Generation of Learning Samples for Historical Handwriting Recognition Using Image Degradation

Andreas Fischer
CENPARMI, Concordia University
Montreal, Canada
an_fisch@encs.concordia.ca

Van Cuong Kieu
LaBRI, University of Bordeaux I
Bordeaux, France
vkieu@labri.fr

Muriel Visani
L3i, University of La Rochelle
La Rochelle, France
muriel.visani@univ-lr.fr

Ching Y. Suen
CENPARMI, Concordia University
Montreal, Canada
suen@encs.concordia.ca

ABSTRACT

Historical documents pose challenging problems for training handwriting recognition systems. Besides the high variability of character shapes inherent to all handwriting, the image quality can also differ greatly, for instance due to faded ink, ink bleed-through, wrinkled and stained parchment. Especially when only few learning samples are available, it is difficult to incorporate this variability in the morphological character models. In this paper, we investigate the use of image degradation to generate synthetic learning samples for historical handwriting recognition. With respect to three image degradation models, we report significant improvements in accuracy for recognition with hidden Markov models on the medieval Saint Gall and Parzival data sets.

Keywords

Historical Documents, Handwriting Recognition, Learning Sample Generation, Synthetic Images, Image Degradation, Hidden Markov Models

1. INTRODUCTION

The interest in handwriting recognition for historical documents has been growing rapidly in recent years [1]. Automatic reading of historical manuscripts would provide access to the contents of a vast amount of digitized historical documents stored at libraries worldwide. However, current handwriting recognition systems are still far from being perfect. Typical problems include the high variability in character shapes, the large number of word classes, and the inability to segment touching and broken characters prior to recognition [12]. For historical documents, an additional problem is caused by the differences in the image quality stemming for instance from faded ink, ink bleed-through, wrinkled and stained parchment. While this problem has been mainly addressed in the context of layout analysis [14], it has also an effect on learning-based recognition systems. Indeed, the morphological character models have to incorporate different degradations at learning stage in order to generalize well to unseen test images.

Image degradation is a promising approach to distort available learning samples in order to enlarge the training set and incorporate more variability in the image quality. Diverse image degradation models have been proposed in document analysis [2], for example barcode printing defects [15], ink bleed-through [17], and "hard pencil noise" [7]. A typical application scenario is the evaluation of document analysis systems under different degrees of distortion, for instance to assess symbol recognition and spotting [3], music score staff removal [22], and handwriting recognition [18].

In our application scenario, image degradation is used to generate additional learning samples. This approach has been pursued in [20, 21] for modern handwriting. Several perturbation models were applied to the handwriting images to distort the shape of the characters and the baseline of the text lines. Promising improvements of the recognition accuracy are reported with synthetic learning samples. However, aiming at modern handwritings, this degradation model does not provide the means to generate typical noise observed in digitized historical manuscripts.

In this paper, we investigate the case of historical handwriting based on three degradation models, namely Kanungo noise [9], a character degradation model [10], and a geometric distortion model [13]. They are applied to binary text line images to generate additional learning samples for hidden Markov model based recognition. On the medieval Saint Gall [4] and Parzival [6] data sets, we demonstrate that the recognition accuracy can be significantly improved with the synthetic learning samples.

The remainder of this paper is organized as follows. First, the data sets are introduced in Section 2. Next, the image degradation models are detailed in Section 3. Section 4 provides information about the recognition system and Section 5 presents the experimental results. Finally, we draw some conclusions in Section 6.

Table 1: Data Sets

	Saint Gall	Parzival
Century	9th	13th
Language	Latin	German
Writers	1	3
Pages	60	47
Text Lines	1,410	4,477

2. DATA SETS

Two medieval data sets of the publicly available IAM-HistDB[1] are used in this paper, namely the Saint Gall database and the Parzival database. Data set statistics are listed in Table 1 and sample images are shown in Figures 1a and 1b. The images illustrate considerable differences in the image quality even within a single manuscript.

