International Arab Conference on Information Technology (ACIT'2016)

Segmentation of text/graphic from handwritten mathematical documents using Gabor filter

Yassine CHAJRI, Abdelkrim MAARIR, Belaid BOUIKHALENE

University Sultan Moulay Slimane, Morocco
yassine.chajri@gmail.com, b.bouikhalene@usms.ma

Abstract: *Most of handwritten mathematical documents contain graphics in addition to mathematical text. Thus, these documents must be segmented into homogenous areas to facilitate their digitization. Text and graphic segmentation from these documents aims at segmenting the document into two blocks: the first contains the texts and the second includes the graphical objects. In this paper, we focus our interest on document segmentation based on the texture and precisely the frequency methods. These methods are ideal to characterize the texture and allow detecting the frequencies and orientations characteristics. Firstly, we present the main steps of our system (pre-processing, features extraction (using Gabor filter), post-processing and text/graphic segmentation). Secondly, we discuss and interpret the results obtained by our system.*

Keywords: *Handwritten mathematical document; segmentation; graphic; Gabor filter; co-occurrence matrix; Haralick features.*

1. Introduction

Traditionally, transmission and information storage were performed using the paper documents. Also, 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. So, the treatment of paper documents must be done in an efficient and integrated manner. The ultimate solution would be a computer deals with a paper document as effectively as it is able to do with other digital media. Document analysis is an area that is interested in solving this big problem and includes all the techniques that can identify texts, graphics or notations and extract information. For us, we focus our interest mainly on segmenting the image of handwritten mathematical document into microstructures (text, graphic) based on the texture attribute. Texture segmentation consists in segmenting an image into blocks (regions) having the same texture characteristics. The choice of using the texture attribute for image segmentation is justified by the fact that the mathematical text can be seen as a texture; while the graphics have a different texture. And also because this approach allows extracting a set of information without any necessary knowledge of the context, semantics or the physical characteristics of the studied image.

In this paper, we present a new approach based on the filter Gabor. This approach consists in applying a bank of Gabor filters to analyze and characterize the image texture. This texture analysis allows us to segment the document image into mathematical text and graphic regions by using K-means clustering algorithm. We present also a comparison between the results obtained by this approach and those obtained by using Grey Level Co-occurrence Matrix (GLCM).

This paper is organized as follows. Section 2 presents some related works. Section 3 presents our approach for text/graphic segmentation. Section 4 describes the co-occurrence matrix based algorithm for text/graphic segmentation. The last section shows and interprets the experimental results.

2. Related Works

The literature offers us several techniques for text-graphic segmentation from document image. Among the most popular approaches we can cite:

- Run Length Smoothing algorithm [1, 2, 3]: consists in applying a double unidirectional smoothing of the image according to two thresholds. In the case of horizontal smoothing, the threshold is equal to the average of inter-word spaces. Against, the threshold in the case of vertical smoothing is equal to the line spacing. The segmentation is obtained by applying the logical operator "and" between two resulting images. But, the run length smoothing method is

sensitive to font-size, character spacing, line and column spacing.

- Recursive Projection Profiles method [3, 4, 5]: based on the projection of the numbers of black pixels in the columns and rows of the image onto horizontal and vertical axes, respectively. This projection generates detectable valleys in the projection profiles. So, the column structure of the document can be extracted easily. The recursive projection profiles is limited to rectangular blocks, and consequently not suited for skewed text.
- Connected Components method [6]: used also in the segmentation text-graphic. The failures of this algorithm are that it is character size dependent, sensitive to inter-line and inter-character spacing and sensitive to the digitizing resolution.
- Voronoi mesh [7, 8]: allows characterizing the regions with complex shapes and how they are disposed.
- Geometrical moments: useful for the documents segmentation as shown in [9].

3. Approach proposed for text and graphic segmentation

This approach is based on Gabor filter and allows segmenting the handwritten mathematical document image into text and graphic regions by performing texture analysis.

