Binary Data Comparison using Similarity Indices and Principal Components Analysis

Nouhoun Kane, Khalid Aznag, Ahmed El Oirrak, and Mohammed Kaddioui
Department of Computer Science, Cadi Ayyad University, Morocco

Abstract: *This work is a study of binary data, especially binary images sources of information widely used. In general, comparing two binary images that represent the same content is not always easy because an image can undergo transformations like: Translation, rotation, affinity, resolution change and scale or change in appearance. In this paper, we will try to solve the translation and rotation problems. For translation case, the similarity indices are used between the image rows or blocks. In the case of rotation, first the coordinate's contours are extracted, then we compute the covariance matrix used in the Principal Components Analysis (PCA) and the corresponding eigen values which are invariant to this type of movement. We also, compare our approach having complexity O(M+N) to Hausdorff Distance (HD) that has complexity of O(M×N) for an M×N image. In our approach, HD is used only to compare distance between 1D signatures.*

Keywords: *Binary images, covariance matrix, similarity index, HD.*

Received December 10, 2013; accepted June 12, 2014; published online March 13, 2015

1. Introduction

To compare two images, we must first represent these images through effective descriptors. These descriptors represent general information on color, texture and shapes of the extracted image [24]. Their choice determines the effectiveness of the method and is a delicate step of indexing [23]. The color histogram is widely used as an indexing descriptor space [10].

Other characterizations of contours are possible such as fourier coefficients, eccentricity, euler number and moment invariants [11, 12, 27]. But, some methods like SIFT and SURF [5, 18] are not suitable for binary features.

On the other hand, characterizations based on the autocorrelation function are used for textured images [19]. However, textures are useful only for textured images which are a special case.

Discriminating descriptors once extracted are arranged to form the signature of the image. Signatures are then used to compare images [20]. This comparison must prove the degree of similarity between images. There are two ways to structure information to form a signature: Global signature and local or partial signature. The histogram as shown in Figure 2 [4] is an example of global signature and techniques used in this paper are also global.

This paper is organized as follows: Section 2 describes the problem of translation between binary images and how we can solve it by calculating a similarity measure between rows of the image. Section 3 presents the rotation problem in a picture and it turned out that the covariance matrix of contour coordinates is an invariant measure according to rotation. In section 4, we compare our proposed approach to Hausdorff based approach.

2. Translation and Similarity Index

Binary data is one of the most common representations of patterns and similarity measures between these types of data are essential in many problems such as: Clustering, classification, etc., Since, jaccard proposed a similarity measure to classify ecological species in 1901, many similarity measures and distances have been proposed in various areas. Implementing appropriate measures has for result more accurate data analysis.

For example, Jaccard similarity measure has been used for classification of ecological species [14]. Binary similarity measures were then applied in biology, anthropology, taxonomy, image retrieval, text retrieval, geology and chemistry [21, 25]. Recently they were actively used to solve identification of fingerprints, iris pictures problems and recognition of manuscripts characters [7, 8, 9]. Many articles discuss their properties and characteristics [13, 15].

The Simple Matching Coefficient (SMC) is a simple similarity index and the base of the most of other indices.

2.1. Simple Matching Coefficient

Given two objects A and B, each with n binary attributes, the SMC coefficient is a useful measure of the overlap that A and B share with their attributes. Each attribute of A and B can either be 0 or 1. The total number of each combination of attributes for both A and B are specified as follow:

- f_{11}: Represents the total number of attributes where A and B both have a value of 1.
- f_{01}: Represents the total number of attributes where the attribute of A is 0 and the attribute of B is 1.

- f_{10}: Represents the total number of attributes where the attribute of A is 1 and the attribute of B is 0.
- f_{00}: Represents the total number of attributes where A and B both have a value of 0.

Each attribute must fall into one of these categories, meaning that $f_{11}+f_{01}+f_{10}+f_{00}=n$.

