Language Independent Thresholding Optimization Using a Gaussian Mixture Modelling of the Character Shapes

Yves Rangoni[1] Joost van Beusekom[2] Thomas M. Breuel[1,2]
[1]Image Understanding and Pattern Recognition (IUPR) Research Group
German Research Center for Artificial Intelligence (DFKI) GmbH
[2]Technical University of Kaiserslautern
D-67663 Kaiserslautern, Germany
{rangoni,beusekom,breuel}@dfki.uni-kl.de

ABSTRACT

One of the first steps in a digitization process is the binarization of the document image. The major further steps like layout analysis, line extraction, and text recognition assume a black and white image as input. Several thresholding methods have been proposed to handle this problem for document images, but few of them take into account the behaviour of the text recognizer. They often rely on parameters that depend on the class of documents. In a large-scale process, neither relying on empirical assumptions nor using a manual tuning is conceivable.

In this paper, we introduce statistical modelling of a suitable binarization for a character recognizer. The model is a mixture of Gaussians that gives the prior of a binarization for having the best suitable transcription afterwards. The training is done on the character level, and tuned specifically for the recognizer. The optimization consists in finding the binarization that produces the best character shapes according to the model.

As opposed to existing methods, the optimization is goal-directed, and is not linked to subjective visual criterions. On the one hand, our method uses high-level character shape information to improve preprocessing, resulting in a language independent system. On the other hand, it can be trained in an unsupervised way, significantly reducing the need for human intervention. We demonstrate the effectiveness of this approach, called *Gaussian Mixture Token Thresholding*, on a subset of the Google 1000 Books dataset containing old documents where we achieve an improvement of more than 10 points compared to a regular binarization.

1. INTRODUCTION

The goal of document binarization is to convert a given greyscale or colour document image into a bi-level representation. The underlying objective is to separate objects, like characters, from the background with the assumption that grey levels of pixels belonging to the two classes are substantially different.

The quality of the binarization plays a major role in document recognition. Indeed, most of the algorithms used during analysis (page orientation, layout analysis, character recognition, etc.) expect a black and white image, and rely on a suitable output of the binarizer. Typically, some binarizations remove the noise and mostly keep the characters, but cause a lot of broken and extremely thin characters (Fig. 1). On the other hand, some other ones may not cut the characters into small pieces when fonts like Times are used, but generate bold-faced and touching characters (Fig. 1). In old documents, the writing from the verso and the noise of the recto appear when the background is complex and the quality of the paper document is poor (Fig. 1).

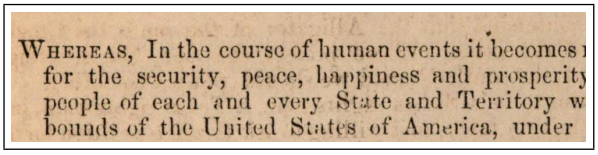


Figure 1: Initial colour image

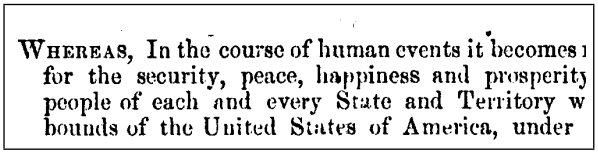


Figure 2: Sauvola's binarization $(W, K) = (40, 0.2)$

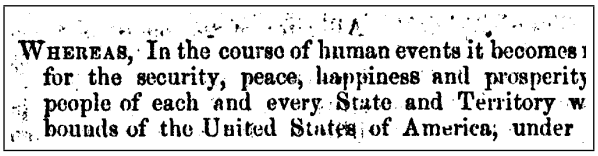


Figure 3: Sauvola's binarization $(W, K) = (40, 0.1)$

Many methods have been proposed, but few of them have been tested on a large heterogeneous and public dataset, and are still not widely used in practice. Although some techniques have been proposed in the literature, regular techniques as Sauvola, Niblack and histogram based thresholding are mostly preferred even for ancient documents, where having a good binarization is still challenging.

MOCR '09, July 25, 2009 Barcelona, Spain

In this paper, we introduce a fully automatic method for finding the best parameters of a well-known binarization technique in order to optimize the performance of a targeted OCR system. The experimental part focuses on the Sauvola's technique, but the proposed framework is still fully valid for any kind of binarization method. We propose a statistical modelling of well-binarized images for a specific line recognizer. The underlying objective is to feed the character recognition engine with the most suitable characters.

A Gaussian mixture is employed to model the well-binarized characters. It is trained with document images that produce a high quality transcription. The Gaussian mixture is then able to estimate the likelihood of a good thresholding for a new page. In fact, the method works on the connected component level (the tokens) and does not make any use of knowledge about the document. It results in a language independent method. We named it *GM-token thresholding*.

The paper is organized as follows. Section 2 describes an overview of the related work in textual document image binarization. Section 3 details the steps of the *GM-token thresholding* method. Then, experiments in section 4 show how it performs on a public dataset of ancient documents. Finally, conclusions and perspectives will be discussed in the last section.

