International Journal of Future Generation Communication and Networking
Vol. 9, No. 8 (2016), pp. 307-316
http://dx.doi.org/10.14257/ijfgcn.2016.9.8.30

Recognition of Handwritten Mathematical Text

Yassine Chajri[1] and Belaid Bouikhalene[2]

[1]*Laboratory of Information Processing and Decision Aids, Faculty of Science and Technology, Beni Mellal, Morocco*

[2]*Department of Mathematics and Computers, Polydisciplinary Faculty, Beni Mellal, Morocco*
[1]*yassine.chajri@gmail.com*, [2]*b.bouikhalene@usms.ma*

Abstract

The recognition of handwritten mathematical documents is a very important field which is interested in identifying the text, graphics or notations and in extracting information.

In this paper the focus will be on the handwritten mathematical text recognition and we are going to present all the necessary steps in our system (preprocessing, text lines segmentation, mathematical symbols segmentation and symbols recognition).

Keywords*: Handwritten mathematical text; recognition; extraction; segmentation; Radon transform.*

1. Introduction

Today organizations are still using a lot of printed or handwritten paper documents that need to be represented in digital form and used in the most efficient manner. Similarly, a large number of books and old documents around the world are kept in the archives and which are threatened of disappearing. It is therefore important to preserve this legacy and make it accessible to everyone and easy to interpret.

The recognition of handwritten documents is an area that is interested in solving this big problem. It includes all the techniques that help to transform characters, words and all handwritten symbols into digital form. The objective of the documents analysis is to identify text, graphics or notations composing an image and more generally to extract information.

The importance of mathematics in all branches of science has led us to focus on the handwritten mathematical text recognition. Due to the different writing styles of mathematical expressions and the large number of mathematical symbols, the recognition of handwritten text has become a very difficult operation.

For this, we present a new system that is able to deal with these difficulties and challenges. This system is based on using the Radon transform in the preprocessing phase and also in the text lines segmentation. Regarding the mathematical symbols recognition, we present a descriptor which is hybridizing between two attributes that characterize the objects (shape and texture). For the first attribute, we are interested in extracting the specific points (Intersection points and Endpoints) that characterize each mathematical symbol skeleton and also in extracting and classifying the lines forming these skeletons (horizontal lines, vertical lines and lines at 45 degrees). Concerning the texture attribute, we are going to use the co-occurrence matrix with Haralick features.

This paper is organized as follows. Section 2 presents some related works. Section 3 offers a description of Radon transform. Section 4 describes the preprocessing techniques used in this system. Section 5 presents the text lines segmentation algorithm. In section 6,

ISSN: 2233-7857 IJFGCN

we describe the proposed algorithm for symbols recognition. The last section shows the experimental results.

2. Related Works

Firstly, we are going to present a set of existing techniques which deal with the subject of text lines segmentation, then the second part will be devoted to the presentation of some shape and texture descriptors presented in the literature.

Concerning the text lines segmentation, the methods presented can be classified as follows: Smearing methods, Grouping methods, Hough transform methods, Projection-based methods, stochastic methods and other methods [1].

Smearing methods have become very efficient in the last decade, like Run Length Smoothing Algorithm which tries to increase the black areas. The principle of this method is very simple: if we find a white space between two black blocs and the distance separating them is within a predefined threshold limit, the white is replaced by black [2,3].

In Grouping methods, the units are aggregated in a bottom-up strategy, these units can be pixels or linked components, blocks or other features like salient points. These entities are linked to build alignments. The joining scheme uses local and global criteria to check local and global consistency [4,2].

Hough transform is used to find straight lines in images. The image is transformed in the Hough domain; the alignments are hypothesized in the Hough domain and the validation is made in the image domain. It assumes that the local maxima correlate with text lines. This method needs an adaptation in the free-style handwritten text and has trouble in detecting curved text lines [5].

