ISYREADET: un sistema integrato per il restauro virtuale di documenti antichi

E. CONSOLE[1], V. BURDIN[2], G. CAZUGUEL[2], S. LEGNAIOLI[3], V. PALLESCHI[3], R. TASSONE[1], A. TONAZZINI[4]

Sommario

ISYREADET (Integrated System for Recovering and Archiving Degraded Texts) è un progetto di ricerca finanziato dalla Commissione Europea il cui obiettivo è stato quello di realizzare un sistema integrato, hardware e software, per il restauro virtuale di documenti storici danneggiati con l'utilizzo di metodi e strumenti innovativi, come camere multispettrali e algoritmi di elaborazione di immagini. Nel corso dei due anni di vita del progetto (2003-2004) il consorzio, formato da cinque PMI (T.E.A. s.a.s., Catanzaro, Art Conservation, Vlaardingen, Atelier Quillet, La Rochelle, Art Innovation, Hengelo, Transmedia Technology, Swansea) e tre enti di ricerca (CNR – Istituto per i Processi Chimico-Fisici, Pisa, CNR – Isituto di Scienza e Tecnologie dell'Informazione, Pisa, ENST – École Nationale Supérieure des Télécommunications, Brest) ha condotto con successo una serie di attività. Le attività previste nella realizzazione del progetto hanno riguardato l'analisi e classificazione delle differenti tipologie di danno possibili, la digitalizzazione dei documenti-test con una camera multispettrale, la selezione di adeguati algoritmi di miglioramento dell'immagine e successiva applicazione, la predisposizione dell'interfaccia grafica user-friendly. Vengono qui illustrati i risultati conseguiti applicando gli algoritmi sviluppati per il restauro virtuale dei documenti.

[1] T.E.A. s.a.s. di E. Console & C., Catanzaro, Italy.

[2] École Nationale Supérieure des Télécommunications, Technopole Brest Iroise, Brest, France.

[3] CNR – Istituto per i Processi Chimico-Fisici, Pisa, Italy.

[4] CNR – Istituto di Scienza e Tecnologie dell'Informazione, Pisa, Italy.

info@teacz.191.it, valerie.burdin@enst-bretagne.fr, vince@ipcf.cnr.it, anna.tonazzini@isti.cnr.it

Abstract

Isyreadet (Integrated System for Recovering and Archiving Degraded Texts) is a research project funded by the European Commission whose aim has been to realize an integrated hardware and software system for the virtual restoring of damaged historical documents using innovative methods and tools, such as multispectral cameras and image processing algorithms. During the two years life of the project (2003-2004) the consortium, formed by five SMEs (T.E.A. s.a.s., Catanzaro, Art Conservation, Vlaardingen, Atelier Quillet, La Rochelle, Art Innovation, Hengelo, Transmedia Technology, Swansea) and three RTD Performers (CNR – Istituto per i Processi Chimico-Fisici, Pisa, CNR – Isituto di Scienza e Tecnologie dell'Informazione, Pisa, ENST – École Nationale Supérieure des Télécommunications, Brest), has successfully carried out a series of activities. The activities provided for the realization of the project have been related to the analysis and the classification of different kind of possible damages, the digitalization of the test documents using a multispectral camera, the selection of suitable image enhancement algorithms and further application, the implementation of the user-friendly graphic interface. Above are shown the outcomes reached by the application of the algorithms for the virtual restoration of the documents.

Introduzione

Il patrimonio culturale europeo è costituito in massima parte da documentazione storica in forma cartacea che, per propria natura, è soggetta a deterioramento e che, con l'andar del tempo, corre il rischio di danneggiarsi irreparabilmente. Grazie allo sviluppo delle tecnologie informatiche oggi è possibile ricorrere a strumenti che consentono di conservare tale patrimonio in modo permanente. Tuttavia la sola scansione ed archiviazione dei testi come semplice immagine digitale può non bastare. Come molto spesso accade i caratteri sono difficilmente leggibili e, a causa del deterioramento del supporto (carta, pergamena, ecc.), frammenti di parola restano nascosti o illeggibili. Si rendono allora necessarie tecniche di elaborazione delle immagini che possono configurarsi come veri e propri interventi di restauro virtuale e che le tecniche fino ad oggi adottate, e solo in lavori sperimentali, non consentono perché ancora troppo costose e complicate al punto da risultarne proibitivo l'utilizzo corrente.

