Document Representation Using Wavelet Decomposition

Nouhoun Kane[1], Ahmed El Oirrak[2]

[1, 2]Faculté des sciences Semlalia, Dept Informatique, LISI, Marrakech, Maroc

[1]nouhounkane@gmail.com; [2]oirrak@yahoo.fr

***Abstract*-In this paper we use Discrete Wavelet Decomposition (DWT) for comparing document. First documents are transformed to ASCII code; then for each level DTW coefficients are extracted from ASCII code.**

Performances are measured using real's documents. A simple comparison is made between DWT and Bag of Word (BOW) representation to show goodness of proposed technique.

Keywords- Component; Text Representation; Discrete Wavelet Transform (DWT); Bag of Word(BOW) Representation

I. INTRODUCTION

Research in automatic document recognition (and documents synthesis) by machine has attracted a great deal of attention over the past five decades.

In general, there are two common frameworks for document representation [1]: a vector space model [2] and a probabilistic model [3, 4]. In Vector Space Model, each document is represented as a vector, which indexes are all words (terms) used in a given text collection. Representations based on frequency of terms do not reflect pertinence of terms in the document. Some terms with low frequencies can play importance in topic of sentences or the document. Otherwise, some terms with high frequencies can be irredundant or noisy.

This article presents an alternative representation in document recognition using wavelet decomposition.

The paper is organized as follow:

In Section II the DWT is introduced; Section III presents the classic representation (BOW) and the proposed one.

II. DISCRETE WAVELET TRANSFORM (DWT)

Wavelets are mathematical functions [5, 6] that cut up data into different frequency components, and then study each component with a resolution matched to its scale. They have advantages over traditional Fourier methods in analyzing physical situations where the signal contains.

Regardless of its scale and magnitude, a function is admissible as a wavelet

If and only if

$$\int_{-\infty}^{\infty} |\psi(t)|^2 dt \prec \infty$$

$$C_\psi = 2\pi \int_{-\infty}^{\infty} \frac{|\Psi(w)|^2}{|w|} dw \prec \infty$$

The continuous wavelet transform of a function s (t) (assumed to have zero mean and finite energy) is defined as

$$S(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \psi' \left(\frac{t-b}{a} \right) s(t) dt$$

ψ' complex conjugate of ψ;

$b \in \Re$ the time shift;

$a \succ 0$ the scale of the analyzing wavelet.

If we denote by $\psi_{a,b}(t) = a^{-1/2}\psi\left(\frac{t-b}{a}\right)$ the continuous wavelet transform is generally expressed with the following integral

$$S(b,a) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} \psi'_{a,b}(t)s(t)dt$$

In the discrete domain, the scale and shift parameters are discretized as $a = a_0^m$ and $b = nb_0$

The analyzing wavelets are also discretized as follow:

$$\psi_{m,n}(t) = a_0^{-m/2}\psi\left(\frac{t-nb_0}{a_0^m}\right)$$

Where *m* and *n* are integer values.

The Discrete Wavelet Transform (DWT) and its inverse transform are defined as follow:

$$S_{m,n} = \int_{-\infty}^{\infty} \psi'_{m,n}(t)s(t)dt$$

$$s(t) = k_\psi \sum_m \sum_n S_{m,n}\psi_{m,n}(t)$$

Where k_ψ is a constant value for normalization.

The wavelet used in this work is Haar wavelet defined as follow:

$$\psi(t) = \begin{cases} 1 & if\ 0 \prec t \le 1/2 \\ -1 & if\ 1/2 \prec t \le 1 \\ 0 & otherwise \end{cases}$$

III. TEXT REPRESENTATIONS

A. BOW Representation

In the context of studying similarities between text documents, several applications and methods are used. Here, we are going to compare two different methods: BOWR and DWT. The BOWR is based on the search of keywords in a text. The presence of a keyword in a text is denoted by 1 and its absence by 0, which gives results in a matrix of 0 and 1.

B. DWT Representation

In these simple experimentation two documents D3 and D4, of size 1035 and 1425 respectively are represented by two DWT of size 249 and 339. First the DWT representation allows discrimination between texts as shown in Figs. 1 and 2. Two different texts have two different representations.