Projection-based methods are primarily used in the segmentation of printed document text, but it can be adapted to handwritten documents. The horizontal projection profile is mainly used in these methods which create a vector containing the sums of pixels of each line. It reaches the purpose by finding the minimum and the maximum. The local maxima of the vector represent the text line while the local minima represents the white area between lines. Horizontal projection has troubles with multi-skewed, curved and fluctuating lines [5,4].

Stochastic method is based on probabilistic algorithm. This method extracts non-linear paths between overlapping text lines, which are collected by using Hidden Markov Modeling (HMM). So the image is divided into little cells, each one representing a state of the hidden Markov model. For best results in segmentation the searching process begins from the left to the right. If we have a touching component, the optimal path will cross this component at points with a minimum of black pixels as possible [4,2]. We can also find Repulsive-Attractive network method, processing of the overlapping and touching components method.

Image analysis or image processing is interested in the description (shape, color, texture) and / or quantification (number, density) of various objects that make up the studied image. But for this work, we will focus on the description of the two visual attributes: shape and texture.

The shape is one of the most used attributes to characterize the objects in an image. For this, Jain and Valaya [6] proposed using the gradients orientation histogram on the contours. Moreover, Ferecatu [7] proposed a descriptor for detecting lines in an image. Hu [8] also proposed nonlinear functions based on geometric moments that are invariant to translation, rotation and scale changes. We find the Fourier descriptors which describe the contour by its frequency components, and we also find the Zernike moments. Oliva and Torralba [9] are based on how the human perceives the overall scene structure.

Like the shape, the texture is a basic feature of the image because it concerns an important component of human vision. However, the notion of texture is less well defined

compared to other visual attributes. It seems fair to say that the problem lies in the definition itself of the texture. Many studies have been proposed in the literature to generalize the notion of texture:

Rosenfeld and Troy [10], Gross [11] and Wu [12] define the texture as a repetitive arrangement of a specific zone. Another definition comes from Unser and Eden [13] where they declare that the texture is a structure with certain spatial properties homogenous and invariant to translation.

Haralick and Shapiro [14] defined texture as the uniformity, density, coarseness, roughness, regularity, intensity and directionality of discrete tonal features and their spatial relationships.

Regarding the methods that are interested in the characterization of texture, Tuceryan [15] presented four families of texture characterization tools: statistical methods, geometric methods, methods based on probabilistic models and frequency methods.

3. Radon Transform

The Radon transform is a mathematical technique developed by the mathematician Johann Radon [16]. This transform converts a function (image) $f(x,y)$ in a series of projections for each angle $\theta \in [0,\pi]$. A projection at a given angle θ is obtained by the linear integration of the function on all parallel lines.
The result is a new image $R(\rho,\theta)$ that can be written mathematically by [17]:

$$R(\rho,\theta)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,\delta(\rho - x\cos\theta - y\sin\theta)\,dx\,dy \quad (1)$$

where:

$$\rho = x\cos\theta + y\sin\theta \quad (2)$$

$\delta()$ is the Dirac delta function.

The Radon transform is frequently used in different areas such as tomography, astronomy and seismology.

4. Preprocessing

Preprocessing is one of the most important steps in any document recognition system because of its ability to remedy the problems associated with scanner quality, scan resolution, type of printed documents, paper quality, fonts used in the text, *etc*.

In our case, this step is more important in solving these problems and the problems related specifically to the handwritten mathematical documents; it is also important in order to have an efficient and robust system able to recognize these documents. For our system, we have chosen to convert our document in black and white and to apply median filter for noise removal.

Skewing of the scanned documents is one of the problems that influence negatively the efficiency of document recognition system. So, the skew detection and correction becomes a mandatory step in the preprocessing phase. To do this, we have chosen to use the Radon Transform.

5. Text Lines Segmentation

One of the first stages in the conception of text recognition system is the text segmentation into lines. This operation is made difficult in the case of handwritten mathematical text. When dealing with such documents, lines segmentation has to solve

some obstacles that are uncommon in modern printed text. Among these difficulties: skew angle of the text lines, overlapping symbols, and adjacent text lines touching.