Il progetto Isyreadet (Integrated System for Recovering and Archiving Degraded Texts), finanziato dalla Commissione Europea con i fondi del V Programma Quadro di Ricerca, Sviluppo Tecnologico e Dimostrazione (1998-2002) si è proposto di realizzare un sistema integrato, hardware e

software per il restauro virtuale e l'archiviazione di documenti danneggiati utilizzando metodi e strumenti innovativi, come camere multispettrali e algoritmi di elaborazione di immagini.

Effettuata un'attenta analisi delle differenti tipologie di danno presenti sui documenti, l'attenzione è stata focalizzata su quei documenti che appaiono degradati a causa della sovrapposizione di "strutture" che interferiscono con il testo principale. A volte queste "strutture", la cui rimozione rappresenta un problema non del tutto risolto, costituiscono elementi di sicuro interesse, come nel caso di filigrane o di palinsesti, che è più opportuno analizzare piuttosto che rimuovere.

Si è proceduto con l'acquisizione delle immagini dei documenti danneggiati mediante una camera multispettrale che ha restituito immagini ad alta risoluzione nello spettro del visibile e del vicino infrarosso.

A tale fine sono stati utilizzati due differenti strumenti, le cui caratteristiche principali sono riassunte nella Tab. 1.

Tab. 1 – Caratteristiche principali degli strumenti utilizzati per l'analisi multispettrale

Caratteristiche	*Camera A*	*Camera B*
Trasferimento dati	interfaccia FDL-PCI	interfaccia Firewire + USB
Caratteristiche del sensore	Front Illuminated Full-Frame Architecture	CCD progressive scan image sensor
Risoluzione del sensore	758(H) ×516(V) Pixels	1360(H) ×1036(V)
Dinamica	14 bits	8/10 bits
Segnale/Rumore	70 db	56 db
Filtri	Rosso Verde Blu Vicino IR Riflessione in UV (400 nm) Banda stretta a 530 nm Banda stretta a 640 nm Banda stretta a 710 nm Falsi colori IR	RGB B W luminance Riflessione in UV Fluoresciensa in UV Vicino IR1 Vicino IR2 Falsi colori IR 1 Falsi colori IR 2
Obbiettivo	50 mm	23 mm
Fuoco	Manual-refresh rate 1Hz	Manual-Continuous refresh rate
Formato immagini	JPG,BMP,GIF TIFF-758×516 pixels, 96 pixels per inch 24/16M BW (1152Kb) 72/4.7×10^{21} Col (3456Kb)	TIFF-1360×1036 pixels, 96 pixels per inch 8/256 BW (1376Kb) 24/16M col (4128 Kb)

Si è riscontrato come l'aumento del numero delle bande spettrali corrisponde ad un notevole miglioramento della separazione dei testi nei documenti, non solo nei palinsesti, ma anche nei manoscritti contenenti effetti di trasparenza o assorbimento degli inchiostri dalla pagina retrostante.

Successivamente, al fine di rimuovere le degradazioni più frequenti, sono state applicate svariate tecniche di miglioramento dell'immagine, basate sui filtri anisotropici, morfologia matematica, decorrelazione dei colori ed altri metodi. Sono state sviluppate procedure semplici e veloci che consentono il miglioramento del testo principale e l'estrazione dal documento di testi nascosti o di particolari tessiture. Gli algoritmi, sviluppati dal CNR-ISTI e dall'ENST, sono stati implementati nel software PREOCR che rappresenta uno dei risultati principali del progetto, semplice da utilizzare anche per gli utenti non esperti, dal momento che il software consente l'applicazione di algoritmi molto specifici senza che ci sia la necessità di avere particolari conoscenze matematiche.

