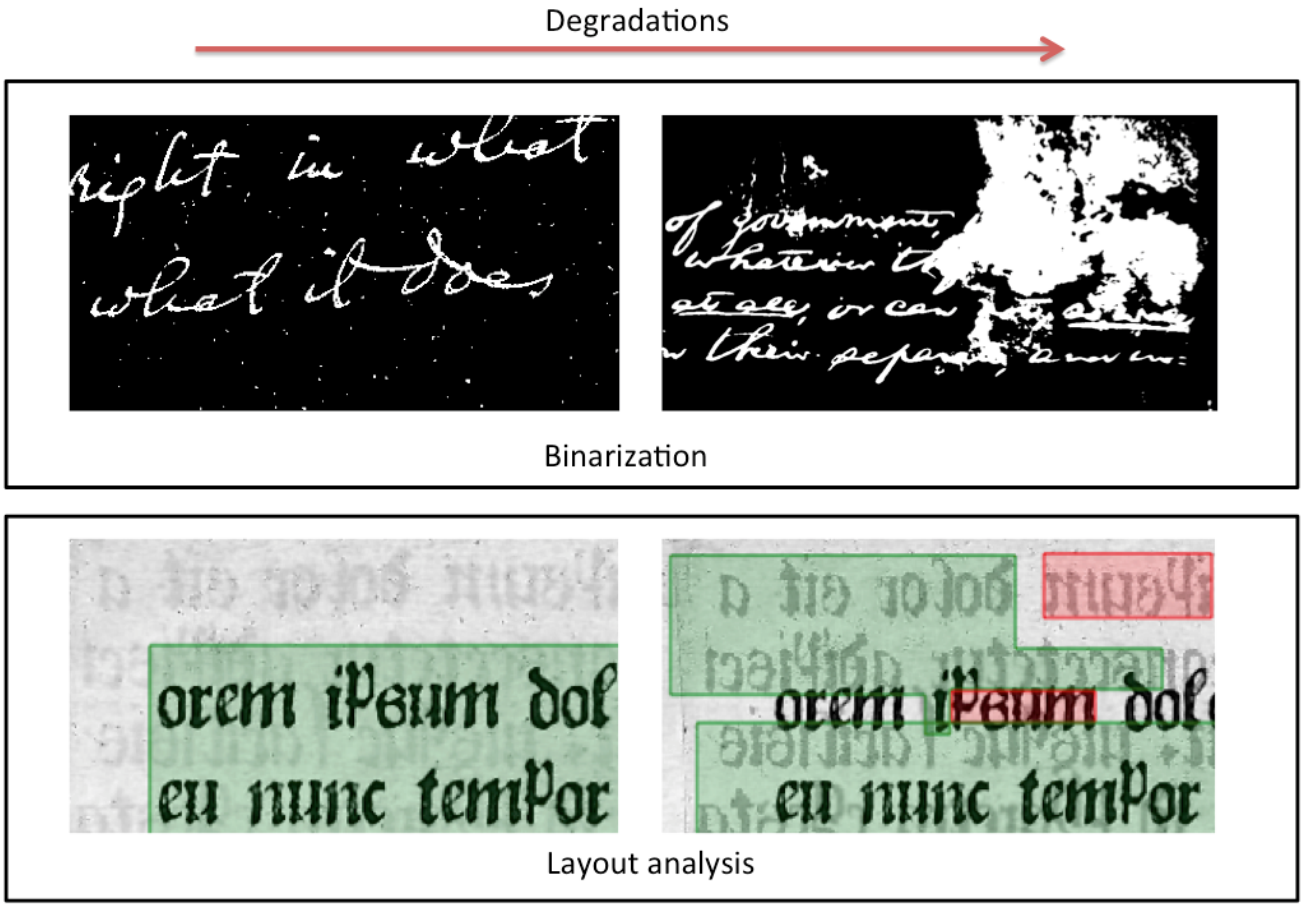
Quality evaluation of ancient digitized documents for binarization performances prediction

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

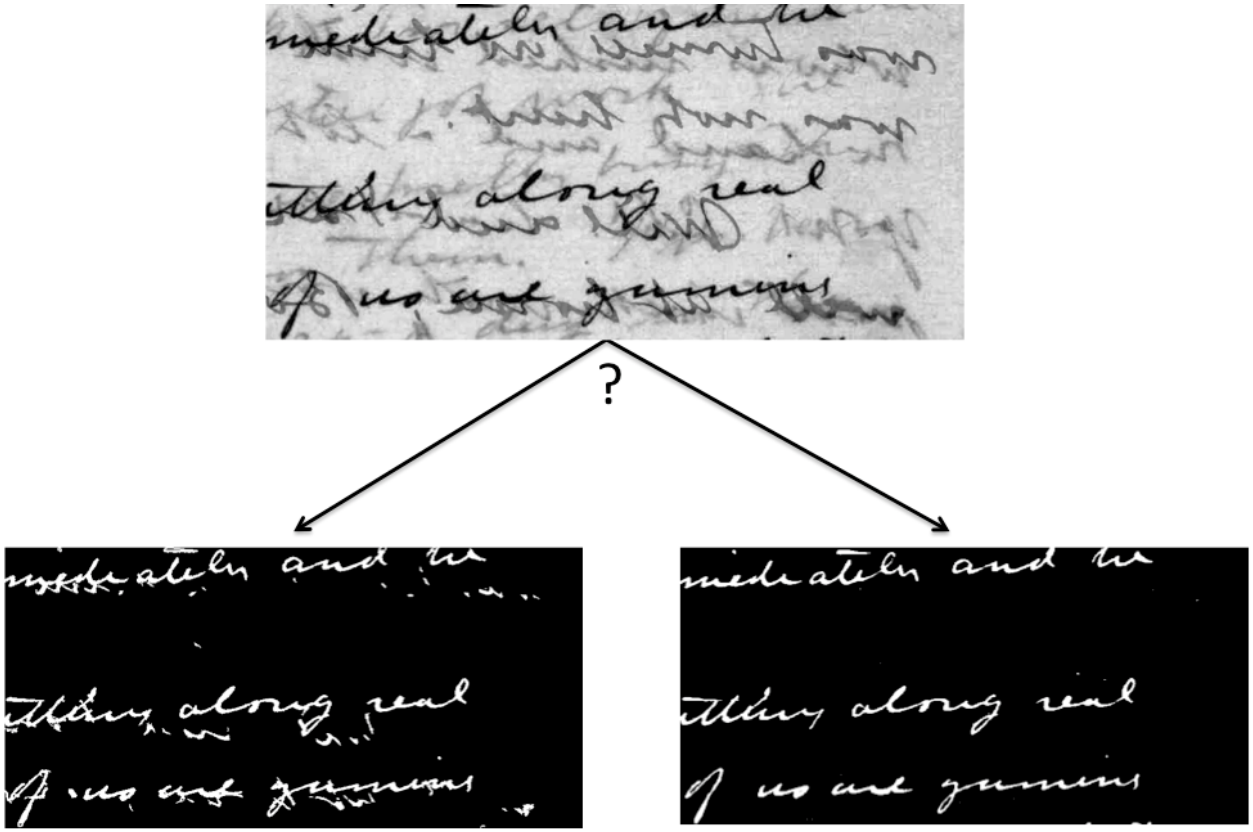
LaBRI (Laboratoire Bordelais de Recherche en Informatique)

August 26, 2013

The image quality impacts algorithms performances.