This index is defined by:

$$SMC = \frac{|matching\ attributes|}{|attributes|} \tag{1}$$

Where $|matching\ attributes|=f_{11}+f_{00}$ represents pixels that are matched (values are 0 or 1) and $|attributes|=f_{01}+f_{10}+f_{11}+f_{00}$ represents the number of pixels on the two rows or vectors to be matched. Thus, we can rewrite the SMC index as follows:

$$SMC = \frac{f_{11}+f_{00}}{f_{01}+f_{10}+f_{11}+f_{00}} \tag{2}$$

2.2. Similarity between Image Rows or Columns using Jaccard Index

Principle used for the Jaccard index is similar to SMC except that no account is considered for pixels with values equal to 0.

$$Jaccard = \frac{|matching\ present\ attributes|}{|attributes\ values\ present|} \tag{3}$$

Where $|matching\ present\ attributes| = f_{11}$

$|attributes\ values\ present| = f_{01}+f_{10}+f_{11}$

In this work we chose to work with Jaccard index as shown in Figures 1 and 2 using Equation 4 because images backgrounds are black. Technique used is as follows:

- We fix a row on image (row containing the center of gravity of the object i.e., white pixels).
- An index of similarity between each row of image and the fixed row is calculated.
- Signature image is formed by the values of the index.

a) Translation along the x axis. b) Translation along the y axis.

Figure 1. Two images that represent the same content with added translations.

a) Similarity histogramme for image in Figure 1-a. b) Similarity histogramme for image in Figure 1-b.

Figure 2. Histograms of similarity indices.

Note, for images with white backgrounds, Jaccard coefficient must be changed because the black color is now the objects within the image, so we have:

$$Jaccard = \frac{f_{00}}{f_{01}+f_{10}+f_{00}} \tag{4}$$

We can also, work with image columns instead of rows.

2.3. Similarity between Image Blocks

In this section, we propose to compute similarity indices between image blocks. Blocks generally represent the neighbourhood of a pixel; this neighbourhood is a rich source of local information, which is not the case for rows that are source of global information.

The steps of proposed technique are as following: Cutting the image into blocks, for two successive blocks we calculate the similarity index. This work is repeated for all blocks of the image in opposite direction. Two successive blocks are defined as follow:

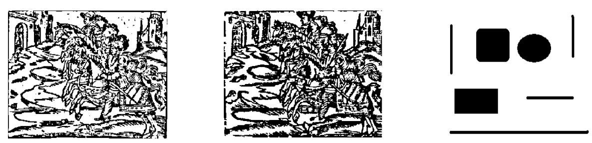
$$\begin{pmatrix} I(i-1,j-1) & I(i-1,j) & I(i-1,j+1) \\ I(i,j-1) & I(i,j) & I(i,j+1) \\ I(i+1,j-1) & I(i+1,j) & I(i+1,j+1) \end{pmatrix}$$

And

$$\begin{pmatrix} I(i-1,j+2) & I(i-1,j+3) & I(i-1,j+4) \\ I(i,j+2) & I(i,j+3) & I(i,j+4) \\ I(i+1,j+2) & I(i+1,j+3) & I(i+1,j+4) \end{pmatrix}$$

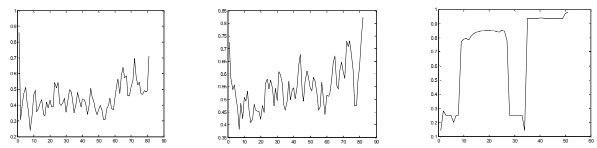
With i=2: M-4 and j=2: M-4. These two blocks are 8×8 neighbourhoods for pixel (i, j) and pixel (i, j+3). Thus, Jaccard index is now defined between two matrices of size 3×3.

In Figure 3, we show 3 images on which we apply similarity blocks to extract signatures shown in Figure 4. The vector of all similarity indices is a signature for the image as shown in Figure 4.



a) First tested image. b) Second tested image. c) Third tested image.

Figure 3. The three images used to test similarity between image blocks: Images in Figures 3-a and 3-b are close to each other while the third one in Figure 3-c is utterly different.



a) Signature for image in Figure 3-a. b) Signature for image in Figure 3-b. c) Signature for image in Figure 3-c.

Figure 4. Signatures for images in Figure 3: Both first signatures (a and b) are similar while the third one (c) is different.

In Table 1, a Hausdorff Distance (HD) is computed between signatures of the three images. Let J_1, J_2 and J_3 denote vectors signatures for images in Figures 3-a, b and c respectively. Then:

Table 1. HD between images signatures.