Tecniche di analisi statistica adottate

Questa fase della ricerca, condotta dall'ISTI per quel che concerne i metodi di decorrelazione e dall'ENST per quel che concerne i filtri anisotropici e la morfologia matematica, ha riguardato la descrizione matematica del fenomeno della sovrapposizione di testi e tessiture che si presenta frequentemente in documenti antichi e lo studio di tecniche basate sulla disponibilità di viste multispettrali per la loro separazione ed estrazione. Le degradazioni causate da tessiture complesse dello sfondo, infiltrazione e trasparenza dell'inchiostro dalla pagina retrostante, macchie, che interferi-

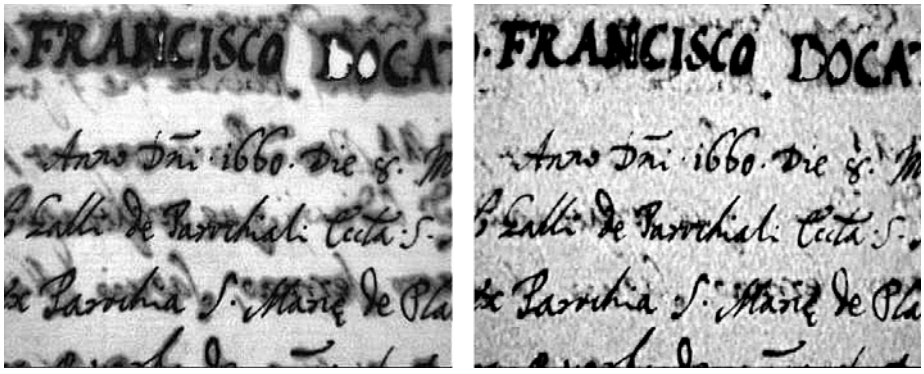

Fig. 1 – Esempio di rimozione di bleed-trough mediante trasformazione dallo spazio di colore RGB allo spazio di colore YES.

scono con il testo principale devono essere rimosse. Alternativamente, filigrane e testi precedenti parzialmente cancellati per un successivo riutilizzo del supporto possono ancora essere visti come strutture interferenti che però costituiscono oggetto di interesse di per se stesse e che devono quindi essere evidenziate ed estratte dal testo principale.

In entrambi i casi, per l'immagine digitale del documento (intesa in senso lato come l'insieme di osservazioni multiple, ottenute con modalità sensoriali diverse, del documento) viene proposto un modello di sovrapposizione (mistura) lineare di varie strutture, o classi, ciascuna caratterizzata da uno spettro di riflettività differente. Ogni classe è considerata uniforme e approssimata da un valor medio in ciascuna componente (canale) dei dati, e questi valori formano una matrice di "mistura" sconosciuta. Ogni pixel in un canale contiene il contributo dalle intensità locali di tutte le classi, moltiplicato per gli elementi di mistura corrispondenti. In formule, si ha:

$$x(t) = As(t) \qquad t = 1, \ldots, T \qquad (1)$$

dove x(t) rappresenta il vettore N-dimensionale della scansione multispettrale del documento al pixel t, s(t) è il vettore delle "quantità" delle M classi sovrapposte nel documento al pixel t, e A è la matrice NxM il cui elemento A_{ij} rappresenta l'indice di riflettività medio della classe i-esima alla j-esima lunghezza d'onda.

Un caso particolare ma di interesse pratico è quello in cui si hanno a disposizione tre viste (tipicamente i canali rosso, verde e blu di un'acquisizio-