[18] is a paper that treated the segmentation of characters from old typewritten documents using Radon Transform and which inspired us to use the Radon transform to segment lines from the mathematical text. The basic idea that served us deeply for using this transform is to exploit its ability to extract lines from the image because of its ability to represent image lines in the form of peaks.

In mathematical document, there is a line of background pixels between two lines. This property will be well exploited for segment the lines using Radon transform. As we can see in Figure 2 and Figure 4 representing Radon transform of mathematical text with 0° to 179° degrees of projection angle, there are some colored spots, which are the peaks of Radon transform and which represent the number of lines in this document.

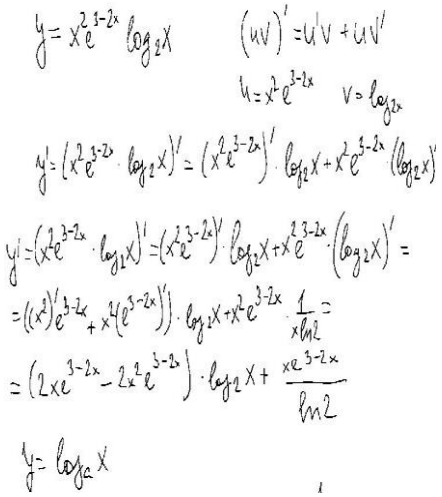

Figure 1. Text Number 1

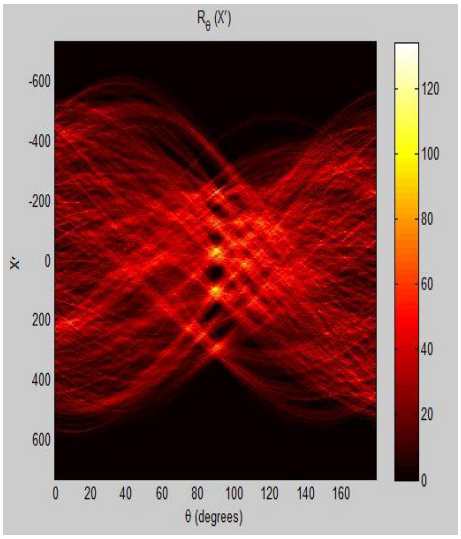

Figure 2. Radon Transform Representation of Text Number 1

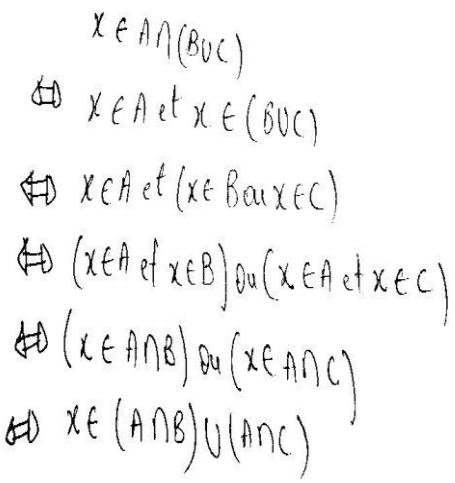

Figure 4. Text Number 2

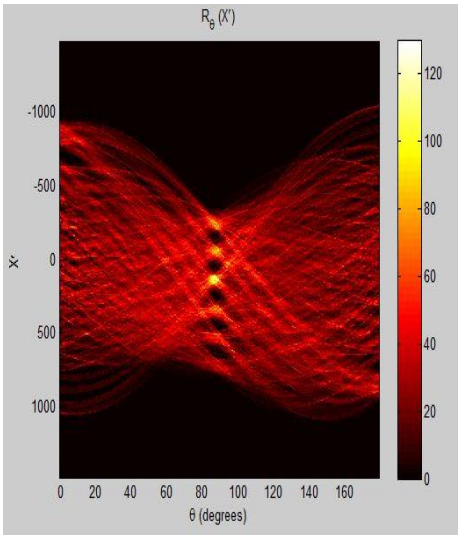

Figure 3. Radon Transform Representation of Text Number 2

After identifying the lines based on the result of the Radon transform, all the text lines can be extracted in a very precise way.