The main steps of this algorithm are given below:

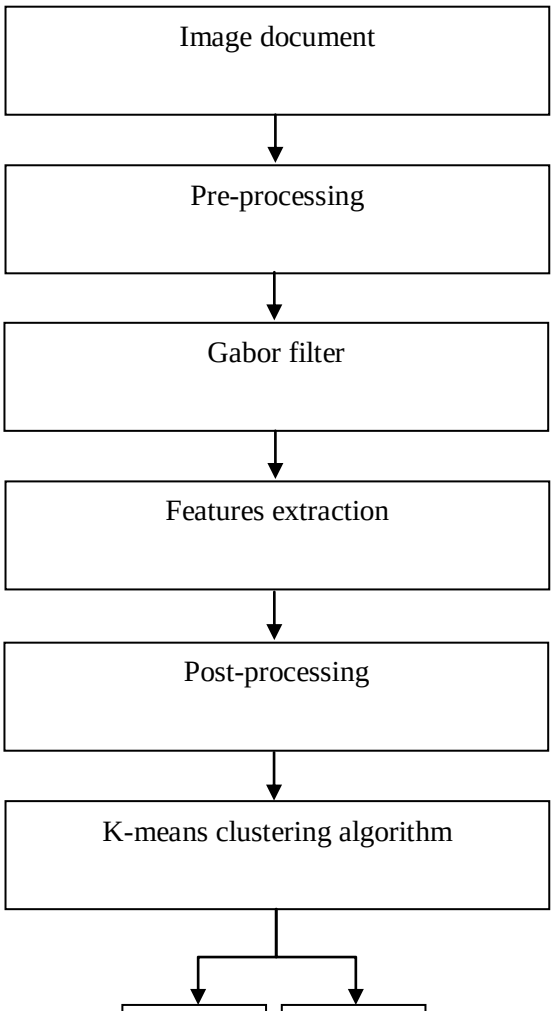

Figure 1. System architecture.

3.1. Pre-processing

Image pre-processing is one of the most important steps in our system because of its ability to remedy the problems associated with the images acquisition. The main techniques of this pre-processing step are presented below (figure 2):

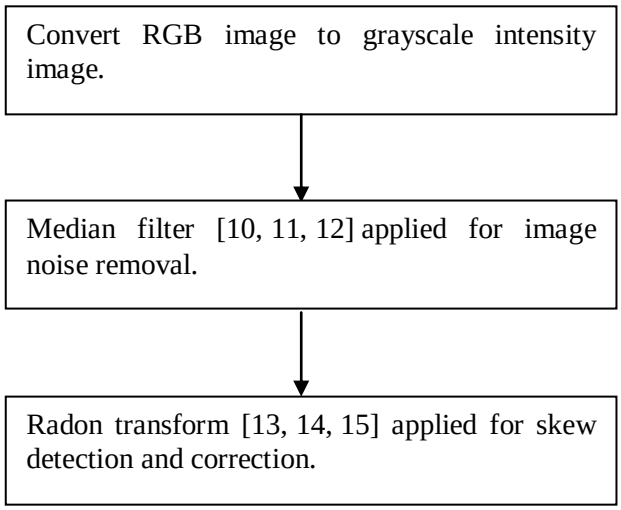

Figure 2. System pre-processing techniques.

3.2. Gabor filter

Most textures can be characterized through the local spatial frequency and orientation information present in the image. The Gabor filters seem well suited for texture analysis because they have been shown to posses optimal localization properties in both spatial and frequency domain. Gabor filters have been used in many applications, such as edge detection, target detection, retina identification, fractal dimension management, image coding and image representation.