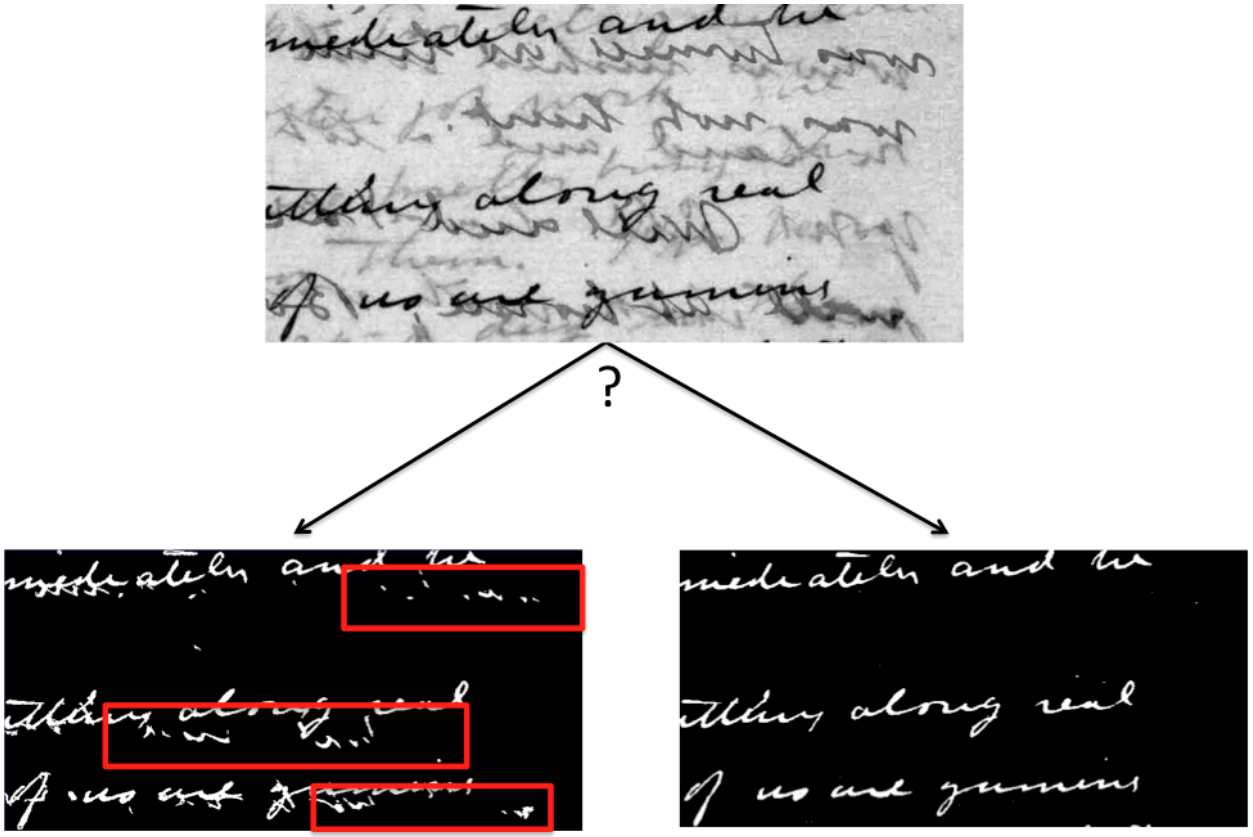
How to choose the best algorithm depending on the image degradation ?



Lu (Winner of ICDAR 2009)

Otsu

How to choose the best algorithm depending on the image degradation ?



Lu (Winner of ICDAR 2009)

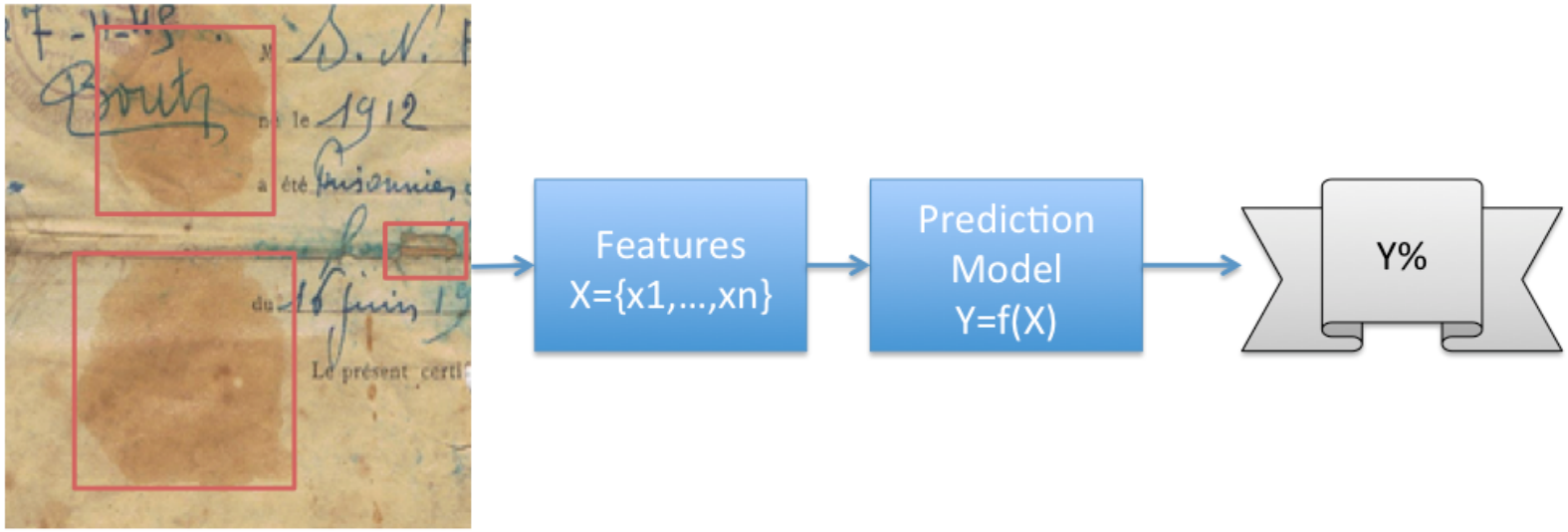
Otsu

Our approach

Predicting algorithm performance.

1. **Identify** and **characterize** degradations in the document image and create **dedicated features**.
2. Use the features to **predict** the algorithm performances.