2.1 Saint Gall Database

The Saint Gall database is based on a medieval Latin manuscript from the 9th century that contains the hagiography *Vita sancti Galli* by Walafrid Strabo. The Abbey Library of Saint Gall, Switzerland, holds a manual copy of the work within the Cod. Sang. 562, which was written by a (probably) single experienced hand in Carolingian script with ink on parchment.

The database currently includes 60 manuscript pages as indicated in Table 1. For further details on this database, we refer to [4].

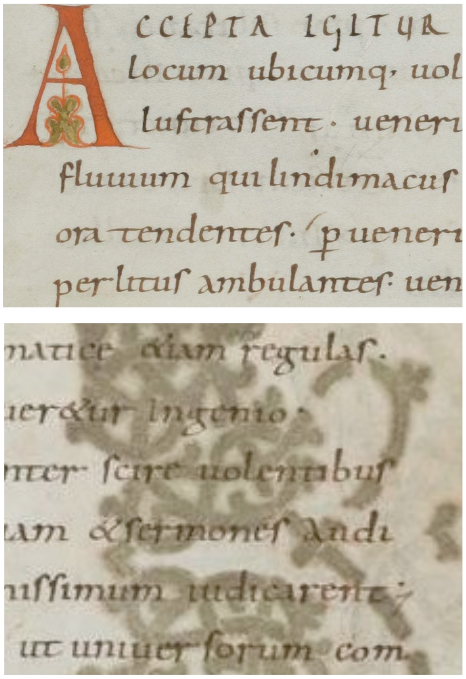
2.2 Parzival Database

The Parzival database is based on a medieval German manuscript from the 13th century that contains the epic poem *Parzival* by Wolfram von Eschenbach, one of the most significant epics of the European Middle Ages. There exist several manual copies of the poem that differ in writing style and dialect of the language. For the Parzival database, the Cod. 857 is considered, which is held by the Abbey Library of Saint Gall, Switzerland. It was written in Middle High German by several writers with ink on parchment using Gothic minuscules.

The Parzival database currently consists of 47 manuscript pages as indicated in Table 1. For further details on this database, we refer to [6].

3. IMAGE DEGRADATION MODELS

In this section, the three image degradation models that are used to generate synthetic learning samples[2] are presented. They are applied to binary text line images after a series of image preprocessing steps. The resulting degradation images have been made publicly available.[3]



(a) Saint Gall Database

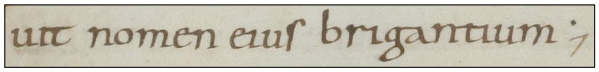
(b) Parzival Database

Figure 1: Data Sets

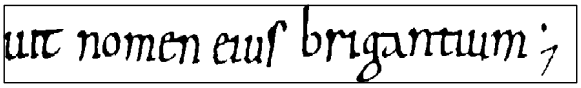
[1] http://www.iam.unibe.ch/fki/databases/iam-historical-document-database

[2] Sometimes also called "semi-synthetic" since the samples are real images distorted with synthetic noise.

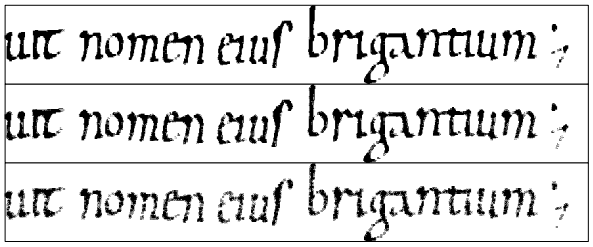
[3] http://www.labri.fr/perso/vkieu/content/Databases/hwr_parzival_saintgal.html