In spatial domain, a two-dimensional even-symmetric Gabor filter can be written as:

$$h(x,y) = exp(-\frac{1}{2}(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}))cos(2\pi u_0 x)$$

Where σ_x and σ_y are the standard deviations of the Gaussian envelope along the x and y directions, respectively, and u_0 is the frequency of the sinusoidal plane wave along the x-direction (0°orientation). A

rotation of the x - y plane by angle θ will result in Gabor filters at orientation θ [16, 17].

3.3. Features extraction

Gabor filters seem well suited in document text and graphic segmentation. But to succeed this segmentation, it is necessary to make a good parameterization of these filters. For this, we were inspired by the work [18]. The authors show how to parameterize a Gabor filter bank. They recommend using the following rule:

For an image with N columns, where N is a power of 2, the frequencies are: $\{1\sqrt{2},2\sqrt{2},4\sqrt{2},\dots,(\frac{N}{4})\sqrt{2}\}$ and for each radial frequency, four filters, with orientations $\theta = 0°, 45°, 90°, 135°$ would be selected. By applying the rule described above, for our case, we will have a bank of 28 filters with:

$f_0=1\sqrt{2}$; $f_1=2\sqrt{2}$; $f_2=4\sqrt{2}$; $f_3=8\sqrt{2}$; $f_4=16\sqrt{2}$; $f_5=32\sqrt{2}$; $f_6=64\sqrt{2}$; $\theta_0 = 0°$; $\theta_1 = 45°$; $\theta_2 = 90°$; $\theta_3 = 135°$;

To extract the Gabor features, it is important to calculate them from each Gabor response matrix. So, each pixel in the input image will be characterized by 28 values.

Let $I(x,y)$ denote a grey-scale image and let $G_{u,v}(x,y)$ denote a Gabor filter defined by its centre frequency f_u and orientation θ_v. The feature extraction procedure can be written as the convolution of the image $I(x,y)$ with the Gabor filter $G_{u,v}(x,y)$ [19].

$$C_{u,v}(x,y) = I(x,y) * G_{u,v}(x,y)$$

As we can notice, the above expression represents the complex convolution which can be well exploited by separating its real and imaginary parts.

$$E_{u,v}(x,y) = Re[C_{u,v}(x,y)]$$
$$O_{u,v}(x,y) = Im[C_{u,v}(x,y)]$$

After the determination of these parameters, we can compute the filter magnitude response $M_{u,v}(x,y)$.

$$M_{u,v}(x,y)=\sqrt{E_{u,v}(x,y)^2 + O_{u,v}(x,y)^2}$$

3.4. Post-processing

The number of features extracted is far too big and which makes the segmentation step very difficult. To overcome this problem, a set of post-processing techniques is required:

- Gaussian low-pass filtering to smooth the Gabor magnitude information.
- Normalize features to be zero mean and unit variance.
- Use Principal Component Analysis to move from a 28 representation of each pixel in the input image into a 1 intensity value for each pixel.

3.5. K-means clustering

A texture based method for document segmentation consists in segmenting the document into homogenous zones which share similar texture characteristics. For example, in a document, the text blocks are assumed to have different texture characteristics to the graphics blocks. In our case, to segment the handwritten mathematical document into text and graphic components, we have chosen to use K-means clustering algorithm. This choice is justified by its efficiency, simplicity and ease of implementation. K-means algorithm splits a set of objects into K clusters. The result of this algorithm is k clusters wherein the objects within each cluster are as close to each other as possible and as far from objects in other clusters as possible.

The main steps of K-means algorithm are:

- Define k centroids, one for each cluster (the better choice is to place them as much as possible far away from each other).
- Calculate the distance between each data point and clusters centroids.
- Assign the data point to the cluster centroid whose distance from the cluster centroid is minimum as compared to all the cluster centroids.
- Update the centroid value by taking the average of all the objects that belong to the group.
- Recalculate the distance between each data point and new obtained cluster centers.
- Repeat the steps (from step 3) until there are no more changes to cluster membership.