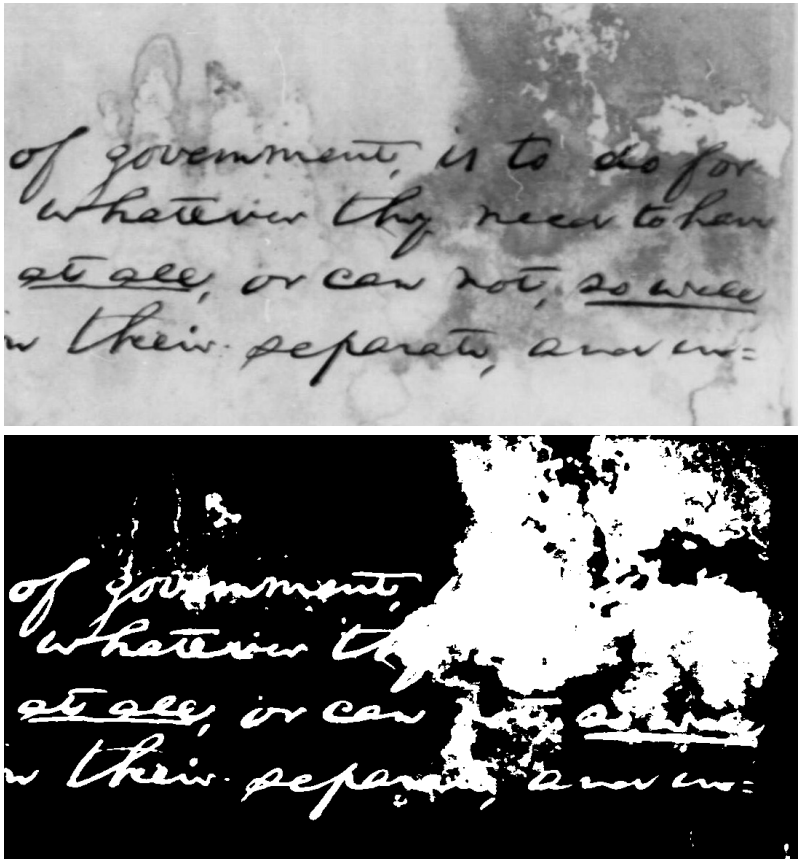
=> **Select** the most effective algorithm for each image.

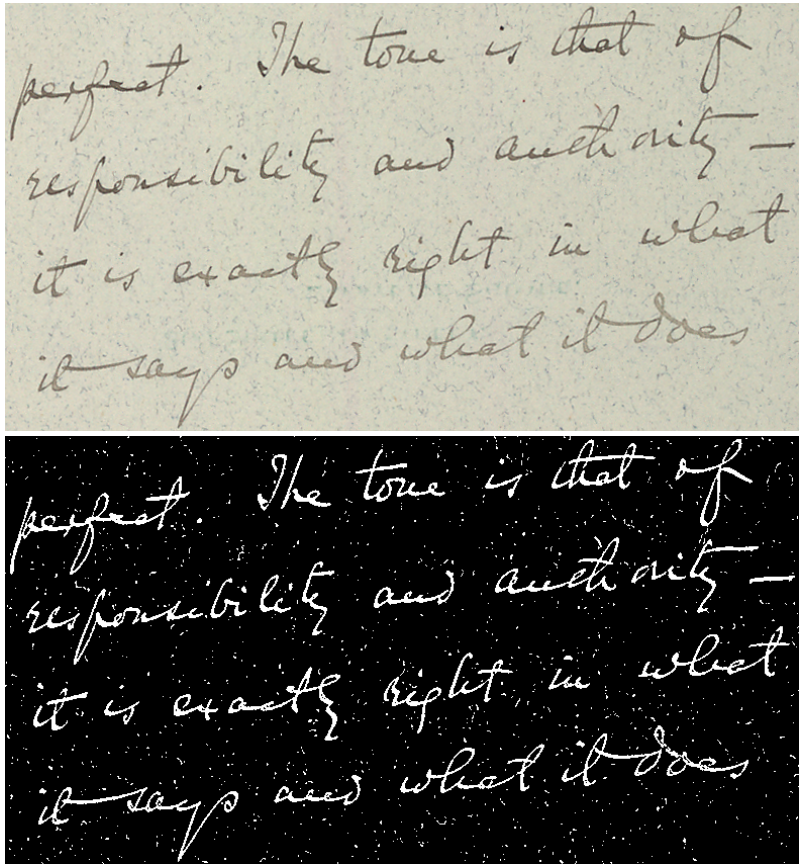


1. Indentify and characterize degradations
   - [Step 1] Algorithms errors and characterization of degradations.
   - [Step 2] Ink, degradations and background pixels extraction.
   - [Step 3] Features definition.

2. Algorithm performances prediction.
   - Prediction model creation and validation.
   - Predicting binarization methods performances.
   - Automatic selection of the best binarization method.

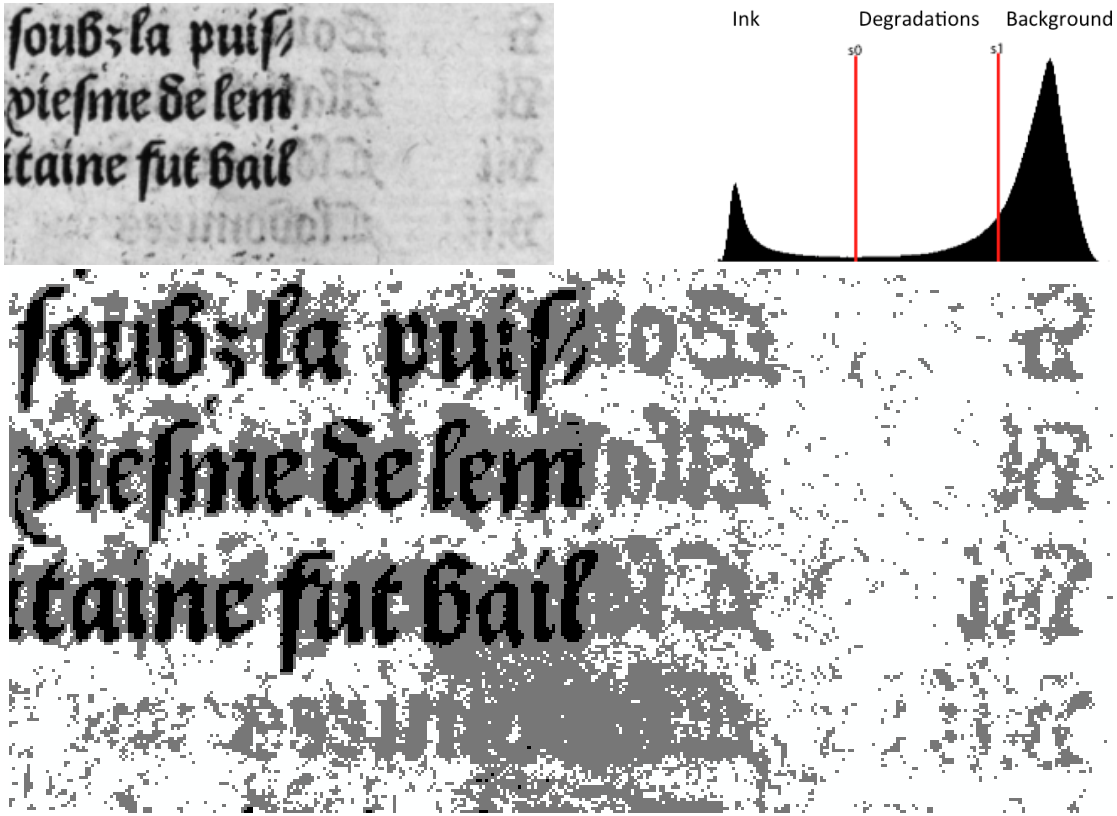
[Step 1] Algorithms errors and characterization of degradations.