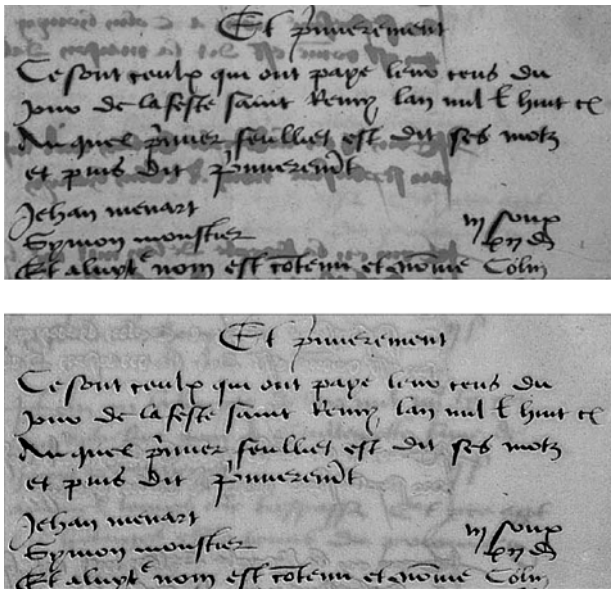

Fig. 2 – Evidenziazione del testo nascosto mediante ortogonalizzazione simmetrica

ne a colori nel visibile) e si presuppone la presenza nel documento di tre classi distinte (ad esempio la tessitura del supporto cartaceo, il testo principale e un solo testo interferente).

Sotto l'ipotesi di uniformità per la riflettività di ciascuna classe, e se gli indici di riflettività sono diversi da classe a classe, la matrice di mistura A sarà non singolare e, almeno in linea di principio, dovrebbe essere possibile recuperare le tre classi distinte applicando l'inversa di tale matrice al vettore RGB. Tuttavia, i colori medi di ciascuna classe sono non noti, e quindi occorre affidarsi a tecniche "cieche" o non supervisionate per separare, estrarre e classificare le diverse classi.

L'operazione di moltiplicare i canali del rosso, del verde e del blu per una matrice invertibile corrisponde alla proiezione in uno spazio di colori diverso, in cui particolari caratteristiche dell'oggetto ripreso possono risultare evidenziate. Infatti la rappresentazione RGB, sebbene sia la più utilizzata nell'elaborazione di immagini, presenta delle limitazioni in termini di massimizzazione del contenuto informativo dell'immagine e quindi in let-

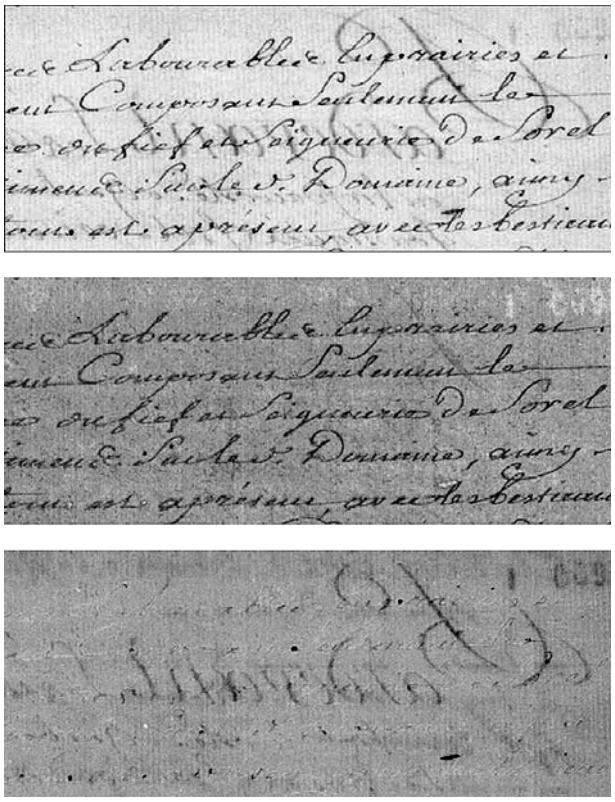

Fig. 3 – Separazione del testo fronte/retro mediante ortogonalizzazione simmetrica.

teratura sono stati proposti differenti spazi di colore, più adatti per compiti di analisi di immagini, quali segmentazione, rilevazione dei bordi, ecc. Un primo approccio adottato è consistito quindi nel verificare l'efficienza di alcuni spazi di colore noti sui documenti di ISYREADET. In particolare, gli spazi YES (Xerox, 1989, Knox et al., 1977) e OHTA (Ohta et. al., 1980) si sono rivelati di una certa utilità per la rimozione del bleed-through in immagini di documenti che si presentano rossastri per effetto della degradazione dell'inchiostro ferro-gallico.