$HD(J_1, J_2)$	$HD(J_1, J_3)$	$HD(J_2, J_3)$
0.00	22.52	23.50

We use the discrete *HD* defined by *HD* (*A*, *B*)= *max*(*h*(*A*, *B*), *h*(*B*, *A*)).

Where $h(A,B) = \max_{a \in A} \min_{b \in B} \|a - b\|$.

We can also use Dynamic Time Warping (DTW) [16] or Chamfer Distance [2].

3. Rotation and PCA

PCA has its source in an article published in 1901 by Karl Pearson [1, 26, 28]. Here, we give a brief description of PCA principle (Suppose $x_1, x_2, ..., x_M$ are $N\times1$ vectors):

- *Step* 1: $\bar{x} = \frac{1}{M}\sum_{i=1}^{M} x_i$
- *Step* 2: Subtract the mean:

$$\Phi_i = x_i - \bar{x}$$

- *Step* 3: Form the matrix:

$$A = [\Phi_1 \Phi_2 \cdots \Phi_M]$$

($N\times M$ matrix), then compute:

$$C = \frac{1}{M}\sum_{n=1}^{M} \Phi_n \Phi_n^T = AA^T$$

(Sample covariance matrix, $N\times N$, characterizes the scatter of the data)

- *Step* 4: Compute the Eigen values of:

$$C: \lambda_1 > \lambda_2 > \cdots > \lambda_N$$

- *Step* 5: Compute the Eigen vectors of:

$$C: u_1, u_2, \cdots, u_N$$

Since, *C* is symmetric, $u_1, u_2, ..., u_N$ from a basic, (i.e., any vector *x* or actually $(x - \bar{x})$, can be written as a linear combination of the Eigen vectors):

$$x - \bar{x} = b_1 u_1 + b_2 u_2 + \cdots + b_N u_N = \sum_{i=1}^{N} b_i u_i$$

Thus, the Principal Component Analysis (PCA) uses a matrix indicating the degree of similarity between variables and then we compute projection matrix for variables in the new space. This symmetric matrix, which includes the variance of variables on the diagonal and elsewhere is called covariance matrix. Covariance measures the degree of independence of two variables.

Under the action of rotation contours coordinates of the second image are related to coordinates contours of the first image by:

$$\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{5}$$

For all points.

$$\begin{Bmatrix} x \\ y \end{Bmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix}_{1 \le i \le n}$$

Where $\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix}$: Denotes contours vectors of second. Image $\begin{Bmatrix} \hat{x} \\ \hat{y} \end{Bmatrix} = \begin{pmatrix} \hat{x}_i \\ \hat{y}_i \end{pmatrix}_{1 \le i \le n}$

The analytical writing is given by:

$$\hat{x} = \cos\theta x_i + \sin\theta y_i$$

$$\hat{y} = -\sin\theta x_i + \cos\theta y_i$$

Let *M* denotes:

$$M = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$$

And

$$\hat{M} = \begin{pmatrix} \hat{x}_i \\ \hat{y}_i \end{pmatrix}$$

The covariance matrix for transformed image is:

$$\hat{V} = \hat{M}\hat{M}^t = (\hat{x}_j \quad \hat{y}_j)\begin{pmatrix} \hat{x}_i \\ \hat{y}_i \end{pmatrix} \tag{6}$$

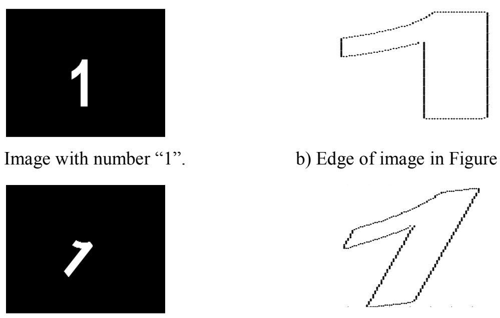
For i=1, ..., *n* and *j*=1, ..., *n* so, we have:

$$\begin{aligned} \hat{x}_i\hat{x}_j + \hat{y}_i\hat{y}_j &= (\cos\theta x_i + \sin\theta y_i)(\cos\theta x_j + \sin\theta y_j) + \\ &\quad (-\sin\theta x_i + \cos\theta y_i)(-\sin\theta x_j + \cos\theta y_j) \\ &= \cos^2\theta x_i x_j + \sin^2\theta x_i x_j + \cos^2\theta y_i y_j + \\ &\quad \sin^2\theta y_i y_j + \cos\theta\sin\theta y_i x_j - \cos\theta\sin\theta y_i x_j \\ &= x_i x_j + y_i y_j \end{aligned}$$

Which proves invariance under rotation.

