A Semantic Approach for Mathematical Expression Retrieval

Zahra Asebriy
Dept of Applied Mathematics and Computer Science Laboratory (LAMAI), Cadi Ayyad University, Marrakesh, Morocco

Said Raghay
Dept of Applied Mathematics and Computer Science Laboratory (LAMAI), Cadi Ayyad University, Marrakesh, Morocco

Soulaimane Kaloun
SAEED, Higher School of Technology, Cadi Ayyad University, Essaouira, Morocco

Omar Bencharef
SAEED, Higher School of Technology, Cadi Ayyad University, Essaouira, Morocco

***Abstract*—Math search or mathematical expression retrieval has become a challenging task. Mathematical expressions are very complex, they are highly symbolic, and they have a semantic meaning that we should respect. In this paper, we propose a similarity search method for mathematical expression based on a multilevel representation of expressions and a multilevel search. We used the K-Nearest Neighbors with three types of distances to evaluate relevance between expressions. In the experimental level, the proposed system significantly outperforms statistical algorithms.**

***Keywords*—Mathematical expression; Retrieval information; MathML; Semantic similarity**

I. Introduction

The fast expansion of information technology and the spread of digital libraries in different domains make search engines necessary to help users to share and to retrieve any information from the web or from numerical libraries.

Now days, search engines can search all kinds of documents including text, image, audio and video. Therefore, documents containing special data such as mathematical expressions, tables, diagrams and drawings cannot be retrieved by classical search engines.

As distinct from text retrieval, retrieving math expressions have been researched for several years. Now it's still in the research stage. There are a few researches in the field of Mathematical Expressions Retrieval (MER) and a few number search engines dedicated to this subject, like MathFind [1], Active Math [2] Wolfram search, Wikipedia search formulas and MathWeb search. Most of these number search engines are based on text retrieval techniques.

Mathematical expressions are highly symbolic and they have their own structures. For example [3]:

- The order of elements in the mathematical expressions has semantic meaning, for example, $\sum \sin(\exp(x))$ and $\sum \exp(\sin(x))$are two completely different expressions with the same elements but do not have the same orders. So it is important to respect the order of the elements in the mathematical search to retrieve the right expression.
- if there are two math expressions $(a+b)$ and $\sqrt{(a+b)}$, the role of sub-expression $(a+b)$ in each expression is different, and if the query is to find the square root of $(a+b)$, the system must consider this particularity and all relevant expressions $\sqrt{(a+b)}$ should be strongly ranked.
- Mathematical equations can be written with different notation but they can have the same semantic meaning for example $(a+b)$ and $(x+y)$ are the same expressions.

Retrieving mathematical formulas with all these constraints requires a system based on semantic representations of the query math expressions. As examples of these representations, There are several common Mathematical Markup Languages: Latex [4] OpenMath [5], ASCII [6] and MathML[7].

MathML is an application of XML (Extensible Markup Language) for encoding notational and semantic structure of mathematical expressions. Actually, it is used by many systems for retrieving math expressions on the web [8, 9, 3].

In this paper, we propose an algorithm to extract features vectors of mathematical expressions represented on MathML and a multilevel search algorithm based on K-Nearest Neighbor's. We are going to use a variety of distances measure, first to evaluate the efficiency of our system and to find the best one for our system.

II. Related Work

Retrieving mathematical equations has attracted much attention from researchers in the past decade, and several related systems or methods on this task have been reported. Currently, several researches have been realized to develop and improve retrieving mathematical equations from the web or in digital library.

The system proposed by Yokoi et al. [10] was a new similarity search scheme for mathematical expressions. They

started by introducing a similarity measure based on Subpath Set and proposed a MathML conversion that is apt for it. The aim of this method used is to return similar equations by measuring the similarity using tree matching techniques and by reforming the structure of content based MathML. Based on their First experiences, they believe that their proposed system has the potential to provide a flexible interface for searching mathematical expressions on the web.