Le trasformazioni da esse operate sul vettore RGB sono le seguenti:

$$\begin{pmatrix} Y(t) \\ E(t) \\ S(t) \end{pmatrix} = \begin{pmatrix} 0.253 & 0.684 & 0.065 \\ 0.5 & -0.5 & 0.0 \\ 0.25 & +0.25 & -0.5 \end{pmatrix} \begin{pmatrix} R(t) \\ G(t) \\ B(t) \end{pmatrix} \tag{2}$$

per la rappresentazione nello spazio YES, e:

$$\begin{pmatrix} O(t) \\ H(t) \\ T(t) \end{pmatrix} = \begin{pmatrix} 0.33 & 0.33 & 0.33 \\ 0.5 & 0.0 & -0.5 \\ -0.25 & 0.5 & -0.25 \end{pmatrix} \begin{pmatrix} R(t) \\ G(t) \\ B(t) \end{pmatrix} \tag{3}$$

per la rappresentazione OHTA, rispettivamente.

Tuttavia tali spazi di colore prevedono l'utilizzo di una matrice di mistura predefinita e fissa, indipendente dal documento in esame.

In linea di principio, se, come appare ragionevole supporre, le classi sovrapposte sono fra loro più "scorrelate" di quanto non lo siano i vari canali, si può ipotizzare che per ogni documento esista una matrice che proietta i canali RGB in un diverso spazio di colore in cui le tre classi appaiono separate. A questo scopo sono state utilizzate tecniche statistiche di separazione cieca e adattiva di sorgenti, che consentono la stima della matrice di mistura che meglio si adatta al documento sotto esame.

Per forzare la decorrelazione statistica sulla base dei dati disponibili, cioè i canali di colore, occorre stimare la matrice di covarianza e diagonalizzarla. Questo è equivalente a ortogonalizzare i diversi canali attraverso l'applicazione di un'opportuna matrice. Il risultato dell'ortogonalizzazione ovviamente non è unico. Sono quindi state valutate sperimentalmente le prestazioni di due strategie differenti, cioè di due diverse matrici di ortogonalizzazione.

La prima, quando applicata ai dati, produce uscite mutuamente ortogonali caratterizzate da massima varianza in ogni direzione principale (approccio alle componenti principali, PCA).