3.1. Experiments 1

In the following experiment as shown in Figure 5, we apply this technique to eliminate rotation and extract two invariant values for each image. Those values are the two largest Eigen values of the covariance matrix as shown in Table 2.



a) Image with number "1". b) Edge of image in Figure 5-a.

c) Image with rotated number "1". d) Edge of image in Figure 5-c.

Figure 5. Two images used in experiments and their corresponding edges.

Table 2. The two largest Eigen values of the covariance matrix for contours.

Two Eigen Values for V	Two Eigen Values for $\hat{V}$
0.0287	0.0390
3.5376	3.4347

3.2. Experiments 2

For synthetic contours Figures 6 and 7, we obtain a perfect result, the following 2D curve was created using the parametric equation:

$$\begin{cases} x(t) = 2\cos(t) \\ y(t) = \sin(t) + \dfrac{1}{2}\sin(5t) \end{cases} \quad t \in [0, 2\pi] \tag{7}$$

a) Curve of Equation 7. b) Rotated curve of Equation 7.

Figure 6. Synthetic contour in Figure 6-a and the same contour in Figure 6-b after applying a synthetic rotation.

a) Covariance matrix of contour in Figure 6-a b) Covariance matrix of contour in Figure 6-b.

Figure 7. 2D display of the covariance matrix for contours.

To test our technique to noise resistance, noise with different percentage is added Figure 8 and values in Table 4 does not change much compared to one illustrated in Table 3.

Table 3. The two largest eigen values of covariance matrices in Figures 7-a and 7-b respectively.

Synthetic Contour	Transformed Contour
6.25	6.25
20.40	20.40

Table 4. The two largest eigen values of the covariance matrix in Figures 8-b and 8-d.

Noisy Contour (20% added)	Noisy Contour (10% added)
6.24	6.28
20.55	20.50

a) Edge with 10% noise added. b) Covariance matrix for edge in Figure 8-a.

c) Edge with 20% noise added d) Covariance matrix for edge in Figure 8-c.

Figure 8. Noisy edges and 2D display of the covariance matrix.

4. Comparison

In the work [3] an approach for binary image comparison without feature extraction was presented. It uses the windowed *HD* in a pixel adaptive way. They measure *HD* not between two full images, but between subimages extracted by a window. It is therefore necessary to define the extent of the *HD* in a window. This amounts to modify the definition of the overall measurement by introducing a restriction to sets of points at the window.

Let *A* and *B* be two bounded sets:

$$HD_w(A,B) = \max\left(h_w(A,B), h_w(B,A)\right)$$

Where *w* denotes a window and:

$$h_w(A,B) = \max_{a \in A} \min_{b \in B} \left\| a - b \right\|$$

The size of the *w* window must be fixed. This can be done by the user or automatically and globally or even automatically and locally according to the local environment.

This distance is not invariant to rotation and under translation Figure 9. We have $HD_w(A, A+V)=|V|$. So, they did not take care of the situation where images might be rotated.

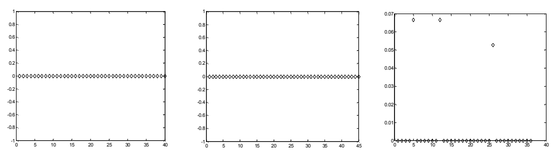
For two binary images, assuming that they have the same resolution and same orientation of object(s) in the images, the map of all local dissimilarity measures i.e., CDL Figure 10 made on different places images constitute the proposed signature in this work.

a) Bold line (left). b) Bold line (right). c) Dot line (right).