[Step 2] Ink, degradations and background pixels extraction

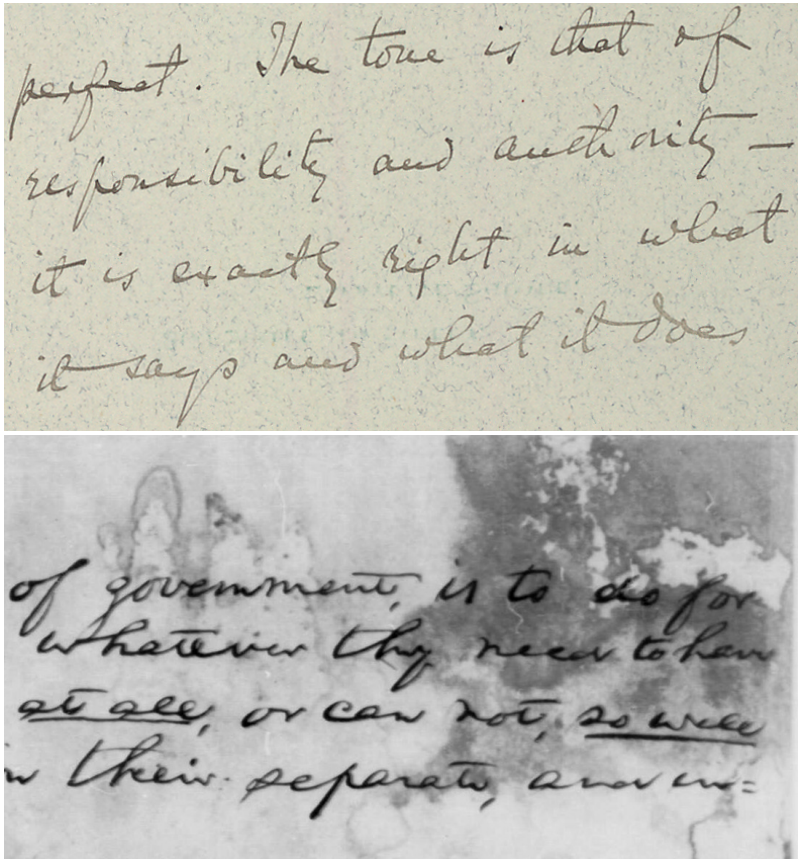
Extraction of 3 layers [MC09] : ink, degradations and background.

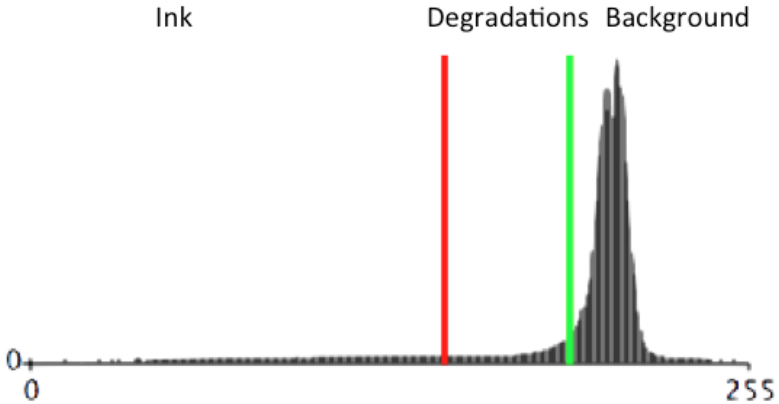


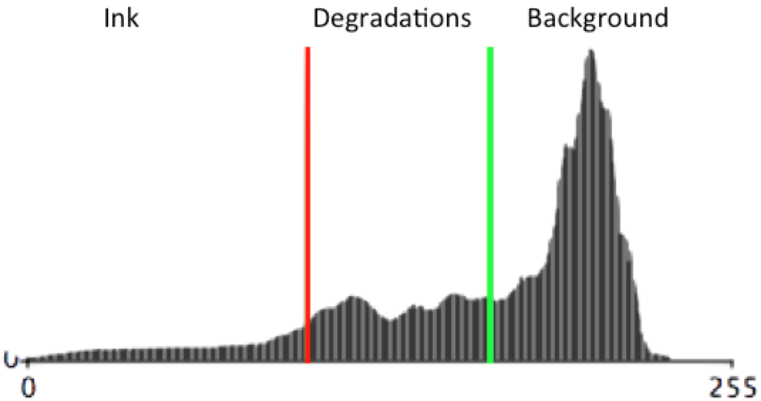
[Step 3] Features definition

15 Global features

Characterization of the overall distribution of the different layers.



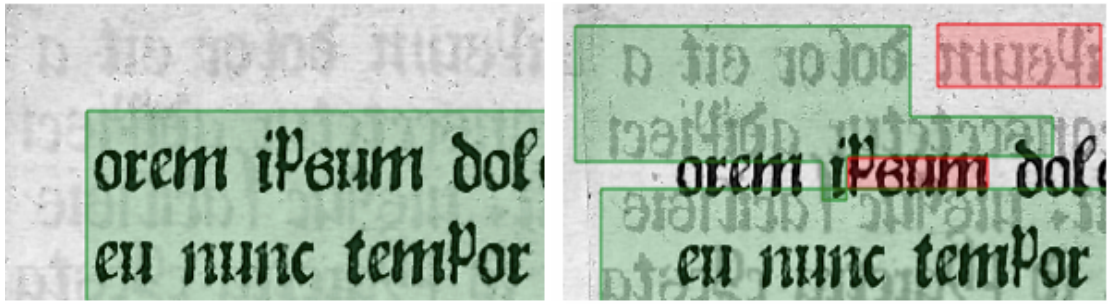




[Step 3] Features definition

15 Global features

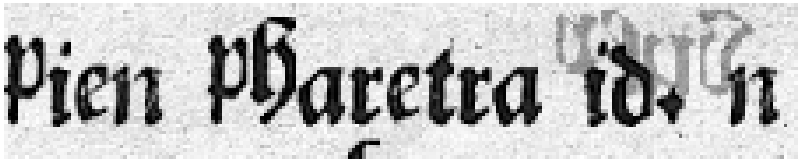
1. Distance between the ink average grayscale and the degradations average grayscale.
2. Distance between the degradations average grayscale and the background average grayscale.

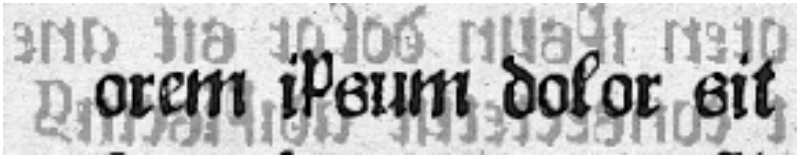


[Step 3] Features definition

15 Global features

The amount of degradation pixels (with proportion to the amount of ink).