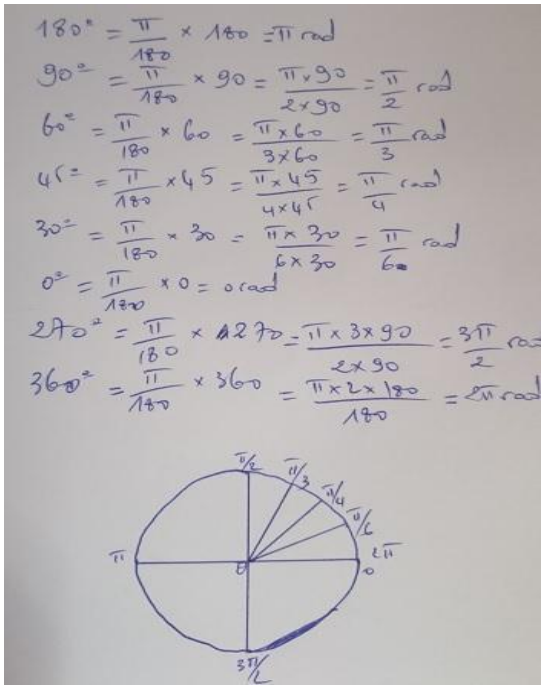

Figure 3. Original image.

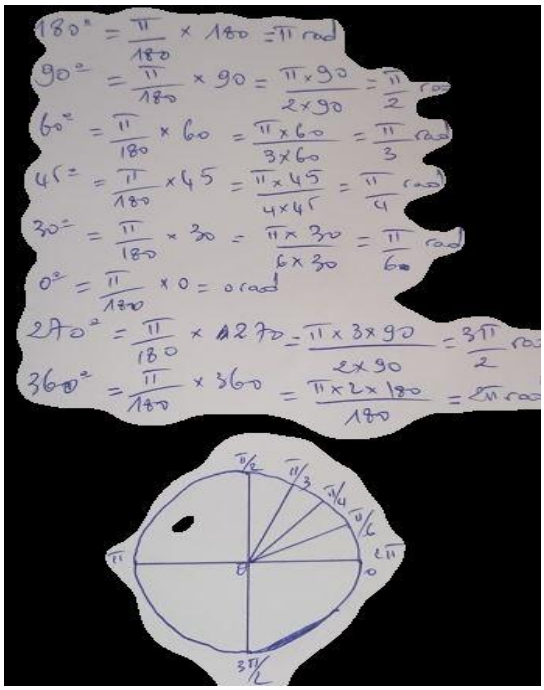

Figure 4. Text and graphic blocks segmentation.

4. Co-occurrence matrix and Haralick features based approach for text and graphic segmentation

The co-occurrence matrix with Haralick features 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 [20]. A GLCM element $p_{d,\theta}(i,j)$ is the joint probability of the gray level pairs i and j in a given direction θ separated by distance of d units. GLCM is a matrix that we can extract several pieces of information. For this, Haralick [21] proposed fourteen descriptors calculated from this matrix.

For this work, we have chosen 5 parameters.

- Energy:

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

- Entropy:

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

- Local homogeneity:

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

- Contrast:

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

- Correlation:

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

4.1. Features extraction

The choice and the determination of inter-pixels distance d and orientation θ is a crucial step which influence negatively or positively the final result. To calculate GLCM features (Energy, Entropy, Local homogeneity, Contrast and Correlation), Multi-distance and multi-direction can be determined. In this work we calculate them by using one distance $d = \{1\}$ and four orientations $\theta = \{0°, 45°, 90°, 135°\}$. So, in vector feature f, for each descriptor, we retained the mean of the values calculated in these four directions.

$$f = \{mean_\theta(eng); mean_\theta(ent); mean_\theta(homloc); mean_\theta(cnt); mean_\theta(corr)\}$$

4.2. Text and graphic blocks segmentation by CLARA clustering algorithm.

CLARA (Clustering Large Application) was developed by Kaufman and Rousseeuw in 1990 whose purpose was to reduce the computation time of PAM [22]. It allows manipulating large vectors and can deals with much larger data sets. CLARA works on several sample of database instead of working on the total population.