Tam T. Nguyen et al. [3] presented a lattice-based approach for mathematical search using Formal Concept Analysis (FCB) which is a powerful data analysis used for information retrieval [11]. This approach involves several phases. In the first time, they extract features from code MathML representation. These features are used to construct a mathematical lattice construction. At the query retrieval phase, the query expression is processed and inserted into math concept lattice, which matches with math expressions concept in the concept lattice to rank the relevant math expressions. The results have shown that the proposed approach has performed better than the conventional best match retrieval technique. Another important advantage of the proposed lattice-based approach lies in its support for the visualization and navigation of search results via a dynamic graph.

In their work [12], S-Q Yang and X-D Tian tried to research and develop special retrieval method. They proposed a maintenance algorithm of mathematical expression index based on Formula Description Structure (FDS), which includes the index item searching, inserting and deleting operations. Moreover, S.Q Yang et al. designed a matching model of mathematical expressions based on Formula Description Structure (FDS) index [13]. For realizing exact matching, the math retrieval attributes were embedded an index in three query modes called global query mode, local query mode and operational query mode.

L. Gao et al. [14] proposed a semantic enrichment technique to retrieve mathematical formulae from web pages and PDF documents with a novel query input interface, which allows users to copy formula queries directly from PDF documents without using formulas with Markup languages. They used a novel indexing and matching to search similar mathematical expressions based on both textual and spatial similarities. The proposed system achieves better performance compared with two representative mathematics retrieval systems.

MathSearch [15] is a formula-based search engine for mathematical information on the internet. In this system, Mathematical formula Query Language (MQL) [16] was designed for expressing and processing query. MQL contains two forms: a character string form (MQLS) and XML form (MQLX). By MQLX and MQLS, semantics query wildcard and combination query can be accomplished in MathSearch.

WikiMirs [17] is a tool to facilitate mathematical formula retrieval in Wikipedia. This system involves several phases. In the first phase, this system normalized Latex formulas of Wikipedia into a unified mode. Then terms were extracted from the normalized presentation tree. These terms reflect the features of the expression through series of processors such as presentation tree parser normalize and term extractor. These extracted terms used to establish an inverted index. In the last step, users query math expressions in latex form were processed with the above steps and retrieved.

III. Feature Extraction

A. Processing of Mathematical expressions

The processing of mathematical expressions as uniform mathematical representation plays an important role in the area of math search systems and digital libraries.

In this research, we use MathML for encoding notational and semantic structure of mathematical expressions (show examples 1&2). Currently, it is used by many systems for retrieving math expressions on the web [8, 9, 3], and by other applications like code MathML translator to Nemeth Braille code [18].

Example 1: Code MathML of math expression $x + y = 2$

```
<math>
<mrow>
<mi>x</mi><mo>+</mo><mi>y</mi><mo>=</mo><mn>2<
    /mn>
</mrow>
</math>
```

Example 2: Code MathML of math expression $\frac{2x^2+\sqrt{1+x}}{x}$

Tapez une équation ici.

```
<math>
<mrow>
<mfrac>
<mrow>
<mn>2</mn><msup>
<mi>x</mi>
<mn>2</mn>
</msup>
<mo>+</mo><msqrt>
<mrow>
<mn>1</mn><mo>+</mo><mi>x</mi></mrow>
</msqrt>
</mrow>
<mi>x</mi>
</mfrac>
</mrow>
</math>
```

B. Extraction process

To find a structural and semantic similarity between mathematical expressions in a big data base or on the web scale level we need a reduced and efficient representation that respects the structural and semantic specification of mathematical formula. First we choose to act with a simple algorithm that counts the number of occurrence of each operator and Math function. In the second algorithm we propose to use a multilevel representation of expressions.

1) Statistical algorithm

It extracts through the MathML code the number of each operator $(+,-,*,/)$, variables, constants, and functions $(log, sin, cos, exp...)$ and stores them in a vector (Fig. 1).