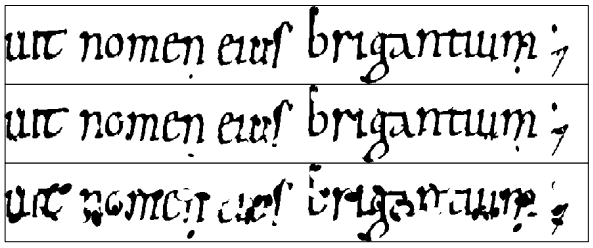
(a) Original



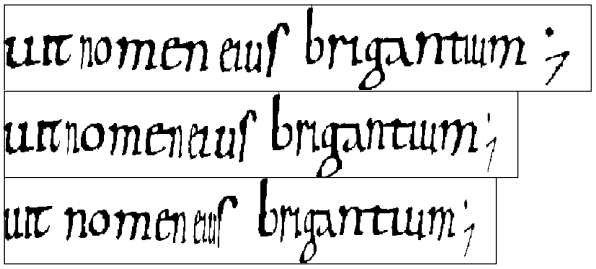
(b) Normalization



(c) Kanungo



(d) Character



(e) Geometric

Figure 2: Different types of image degradation on the Saint Gall database.

3.1 Preprocessing

The first step in preprocessing scanned manuscript images for handwriting recognition consists of layout analysis and text line extraction. In order to focus on handwriting recognition, we work directly with pre-segmented text line images in this paper. They were extracted with a semi-automatic procedure detailed in [5].

Next, the text foreground is determined by means of binarization. For the Saint Gall database, Sauvola's method is employed [4] and for the Parzival database, local edge enhancement is performed with Difference of Gaussians before applying a global threshold [5].

Finally, the binary images undergo a series of normalization operations in order to standardize their appearance before recognition. Normalization includes skew correction as well as a vertical division into an upper, middle, and lower region. An exemplary normalization result is shown in Figures 2b and 3b. For further details, we refer to [5].

3.2 Kanungo Noise

The first degradation model was proposed by Kanungo *et al.* in [9] and was validated in [8]. The model distorts character edges by using a nonlinear local selection process and a morphological closing operation. First, the selection process flips the value of some foreground and background pixels (foreground pixel changes to background and vice versa) according to a probability function. The probability function depends on a distance transform as follows:

- Foreground pixels: $p = \alpha_0 e^{-\alpha d} + \eta$
- Background pixels: $p = \beta_0 e^{-\beta d} + \eta$

where d is the distance of a pixel to the nearest character edge. The two input parameters α and β control the number of flipped pixels. The three parameters α_0, β_0, and η are considered as the initial inputs. Secondly, a morphological closing operation is applied to smooth the character edges. This model works well with binary images and is widely used to simulate the presence of noise for assessing the performances of different document analysis methods.

We consider three degradation levels with increasing distortion effect:

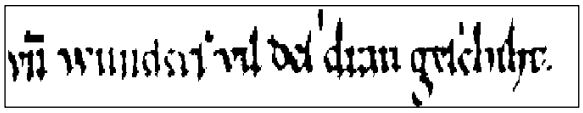
- Level 1: $\alpha = \beta = 7$, $\alpha_0 = \beta_0 = 1$, $\eta = 0$
- Level 2: $\alpha = \beta = 5.5$, $\alpha_0 = \beta_0 = 1$, $\eta = 0$
- Level 3: $\alpha = \beta = 4.5$, $\alpha_0 = \beta_0 = 1$, $\eta = 0$

The effects of the different degradation levels are illustrated in Figures 2c and 3c. The number of flipped pixels ranges from 100-200 for level 1, 400-800 for level 2, and 800-1600 for level 3.

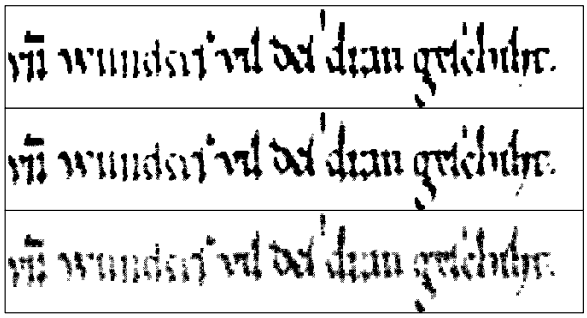
3.3 Character Degradation

The second degradation model is the character degradation model proposed in [10], which is specifically designed to

