An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents

Gabriel Pereira e Silva
DES – CTG – UFPE
Recife, PE, BRAZIL
gfps.cin@gmail.com

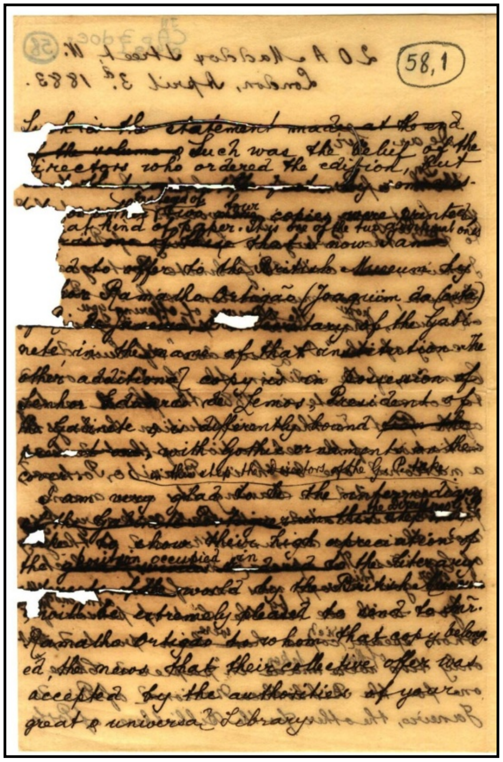
Rafael Dueire Lins
DES – CTG – UFPE
Recife, PE, BRAZIL
rdl.ufpe@gmail.com , rdl@ufpe.br

***Abstract* — Automatic optical character recognition is an important research area in document processing. There are several commercial tools for such purpose, which are becoming more efficient every day. There is still a lot to be improved, in the case of historical documents, however, due to the presence of noise and degradation. This paper presents a new approach for enhancing the character recognition in degraded historical documents. The system proposed consists in identifying regions in which there is information loss due to physical document degradation and process the document with possible candidates for the correct text transcription.**

Keywords* — *historical documents, OCR, character recognition, physical noises.

1. Introduction

There is today a huge quantity of paper legated historical documents. Such documents, accumulated over the centuries, contain important information that is the memory of mankind. Most of those documents were not digitized yet, thus the access to them is limited to very few specialists in libraries or museums. There is a wide effort throughout the world in making historical documents [4], books, and even conference proceedings [2] available in the world-wide-web. Such effort is only possible using automatic processing for image enhancement and transcription. The OCR systems available today do not yield satisfactory results for historical documents, as they are very sensitive to a wide range of noises [1]. The taxonomy for noises in paper documents proposed in reference [3], asserts that in historical documents there is a prevalence of physical noises originated either by the natural paper degradation in unsuitable storage conditions. Historical documents often exhibit back-to-front interference [4] (also known as bleeding or show-through [9]), paper aging, washed-out ink, folding marks, stains, and torn-off parts, as one may observe in the document shown in Figure 1, a hand written letter from Joaquim Nabuco, a Brazilian statesman, writer, and diplomat, one of the key figures in the campaign for freeing black slaves in Brazil (b.1861-d.1910), kept by the Joaquim Nabuco Foundation [18] a social science research institute in Recife, Brazil.



58,1

~~Such is the statement made at the end~~
~~of the volume~~. Such was the belief of the
directors who ordered the edition, but [illegible]
[illegible]
[illegible] of four copies were printed
at kind of paper. This is one of the two [illegible] ones
[illegible] that I now [illegible]
[illegible] to offer to the British Museum by
[illegible] Ramalho Ortigão (Joaquim da Costa)
[illegible] of the Gabi-
nete in the name of that institution; the
other additional copy is in possession of
Senhor Eduardo de Lemos, President of
the Gabinete & is differently bound [illegible]
[illegible], with Gothic ornaments in the
corners [illegible] the directors of the Gabinete
I am very glad [illegible] the [illegible]
[illegible]
[illegible] to show their high appreciation of
the position occupied [illegible] the literary
[illegible] the world by the British [illegible]
will be extremely pleased to send to Mr.
Ramalho Ortigão to whom that copy belong-
ed, the news that their collective offer was
accepted by the authorities of your
great & universal Library