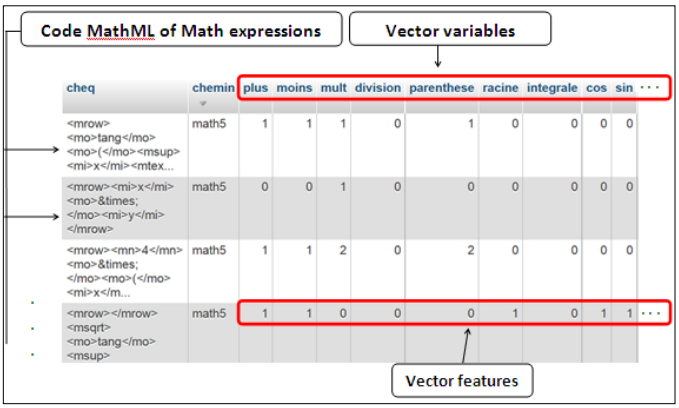


Fig. 1. Example of extracted vectors applied to dataset examples, using algorithm 1

The statistical approach is fast and reduced and in many cases it can detect real similarity between mathematical expressions. On the other hand using just the number of occurrence of each expression can produce false similarity detection like the case in examples 1 and 2 in Fig. 2.

Example 1: $1 + e^{x(x+1)}$ and $x(x+3) + e^2$

Example 2: $\sqrt{y+1}$ and $x + \sqrt{1}$

Fig. 2. Example of false similarity obtained using statistical algorithm

It's clear that there is no semantic similarity in the two examples of Fig. 2. As a result we need an algorithm also fast and more efficient regarding semantic similarity.

2) Proposed method

In order to define different levels for each math formula, we need to convert all mathematical equations into MathML code.

The first level was established by searching all main operators (+, -, *, /) linking all brackets and functions (trigonometric, logarithmic and algebraic) which all expressions into brackets and arguments of functions were defined and replaced by the term "***exp***". For example:

$\sqrt{x+1} + x^2 ln(x^2+1)$ Become $\sqrt{exp} + exp\ ln(exp)$

The values of each ***"exp"*** are stored to be used in the second level.

The vector of level 1 is the outcome of the statistical algorithm applied to the reduced expression. What gives: one $\sqrt{}$, one ln, and 3 exp. So the representative vector of level one becomes:

(3,0,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0)

These features were extracted and stored in the vectors $\boldsymbol{Vl_1}$. In the 2nd level, we treated each "*exp*" by repeating the same procedure used in the first level and we stored the extracted features in the vectors $\boldsymbol{Vl_2}$. We continued applying this method until obtained all levels and all vectors $\boldsymbol{Vli(Vl_1, Vl_2,...,Vl_n)}$ for each equation. In practice we use only 3 levels.

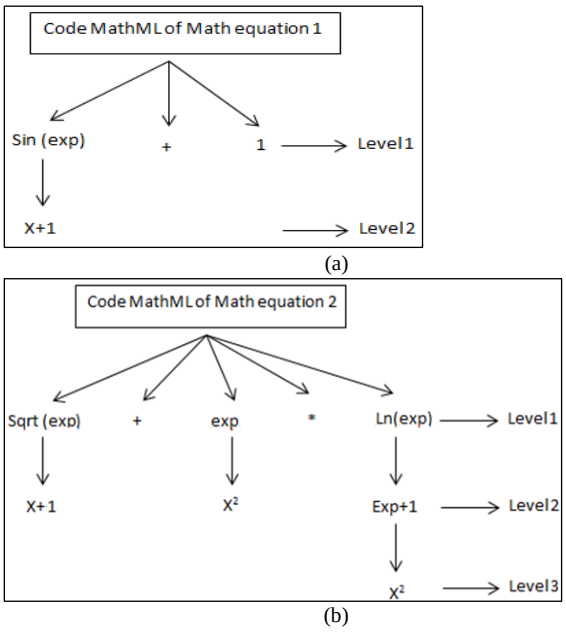


Fig. 3. (a) Code MathML of Math equation $\sin(x+1)+1$ by levels, (b) Code MathML of Math expression $\sqrt{x+1} + x^2\ln(x^2+1)$ by Levels

IV. Retrieving Results

The proposed method was based on the structural and semantic multilevel similarity between mathematical equations. The similarity degree was obtained based on all representation levels (Fig 3a, 3b).