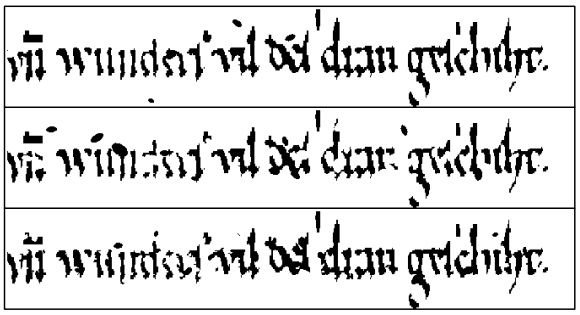
(a) Original



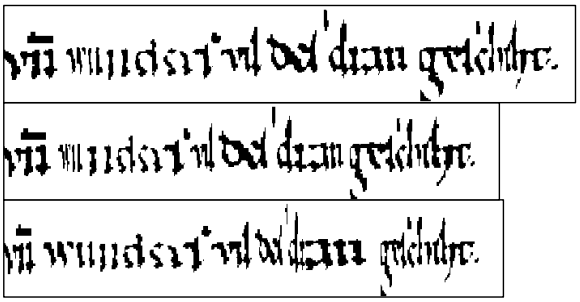
(b) Normalization



(c) Kanungo



(d) Character



(e) Geometric

Figure 3: Different types of image degradation on the Parzival database.

mimic typical noise regions in ancient documents such as ink smudges and white specks or streaks due to the age of the manuscripts. This kind of noise mostly appears in the neighborhood of the characters and can lead to touching and broken characters.

As detailed in [10], this noise can be simulated in three steps. First, the seed-points (centers of noise regions) are selected as the flipped pixels resulting from the nonlinear local selection process proposed by Kanungo *et al.* [9]. Secondly, at each seed-point, a noise region is defined with three properties: shape, size, and direction. For example, an elliptic noise region has two axes (minor and major) which depend on the gradient vector $\overrightarrow{v}$ at its center and on two input parameters a_0 and g. The parameter a_0 controls the size of the elliptic region and g its "flatness". The size parameter a_0 determines whether noise regions are more likely to break, connect, or modify character strokes. Finally, in order to obtain degraded regions as realistic as possible, the gray value of the pixels inside each noise region is smoothed with a Gaussian function. Since we are working with binary images, this last step is omitted.

Again, we define three degradation levels with increasing distortion effect:

- Level 1: $\alpha = \beta = 8.5$, $\alpha_0 = \beta_0 = 1$, $\eta = 0$ $g = 0.6$, $a_0 = 5$
- Level 2: $\alpha = \beta = 7$, $\alpha_0 = \beta_0 = 1$, $\eta = 0$ $g = 0.6$, $a_0 \in \{3, 4, \ldots, 7\}$ (random selection)
- Level 3: $\alpha = \beta = 6.2$, $\alpha_0 = \beta_0 = 1$, $\eta = 0$ $g = 0.6$, $a_0 \in \{3, 4, \ldots, 10\}$ (random selection)

The effect is shown in Figures 2d and 3d. The number of pixels in the elliptic noise regions ranges from 50-100 for level 1, 100-200 for level 2, and 200-400 for level 3.

3.4 Geometric Distortion

The third degradation model simulates a geometric surface deformation. It was proposed by J. Liang *et al.* in [13] to evaluate their restoration algorithm. First, this model maps every point in the plane of the original image to a point on a curved surface. The curved surface is then divided into planar strips (*e.g.* quadrilaterals) that are mapped back to the image plane.

We employ this degradation model with sinusoidal and parabolic surfaces. The distortion levels are controlled by the amplitude a and wavelength λ of the surface:

- Level 1: sinusoidal surface, $a = 10$, $\lambda = 0.5$
- Level 2: sinusoidal surface, $a = 15$, $\lambda = 0.25$
- Level 3: parabolic surface, $a = 20$, $\lambda = 0.2$

The effect is illustrated in Figures 2e and 3e.

4. RECOGNITION SYSTEM

In this section, we briefly describe the hidden Markov model based recognizer used in the experiments. For a full description, we refer to [16, 6].