The main steps of this algorithm are as follows:

- Draw multiple samples of the data set.
- Apply PAM to each sample.
- Return the best clustering.

A sample shall be selected randomly in order to well represent the original data [23].

5. Results and interpretation

We have prepared a dataset-image of handwritten mathematical documents. This dataset contains 100 images which are composed by text and graphic components. To evaluate the performance of our approach, three performance indices are calculated: Image Segmentation Rate (ISR), Graphic Extraction Rate (GER) and Text Extraction Rate (TER).

$$ISR = \frac{Number\ of\ images\ correctly\ segmented}{Number\ of\ images}$$

$$GER = \frac{Number\ of\ graphic\ blocks\ \ correctly\ extracted}{Number\ of\ graphic\ blocks}$$

$$TER = \frac{Number\ of\ text\ blocks\ \ correctly\ extracted}{Number\ of\ text\ blocks}$$

Results obtained (table 1, table 2 and table 3) show that our approach can correctly segment most of the images into homogeneous zones (text and graphic) with very important precision.

Table 1. Results obtained by Gabor filter algorithm and K-means clustering algorithm.

GER	TER	ISR
93 %	97 %	92 %

Table 2. Results obtained by Gabor filter algorithm and CLARA clustering algorithm.

GER	TER	ISR
87 %	94 %	84 %

Table 3. Results obtained by co-occurrence matrix and Haralick features and K-means clustering algorithm.

GER	TER	ISR
85 %	90 %	81 %

Table 4. Results obtained by co-occurrence matrix and Haralick features and CLARA clustering algorithm.

GER	TER	ISR
80 %	88 %	78 %

The figures below (figure 5, figure 6 and figure 7) present the difference between both algorithms in terms of blocks segmentation.

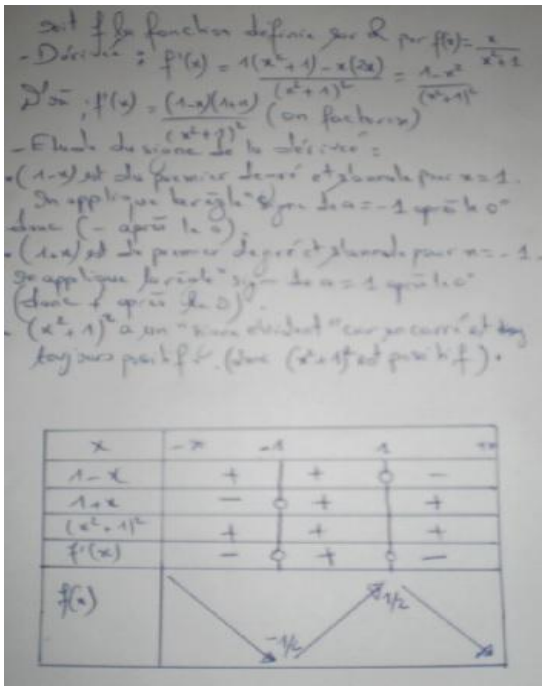

Figure 5. Original image.

