentissero di formulare ipotesi di ricerca diverse da quelle comunemente formulate dai paleografi.
Sono stati selezionati sulle immagini alcuni grafi secondo criterio casuale, utilizzati come modelli per le query poi lanciate sulle immagini. Sono state quindi analizzate le istanze restituite dalle diverse interrogazioni, delle quali si descrive di seguito, per necessità di sintesi, il solo risultato relativo al grafo "psi":

- positivi (grafi omografi al psi): 75%, di cui 50% nel ms A e 25% nel ms B
- falsi positivi: 25%, di cui 5% nel ms A e 20% nel ms B;

da attenta analisi dei falsi positivi sono risultati i seguenti elementi di interesse:

- grafi quasi omografi al "psi": 20%, di cui 5% in ms A e 15% in ms B, tutti riproducenti la lettera "phi", con tratto delle curvature perfettamente sovrapponibile alle corrispondenti del "psi";
- grafi approssimativamente omografi: 5%, tutti in ms B, tutti riproducenti la lettera "y", con alcuni tratti delle curvature sovrapponibili alle corrispondenti del "psi".

Le istanze positive possono essere di per sé sufficienti per elaborare un'ipotesi di ricerca finalizzata a dimostrare che, diversamente da quanto a oggi comunemente riconosciuto, i due manoscritti possano essere opera del medesimo amanuense. Ha invece costituito risultato del tutto inatteso la restituzione di un'ampia percentuale di grafi "falsi positivi" aventi tratti del tutto omografi rispetto al grafo scelto come modello. Un dato, questo, che renderebbe quasi inevitabile sia intraprendere ricerche più complesse e approfondite, anche "analogiche", finalizzate a verificare l'ipotesi di cui sopra, sia formulare altre ipotesi, quali:

- che i due manoscritti siano stati prodotti da mani diverse nello stesso scriptorium, nel quale però si utilizzava un canone estremamente rigido;
- che siano stati prodotti dalla stessa mano in tempi diversi e in scriptoria differenti, nei quali si utilizzava il medesimo canone ma con alcune leggere varianti;
- che il medesimo canone di particolare rigore sia stato utilizzato in un determinato scriptorium con leggerissime modifiche nel corso del tempo (secoli?), ovviamente da amanuensi diversi.

3. Conclusioni

In questo documento abbiamo descritto le caratteristiche di ICRPad M-Evo, un sistema brevettato di graphic matching per il riconoscimento digitale dei manoscritti che propone un nuovo approccio alla ricerca e al recupero delle informazioni di contenuto nelle biblioteche digitali. Questo approccio si basa sull'applicazione ai data humanities del quarto paradigma dei data science per lo sviluppo della conoscenza nel campo scientifico, che è alla base dell'informatica scientifica. Il processo di formazione si basa sull'algoritmo di corrispondenza descritto, che utilizza il riconoscimento della forma senza alcun processo di segmentazione. Si seleziona una regione appropriata che automaticamente crea il modello grafico da utilizzare per la ricerca all'interno di data base di immagini.

Riferimenti bibliografici

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).

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).

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.

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).

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

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.

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).

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).

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).

Herzog R., Neumann B., & Solth A. (2011). Computer-based Stroke Extraction in Histori-

cal Manuscripts, Manuscript Cultures. Newsletter No. 3, (pp. 14-24).

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).

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

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

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

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).

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.

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).

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).

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.

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

Autori

Nicola Barbuti nicola.barbuti@uniba.it

Ricercatore Universitario Confermato in Archivistica, Bibliografia e Biblioteconomia presso il Dipartimento di Studi Umanistici (DiSUM) dell'Università degli Studi di Bari Aldo Moro. Svolge attività di ricerca e docenza in scienze biblioteconomiche e dell'informazione, digital cultural heritage, digital humanities. È Responsabile scientifico UNIBA nella Scuola a Rete Nazionale DiCultHer. È Coordinatore del Polo Apulian DiCultHer. È co-inventore del software ICRPad.

Stefano Ferilli stefano.ferilli@uniba.it

Professore Associato per il settore INF/01 presso il Dipartimento di Informatica dell'Università degli Studi di Bari Aldo Moro. Attualmente è Direttore del Centro Interdipartimentale di Logica ed Applicazioni. La sua attività scientifica si focalizza su temi inerenti l'acquisizione automatica di conoscenza espressa in formalismi simbolici, in particolare sui fondamenti logici ed algebrici dell'apprendimento automatico di concetti e sul confronto di descrizioni, elaborando modelli e metodi per la loro applicazione, fornendone realizzazioni ed applicazioni a domini del mondo reale. Collabora all'implementazione del software ICRPad.

Tommaso Caldarola t.caldarola@dabimus.com

Senior Software Architect esperto nella definizione e implementazione di procedure per il controllo di qualità per il system testing, per la scrittura di documentazione tecnica, per le modalità di bug trace e object management per una corretta gestione delle componenti sw finalizzata a facilitarne il riuso, gestione dei processi di configuration, patching & versioning management. È co-inventore del software ICRPad.