2. RELATED WORK

Several approaches for binarizing greyscale or colour documents have been proposed in literature [15]. They can be broadly divided into global and local methods. Global binarization methods try to find a single threshold value for binarizing the whole page. Each pixel in the document image is assigned to page foreground or background based on its grey value. Global methods are computationally inexpensive and they give good results for office-scanned documents. However, if the illumination over the document is not uniform, i.e. in the case of camera-captured documents, they fail to binarize the document. Local methods try to overcome this problem by computing thresholds for each pixel individually, using information from the local neighbourhood of the pixel. They are able to achieve good results even on severely degraded documents, but they are often slow since the computation of image features from the local neighbourhood is to be done for each image pixel. Without describing in detail all the key points and drawbacks of each of them, several comparison surveys like [19, 13, 15, 10, 2] suggest that Sauvola's binarization method [14] outperforms the other local thresholding techniques; whereas Otsu's method [11] works best among the global techniques.

Most of the binarizer, as the well-established Sauvola's method, depends on parameters that are influencing its performance greatly. Correct parameter values are not straightforward to find, especially for most of the local techniques. Usually, subjective evaluations employ humans who tune the parameters according to their perceptual impression. This kind of manual procedure can be sufficient for a small office application, but it is not suitable for achieving at the same time a high accuracy and a fast processing for a large range of heterogeneous documents. For a high volume scanning environment, an automatic method is required.

Some techniques, as proposed by [21], [7] or [2], have been developed to obtain these right parameters by optimizing a criterion, i.e. an edge detection, which should quantify how suitable a thresholding is. There are at least two main drawbacks in such a technique. First, the criterion to optimize does not necessary imply that the resulting settings will be the best for the recognizer [9]. Second, the most advanced ones need also initial settings or initial assumptions to work. Even if techniques using a training [22] could tackle this problem by learning them, they belong to the same class of document and are less efficient when used with unseen documents. On top of that, the overhead of extra computations is sometimes not justified compared to the gain of performance.

3. GM-TOKEN THRESHOLDING

3.1 Overview of the method

The token driven thresholding technique with a Gaussian mixture modelling (GMTT) uses character similarity to evaluate the right binarization for a document image. The high-level idea is to model the characters of an image that produce the best results with the text recognizer engine (let call it OCR). The training of a Gaussian mixture on perfectly binarized images creates a yardstick dictionary that is employed during the estimation step. Several binarizations are computed for a new input image and the correct one is the one that outputs the most similar characters to the dictionary.

The main advantage of this approach is the use of an unsupervised training. There is no problem of transcription ground-truth generation, and the learning is performed for a character engine, and not for a particular class of document. When the model is trained, it can be applied on any kind of document. If the binarization step really fails to find a good thresholding, it means that the document is not appropriate for the OCR, e.g. Arabic script used with a Latin OCR. This adaptive thresholding can both produce an optimized thresholding for the targeted OCR, and also warn other intermediate processes (layout analysis, line extraction, etc.) that the document needs a specific processing.

By looking only at connected components instead of character, we produce at the same time a very time efficient and a language independent method. Of course, as the method uses high-level character shape information to improve preprocessing, it is not directly script independent. But as the OCR engines work with a fixed and known alphabet, the generation of a good dictionary is easy to set out. Last, but not least, most of the computations needed during the estimation of the thresholdings are required, or at least useful, and valid for further steps of the recognition flow.

The method needs first to evaluate and model a good binarization for a document. Given an image X, binarized with a function b, composed of n connected components x_i. The score s of $b(X)$ knowing a correct binarization B^* is approximated by (1):

$$s(b(X), B^*) = \sum_{x_i \in b(X)} s(x_i, B^*) \quad (1)$$

We chose a Gaussian mixture as a probability density function to model s. The mixture is a linear combination of k Gaussians $\mathcal{G}_j$ with their weights w_j. Each Gaussian of a normal distribution $\mathcal{N}_j$ with means μ_j and standard

deviations σ_j (2):

$$s(x_i, B^*) = \sum_{j=1}^{k} w_j \mathcal{G}_j(x_i) = \sum_{j=1}^{k} w_j \mathcal{N}(x_i, \mu_j, \sigma_j) \quad (2)$$

We have two objectives now: Finding a good model for a perfect thresholding B^*, and finding what a perfect thresholding is, relative to the OCR output quality. When solved, the GMTT consists in finding the best binarization b_i produced by several binarizers $\{b_1, \ldots, b_m\}$ where i is defined by (3):

$$\underset{j}{\operatorname{argmax}}\, s(b_j(X), B^*) \quad (3)$$

3.2 Construction of a ground truth

The training of the Gaussian mixture requires perfectly binarized characters. A ground-truth dataset is then needed. We want to model the character shapes that are suitable for the recognizer. The first straightforward idea is to generate synthetic characters. It is a good start but it might be suboptimal since the OCR can behave better on other shapes. We decided to feed the training with real characters, coming from scanned documents. Starting from greyscale images, our approach tries different binarizations and finds which black and white images produce the best OCR results. It implies running the OCR on every binarized image. If a transcription ground truth is provided using an edit distance based scoring allows to rank the binarizations. We chose a more generally applicable approach which consists in evaluating the ratio of Words In Dictionary *wid* of the transcription given by the OCR. Producing the best binarization b^* of X is equivalent to solve (4):

$$b^*(X) = b_j(X), j = \underset{i}{\operatorname{argmax}}\, wid(b_i(X)) \quad (4)$$

If the set of b_i is finite, an exhaustive search is enough. Applied on several X, we can generate several well-binarized images $b^*(X)$ (like Fig. 4), to feed the ground truth with black and white pages of suitable characters and having a model for B^*.