4.1 Features

Using a sliding analysis window with a width of 1 pixel moving from left to right, the binary text line images are represented by a sequence of feature vectors. Each window captures nine geometric features including the center of gravity, the second order moment, the fraction of black pixels, the contour positions, the deviation at the contours, and the number of black-white transitions. For more details, we refer to [16].

4.2 Training

Morphological character models consist of a series of states arranged in a linear topology. Each state captures the distribution of the sliding window features with a mixture of Gaussians with diagonal covariance matrices. Training is performed with the Baum-Welch algorithm [19] based on known transcriptions of text line images.

We employ a training procedure that gradually raises the number of Gaussians G. Starting with $G = 1$, the training set is processed I times in the first training epoch. Then, the number of Gaussians is raised to $G = 2$ by splitting and the training set is again processed I times in this second training epoch. This procedure is continued until a maximum number of Gaussians is reached.

4.3 Recognition

Two goal functions are optimized during recognition with the Viterbi algorithm [19]. First, the best match of the image with the trained morphological character models is established. Secondly, the most likely sequence of words is found with respect to a bigram language model that describes the probability of a word given the preceding word. The two goal functions are balanced with a grammar scale factor and a word insertion penalty [24].

We estimate the word bigrams on all training and validation text lines and smooth them with the modified Kneser-Ney method [11] to cope with unseen word bigrams. Furthermore, we use a closed vocabulary that contains all words of the database, even those of the test set. It is a common practice to establish comparable results without out-of-vocabulary words. For the Saint Gall database, the recognition vocabulary contains $5,762$ words and for the Parzival database, the vocabulary contains $4,934$ words.

5. EXPERIMENTAL EVALUATION

We have conducted an experimental evaluation of the proposed image degradation models on the Saint Gall and Parzival data sets (see Section 2).

All three models, namely Kanungo noise, character degradation, and geometric distortion, are applied to the training set individually to create one distorted version of each learning sample. In effect, the size of the training set is doubled. It is used to train a new recognition system, which is then compared with the reference system trained on the original training set.

Table 2: Database Setup

	Saint Gall	Parzival
Training	468	2237
Validation	235	912
Test	707	1328

Table 3: Recognition Accuracy

	Saint Gall	Parzival
Reference	88.99	83.89
Kanungo	90.15 +1.16	86.95 +3.06
Character	90.42 +1.43	85.37 +1.48
Geometric	89.25 +0.26	85.66 +1.77
Combined	90.81 +1.82	87.12 +3.23

In addition to an individual model evaluation, we also combine up to two noise models by adding their degradation images to the same training set, which is then three times larger than the original training set.

5.1 Setup

First, the text lines of both data sets are split into disjoint sets for training the HMM, validation of system parameters, and testing the final performance. The number of text lines is indicated in Table 2 for each set.

For the image degradation models, the best degradation level $L \in \{1, 2, 3\}$ is determined on the validation set. Also, the best pair of degradation models for the model combination is found with respect to the validation accuracy.

System parameters of the HMM recognizer, which are optimized on the validation set, include the number of training iterations $I \in \{2, 3, 4, 5\}$ per training epoch and the final number of Gaussian mixtures $G \in \{5, 10, \ldots, 30\}$. The number of HMM states and the language model parameters have been adopted from previous experiments [23].

5.2 Results

The word accuracy results are indicated in Table 3 for each individual image degradation model as well as for the best combination of two degradation models. Next to the word accuracy, the improvement over the reference system is indicated. With the exception of geometric distortions on the Saint Gall database, 7 out of 8 improvements over the reference system are statistically significant (t-test, $\alpha = 0.05$).