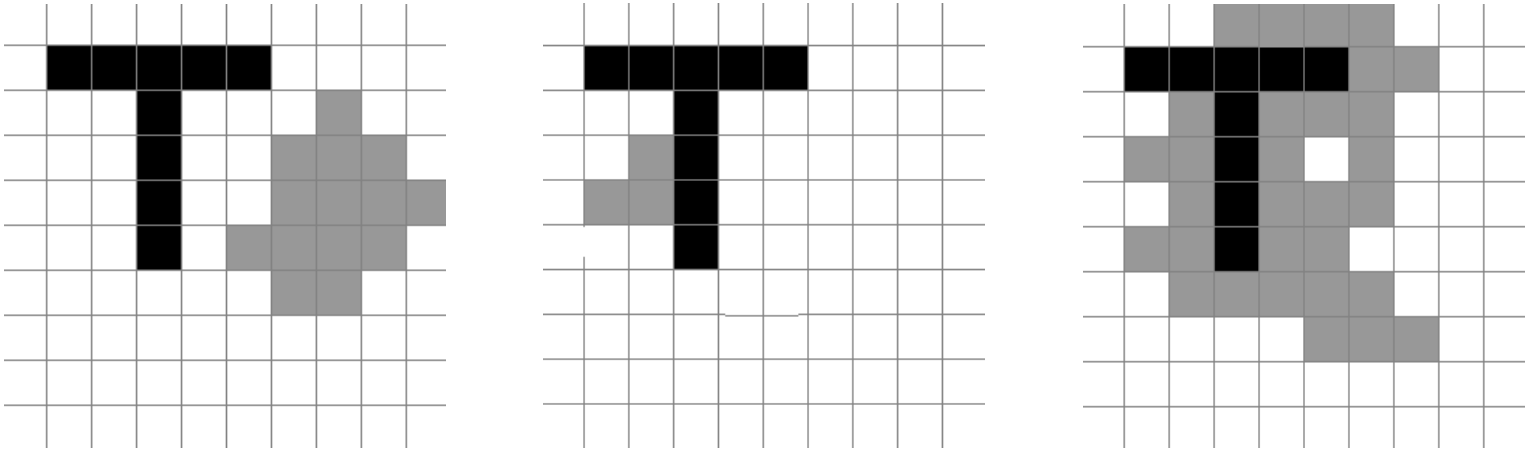




[Step 3] Features definition

3 Local features

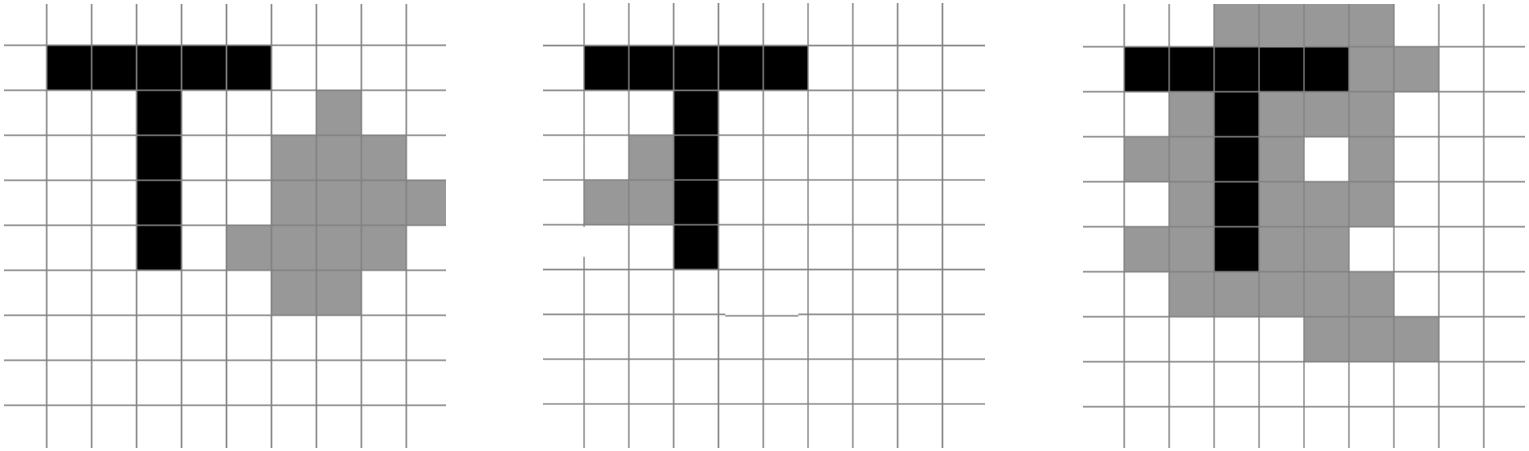
Localization of degradations pixels in regards to the localization of ink pixels.



[Step 3] Features definition

3 Local features

Localization of degradations pixels in regards to the localization of ink pixels.

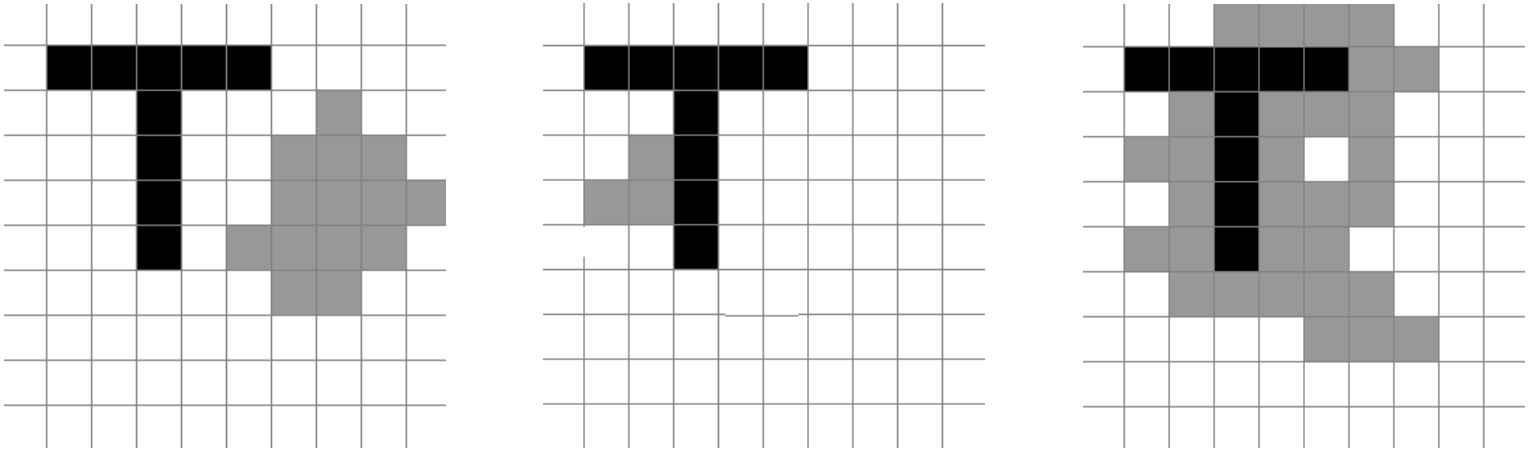


- Amount of degradation CCs *not connected* to an ink CC.

