Ancient documents bleed through evaluation and its application for predicting OCR error rates

V. Rabeux, N. Journet, J.P. Domenger

LaBRI Laboratoire Bordelais de Recherche en Informatique France (Bordeaux)

March 14, 2012

The project

Quality evaluation of very old document images by providing meta-data caracterizing a document's defects.

Why document image quality evaluation (instead of just restoration) ?

- Avoid (simplify) manual analysis of quality,
- Drive restoration algorithm, and avoid restoration of images that don't need it,
- Predict OCR error rates and other processes,

What this presentation aims to

ette de la chait de la plase on lon trent le marcie aux chenando vaches et pourceaulo et aultres lestes sant les la porte des tiligno. folio i son ette de la chait du terrie: que les habitans duncere ont fennt les la cono chulendesmas. folio donos paulo and a malo folio donos paulo and a malo and a malo folio donos paulo and a malo and a malo folio donos paulo and a malo and a malo and a malo folio donos paulo and a malo and a malo and a malo folio donos paulo and a malo and a malo folio donos paulo and a malo and a malo folio donos paulo and a malo and a malo folio donos paulo and a malo and a malo folio donos paulo and a malo and a malo folio donos paulo and a malo and a malo folio donos paulo a malo a malo and a malo folio donos paulo and a malo and a malo folio donos paulo a malo a malo and a malo folio donos paulo a malo a malo and a malo folio donos paulo a malo a malo donos paulo donos paulo donos paulo donos paulo donos paulo donos paulos paulo donos paulos paulos paulos paulos paulos paulos paulos paulo donos paulos paul

Figure: An old french document with bleed through

- Propose measures able to evaluate bleed through,
- Illustrate the measures accuracy by **predicting** the OCR error rate.

Measuring step by step

Our approach is composed of several steps :

- recto and verso registration,
- 2 identification of ink and bleed through pixels,
- Image in the second second

After the computation :

The bleed through's page is characterized by six measures.

The verso and recto images are binarized and registered :



★ Ξ →

The verso and recto images are binarized and registered :



V. Rabeux, N. Journet, J.P. Domenger Ancient documents bleed through evaluation and its application 5/1

• • = • • =

The verso and recto images are binarized and registered :



• • = • • =

The verso and recto images are binarized and registered :



V. Rabeux, N. Journet, J.P. Domenger Ancient documents bleed through evaluation and its application 5/1

.



Bleed through :

-

a.

Bleed through characteristics

cunangdifte lan mil taos case faivante drivnens pan lefquelles greatfron bonfart alose chyftelam Anice senora er a tenora laton quil anoir faitte en interettes ei lafortente contre lujo define oft faute menton, parmi lefquelles lettes e tenoraat infines dra dutte barroluton vivier fel didt fahr maio Julion i djeuelle fon en anversim er finness felles die son fel en latute panofte duite omnete en date lan mil nois cene forsa neuf le famet apse la fitte fame. Andres vante lefquelle lettere tenoraan er expositione fon latures andre ford A trefillutre Prince Henry de Nauarre, Françoys deferpt fon treshumble, & trefobeyffant feruiteur, Salut, & felicité perperuelle, Vous eftes deuemet

.

aduerty par la leçon des Littres lainétz (Prince trefullutre) que noz premiers peres effoyer vestus de fueilles & de peaux, pour couprir la nudité de leur corps seulement: mais peu à peu, croisfant duce l'aa-

Figure: Bleed through characteristics : a.**intensity**, b.**quantity** and c.**location**.

Metrics 1 and 2 : Bleed through intensity

Measures the bleed through intensity in relation to the background or ink.



Figure: OCR errors due to the bleed through intensity (green zones correspond to recognized text; red zones correspond to unrecognized text)

Metrics 1 and 2 : Bleed through intensity

Measures the bleed through intensity in relation to the background or ink.

• Distance to the ink :

$$\mathcal{MI}_i = \frac{\mu_{T_r} - \mu_{I_r}}{255}$$

- bleed through is close to ink : 0,
- bleed through is far from ink : 1.

Metrics 1 and 2 : Bleed through intensity

Measures the bleed through intensity in relation to the background or ink.

• Distance to the ink :

$$\mathcal{MI}_i = \frac{\mu_{\mathcal{T}_r} - \mu_{\mathcal{I}_r}}{255}$$

• Distance to the background :

$$\mathcal{MI}_b = \frac{\mu_{B_r} - \mu_{T_r}}{255}$$

- bleed through is close to ink : 0,
- bleed through is far from ink : 1.
- bleed through is close to background : 0,
- bleed through is far from background : 1.

4 3 6 4 3

Metrics 1 and 2 : Bleed through intensity

Measures the bleed through intensity in relation to the background or ink.



V. Rabeux, N. Journet, J.P. Domenger Ancient documents bleed through evaluation and its application 8/1

Third metric : the bleed through quantity

Measure the quantity of bleed through in relation to the quantity of ink.

porem ipsum dolor sit Pien Pharetra id. n

Figure: Variable bleed through quantity

Third metric : the bleed through quantity

Measure the quantity of bleed through in relation to the quantity of ink.