Table 4 lists the relative word error reduction for all models. The reductions are also shown in Figure 4. The synthetic learning samples realize an error reduction of up to 16.53% on the Saint Gall database, achieving an accuracy of 90.81%. On the Parzival database, an error reduction of up to 20.05% is achieved and an accuracy of 87.12% is reported. These re-

Table 4: Error Reduction

	Saint Gall	Parzival
Kanungo	10.54	18.99
Character	12.99	09.19
Geometric	02.36	10.99
Combined	16.53	20.05

Table 5: Degradation Levels

	Saint Gall	Parzival
Kanungo	1	1
Character	2	1
Geometric	1	3

sults clearly demonstrate the benefit of generating synthetic learning samples using image degradation.

When comparing the individual models, the character degradation model and the Kanungo noise tend to outperform the geometric distortion. Table 5 lists the optimal degradation levels of the different models. In 4 out of 6 cases, the lowest degradation level was preferred.

For both data sets, the combination of noise models achieves the best result. For the Saint Gall database, the optimal pair of models was $(Character, Geometric)$ and for the Parzival database, the optimal pair was $(Kanungo, Character)$. The combination results demonstrate that multiple sources of degradations are desirable for the synthetic generation of learning samples.

6. CONCLUSIONS

In this paper, we have investigated image degradation models to generate synthetic learning samples for historical handwriting recognition. The three proposed degradation models include Kanungo noise, character degradation, and geometric distortion. They have been applied to distort binary text line images for training a hidden Markov model based recognizer.

On the medieval Saint Gall and Parzival data sets, significant error reductions could be achieved with the synthetic learning samples. For closed-vocabulary recognition, the best word accuracy results were 90.81% for the Saint Gall database and 87.12% for the Parzival database. This corresponds to an error reduction over the baseline system of 16.53% and 20.05%, respectively. For both data sets, the best results were achieved by a combination of different degradation models.

In the context of historical manuscripts, the importance of learning samples cannot be overemphasized. Especially when bootstrapping a new handwriting recognition system for a new kind of historic script, every learning sample counts. Image degradations have shown a promising potential to incorporate more variability in the training set, which leads

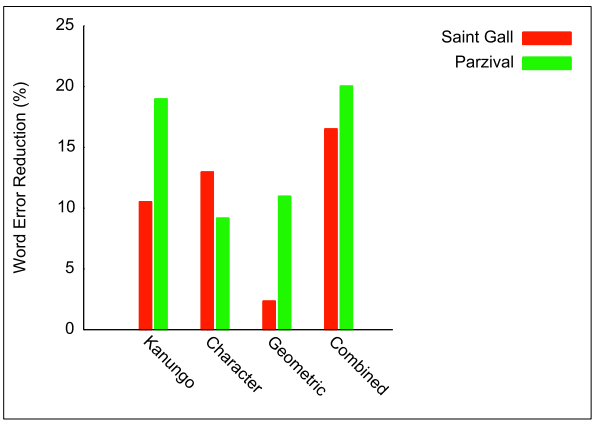


Figure 4: Error Reduction.

to a better generalization capability of the recognizer.

Future work includes improvements in the degradation models, in particular the investigation of realistic 3-dimensional distortions of the writing support. The goal is to create realistic manuscript images that incorporate several types of degradations.

7. ACKNOWLEDGMENTS

This work has been supported by the Swiss National Science Foundation fellowship project PBBEP2_141453 and the French National Research Agency DIGIDOC project.

8. REFERENCES

[1] A. Antonacopoulos and A. Downton. Special issue on the analysis of historical documents. *Int. Journal on Document Analysis and Recognition*, 9(2):75–77, 2007.

[2] H. S. Baird. The State of the Art of Document Image Degradation Modeling. In *In Proc. 4th Int. Workshop on Document Analysis Systems*, pages 1–16, 2000.

[3] M. Delalandre, E. Valveny, T. Pridmore, and D. Karatzas. Generation of Synthetic Documents for Performance Evaluation of Symbol Recognition & Spotting Systems. *Int. Journal on Document Analysis and Recognition*, 13(3):187–207, 2010.

[4] A. Fischer, V. Frinken, A. Fornés, and H. Bunke. Transcription alignment of latin manuscripts using hidden Markov models. In *Proc. 1st Int. Workshop on Historical Document Imaging and Processing*, pages 29–36, 2011.