[Step 3] Features definition

3 Local features

Localization of degradations pixels in regards to the localization of ink pixels.

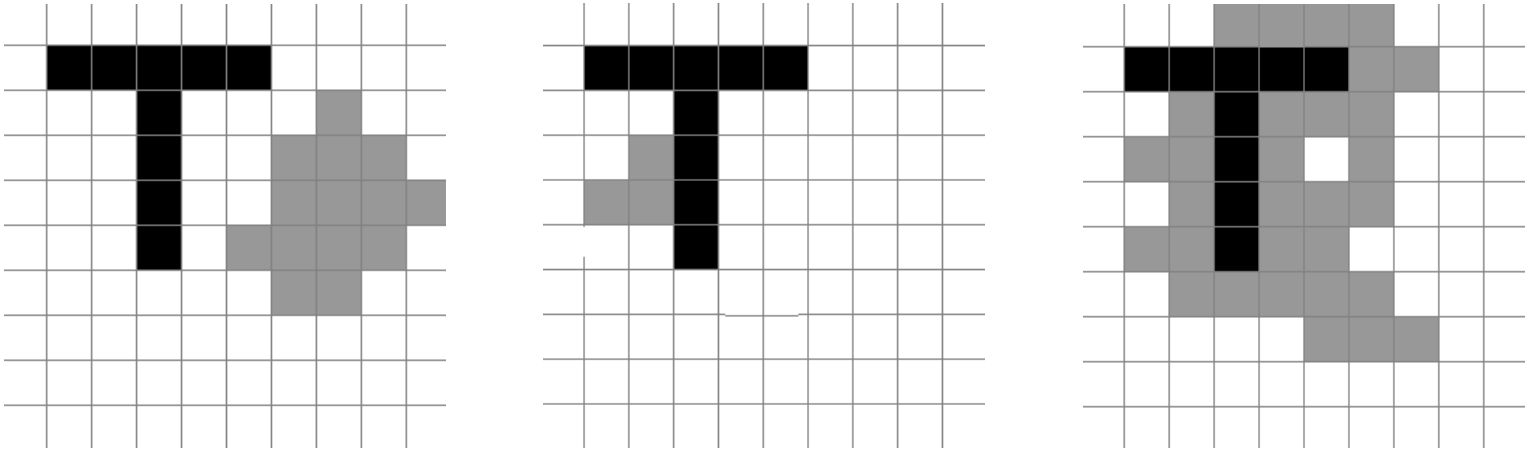


- Amount of degradation CCs *not connected* to an ink CC.
- Amount of degradation CC *connected* to an ink CC.

[Step 3] Features definition

3 Local features

Localization of degradations pixels in regards to the localization of ink pixels.

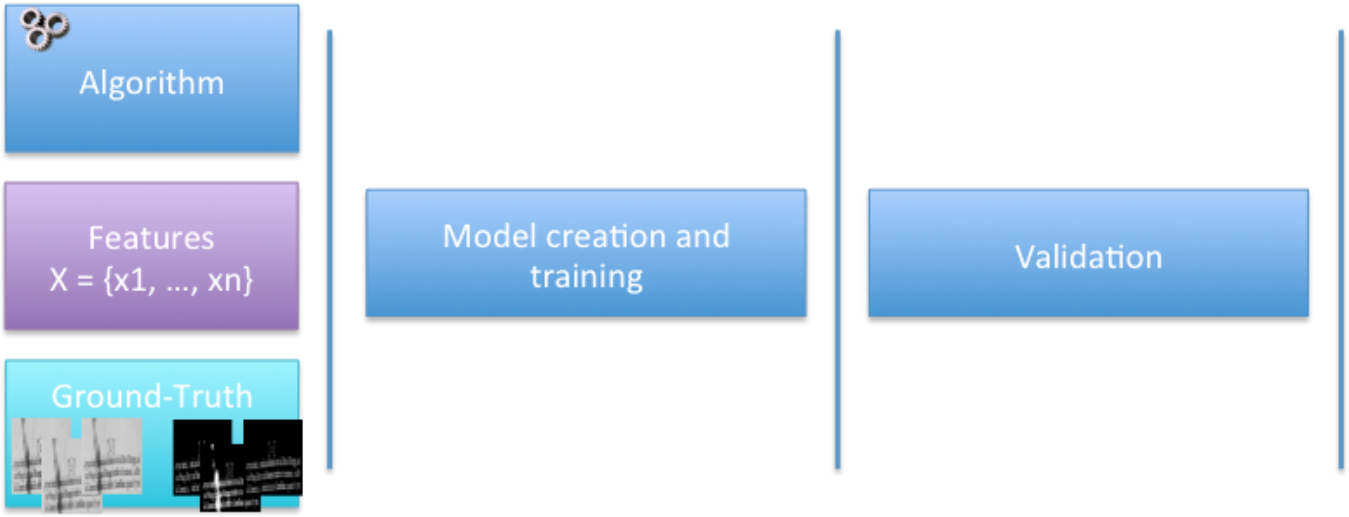


- Amount of degradation CCs *not connected* to an ink CC.
- Amount of degradation CC *connected* to an ink CC.
- Distortion of an ink CC (when connected to a degradation CC).