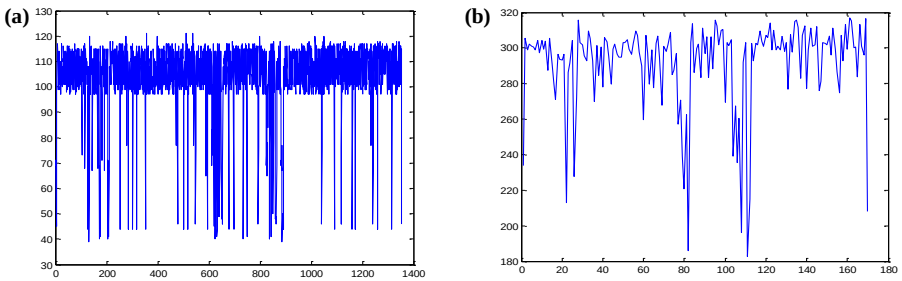

Fig. 1 ASCII code representation of Document_3 (D3) and Wavelet of order 2

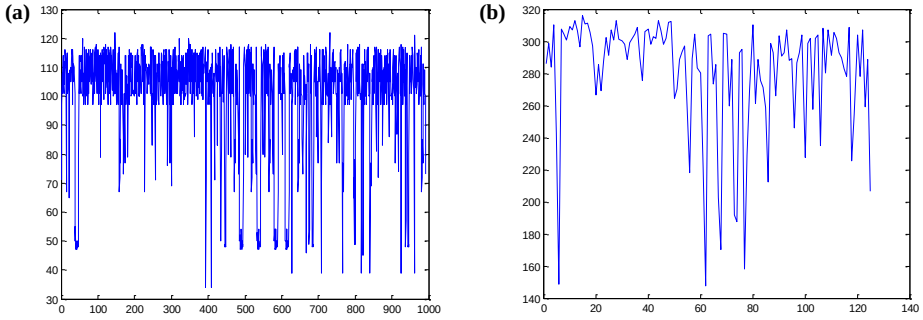

Fig. 2 ASCII code representation of Document_4 (D4) and Wavelet of order 2

C. *Comparison Between DWT and BOW Representation*

In the following, we will try to make a short comparison between BOWR and DWT. We chose five documents D1, D2, D3, D4 and D5 containing almost the same keywords. We chose the following keywords: “Information”, “System”, “Network” and “wireless” We entered the documents and the keywords in our web application developed in PHP programming language for this experience. Here are the results:

	D1	D2	D3	D4	D5
information	0	0	1	1	1
system	1	0	0	0	0
network	1	1	1	1	1
wireless	0	1	1	1	0

The word “information” is found in documents D3, D4 and D5.

The word “system” is found only in D1.

The word “network” is found in documents D1, D2, D3, D4 and D5.

The word “wireless” is present in documents D2, D3 and D4.

To measure the degree of similarity, we propose here to use of the Jaccard index. This one is useful for studying similarities between objects made of binary attributes.

Let A and B two sets of binaries, the index Jaccard is defined as follow:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

The Jaccard index between our documents is shown in the table below:

	D1	D2	D3	D4	D5
D1	1.00	0.33	0.25	0.25	0.33
D2	0.33	1.00	0.67	0.67	0.33
D3	0.25	0.67	1.00	1.00	0.67
D4	0.25	0.67	1.00	1.00	0.67
D5	0.33	0.33	0.67	0.67	1.00

According to our dictionary {“information”, “system”, “network”, “wireless”}, documents D3 and D4 are identical because the Jaccard index is 1, D2 and D3 are similar as the Jaccard index is 0.67 which is close to 1, etc. While Fig. 1b and Fig. 2b illustrate two different graphs for D3 and D4. Thus the BOWR representation gives an idea about the key words in the document but the content can be different from one document to another, even if they have the same keywords.

Representation using wavelet is more accurate if we limit ourselves to higher coefficients of the transformation; information is lost if we go down to the lower levels of the transformation.

Better representation is to combine the two representations to get an idea of the general structure and content of the document.

IV. CONCLUSION

This research describes a new method for document recognition using wavelet transform features.

In future research will be also performed on spectral analysis, as has been performed in wavelet domain for document analysis. In principle, this research is a modification of the previous methods which is applied for speech recognition. The differences between the present recognition method and the previous method lie in the features selected for analysis and in the length of the period for extracting the wavelet features.

The number of levels of wavelet decomposition and the type of decomposition are different from the previous methods applied for speech recognition.

REFERENCES

[1] S. Doan. A fuzzy-based approach to text representation in text categorization. In Proceeding of 14th IEEE Int'l Conference onn Fuzzy Systems - FUZZ-IEEE 2005, pp. 1008–1013, Nevada, U.S., 2005.

[2] Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing.Communications of the ACM, vol. 18, iss. 11, pp. 613–620, 1975.

[3] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In Proceedings 10th European Conference on Machine Learning (ECML), pp. 137–142, 1998.

[4] I. Moulinier and J. G. Ganascia. Applying an existing machine learning algorithm to text categorization. In S.Wermter, E. Riloff, and G. Schaler, editors, Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing, pp. 343–354. Springer-Verlag, Heidelbeg, 1996.

[5] Amara Graps, An introduction to wavelets, IEEE Computational Science and Engineering, vol. 2, iss. 2, pp. 51–60, 1995.

[6] Shivesh Ranjan, Exploring the Discrete Wavelet Transform as a Tool for Hindi Speech Recognition, International Journal of Computer Theory and Engineering, vol. 2, iss. 4, pp. 1793–8201, Aug., 2010.