Figure 9. Image A: A bold line on the left Image B: A bold line to the right. Image C: A dots line on the right.

a) CDL between Figures 9-a and c. b) CDL between Figures 9-a and c.

Figure 10. CDL between the images A and B. There are two distance values in the CDL: 0 (blue) and 12 (pink), which is the value of the overall HD. CDL between image A and the image C.



a) Signature for image in Figure 9-a. b) Signature for image in Figure 9-b. c) Signature for image in Figure 9-c.

Figure 11. Jaccard similarity indices between the image rows for image A (Figure 9-a), image B (Figure 9-b) and image C (Figure 9-c).

Using similarity indices as shown in Figure 11, we computed signatures of images in Figure 9. We can see that both images a and b have the same signatures (zero in this case), while signature for image c has three non-nulls (different from zero) values.

CDL in Figure 10 cannot tell us if images in Figure 9 are the same or not as it gives 2 lines Figures 10-a and b).

5. Conclusions

This work was the opportunity to expose problems related to automatic recognition of individual binary images. There are two main techniques for extracting a signature from binary images: The first one uses a similarity index between rows or blocks of the image, the second uses covariance matrix to eliminate the rotation transformation effect.

In the paper of Baudrier *et. al*. [3] we can see that the HD_w is not invariant according to translation. We also, proved in this work that HD_w is invariant according to rotation using covariance matrix and its Eigen values.

In future work, as each row or column in binary image can be considered as 1D vector (so we have N or M binary vector), we can reuse algorithm such as BRIEF, ORB or BRISK [6, 17, 22] categorized as binary valued features and are reserved to binary dataset.

References

[1] Aldrich J., "Karl Pearson's Biometrika: 1901-36," *Biometrika*, vol. 100, no. 1, pp. 3-15, 2013.

[2] Barrow H., Tenenbaum J., Bolles R., and Wolf H., "Parametric Correspondence and Chamfer Matching: Two New Techniques for Image Matching," available at: http://ijcai.org/Past%20Proceedings/IJCAI-77-VOL2/PDF/024.pdf, last visited 1997.

[3] Baudrier E., Millon G., Nicolier F., and Ruan S., "A Fast Binary-Image Comparison Method with Local Dissimilarity Quantification," *in Proceedings of the 18th Conference on Pattern Recognition*, Hong Kong, pp. 216-219, 2006.

[4] Brunelli R. and Mich O., "Histograms Analysis for Image Retrieval," *Pattern Recognition*, vol. 34, no. 8, pp.1625-1637, 2001.

[5] Bay H., Tuytelaars T., and Van L., "SURF: Speeded up Robust Features," *in Proceedings of the 9th European Conference on Computer Vision*, Graz, Austria, pp. 404-417, 2006.

[6] Calonder M., Lepetit V., Strecha C., and Fua P., "BRIEF: Binary Robust Independent Elementary Features," *in Proceedings of the 11th European Conference on Computer Vision*, Crete, Greece, pp. 778-792, 2010.

[7] Cha H. and Srihari N., "A Fast Nearest Neighbor Search Algorithm by Filtration," *Pattern Recognition*, vol. 35, no. 2, pp. 515-525, 2000.

[8] Cha S., Tappert C., and Srihari S., "Optimizing Binary Feature Vector Similarity Measure using Genetic Algorithm," *in Proceedings of the 7th International Conference on*, Edinburgh, Scotland, pp. 662-665 , 2003.

[9] Cha S., Yoon S., and Tappert C., "Enhancing Binary Feature Vector Similarity Measures," available at: http://digitalcommons.pace.edu/cgi/viewcontent.cgi?article=1017&context=csis_tech_reports, last visited 2006.

[10] Ciocca G. and Schettini R., "Using a Relevance Feedback Mechanism to Improve Content-based Image Retrieval," *in Proceedings of the 3rd International Conference, VISUAL'99 Amsterdam*, pp. 107-114, 1999.

[11] Derrode S. and Ghorbel F., "Robust and Efficient Fourier-mellin Transform Approximations for Invariant Grey-level Image Description and Reconstruction," *Computer Vision and Image Understanding*, vol. 83, no. 1, pp. 57-78, 2001.

[12] Hu K., "Visual Pattern Recognition by Moment Invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179-187, 1962.