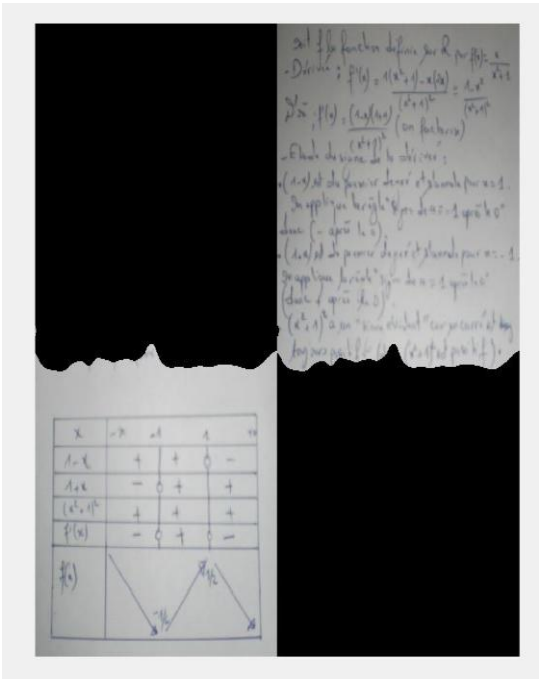

Figure 6. Text and graphic blocks segmentation by Gabor filter.

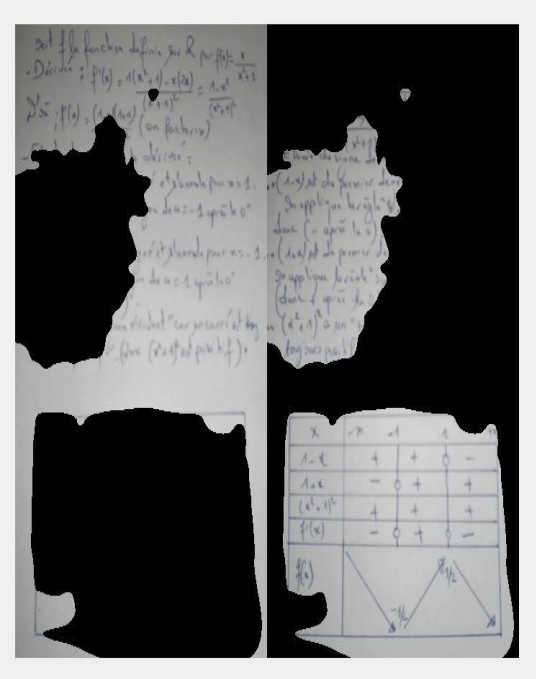

Figure 7. Text and graphic blocks segmentation by co-occurrence matrix and Haralick features.

Regarding the comparison between both algorithms in terms of processing time, we can conclude that generally the segmentation based on the texture requires an important processing time. But, the post-processing techniques that we have made allowed reducing significantly this time.

The main findings of this experiment are:

- Extraction rate of the graphic blocks is lower than that of text blocks for both algorithms.
- Image segmentation rate of the algorithm based on Gabor filter is much higher than that obtained by co-occurrence matrix.
- CLARA algorithm is less efficient than K-means for both text and graphic blocks extraction.
- Image segmentation rate using K-means algorithm is much higher than that obtained by CLARA clustering algorithm.
- Both algorithms require relatively the same processing time.

6. Conclusions

The objective of this work was to create a system which allows segmenting the handwritten mathematical document into two blocks: the first is supposed to contain mathematical text and the second is devoted to the graphical objects. For this, we have presented an approach based on the texture characteristics and more precisely the frequencies and orientations characteristics. This approach is based on the principal steps presented below:

- Pre-processing techniques.
- Filtering the document image by a bank of Gabor filter.
- Calculation of the features vector.
- Post-processing techniques.
- Segmentation of the image into text and graphic zones.

Finally, we have presented a comparison of the results obtained by this algorithm with the results obtained by another algorithm based on the co-occurrence matrix and Haralick features.

References

[1] F. M. Wahl, K. Y. Wong, and R. G. Casey, "Block segmentation and text extraction in mixed text/image documents," *Computer Graphics and Image Processing*, vol. 20, pp. 375-390, 1982.

[2] P. Chauvet, J. Lopez-Krahe, E. Taflin and H. Maître, "System for an intelligent office document analysis, recognition and description," *Signal Processing*, vol. 32, no. 1-2, pp. 161-190, 1993.

[3] D. Wang and S. N. Srihari, "Classification of newspaper image blocks using texture analysis," *Computer Vision, Graphics and Image Processing*, no. 47, pp. 327-352, 1989.