3.3 Learning the model

The Gaussian mixture model is trained by the expectation-maximization algorithm [6]. In order to have feature vectors of the same size, the connected components are resized into a fixed square image. We call blobs the resized connected components and continue to name them x_i. Not all of the connected components are taken into account to create the blobs. Firstly, a filtering is performed by looking at valid blobs. They are said valid when they have similar widths and heights to the medians of all the components. They have to be also not too tiny (less than 8 pixels), otherwise they are filtered out.

It results in blobs representing almost only "real" characters. Once this set of good-looking blobs is created, the training can be performed. We started the training with a synthetic dictionary of filtered blobs: The μ_j contain all the characters recognizable by the OCR (Fig. 3.3), the initial σ_j matrices are set to an identity matrix (and stay to diagonal ones).

After convergence of the EM algorithm, we obtain the final means (Fig. 3.3) and standard deviations (Fig. 3.3). The weights of the mixture are determined after the training. They are set to the frequencies of the ground-truth blobs that are the closest to the corresponding Gaussian, so that high weights are associated with frequent blobs. At this point, (1) can be evaluated for any binarization.

3.4 Optimizing Sauvola's binarization

In the experimental part, we are using the Sauvola's method [14]. The threshold T for each pixel (i, j) in the image depends on local variance of the pixel neighbourhood (5). The local mean $m_W(i, j)$ and standard deviation $\sigma_W(i, j)$ are computed on a window of size $W \times W$ around the pixel with bias K.

$$T_{W,K}(i, j) = m_W(i, j) \cdot \left(1 + K\left(\frac{\sigma_W(i, j)}{128} - 1\right)\right) \quad (5)$$

The main drawback of the method is to set correctly (W, K). Sauvola et al. [14] propose $(15, 0.5)$. Sezgin et al. [15] or Trier et al. [18] have found $(15, 0.2)$ to work better. Another study by Badekas et al. [1] suggested yet another value pair of $(14, 0.34)$. Even if some settings are valid on average, they are not optimal for each image and depend on the targeted application like in our case OCR. Our set of binarizers b_i will be Sauvola binarizers with different values for W, K.

3.5 Estimation of the best binarization

Once the training is done, we have a model that can estimate if a black and white image contains likely connected component for the OCR. If the training was done on a heterogeneous corpus the method is optimized for a script but it is language independent. Solving equation (3) is now straightforward. For a grey image X, each binarizer b_i is applied and the Gaussian mixture is evaluated on the blobs of $b_i(X)$.

Concretely, we are testing equation (5) on X with different values for W, K, the b_i are representing now different couple W, K). We assume having no knowledge about the behaviour of (W, K), and the variables are considered as discrete. We make use the integral images [16] to reduce the runtime, in such a way that the binarization a page is fast and most of the computations done for one W are valid for any K.

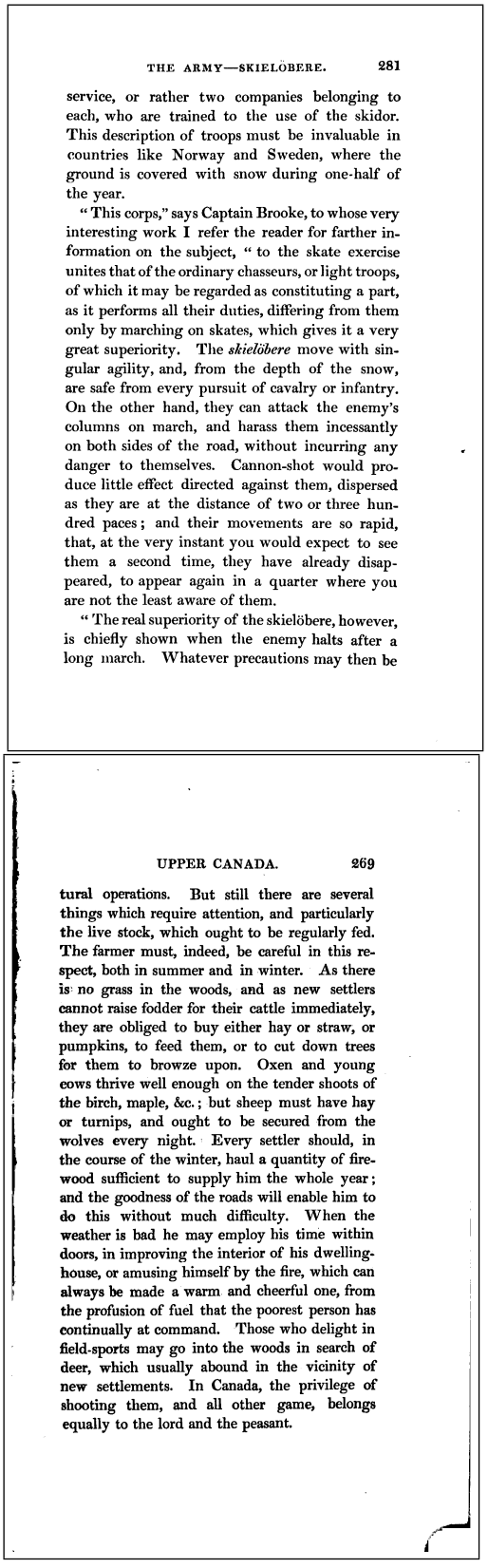
As different binarizations may produce different connected components, the filtering outputs different numbers of candidates. We retain the ratio $rej(b_i(X))$ of rejected components with the total number of connected components. The new objective function derived from (1) is (6), it penalizes binarizations producing too many small components.