[13] Hubalek Z., "Coefficients of Association and Similarity, based on Binary (Presence-Absence) Data: An Evaluation," *Biological Reviews*, vol. 57, no. 4, pp. 669-689, 1982.

[14] Jaccard P., "Étude Comparative De La Distribuition Florale Dans Une Portion Des Alpes Et Des Jura," *Bull Soc Vandoise Sci Nat,* vol. 37, pp. 547-579, 1901.

[15] Kuhns L., "The Continuum of Coefficients of Association," *in Proceedings of Statistical Association Methods for Mechanized Documentation, National Bureau of Standards*, Washington, pp. 33-39, 1965.

[16] Lemire D., "Faster Retrieval with a Two Pass Dynamic Time Warping Lower Bound," *Pattern Recognition*, vol. 42, no. 9, pp. 2169-2180, 2009.

[17] Leutenegger S., Chli M., and Siegwart R., "BRISK: Binary Robust Invariant Scalable Keypoints," *in Proceedings of IEEE International on Computer Vision*, Barcelona, pp. 2548-2555, 2011.

[18] Lowie G., "Distinctive Image Features from Scale-Invariant Keypoints," *the International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[19] Maddess T., Nagai Y., James C., and Ankiewcz A., "Binary and Ternary Textures Containing Higher-Order Spatial Correlations," *Vision Research*, vol. 44, no. 11, pp.1093-1113, 2004.

[20] Philipp S., Vieira B., and Sanfourche M., "Fuzzy Segmentation of Color Images and Indexing of Fuzzy Regions," *in proceedings of Conference on Colour in Graphics, Imaging, and Vision*, Poitiers, pp. 507-512, 2002.

[21] Rehab D. and Rania A., "A Hierarchical K-NN Classifier for Textual Data," *the International Arab Journal of Information Technology*, vol. 8, no. 3, pp. 251-259, 2011.

[22] Rublee E., Rabaud V., Konolige K., and Bradski G., "ORB: An Efficient Alternative to SIFT or SURF," *in Proceedings of IEEE International*

Conference on Computer Vision, Barcelona, pp. 2564-2571, 2011.

[23] Sameer A., Rangachar K., and Ramesh J., "A Survey on the Use of Pattern Recognition Methods for Abstraction, Indexing and Retrieval of Images and Video," *Pattern Recognition*, vol. 35, no. 4, pp. 945-965, 2002.

[24] Smeulders A., Worring M., Santini S., Gupta A., and Jain R., "Content based Image Retrieval at the end of the Early Years," *IEEE Transcrition on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349-1380, 2000.

[25] Smith R. and Chang F., "Automated Binary Texture Feature Sets for Image Retrieval," *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlantic, pp. 2239- 2242, 1996.

[26] Theodore P. and Karl P., *The Scientific Life in a Statistical Age*, Princeton University Press, 2005.

[27] Veltkamp R. and Hagedoorn M., "State-of-the-art in Shape Matching," *Technical Report UU-CS-1999-27*, Utrecht University, the Netherlands, 1999.

[28] Walker M., "The Contributions of Karl Pearson," *Journal of the American Statistical Association*, vol. 53, no. 281, pp. 11-22, 1958.

Nouhoun Kane is a PhD student at the Department of Computer Science, Faculty of Science Semlalia, Cady Ayyad University, Morocco. His current research interests are signal, text and image processing.

Khalid Aznag Is a PhD student at the Department of Computer Science, Faculty of Science Semlalia, Cady Ayyad University, Morocco. His current research interests are 2D and 3D images and 2D curve.

Ahmed El Oirrak joined Cadi Ayyad University, Morocco, in 1999, first as an assistant professor, and received the Doctorate and Habilitation in signal processing from the Mohammed V University, Morocco, in 2001 and University Cadi Ayyad, Morocco, in 2010 respectively. He is presently a PH professor with the Faculty of Sciences of Marrakech Semlalia. His research interests include image processing, pattern recognition and their applications. He is the author of more than 20 publications.

Mohammed Kaddioui is a full professor of computer science at the Department of Computer Science, Faculty of Science Semlalia, Cadi Ayyad University, Morocco. His major field of study is information processing and management and computer graphics.