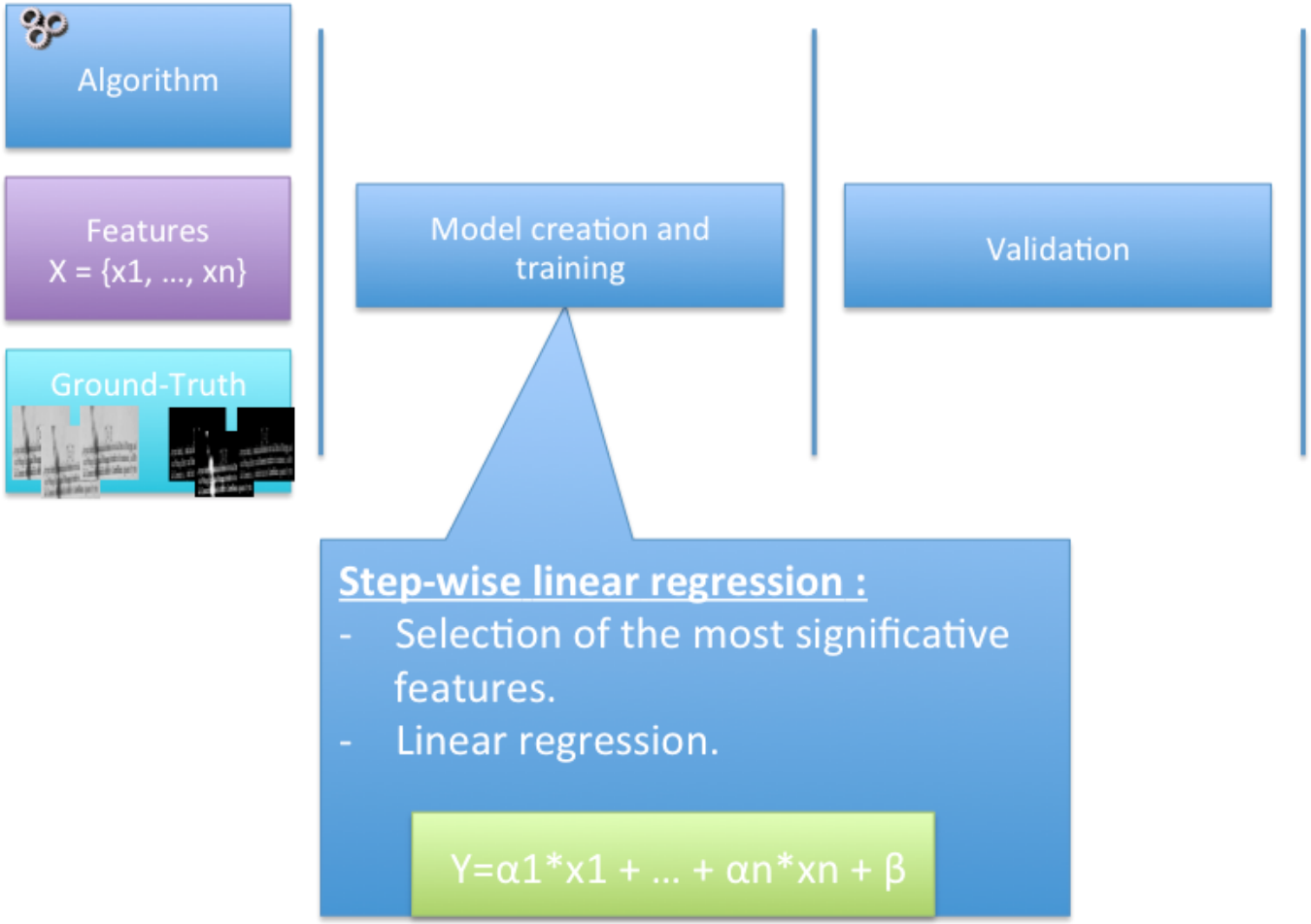
1. Indentify and characterize degradations
   - [Step 1] Algorithms errors and characterization of degradations.
   - [Step 2] Ink, degradations and background pixels extraction.
   - [Step 3] Features definition.

2. Algorithm performances prediction.
   - Prediction model creation and validation.
   - Predicting binarization methods performances.
   - Automatic selection of the best binarization method.

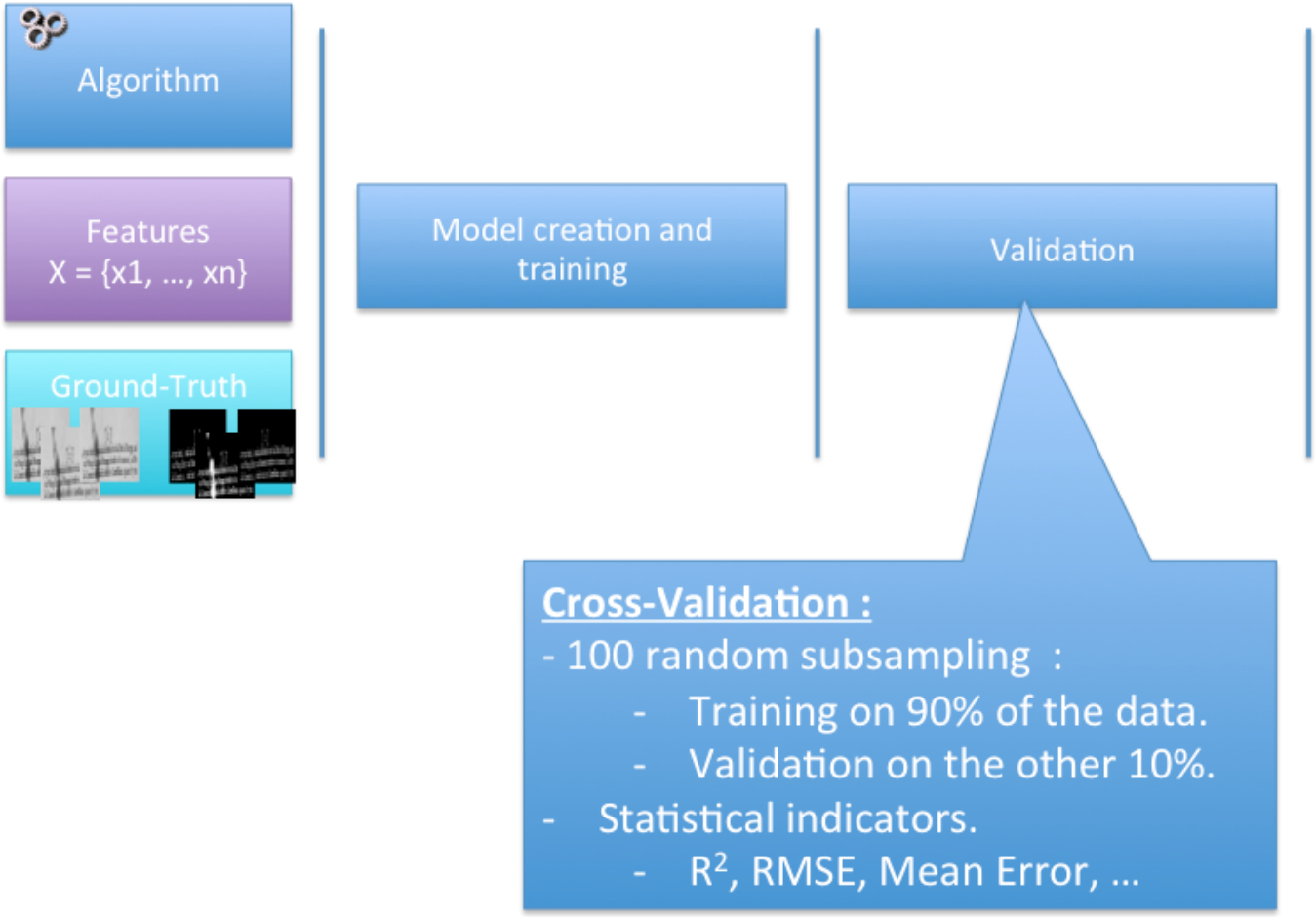
The overall workflow : Model creation and validation.



The overall workflow : Model creation and validation.



The overall workflow : Model creation and validation.