$$s(b_i(X), B^*) = rej(b_i(X))^2 \prod_{i=1}^{n} s(x_i, B^*) \quad (6)$$

4. EXPERIMENTS AND RESULTS

There are a few large and publicly available datasets with greyscale images of complex documents. To evaluate the approach in a challenging application area, we chose the Google 1000 Books dataset [20], which contains scans of old books. From the first 770 volumes, each inner page has been picked. After removing the blank pages, the final subset contains 740 documents. The original images are in colour, we convert them in greyscale by averaging the three colour channel values.

The text ground truth given with the G1000 contains many errors, especially for pages where optimizing the bina-



THE ARMY—SKIELÖBERE. 281

service, or rather two companies belonging to each, who are trained to the use of the skidor. This description of troops must be invaluable in countries like Norway and Sweden, where the ground is covered with snow during one-half of the year.

"This corps," says Captain Brooke, to whose very interesting work I refer the reader for farther information on the subject, "to the skate exercise unites that of the ordinary chasseurs, or light troops, of which it may be regarded as constituting a part, as it performs all their duties, differing from them only by marching on skates, which gives it a very great superiority. The *skielöbere* move with singular agility, and, from the depth of the snow, are safe from every pursuit of cavalry or infantry. On the other hand, they can attack the enemy's columns on march, and harass them incessantly on both sides of the road, without incurring any danger to themselves. Cannon-shot would produce little effect directed against them, dispersed as they are at the distance of two or three hundred paces; and their movements are so rapid, that, at the very instant you would expect to see them a second time, they have already disappeared, to appear again in a quarter where you are not the least aware of them.

"The real superiority of the skielöbere, however, is chiefly shown when the enemy halts after a long march. Whatever precautions may then be

UPPER CANADA. 269

tural operations. But still there are several things which require attention, and particularly the live stock, which ought to be regularly fed. The farmer must, indeed, be careful in this respect, both in summer and in winter. As there is no grass in the woods, and as new settlers cannot raise fodder for their cattle immediately, they are obliged to buy either hay or straw, or pumpkins, to feed them, or to cut down trees for them to browze upon. Oxen and young cows thrive well enough on the tender shoots of the birch, maple, &c.; but sheep must have hay or turnips, and ought to be secured from the wolves every night. Every settler should, in the course of the winter, haul a quantity of firewood sufficient to supply him the whole year; and the goodness of the roads will enable him to do this without much difficulty. When the weather is bad he may employ his time within doors, in improving the interior of his dwelling-house, or amusing himself by the fire, which can always be made a warm and cheerful one, from the profusion of fuel that the poorest person has continually at command. Those who delight in field-sports may go into the woods in search of deer, which usually abound in the vicinity of new settlements. In Canada, the privilege of shooting them, and all other game, belongs equally to the lord and the peasant.

Figure 4: Binarization of two G1000 pages producing a high quality transcription with the OCR

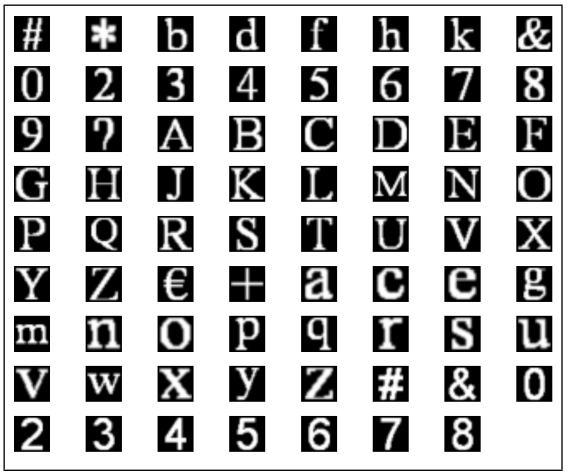


Figure 5: Artificial initial dictionary of means

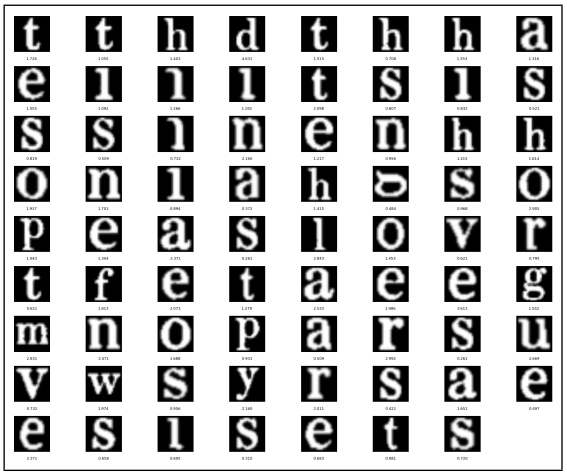


Figure 6: Final means created by the EM algorithm from the initial dictionary and identity sigma matrices