Figure 1. Historical document from Joaquim Nabuco´s bequest

If a document has the physical noises one finds in the document of Figure 1, one stands very little chances of having any success in automatic image-into-text transcription with any commercial system today [6][7][8]. The idea of the system proposed here is to look for torn-off regions or holes and try to "complete" such areas with possible images in such a way as to maximize the probability of the correct transcription of the word as a whole. This way, instead of performing character-to-character recognition as used in conventional OCR tools, the system proposed here

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.117

infers a set of possible words and chooses the one with the highest probability to occur. Incomplete words in holes or torn off areas are completed taking into account the parts left of characters completing them with characters that may possibly "fit" the remaining parts. The OCR drives the choice of the most suitable part to fill in the holes. The choice of the most probable word may be helped by using a dictionary of terms already recognized in the document or in the file as a whole.

The structure of this paper is as follows. Section 2 describes the pre-processing stages of historical document handling. Section 3 presents the automatic image into text transcription system. Section 4 discusses the results obtained and draws lines for future work.

2. Image Pre-processing

The direct application of OCR to noisy images of documents tends to yield poor quality transcription [1]. Thus, to enhance the quality of the images of historical documents for transcription the pre-processing scheme presented in Figure 2 is applied.

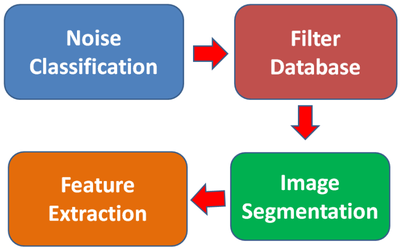


Figure 2. Pre-processing scheme

The pre-processing scheme presented in Figure 2 encompasses four modules. The first module performs "Noise Classification" it analyses the document and a neural classifier similar to the one described in reference [5], but specially tailored for historical documents detects which noises are present in a document. Besides that, the "intensity" of the back-to-front interference (also known as bleeding or show-trough) is measure. Determining the intensity of the back-to-front interference is an important step for the adequate filtering out of such noise in the second module, the "Filter Database". The third module, the "Image Segmentation" is responsible for segmenting the document in lines of text, and those, at their turn, into characters. The last module of the pre-processing scheme performs "Feature Extraction" in which the features of the characters spotted in the previous step are classified.

2.1 Noise Classification and Filtering

The first pre-processing phase is responsible for the automatic generation of "noise maps" that may be present in the document image. The noise classifier described [5] was tuned and retrained to work with the kinds of noises more often found in historical documents such as detection of noisy borders, skew, incorrect image orientation, blur and back-to-front interference. In the case of the last noise the global classification is the result of three cascaded classifiers that detect the strength of the interference and classifies it in "light", "medium", and "strong". In the case of "blur" detection, the classifier works in a similar fashion to the back-to-front one. Figure 3 presents sketches the noise classifier "architecture".

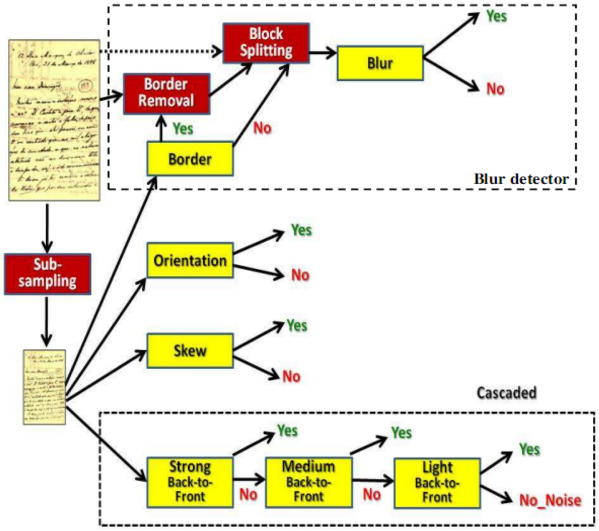
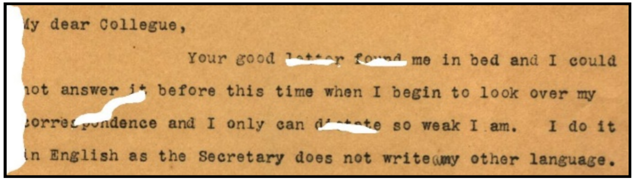


Figure 3. Noise classifier "architecture"

Besides those noises presented in the architecture above a new classification feature was added to detect the presence of holes and thorn-off regions, such as the noises shown in Figure 04.