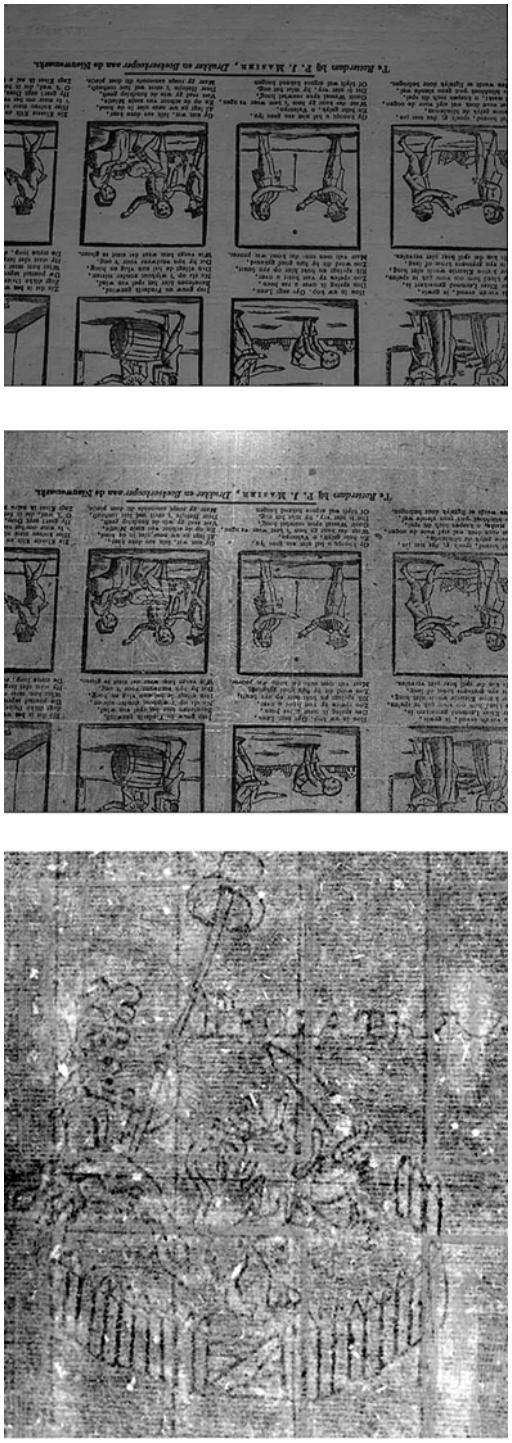

Fig. 4 – Evidenziazione ed estrazione della filigrana mediante ICA.

La seconda è una matrice simmetrica che produce un insieme di vettori ortogonali di norma unitaria che, rispetto agli output della PCA, sono ruotati attraverso la moltiplicazione, da sinistra, per una matrice ortogonale.

Un passo ulteriore è la cosiddetta analisi alle componenti indipendenti, o ICA, in cui, oltre alla decorrelazione, è richiesta la mutua indipendenza fra i canali in uscita (Hyvarinen, 2001, Tonazzini et al., 2004a)

Con questi approcci sono stati ottenuti risultati migliori di quelli forniti dalla proiezione in spazi di colore noti (Tonazzini et al. 2004b, Tonazzini et al. 2004c). È da rimarcare che tali tecniche sono completamente "cieche", in quanto non richiedono intervento da parte dell'operatore, e si adattano automaticamente allo specifico documento trattato.

Un ulteriore, significativo vantaggio di queste tecniche è anche da ricondursi alla possibilità di trattare problemi di dimensione anche superiore a tre, con l'unico vincolo che il numero di canali deve essere maggiore o uguale al numero delle classi. Questo consente un pieno sfruttamento di viste multispettrali/iperspettrali del documento sia nel visibile che nel non visibile, ad esempio l'infrarosso e l'ultravioletto, già di per sé in grado di migliorare la leggibilità di certi tipi di inchiostro e/o di attenuare particolari degradazioni come macchie di umidità. Analogamente, testi fronte e retro che risultano interferenti nel recto e nel verso di una pagina, acquisiti in scala di grigio, possono essere modellati come una mistura 2×2 e trattati efficacemente con tecniche di ortogonalizzazione simmetrica, equivalente, in questo caso, all'ICA (Tonazzini et al., 2006).

In alcuni casi gli algoritmi utilizzati, a differenza di altri proposti in letteratura e basati sulla vista del fronte e del retro della pagina, non richiedono la preliminare registrazione delle immagini.

La qualità dell'immagine può essere ulteriormente migliorata applicando tecniche basate su filtri anisotropici e sulla morfologia matematica.

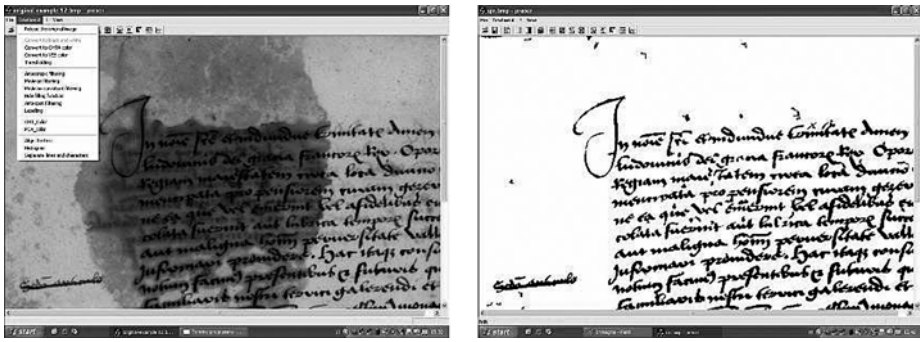

Fig. 5 – Applicazione della morfologia matematica: Rimozione delle macchie

I filtri a diffusione anisotropica discendono dalla teoria della diffusione dei fluidi e nell'elaborazione di immagini la loro applicazione consente di evidenziare aree omogenee dell'immagine, preservando e/o accentuando i contorni, determinando così il miglioramento della qualità della stessa (Perona e Malik, 1990).