[4] G. Nagy, S. Seth, and M. Viswanathan, "A prototype document image analysis system for technical journals," *Computer*, vol. 25, no. 7, pp. 1022, 1992.

[5] M. Krishnamoorthy, G. Nagy, S. Seth, and M. Viswanathan, "Syntactic segmentation and labeling of digitized pages from technical journals," *IEEE Trans. Pattern Anal. and Machine Intell*, vol. 15, no. 7, pp. 737-747, 1993.

[6] L. A. Fletcher and R. Kasturi, "A robust algorithm for text string separation from mixed text/graphics images," *IEEE Trans. Pattern Anal. and Machine Intel l*, vol. 10, Nov. 1988.

[7] M. Tuceryan and Anil K. Jain, "Texture segmentation using voronoi polygons," *IEEE Trans. Pattern Anal. Mach. Intell*, 12(2) :211–216, 1990.

[8] J.M. Chassery and M. Melkemi, "Diagramme de voronoï appliqué à la segmentation d'images et à la détection d'évènements en

imageries multi-sources," *Traitement du Signal*, 8(3) :155–164, 1991.

[9] M. Tuceryan, "Moment-based texture segmentation," *Pattern recognition letters*, 15(7):659–668, July 1994.

[10] S. Mehta and S. Dhull, "Fuzzy based median filter for gray-scale images," *International Journal of Engineering Science and Advanced Technology*, vol.2 no.4, 975-980.

[11] Kh.M. Singh, "Fuzzy rule based median filter for gray-scale images," *Journal of Information Hiding Multimedia Signal Process*. 2011; 2(2).

[12] R. Mehra and R. Verma, "Median Filter for Noise Removal using Particle Swarm Optimization," *International Journal of Computer Applications* (0975 – 8887) Volume 138 – No.4, March 2016.

[13] M. Miciak, " Character Recognition Using Radon Transformation and Principal Component Analysis in Postal Applications," *Proceedings of the International Multiconference on Computer Science and Information Technology*, pp. 495 – 500.

[14] C. Hoilund, "The Radon Transform," *Aalborg University*, VGIS, 07gr721 November 12, 2007.

[15] M. D'Acunto, A. Benassi, D. Moroni and O. Salvetti, "3D image reconstruction using Radon transform," *Signal, Image and Video Processing*, January 2016, Volume 10, No 1, page 1–8.

[16] A. K. Jain and F. Farrokhnia, "Unsupervised Texture Segmentation Using Gabor Filters," *Pattern Recognition*, 24(12):1167 - 1186, 1991.

[17] A. K. Jain, S.K Bkattacharjee and Y. Chen, "On Texture In Document Images," Computer Vision and Pattern Recognition, Proceedings CVPR '92. IEEE Computer Society Conference, 677 – 680, 1992.

[18] A.K. Jain and S. Bhattacharjee, "Text segmentation using Gabor filters for automatic document processing," *Machine Vision and Applications*, Vol.5 no.3,1992.169-184.

[19] V. Štruc and N. Pavešić, "From Gabor Magnitude to Gabor Phase Features: Tackling the Problem of Face Recognition under Severe Illumination Changes".

[20] J. F. Haddon and J. F. Boyce, "Cooccurrence matrices for image analysis", *IEEE Electronic and Communications Engineering Journal*, 5(2):71–83, 1993.

[21] R. M.Haralick, K.Shanmugam and I.Dinstein, "Textural features for classification", *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, nov 1973.

[22] A. Tuan, I.Hanoi, "Réduction de base de données par la classification automatique", *Rapport de stage. Institut de la francophonie pour l'informatique (IFI).* 2004.

[23] L. Kaufman, P. J. Rousseeuw, "Finding Groups in Data: An introduction to cluster analysis", *Wiley Series in Probability and Mathematical Statistics*, 340 p,1990.