My dear Collegue,

Your good letter found me in bed and I could not answer it before this time when I begin to look over my correspondence and I only can dictate so weak I am. I do it in English as the Secretary does not write my other language.

Figure 04. Image from Nabuco´s bequest with thorn-off regions and holes.

The new classifier besides making use of the set of features presented in reference [5], needs two new ones for better classification performance:

- Mean edge value (sharper edges have higher values);
- Mean image saturation.

One hundred images with thorn-off regions and holes were used in the tests performed here (10 original and 90 synthetic images). Besides those images another hundred images without such noises were also tested. The 200 images were split into blocks totaling 2,053 "damaged" blocks and 18,000 "perfect" ones. The result of block classification appears on Table 1, where the entry "Holes" stands for blocks that cover thorn-off regions and holes altogether. One may observe that the accuracy of the classifier is higher than 98%, which may be considered excellent.

Table 1 – Confusion Matrix for hole detection.

Classes	Holes	No_holes	Accuracy
Holes	2,017	36	98.2%
No_holes	78	17,922	99.5%

2.3 Segmentation

The character segmentation algorithm is based on the one in reference [10], suitably modified to handle degraded areas (thorn-off and with holes). The size of the characters in a block classified as having a hole takes into account the size the surrounding characters of the block under observation. Figure 05 top presents an example of the direct segmentation and at the bottom part the result of segmentation taking into account block classification and the size of the segmented characters in the surrounding areas.

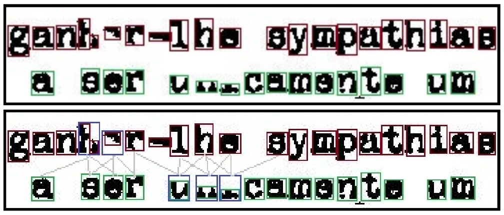


Figure 05. Character segmentation. **(Top)** Direct. **(Bottom)** Taking into account block classification.

2.3 Feature Extraction

One of the most steps in the development of systems for automatic character recognition is the choice of the feature set to be compared. Such features should attempt to be as disjoint as possible to allow good discrimination. The technical literature [20][21][22][23] points at several successful feature sets for character recognition. The choice in this research was for extracting features in specific areas (zones) of the image of characters of the Latin alphabet (A, B, C..., Z). The set of features selected is generated by:

- Geometric Moments [20][23];
- Concavity Measurements [22];
- Shape Representation of Profile [21].

In this research 26 classes of characters were used, each of them with 3,100 character images, from which 1,500 are from NIST database [17], 100 were originated from the set of documents used for testing and the remaining 1,500 were obtained by random "erasure" of parts of characters. Tests were performed using three subsets for training, validation and benchmarking, which correspond to: 25%, 25%, and 50%, of the total, respectively.

3. The Automatic Transcription System

The transcription system presented here is based on how the human brain identifies and processes visual information. One of the most accepted theories about character recognition is that the brain looks only at specific parts of [11]. Thus, the literature presents several papers which adopt zoning as a classification strategy. From the features obtained in each image "zone" it is possible to organize meta-classes to generate sets of words that may "fit" the degraded areas of a document.

3.1 Zoning Mechanism

"Zoning" may be seen as splitting a complex pattern in several simpler ones. In the case of degraded texts, the concern of this paper, this becomes an important discrimination basis amongst classes, as the "real" information is limited only to some classes. Some researchers propose only the "empirical" zoning [11][12][14], in which each character is represented by a rectangle Z, that may assume several different formats, such as the ones presented in Figure 6.
Other researchers propose methods of automatic zoning [13]. This work adopted the strategy of empirical zoning, looking at the best combination of zones targeting at obtaining the best meta-class formation for the degraded characters.