La morfologia matematica (Matheron, 1975, Serra, 1982) è fondata sulla teoria degli insiemi e ha come obiettivo quello di esaminare la struttura geometrica di un'immagine al fine di rendere evidenti le sue connessioni topologiche con un elemento di confronto. Le connessioni dipendono dalla geometria della struttura da evidenziare e dalla sua posizione all'interno dell'immagine da esaminare.

Per estrarre informazioni da un'immagine binaria $A \subseteq E$ (dove E rappresenta l'insieme di tutte le possibili immagini di dimensione nota) si ricorre ad un'immagine più piccola B (detta elemento strutturante) e si applicano degli operatori che agiscono su ogni punto $b \in A$. Gli operatori fondamentali della morfologia sono la dilatazione e l'erosione.

Nel caso specifico si è fatto ricorso alla morfologia matematica per rimuovere le macchie presenti nell'immagine e "pulire" virtualmente il documento considerato.

Conclusioni

Il risultato principale dell'attività di ricerca è consistito nell'implementazione del software PREOCR, di cui attualmente è disponibile una versione prototipale corredata di manuale in italiano, inglese e francese, che può essere richiesta al coordinatore del progetto (T.E.A. sas, info@teacz.191.it). Questo applicativo consente di svolgere una serie di operazioni di trattamento delle immagini finalizzate al miglioramento delle stesse in funzione di un'eventuale applicazione di processi di OCR. Al momento il software lavora utilizzando esclusivamente immagini in formato bmp. Sono state implementate tutte le tecniche precedentemente illustrate e specificamente le trasformazioni spazio-colore (OHTA, YES), i metodi di decorrelazione (ORT, PCA), i metodi di sogliatura (Thresholding), le tecniche di filtraggio (filtro di diffusione anisotropica, filtro min/max, filtro di curvatura min/max, filtro hole filling, rimozione delle macchie). Tra gli obiettivi futuri sta l'ingegnerizzazione del software in modo da ottimizzare maggiormente le sue potenzialità e funzionalità.

Bibliografia

Xerox System Institute, 1989, *Color Encoding Standard*. Xerox Corp., Palo Alto, Calif. 1989, pp. 3-1.

Ohta Y., Kanade T., Sakai T., 1980, *Color Information for Region Segmentation Computer*. Computer Vision, Graphics and Image Processing,13 222-241.

Knox K., Johnston R., Easton R.L., 1977, *Imaging the Dead Sea Scroll*. Optics & Photonics News, Vol. 31, August.

Hyvarinen J. Karhunen J. Oja E., 2001, *Independent Component Analysis*. John Wiley, New York.

Tonazzini A., Bedini L., Salerno E., 2004a, *Independent component analysis for document restoration*. IJDAR, Vol. 7, pp. 17-27.

Tonazzini A., Salerno E., Mochi M., Bedini L., 2004b, *Bleed-through removal from degraded documents using a color decorrelation method*. Lecture Notes in Computer Science, Vol. 3163, pp. 229-240.

Tonazzini A., Salerno E., Mochi M., Bedini L., 2004c, *Blind Source Separation techniques for detecting hidden texts and textures in document images*. Lecture Notes in Computer Science, Vol. 3212, Image Analysis and Recognition, Part II, pp. 241-248.

Tonazzini A., Salerno E., Bedini L., 2006, *Fast correction of bleed-through distortion in grayscale documents by a Blind Source Separation technique*. IJDAR, published online 9 March.

Perona P., Malik J., 1990, *Scale-Space and Edge Detection Using Anisotropic Diffusion*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, No. 7. pp. 629-639. July.

Matheron G., 1975, *Random Sets and Integral Geometry*. John Wiley.

Serra J., 1982, *Image Analysis and Mathematical Morphology*. Academic Press, London.