[5] A. Fischer, E. Indermühle, H. Bunke, G. Viehhauser, and M. Stolz. Ground truth creation for handwriting recognition in historical documents. In *Proc. 9th Int. Workshop on Document Analysis Systems*, pages 3–10, 2010.

[6] A. Fischer, A. Keller, V. Frinken, and H. Bunke. Lexicon-free handwritten word spotting using character HMMs. *Pattern Recognition Letters*, 33(7):934–942, 2012.

[7] D. D. Jian Zhai, Liu Wenyin and Q. Li. A Line Drawings Degradation Model for Performance Characterization. In *Proc. 7th Int. Conf. on Document Analysis and Recognition*, pages 1020–1024, 2003.

[8] T. Kanungo, R. Haralick, H. Baird, W. Stuezle, and D. Madigan. A statistical, Nonparametric Methodology for Document Degradation Model Validation. *IEEE Trans. PAMI*, 22(11):1209–1223, 2000.

[9] T. Kanungo, R. M. Haralick, and I. Phillips. Global and Local Document Degradation Models. In *Proc. 2nd Int. Conf. on Document Analysis and Recognition*, pages 730–734, 1993.

[10] V. Kieu, M. Visani, N. Journet, J. P. Domenger, and R. Mullot. A Character Degradation Model for Grayscale Ancient Document Images. In *Proc. 21st Int. Conf on Pattern Recognition*, pages 685–688, 2012.

[11] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 181–184, 1995.

[12] A. L. Koerich, R. Sabourin, and C. Suen. Large vocabulary off-line handwriting recognition: A survey. *Pattern Analysis and Applications*, 6:97–121, 2003.

[13] J. Liang, D. DeMenthon, and D. S. Doermann. Geometric Rectification of Camera-Captured Document Images. *IEEE Trans. PAMI*, 30(4):591–605, 2008.

[14] L. Likforman-Sulem, A. Zahour, and B. Taconet. Text line segmentation of historical documents: A survey. *Int. Journal on Document Analysis and Recognition*, 9(2):123–138, 2007.

[15] R. Loce and W. Lama. Halftone Banding due to Vibrations in A Xerographic Image Bar Printer. *Journal of Imaging Technology*, 16(1):6–11, 1990.

[16] U.-V. Marti and H. Bunke. Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 15:65–90, 2001.

[17] C. M. Moghaddam R.F. Low Quality Document Image Modeling and Enhancement. *Int. Journal on Document Analysis and Recognition*, 11(4):183–201, 2009.

[18] M. Mori, A. Suzuki, A. Shio, and S. Ohtsuka. Generating New Samples from Handwritten Numerals Based on Point Correspondence. In *Proc. 7th Int. Workshop on Frontiers in Handwriting Recognition*, pages 281–290, Amsterdam, Netherlands, 2000.

[19] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, 1989.

[20] T. Varga and H. Bunke. Effects of Training Set Expansion in Handwriting Recognition Using Synthetic Data. In *Proc. 11th Conf. of the Int. Graphonomics Society*, pages 200–203, 2003.

[21] T. Varga and H. Bunke. Generation of Synthetic Training Data for an HMM-based Handwriting Recognition System. In *Proc. 7th Int. Conf. on Document Analysis and Recognition*, pages 618–622, 2003.

[22] M. Visani, V. Kieu, A. Fornés, and N. Journet. The ICDAR 2013 Music Scores Competition: Staff Removal. In *Int. Conf. on Document Analysis and Recognition, accepted for publication*, 2013.

[23] M. Wüthrich, M. Liwicki, A. Fischer, E. Indermühle, H. Bunke, G. Viehhauser, and M. Stolz. Language model integration for the recognition of handwritten medieval documents. In *Proc. 10th Int. Conf. on Document Analysis and Recognition*, volume 1, pages 211–215, 2009.

[24] M. Zimmermann and H. Bunke. N-gram language models for offline handwritten text recognition. In *Proc. 9th Int. Workshop on Frontiers in Handwriting Recognition*, pages 203–208, 2004.