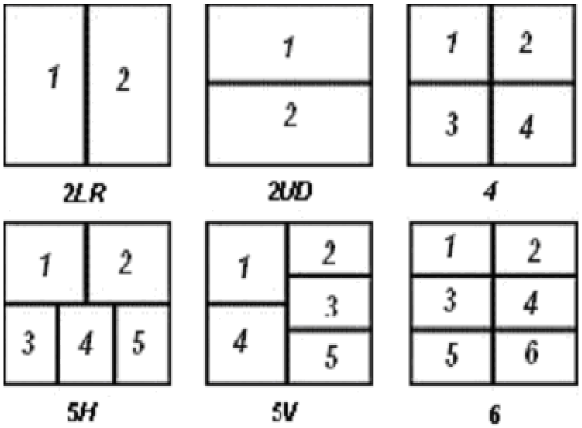


Figure 6. Z = 2LR, 2UD, 4, 5H, 5V and 6.

3.3 The Creation of Meta-classes

Meta-classes are called the “classes of classes”. Such approach targets at reducing the complexity of character recognition . In the specific case of this work, the term meta-class is used to express the set of possible words that may be generated from the block of characters with information loss. For such a purpose, a SOM [15] network was trained with different configurations and the Euclidean distance was used for group and map formation. After the SOM was trained another technique, the treeSOM, is applied to find the best cluster [16]. Such algorithm is adjusted through the choice of randomly chosen thresholds in the interval [0,1] to form different clusters for a given network. The experiments performed in this work took the following values of threshold: 0.6, 0.4, 0.3, 0.25 and 0.1. Each map of the SOM network builds a different cluster and the best amongst them is the one that minimizes the value of ε in Equation 01.

$$\varepsilon = m\frac{\sum_{C\in \mathrm{E}} Xc}{\sum_{A,B\in \mathrm{E}} \delta_{AB}} \qquad (01)$$

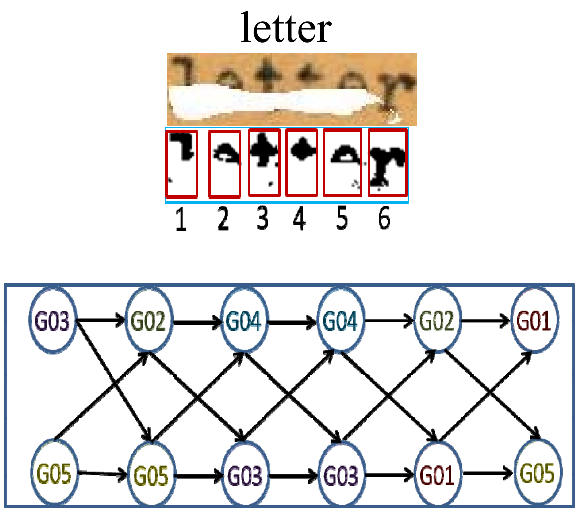
The distance between the groups A and B, denoted by δ_{AB} is equal to the average between the pairs of all elements of group A and all elements of group B. The density of the group denoted by Xc, is the average between the distances of all elements in the same group. The number of groups is m .

For all tests performed the best SOM network was the one that presented a 4x5 mesh configuration and threshold of 0.3. The clusters formed from this network are presented in Table 02.

Table 02 – The best clusters obtained

Clusters	Meta-classes
G 01	m, n, r, v, y, w
G 02	c, e, o, p, u
G 03	b, d, l, k, s
G 04	f, i, j, t, z
G 05	a, g,h,q,x

For the formation of the “candidate” words one observes for each character the two clusters with the highest activation. This way it is possible to build a graph with which one generates all possible combinations of the characters in the defined clusters, as shown in the graph of Figure 7. After the graph completion, one sweeps the graph and forms all possible words. From those one sieves the valid ones by dictionary look-up



G03;G02;G04;G04;G02;G01
G03;G02;G03;G04;G01;G05
...

Figure 7. Example of the generation of graphs for the generation of words.