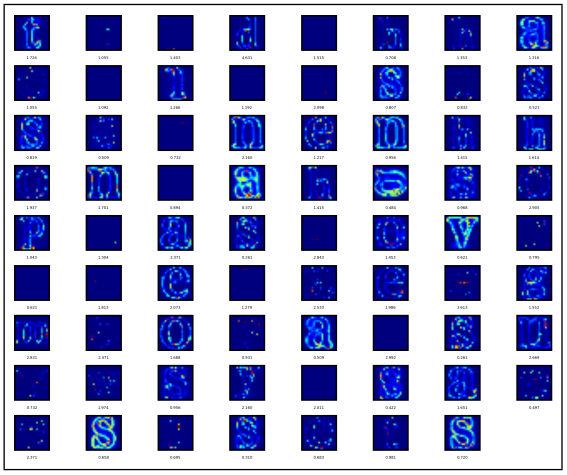


Figure 7: Final standard deviations created by the EM algorithm from the initial dictionary and identity sigma matrices

rization produces large and interesting improvements (the supplied ground truth has been generated by an OCR). Therefore, it can be inappropriate to rely on them. As argued in 3.2, we preferred to work with the 'words in dictionary' ratio instead, which is a more robust and general scheme for datasets without ground truth at all. We kept 10 documents for the training, all the others are used for the evaluation.

The open source project OCRopus 0.3.1 [4] has been used to run the experiments with Tesseract 2.0.3 [17] as the character recognition engine.

4.1 Pre-Evaluation with OCR

To be able to evaluate the GMTT method, we have generated the best black and white version of each page of the corpus. We tested nine W and nine K for Sauvola in the ranges $[10, 90]$ and $[0.1, 0.9]$ respectively. The couple (W^*, K^*) obtaining the best word-in-dictionary ratio is determined for each page. Ten of them are kept for training the GMMT.

The first experiment, fully performed with the OCR, is designed to show how the binarizations behave on a heterogeneous corpus. We also tested a binarizer by range, represented as $(W, K) = (0, 0)$ and the Otsu's binarizer at $(0, 1)$. It can be seen (Fig. 4.1) that values suggested by different authors are far from being optimal on a new corpus of data.

There is no best unique parameter set that can perform well on any kind of documents. Contrary to what we found in the literature, $(W, K) = (15, 0.2)$ or $(W, K) = (15, 0.3)$ for Sauvola is really not optimal for the Google 1000 books dataset. In some cases, even Otsu or the simple binarizer can outperform it (also reported in [8]). As reported in the literature, Sauvola is more affected by K than W, most of the authors set W to a small value around 15 pixels, but higher values seem to work better.

In a previous work [12], we used a small subset of 10 lines in order to not have to perform the OCR each time on the full page. We found that $K = 0.3$ is a good value in general but we still had larger W (4.1). The experimentations showed that the best unique binarizer is then $(50, 0.3)$ (Fig. 4.1). The global behaviour of making a choice with a small subset of lines tends to choose lower value for W and K but keeps a good *wid*.

According to what the literature proposes, we will refer from now to two couples of parameters: $(10, 0.3)$ and $(20, 0.3)$.

4.2 Evaluation of GMTT

Ten pages, such as in Fig. 4, have been used to generate a ground truth for training GMTT as explained in (3.2). The mixture model, composed of 71 Gaussians, was trained as described in (3.3). The connected components smaller than 8 pixels in width or height are automatically filtered out. The remaining ones are scaled into blobs of 21 by 21 pixels. The training generated means and sigmas already presented in Fig. 3.3,3.3. We make use of the Equation (6) for scoring the binarizations on the test pages. During the evaluation, at most 512 tokens are randomly picked from the image, even if the page contains more.

The Figure 9 presents the distribution of the Sauvola's parameters found by the GMTT method. As already seen with the OCR-based method with a small subset of lines, an approximation tends to choose smaller values for W and K.

Nevertheless, choosing different (W, K) does not necessary imply producing an inappropriate binarization. The Figure 10 gives the boxplots of the two reference values for Sauvola versus the GMMT. It can be shown that, without any knowledge on the document, the GMTT is able to choose in average a suitable binarization. Compared to a fixed Sauvola with $(W, K) = (10, 0.3)$, the recognition improvement is 11.2 points (Fig. 4.2). The GMTT method achieves a *wid* of 44.5% whereas Sauvola $(10, 0.3)$ reaches only 33.2% (Fig. 4.2).

We get in average better results than the widely used fixed Sauvola. Compared to a previous work, he have less accuracy than an OCR based method [12] but when the OCR-based method is restricted to a small subset of 5 lines GMTT is still better and largely faster (more than 4 times faster).