• $\mathcal{M}\mathcal{Q}$: bleed through quantity ratio,

$$\mathcal{MQ} = \frac{\|T_r\|}{\|I_r\|}$$

• $\mathcal{MQ} > 1$: more bleed through than text.

Third metric : the bleed through quantity

Measure the quantity of bleed through in relation to the quantity of ink.

porem i peum dolor sit
$$MQ = 1.73$$

pien pharetra id. II $MQ = 0.3$

Metrics 4, 5 and 6 : the bleed through location

Measure the bleed through location impact on letters.



Figure: Locations of a bleed through component.

Metrics 4, 5 and 6 : the bleed through location

Measure the bleed through location impact on letters.

• \mathcal{MA} : components added by bleed through,

$$\mathcal{MA} = \frac{\|\overline{\mathcal{TC}}\|}{\|\underline{I_v}\|}$$



• • = • • = •

• MA = 1: No verso component overlap recto's components.

Metrics 4, 5 and 6 : the bleed through location

Measure the bleed through location impact on letters.

• \mathcal{MS} : letters having their shape modified by a bleed through component,

$$\mathcal{MS} = \frac{\|\mathcal{TC}\|}{\|I_r\|}$$



伺 ト イヨト イヨト

• MS = 1: All recto's letters overlap with a bleed through component.

Bleed through measure

\mathcal{MA} and \mathcal{MS} example :

$$\mathcal{MA} = 1 , \mathcal{MS} = 0 \text{ (no overlaps)}.$$

$$\mathcal{MA} = 0, \mathcal{MS} = 1 \text{ (E and K does overlap)}.$$

伺 ト イヨト イヨト

э





Figure: MSG: measures the mean component expansion (in terms of component area).

• • = • • = •





Figure: MSG: measures the mean component expansion (in terms of component area).

• • = • • = •

Approach protocol

Metrics relevance :

Analyze bleed through metrics in correlation to the OCR error rate.

For a given OCR and dataset :

- Measures computation (6 measures / page),
- OCR runs and error rates computation,
- Inear regression multivariate and sequential (leads to a prediction model).
- Statistical validation of the model .

otent illeum dotot eff a nertion substanting faith ng dit, uioque acificie forem in Planetra esceta s tellus ciun laonet le o int aliquet cet metue et felies figuam In dun a usto plantra ornan ac ut maint i maint in orem iboum dolor sit amet conse ctetur adiliscing efit. t tincidun t sofficitudin enim nec Porttitor

orem Heam bolor eil amet consectetur addieseing effi, mie en erat en nunc tembor inderdiet a a figula, maditur blandit tincidunt d ofor a borta turbie bulbinar in, ulla fucture molectie nedue eget p ulbinar, ordi cureue bulbinar mi, fiduam nec erat esd facus fra Bida gradida ut et nulla, ordi plaretta egesta alfunet, n dofor t efficientificien ul ultricies da accumean en fidero, nteger teffne mi ulfamorber eget tincdunt at bibendum non miei, uiedue imferdie t acun at leo fundant trietibue, nece dismeien una eget aufine solatese.

orem iPsum bofor eit amet consectetur abiPiscin 9 efit, niedue fanci fue efementum eastitie- nae ut nuffa orci 5ief mafesuado teffinenae conceduat nune duis effenentum fodoritie fectus biaun m eem duis curene niei 5iefit eed fefie, niedue en aren non nief ac curnean babibue 5itat eit amet manie, neBenbisee Stetifulum nu na facinia Portitor Sidicate foe citt 5oftwara ante a bonne mi fisufa di maesa, nie Peffenteebue 5irStrat eem 5itat eastitie ea Pien Pharetta de n bac babitaese Platea dictumet, accenae consuc ornare fe

orem iPerum dalar sit 1100rem iPerum dalar sit amet consectetur ad 1100 and consectetur ad iPerum 2000 and a presenting sit. All an magna tortar, estis materia addissing rutar list faareet in grapho rum Porta, cas temba a sit consattise rae us suscitit qualitat a cent accumean rutar presentit qualitat a cent accumean rutar presentit qualitat a cent accumean rutar

V. Rabeux, N. Journet, J.P. Domenger

forem SulPutate in. nte ger non

Ancient documents bleed through evaluation and its applicatio 16/1

Results

Statistical model accuracy



Coefficient of determination $R^2 = 0.99$, Standard error (RMSE) = 7,5

ABBYY Coefficient of determination $R^2 = 0.97$, Standard error (RMSE) = 12,77

Statistical validation

Correlation coefficient between ground truth and the prediction :

Bleed through has a strong effect on the OCR process. High prediction accuracy if bleed through is the only defect.

0.99

V. Rabeux, N. Journet, J.P. Domenger Ancient documents bleed through evaluation and its application17/1

ABBYY

1.006

Perspectives :

Conclusion :

- We proposed measures in order to evaluate bleed through,
- We demonstrated that the bleed through has a strong effect on OCR results.
- The model accuracy shows the relevance of our measures.

Perspectives :

- We can not predict OCR results with just bleed through measures.
- Characters quality and noise evaluation, on very old documents.

Thank you !

V. Rabeux, N. Journet, J.P. Domenger Ancient documents bleed through evaluation and its applicatio19/1