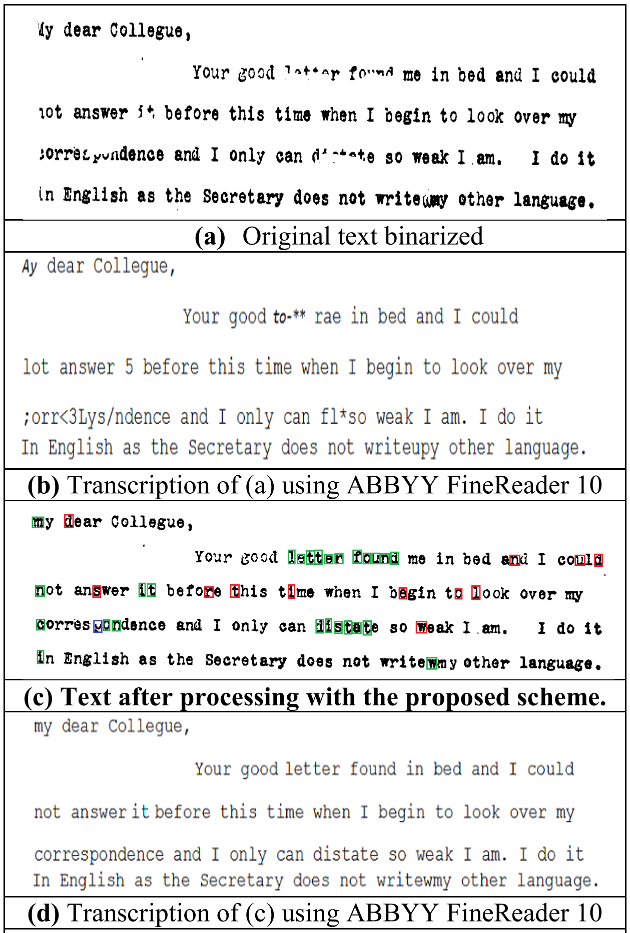
5. Experimental Results

The proposed method was tested with a file of 100 historical documents kept by the Fundação Joaquim Nabuco [18]. The documents were digitized using a flatbed scanner with 200 dpi. Such letters were transcribed and corrected yielding 13,833 words. From those, 3,814 present physical noises (holes and thorn-off parts) that implied in information losses reaching a maximum of 70% of the original word. To access the result of the improvement in recognition rate after word correction the test documents were transcribed with the commercial OCR tool ABBYY FineReader 10 Professional Editor. Figure 8 presents a sample of the results obtained. In part (c) of Figure 8 one sees in green the inserted characters, in red the ones used as models for substitution and blue is no model found in the document area.

A quantitative analysis was performed to evaluate the improvement in transcription rate presented in this work. The same one hundred documents were transcribed twice by the tool ABBYY FineReader 10 Professional Editor and the results for the 3,582 with physical noises is shown in Table 03, classified between “Correct”, “Incorrect” and “Not-recognized” words.

Table 03 – Transcription results in degraded areas.

	FineReader	New Method
Correct	1949	2752
Incorrect	895	646
Not-recognized	738	184
Accuracy	54.5%	78.2%



My dear Collegue,

Your good letter found me in bed and I could
not answer it before this time when I begin to look over my
correspondence and I only can dictate so weak I am. I do it
in English as the Secretary does not write my other language.

(a) Original text binarized

Ay dear Collegue,

Your good to-** rae in bed and I could
lot answer 5 before this time when I begin to look over my
;orr<3Lys/ndence and I only can fl*so weak I am. I do it
In English as the Secretary does not writeupy other language.

(b) Transcription of (a) using ABBYY FineReader 10

my dear Collegue,

Your good letter found me in bed and I could
not answer it before this time when I begin to look over my
correspondence and I only can distate so weak I am. I do it
in English as the Secretary does not writewmy other language.

(c) Text after processing with the proposed scheme.

my dear Collegue,

Your good letter found in bed and I could
not answer it before this time when I begin to look over my
correspondence and I only can distate so weak I am. I do it
In English as the Secretary does not writewmy other language.

(d) Transcription of (c) using ABBYY FineReader 10

Figure 8. Transcription of the document shown in Figure 4 without and with the scheme presented here.

6. Conclusions

This paper presents a scheme for improving the correct transcription rate of historical documents damaged by physical noises such as thorn-off regions and holes originated by punches (i.e. filing and staples) and worms, etc. The scheme automatically detects the damaged areas and when the text is segmented they are associated with blocks that are "replaced" with characters of the same group of features that increase the probability of being the original "damaged" character. Such replacement yields a class of possible lexemes that stand as "candidates" for the original word. Dictionary look-up decides which word is to be chosen as the most likely transcription.

The processing strategy presented here was tested in a batch of one hundred typewritten historical documents totaling over 12,000 words, of which 3,500 were "damaged". The results obtained with ABBYY FineReader 10 Professional Editor and a gain in the correct transcription rate of about 25% was observed.