Even is the overall improvement is obvious, for some documents, the GMTT method is sometimes making a wrong decision as it can be seen on Fig. 11). The GMTT, as most of the pixel-based methods, suffers from mixed text/image pages. To avoid using region classification or layout analysis, the blobs are not always picked form the real characters. If the page contains more pictures than text, then GMTT tends to binarize the image too hard (Fig. 12). On the other hand, the $s(b^*(X), B^*)$ for that cases is largely small compared to the regular ones. In any case, a small $s(b^*(X), B^*)$ implies a small *wid* afterwards. The score can be used for rejection or at least to warn for a bad final binarization.

4.3 Computation Time

If the aim of the binarization is page recognition, the GMTT method requires few extra computations. Binarizing an image with Sauvola using integral images is fast. Extracting connected component, filtering, fitting into blobs are basic operations and are also fast. Evaluating a binarization with the resulting tokens depends on their number and their size. With our settings, evaluating one binarization takes around 1 or 2 seconds, without implementing any optimization. Evaluating the 81 binarizations takes less than two minutes per page on a 2Ghz PC.

But several speed-ups can be definitively employed. The first one is to reduce the search space by focusing more on K than W and use some standard parameters optimization techniques. Another idea consists in extracting the connected components only once, then applying the binarizations and evaluation (with Equation (1)) on them. Finally, the optimization can be performed only on difficult documents: If an initial binarization outputs good score with GMMT, there is no need to continue the procedure. The optimization can be executed only if low $s(b^*(X), B^*)$ is noticed, and ran until reaching an acceptable one.

5. CONCLUSIONS

We presented a statistical motivated method to optimize binarization of document images. The Gaussian Mixture Token Thresholding (GMTT) consists in learning a model for an OCR system based on suitable input images for itself. Once the model is instanced, we dispose of a fast function that gives the prior of a good binarization for the targeted OCR. It results in a method which does not rely on a specific class of documents. Then, finding a good binarization for a page consists in evaluating some binarizer and simply selecting the one with the highest prior. Even if the experimental part was focused on Sauvola's binarization, any kind of binarizer with parameters or set of binarizers can be quickly

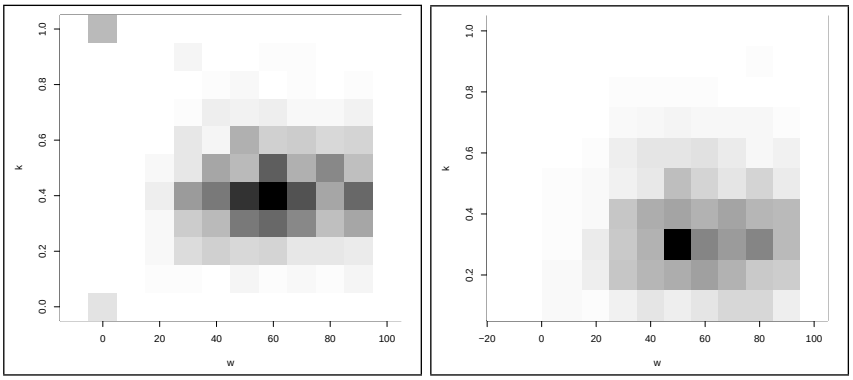


Figure 8: Overview of the best binarizers for G1000. Each best (W, K) contributes to the darkness of a square at that position. $(60, 0.4)$ seems to be the best unique value for this dataset (On the left). On a subset of 10 lines, lower values are observed (On the right).

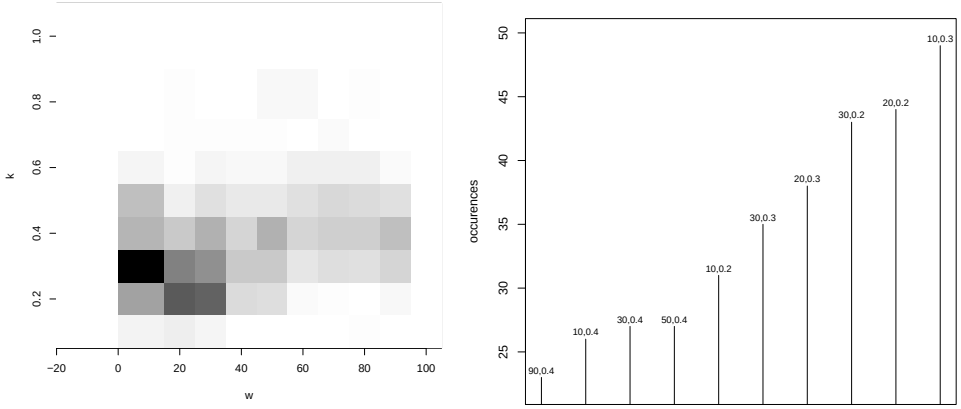


Figure 9: Sauvola's parameters obtained with the GMTT method. The best couple for the G1000 dataset is $(10, 0.3)$