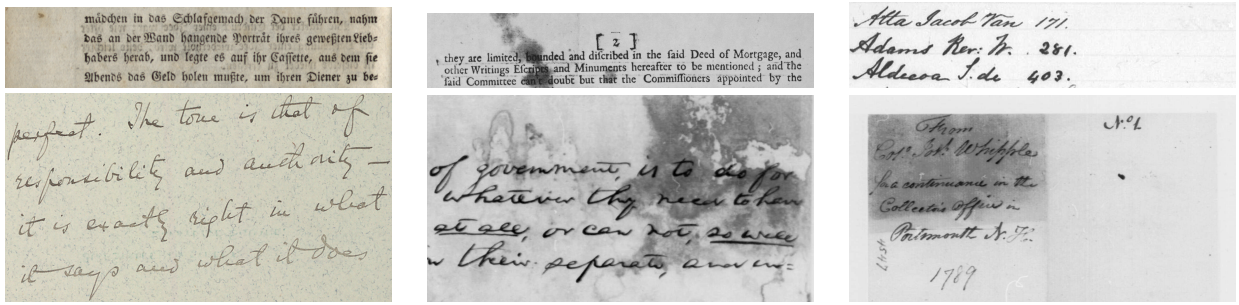
Selected binarization methods

11 binarization methods selected :

1. Globals :
   - Kittler [KI85], Otsu [Ots75], Ridler [C+78], Kapur [KSW85], Li [LT98], Sahoo [SWY97], Shanbag [Sha94]
2. Locals :
   - Bernsen [Ber86], White [WR83], Sauvola [SP00]
3. ICDAR 2009 winner : Lu [SLT11]

The training and validation dataset.

- Ground Truth (DIBCO & H-DIBCO) :



- 36 document images.
- Performances measured with the F-Score.
- Well distributed on the dataset and the set of binarization methods :
  - mean : 0.6; min : 0.1; max : 0.9.

Example : the Sauvola prediction model

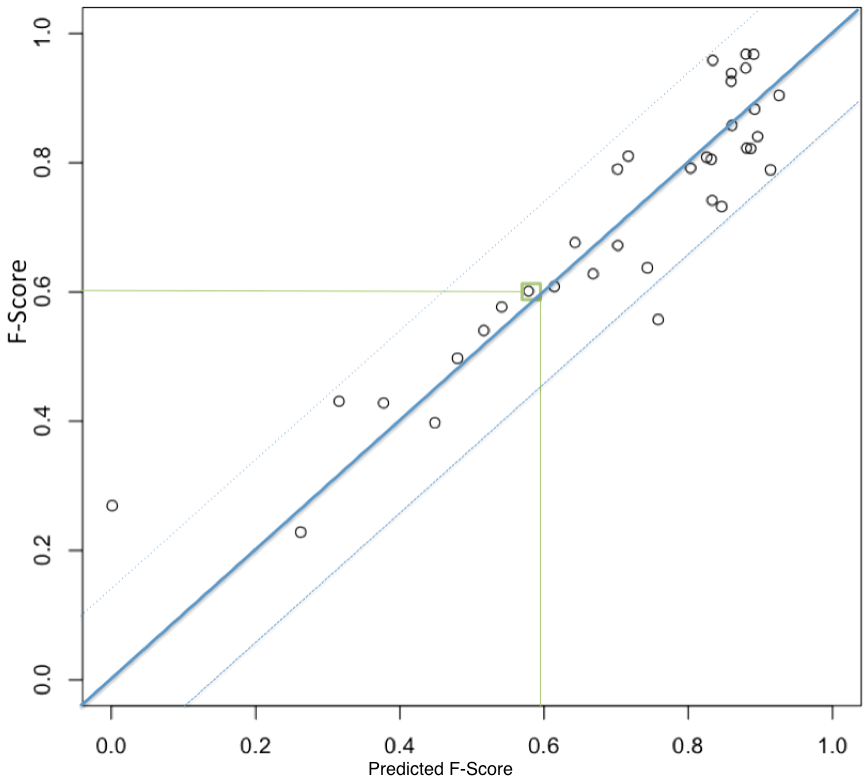
Selected features :

- Distance to the ink,
- Amount of degradations,
- Ink distribution,
- CCs not connected to the ink.

Cross-Validation (means) :

- R^2 : 0.99,
- Coefficient : 1.0007.
- Mean error : 10%

Object lesson : The Sauvola accuracy prediction model.



Prediction model results

Binarization method	Number of selected features	Mean Error
Sauvola	7	10%
Otsu	6	5%
Lu	7	4%
Bernsen	6	6%
Kapur	5	2%
Kittler	7	5%
Li	8	11%
Riddler	4	5%
Sahoo	6	5%
Shanbag	7	6%
White	7	7%

Prediction models accuracy

- Consistent selection of the most significant descriptors.
- About 5.6% of average error on the overall set of models.

Automatic selection of the best binarization method.

Lu (ICDAR 2009 Winner) :

Mean F-Score	Min F-Score
0.89	0.21

Automatic selection of the best binarization method :

Ground-Truth (best case) :

Mean F-Score	Min F-Score
0.91	0.77

Using the prediction models :

Mean F-Score	Min F-Score
0.90	0.61

Conclusion

- Close to the best case.
- **Good detection of difficult images**.

Thank you !

@vrabeux @AnrDigidoc

Fork us on Bitbucket :

https://bitbucket.org/digidoc
https://bitbucket.org/vrabeux/qualityevaluation

E-mail :

vincent.rabeux@labri.fr

J. Bernsen, *Dynamic thresholding of gray level images*, ICPR: Proc. Intl. Conf. Patt. Recog (1986), 1251–1255.

R.T.W. Calvard et al., *Picture thresholding using an iterative slection method*, IEEE Transactions on Systems Man and Cybernetics **8** (1978), no. Aug, 630–632.

J. Kittler and J. Illingworth, *On threshold selection using clustering criteria.*, Systems, Man and Cybernetics **15** (1985), no. 5, 652–654.

J.N. Kapur, P.K. Sahoo, and A.K.C. Wong, *A new method for gray-level picture thresholding using the entropy of the histogram*, Computer vision, graphics, and image processing **29** (1985), no. 3, 273–285.

CH Li and PKS Tam, *An iterative algorithm for minimum cross entropy thresholding*, Pattern Recognition Letters **19** (1998), no. 8, 771–776.

R.F. Moghaddam and M. Cheriet, *Low quality document image modeling and enhancement*, International journal on document analysis and recognition **11** (2009), no. 4, 183–201.

N. Otsu, *A threshold selection method from gray-level histograms*, Automatica **11** (1975), 285–296.

A.G. Shanbhag, *Utilization of information measure as a means of image thresholding*, CVGIP: Graphical Models and Image Processing **56** (1994), no. 5, 414–419.

B. Su, S. Lu, and C.L. Tan, *Combination of document image binarization techniques*, International Conference on Document Analysis and Recognition, IEEE, 2011, pp. 22–26.

J. Sauvola and M. Pietikäinen, *Adaptive document image binarization*, Pattern Recognition **33** (2000), no. 2, 225–236.

P. Sahoo, C. Wilkins, and J. Yeager, *Threshold selection using renyi's entropy*, Pattern recognition **30** (1997), no. 1, 71–84.

J.M. White and G.D. Rohrer, *Image thresholding for optical character recognition and other applications requiring character image extraction*, IBM Journal of Research and Development **27** (1983), no. 4, 400–411.