Figure 5. Line Extracted from Text Number 1

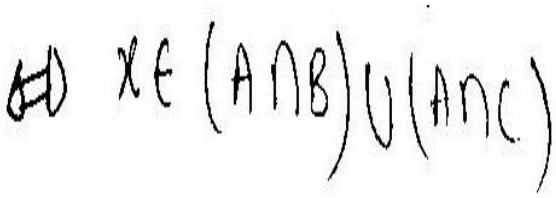

Figure 6. Line Extracted from Text Number 2

When the document lines detection and segmentation steps are completed, it remains now to extract all the symbols that form each line.

In this phase, we focus on scanning the image pixel-by-pixel from top to bottom and left to right, in order to identify connected pixel regions. In other words, we focus on identifying the regions of adjacent pixels which share the same intensity values. This procedure makes us segment all the symbols in each line of the document with a very high precision.

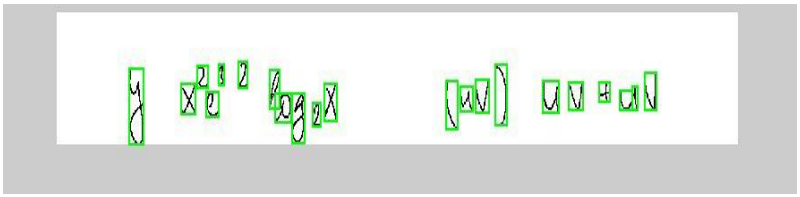

Figure 7. Symbols Segmentation (Text Number 1)

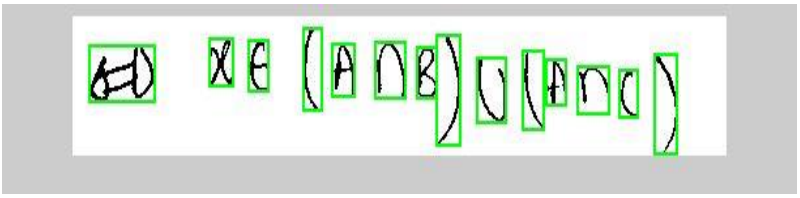

Figure 8. Symbols Segmentation (Text Number 2)

6. Handwritten Mathematical Symbols Recognition

The main idea is to combine two methods that are still used in the field of pattern recognition. The first method is based on the extraction of the geometric characteristics to focus on the difference between the various forms of symbols and the second takes as objective the study of the texture to distinguish images.

6.1. Geometric Characteristics

Our method aims at exploiting of the geometrical characteristics of each symbol skeleton. This technique is based on different types of lines that form each character and is also based on the special points (endpoints and intersection points).

It begins by dividing the image into equal size zones, then the features extraction process will be applied on each of these zones and not on the whole image. This provides more detailed information on the symbols skeletons. After finishing this image zoning step, the algorithm begins to extract the endpoints, intersection points and also the different lines existing in each zone.

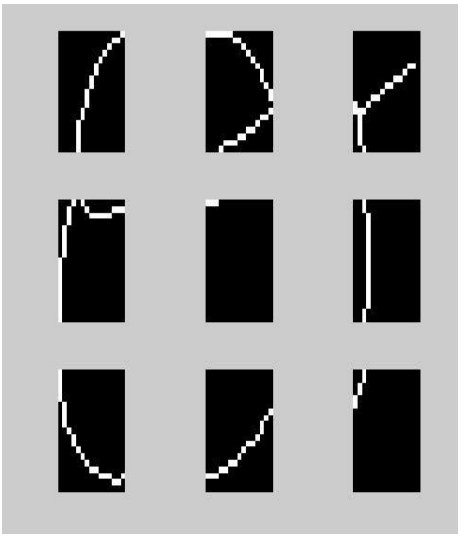

Figure 9. Symbol after the Zoning Step

Among the advantages of this algorithm is its ability to distinguish between:

- Horizontal lines
- Vertical lines
- Lines at 45 degrees

When the lines extraction and their types distinction steps are completed, the vector features for each zone is formed by:

- Number of endpoints
- Number of intersection points
- Number of horizontal lines
- Number of vertical lines
- Number of lines at 45 degrees

6.2. Texture

Texture analysis (structural approach, statistical approach) is an important field of image processing, computer vision and pattern recognition. In this work, the focus will be on statistical methods that treat the texture as a deterministic stochastic process. They model the usual qualitative concepts of texture: "granularity, contrast, uniformity, repeatability, fragmentation, orientation, *etc*."

To meet our needs, we chose to use the co-occurrence matrix with Haralick features which still remains one of the most interesting techniques in the texture analysis domain.

The co-occurrence matrix is based on the joint probability of pixels distribution in the image [19]. The element $p_{d,\theta}(i,j)$ of the co-occurrence matrix (COM) defines the frequency of occurrence of gray levels couples i and j for pixels couples separated by a distance d according to the direction θ.

To extract information, Haralick *et al*. [20] proposed fourteen descriptors calculated from the COM. These descriptors reduce the information contained in the COM and allow a better discrimination between different types of textures.

In this work, we have used the Haralick parameters described below:

- Energy:

$$\text{eng} = \sum_{i=0}^{n}\sum_{j=0}^{n} p_{d,\theta}(i,j)^2$$

- Entropy:

$$\text{ent} = -\sum_{i=0}^{n}\sum_{j=0}^{n} p_{d,\theta}(i,j)\log p_{d,\theta}(i,j)$$

- Local homogeneity:

$$\text{homloc} = \sum_{i=0}^{n}\sum_{j=0}^{n} \frac{1}{1+(i-j)^2} p_{d,\theta}(i,j)$$

- Contrast:

$$\text{cnt} = \sum_{i=0}^{n}\sum_{j=0}^{n} (i-j)^2 p_{d,\theta}(i,j)$$

- Correlation:

$$\text{corr} = \frac{\sum_{i=0}^{n}\sum_{j=0}^{n} ij\, p_{d,\theta}(i,j) - \mu_i \mu_j}{\sigma_i \sigma_j}$$

7. Results and Interpretation

These results concern two different experiments: the first is applied to the symbols presented in this paper and the second is applied to a set of 1000 symbols that were randomly chosen from the dataset [21,22].

Regarding the mathematical symbols presented in the figures, our system recognized all symbols without exception. The second experiment was done on a dataset which gathers Latin and Arabic symbols and has given us very important results.

The tables below show the results obtained in this experiment by using the proposed algorithm.

Table 1. Symbols Recognition Rate by using Geometric Characteristics Algorithm

	Geometric Characteristics algorithm	
	ANN	SVM
Arabic symbols	76%	82%
Latin symbols	80%	89%
Arabic and Latin symbols	70 %	81 %

Table 2. Symbols Recognition Rate by using the Hybrid Algorithm and without Zoning Step

	Hybridization algorithm	
	ANN	SVM
Arabic symbols	90%	94%
Latin symbols	90 %	95 %
Arabic and Latin symbols	89 %	94 %

Table 3. Symbols Recognition Rate by using the Hybrid Algorithm and the Zoning Step

	Hybridization algorithm	
	ANN	**SVM**
Arabic symbols	**92 %**	**97%**
Latin symbols	**93 %**	**98 %**
Arabic and Latin symbols	**90 %**	**97 %**

8. Conclusion

The recognition of handwritten mathematical documents still remains a challenging task. It requires solving some obstacles such as: skew angle of the text lines, different written styles, overlapping symbols, and adjacent text lines touching. In this paper, we presented our approach for handwritten mathematical text recognition which begins with a set of preprocessing techniques and we described our vision for mathematical text lines segmentation which is based on Radon transform. We presented also a new approach for handwritten mathematical symbols recognition based on a hybrid algorithm (shape and texture).