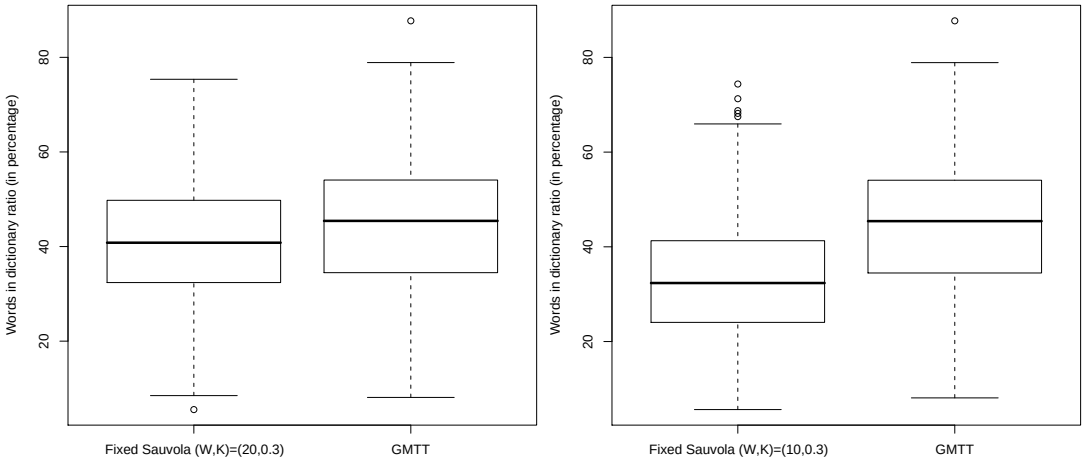


Figure 10: Average quality improvements in words in dictionary, higher values are better. GMTT, on the right of each graph, is compared a fixed Sauvola. The mean improvements are respectively 3.4 and 11.2 points

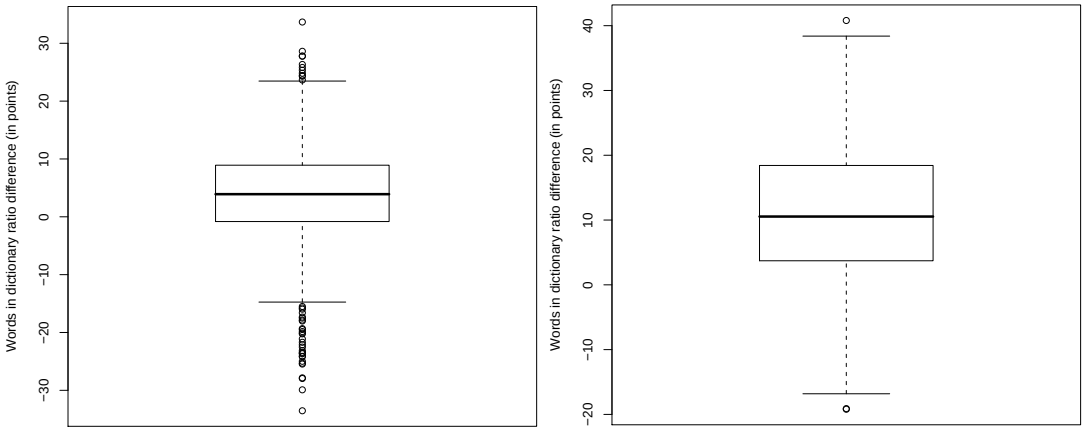


Figure 11: Average difference of words in dictionary per page, higher values are better. For each page, the *wid* of fixed Sauvola is subtracted to the *wid* of GMTT. Negative values mean that $GMMT$ was worse than fixed Sauvola, and positive values mean that GMTT was the best. On the left, fixed Sauvola $(20, 0.3)$ **vs. GMTT, on the right, fixed Sauvola** $(10, 0.3)$ **vs. GMTT**

evaluated, and optimized for character recognition.

Tested on a subset of the G1000 database, the GMTT outputs similar results than an OCR-based method, but GMTT is faster and does not rely on the content of the document and it is language independent. As GMTT uses basic image operations and a simple evaluation method, many algorithmic and implementation optimizations can be used to speed up the process.

The database we tested on was already scanned fairly uniformly and post-processed as part of putting together the document collections so the potential for improvement was limited. For real-world document collections, scanned or captured under a much wider variety of conditions, we expect that our method will yield more significant improvements relative to other preprocessing methods.

6. REFERENCES

[1] E. Badekas and N. Papamarkos: "Automatic Evaluation of Document Binarization Results", *Progress in pattern recognition, image analysis and applications*, vol.3773, pp.1005-1014, 2005

[2] E. Badekas and N. Papamarkos: "Estimation of proper parameter values for document binarization", *International Conference on Computer Graphics and Imaging*, no.10, track 600-037, 2008.

[3] T. M. Breuel: "Robust least square baseline finding using a branch and bound algorithm", *Proc. SPIE Document Recognition and Retrieval IX*, pp.20-27, 2002.

[4] T. M. Breuel: "The OCRopus Open Source OCR System", *Proceedings SPIE DRR XVI*, 2008.

[5] S. F. Chen and J. Goodman: "An empirical study of smoothing techniques for language modeling", *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pp.310-318, 1996.

[6] A. P. Dempster, N. M. Laird and D. B. Rubin: "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol.39, no.1, pp.1-38, 1977.