References

[1] R.D. Lins, G. F. P Silva. Assessing and Improving the Quality of Document Images Acquired with Portable Digital Cameras. In: ICDAR 2007, Curitiba. IEEE Press, 2007. v. 2. p. 569-573.

[2] R.D. Lins, G. F. P Silva, G. Torreao . Content Recognition and Indexing in the Livememory Platform. LNCS, v. 6020, p. 224-230, 2010.

[3] R. D. Lins. A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises. In: ICIAR 2009. Springer Verlag, 2009. v. 5627. p. 844-854.

[4] R. D. Lins, *et al*. An Environment for Processing Images of Historical Documents. Microprocessing & Microprogramming, pp. 111-121, North-Holland, 1994.

[5] R. D. Lins, G. F. P. Silva, S. Banergee, A. Kuchibhotla, M. Thielo. Automatically Detecting and Classifying Noises in Document Images. SAC 2010, ACM Press, 2010. v. 1. p. 33-39.

[6] R. Farrahi, R. Moghaddam and M. Cheriet. Application of multi-level classifiers and clustering for automatic word spotting in historical document images.ICDAR 2009, IEEE Press, 2009. p. 511-515.

[7] S. Pletschacher, J. Hu and A. Antonacopoulos. A New Framework for Recognition of Heavily Degraded Characters in Historical Typewritten Documents Based on Semi-Supervised Clustering. ICDAR 2009, IEEE Press, 2009. p. 506-510.

[8] V. Kluzner, A. Tzadok, Y. Shimony, E. Walach, and A. Antonacopoulos. Word-Based Adaptive OCR for Historical Books. ICDAR 2009, IEEE Press, 2009. p. 501-505.

[9] G. F. P. e Silva; R.D. Lins, S. Banergee, A. Kuchibhotla, M. Thielo. A Neural Classifier to Filter-out Back-to-Front Interference in Paper Documents, ICPR 2010, Istanbul. 2010.

[10] D. M. Oliveira, R. D. Lins, G. Torreão, J. Fan, M. Thielo. A New Algorithm for Segmenting Warped Text-lines in Document Images, ACM-SAC 2011, ACM Press, 2011.

[11] C.Y. Suen, J. Guo, Z.C Li, Analysis and Recognition Of Alphanumeric Handprints by parts, IEEE Transactions on Systems,Man, and Cybernetics, N. 24, p. 614-631, 1994.

[12] Z.C. Li, C.Y. Suen, J. Guo, A Regional Decomposition Method for Recognizing Handprinted Characters, IEEE Transactions on Systems, Man, and Cybernetics, N. 25, p. 998-1010, 1995.

[13] P. V. W. Radtke, L.S. Oliveira, R. Sabourin, T. Wong, Intelligent Zoning Design Using Multi-Objective Evolutionary Algorithms, ICDAR2003, p.824-828, 2003.

[14] C. O. A. Freitas, L.S. Oliveira, S.B.K. Aires, F. Bortolozzi, Zoning and metaclasses for character recognition. ACM–SAC 2007. P. 632-636, 2007.

[15] T. Kohonen, Self-Organizing Maps, Springer Series in Information Sciences,Springer, second edition, vol. 30, 1997.

[16] E.V. Samsonova, J.N. Kok, A.P. IJzerman, TreeSOM: cluster analysis in the self-organizing map, Neural Networks, N. 19, p. 935-949, 2006.

[17] NIST Scientific and Tech. Databases http://www.nist.gov/data/.

[18] Fundação Joaquim Nabuco Fundaj, http://www.fundaj.gov.br/.

[19] ABBYY FineReader 10 Professional Editor, http://finereader.abbyy.com/.

[20] C. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: benchmarking of state-of-the-art techniques", Pattern Recognition, 36(10):2271-2285, 2003.

[21] C. Liu, Y. Liu, and R. Dai, "Preprocessing and statistical/ structural feature extraction for handwritten numeral recognition", Progress of Handwriting Recognition, A.C. Downton and S. Impedovo eds., World Scientific, 1997.

[22] L. Oliveira, R. Sabourin, F. Bortolozzi, and C. Suen, "Automatic recognition of handwritten numerical strings: A recognition and verification strategy", IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(11):1438-1454, 2002.

[23] M. Hu, "Visual pattern recognition by moment invariants", IEEE Transactions on Information Theory, 8(2):179-187, 1962.