Acknowledgments

We are very grateful to the referees for their helpful comments and remarks. Also, we express our deepest gratitude to Prof. Dr. Khalid Chaouch for his invaluable contribution in writing this paper. We thank them very much.

References

[1] D. Brodic, Z. N. Milivojevic and D. R. Milivojevic, "Approach to the Improvement of the Text Line Segmentation by Oriented Anisotropic Gaussian Kernel", electronics and electrical engineering, 2012. No. 2(118).
[2] D. Brodic, "text line segmentation with water flow algorithm based on power function", Journal of electrical engineering, vol. 66, no.3, 2015, 132–141.
[3] G. Louloudis, B. Gatos, I. Pratikakis and C. Halatsis, "Text line and word segmentation of handwritten documents", Pattern Recognition 42 (2009) 3169 – 3183.
[4] L. Likforman-Sulem, A. Zahour and B. Taconet, "Text Line Segmentation of Historical Documents: a Survey".
[5] A. Nicolaou and B. Gatos, "Handwritten Text Line Segmentation by Shredding Text into its Lines", 2009 10th International Conference on Document Analysis and Recognition.
[6] A. Jain and A. Vailaya, "Image retrieval using color and shape", Pattern Recognition, vol.29, n°8, 1996.
[7] M. Ferecatu, "Image retrieval with active relevance feedback using both visual and keyword-Based descriptors", Université de Versailles Saint-Quentin-En-Yvelines. Thèse de Doctorat, 2005.
[8] M.K. Hu, "Visual Pattern Recognition by moment invariants", IRE Transaction on Information Theory, Volume 8, n°2 : 179-187, 1962.
[9] O. Aude and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope", International Journal of Computer Vision, 42(3): 145_175,2001.
[10] A. Rosenfeld and E.B. Troy, "Visual texture analysis", Conference Record for Symposium on Feature Extraction and Selection in Pattern Recognition, pages 115–124, 1970.
[11] T. R. Gross, "Code optimization of pipeline constraints",Technical report, Computer Systems Laboratory, Stanford University, 1983.

[12] C.M. Wu, Y.C Chen and K.S. Hsieh, "Texture features for classification of ultrasonic liver images", IEEE transactions on medical imaging, 11(2):141–152, 1992.
[13] M. Unser and M. Eden, "Multiresolution feature extraction and selection for texture segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(7):717– 728,1989.
[14] R.M. Haralick and L.G. Shapiro, "Glossary of computer vision terms", Pattern Recognition, 24(1):69–93, 1991.
[15] M. Tuceryan and A. K. Jain, "Texture analysis", The Handbook of Pattern Recognition and Computer Vision, pages 207-248, 1998.
[16] J. Radon, "On the determination of functions from their integral values along certain manifolds", IEEE Transactions on Medical Imaging Vol.5 No.4 pages 170-176, 1986.
[17] C. Hoilund, "The Radon Transform", Aalborg University, VGIS, 07gr721 November 12, 2007.
[18] A. A. Desai, "Segmentation of Characters from old Typewritten Documents using Radon Transform", International Journal of Computer Applications (0975 – 8887) Volume 37– No.9, January 2012.
[19] J. F. Haddon and J.F. Boyce, "Cooccurrence matrices for image analysis",IEEE Electronic and Commun ications Enginering Journal, 5(2):71–83, 1993.
[20] R. M. Haralick, K. Shanmugam and I. Dinstein, "Textural features for image classification",IEEE Trans actions on Systems, Man, and Cybernetics, SMC-3(6):610–621, nov 1973.
[21] Y. Chajri and B. Bouikhalene, "Handwritten mathematical symbols dataset", Data in brief, vol. 7, June (2016), pp. 432–436.
[22] Y.Chajri and B.Bouikhalene, "Handwritten Mathematical Expressions Recognition", International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol.9, No.5 (2016), pp.69-76.