[7] B. Gatos, I. Pratikakis and S.J. Perantonis: "Improved Document Image Binarization by Using a Combination of Multiple Binarization Techniques and Adapted Edge Information" *International Conference on Pattern Recognition*, pp.1-4, 2008.

[8] J. He, Q. D. M. Do, A. C. Downton and J. H. Kim: "A Comparison of Binarization Methods for Historical Archive Documents", *International Conference on Document Analysis and Recognition*, vol.1, no.8, pp.538-542, 2005.

[9] Y. Li, D. Lopresti, G. Nagy and A. Tomkins: "Validation of Image Defect Models for Optical Character Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp.99-108, 1996.

[10] F. Jiménez-López, D. Cuesta-Frau, J. Linares-Pellicer and P. Micó-Tormos: "A comparative study of local thresholding methods for document image binarization", *Machine Graphics and Vision International Journal* vol.15, no.3, pp.439-450, 2006.

[11] N. Otsu: "A threshold selection method from gray level histograms", *IEEE Trans. Systems, Man and Cybernetics*, vol.9, pp.62-66, 1979.

[12] Y. Rangoni, F. Shafait and T. M. Breuel: "OCR Based Thresholding" *Conference on Machine Vision Applications*, 2009.

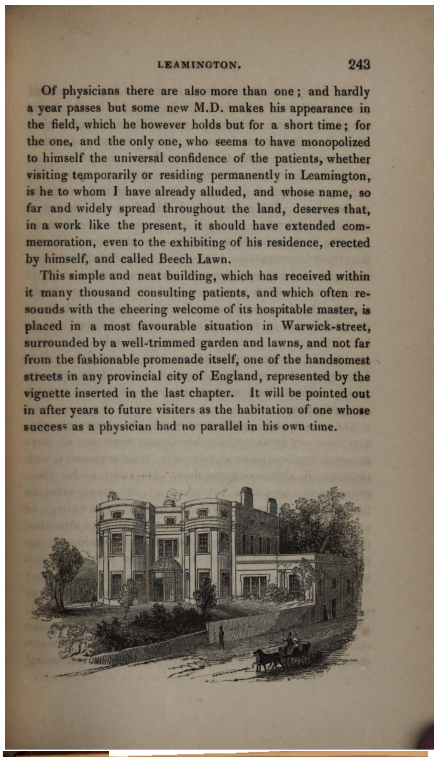
[13] Y. Rangsanseri and S. Rodtook: "Comparative Study of Thresholding Techniques for Gray-Level Document Image Binarization", *International Conference on Electrical and Electronic Technology* vol.1, no.10, pp.152-155, 2001.

[14] J. Sauvola and M. Pietikäinen: "Adaptive document image binarization", *Pattern Recognition*, vol.33, no.2, pp.225-236, 2000.

[15] M. Sezgin and B. Sankur: "Survey over image thresholding techniques and quantitative performance evaluation", *Electronic Imaging*, vol.13, pp.146-165, 2004.

[16] F. Shafait, D. Keysers and T. M. Breuel: "Efficient implementation of local adaptive thresholding techniques using integral images", *Document Recognition and Retrieval XV*, vol.6815, 2008.

[17] R. Smith: "An Overview of the Tesseract OCR



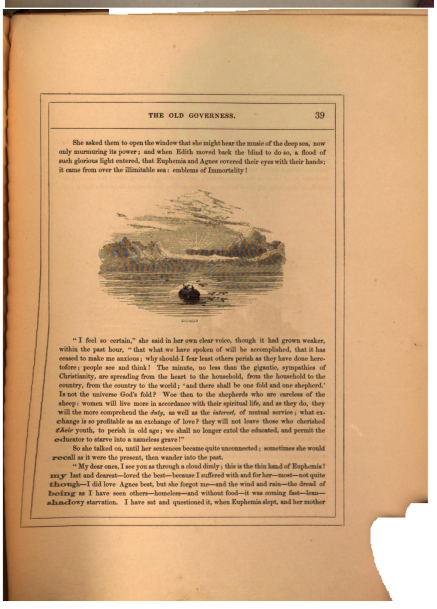


Figure 12: The two worst cases for GMTT

Engine", *International Conference on Document Analysis and Recognition*, vol.2, no.9, pp.629-633, 2007.

[18] O. D. Trier and A. K. Jain: "Goal-directed evaluation of binarization methods", *Pattern Analysis and Machine Intelligence*, vol.17, no.12, pp.1191-1201, 1995.

[19] O. D. Trier and T. Taxt: "Evaluation of Binarization Methods for Document Images", *Pattern Analysis and Machine Intelligence*, vol.17, no.3, pp.312-315, 1995.

[20] L. Vincent: "Google book search: document understanding on a massive scale", *International Conference on Document Image Analysis*, pp.819-823, 2007.

[21] Y. Yitzhaky and E. Peli: "A method for objective edge detection evaluation and detector parameter selection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.25, no.8, pp.1027-1033, 2003.

[22] Y. Zhu: "Augment Document Image Binarization by Learning", *International Conference on Pattern Recognition*, no.19, pp.1-4, 2008.