After defining the expression levels and vectors $\boldsymbol{Vli}$, in the first time we retrieve all expressions that have similar vectors $\boldsymbol{Vl_1}$ to our math query. Then we move to the second level for only these equations already retrieved in the first level. The same procedure, used in the first level, was repeated to recover expressions that have similar vectors $\boldsymbol{Vl_2}$. We continue with the same procedure in level 3.

The K-Nearest-Neighbor (KNN) algorithm was used in this phase to retrieve math equation. It is a non-parametric lazy learning algorithm [19] and is an instance-based learning algorithm that uses a distance function of pairs of observation. KNN was based on the measurement of the distance to search a similarity between the query math equation and those of database. This distance is calculated using one of the following measures:

- Euclidean distance:

$$d(Vlx, Vly) = \sqrt{\sum_{i=1}^{n} (Vlx_i - Vly_i)^2}$$

Where: Vlx and Vly are two features vectors of two mathematical equations;

n is the number of attributes (vector size)

- Minkowski distance:

$$d(Vlx, Vly) = (\sum_{i=1}^{n} |Vlx_i - Vly_i|)^{\frac{1}{p}}$$

Euclidean distance is a special case for p=2 of Minkowsky distance

- P=1 we obtain the Manhattan distance:

$$d(Vlx, Vly) = \sum_{i=1}^{n} |Vlx_i - Vly_i|$$

- P=∞ we obtain the Chebychev distance:

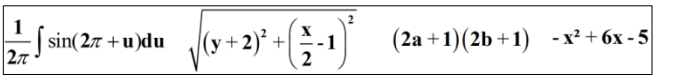
$$d(Vlx, Vly) = \max_i |Vlx_i - Vly_i|$$

- Correlation distance:

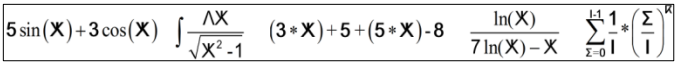
$$d(Vlx, Vly) = 1 - \frac{\sum_{i=1}^{n}(Vlx_i - \overline{Vlx_i})(Vly_i - \overline{Vly_i})}{\sqrt{\sum_{i=1}^{n}(Vlx_i - \overline{Vlx_i})^2 (Vly_i - \overline{Vly_i})^2}}$$

V. Experimental Results

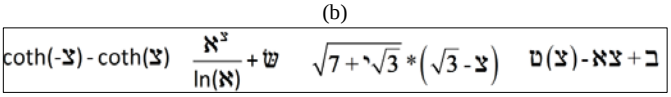
To evaluate the effectiveness of the proposed system we create a dataset of mathematical expression. The dataset is constructed using MathType. MathType is an interactive tool for authoring mathematical material, In the Microsoft Word or Power Point. There is MathType Ribbon Tab to facilitate editing, inserting and math equations creation. The dataset elements can be easily converted to MathML or Latex. In this set we have created 6925 mathematical expressions using symbols from five languages Latin, Arabic, Tifinagh [20,21], Hebrew and Japanese(Fig. 4 a, b, c, d, e). For each language, we have written 1385 different types of math expressions such as polynomial, algebraic, statistic, trigonometric and logarithmic.



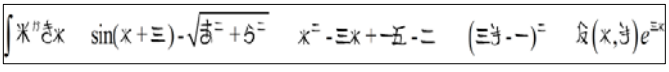
(a)



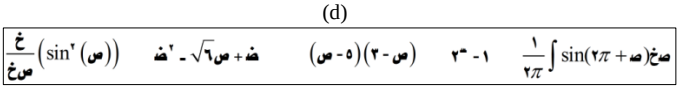
(b)



(c)



(d)



(e)

Fig. 4. (a) Math expressions in Latin; (b) Math expressions in Tifinagh; (c) Math expressions in Hebrew; (d) Math expressions in Japanese; (e) Math expression in Arabic

In this subsection, we present the results of the proposed system using Euclidean distance, Minkowski distance and Correlation distance. We compared our results to the statistical approach.

We found difficult to evaluate results using recall and precision evaluation using a similarity measure as proposed by T.T Nguyang et al. [3], in their paper the similarity measure between two expressions *E1* and *E2* represented by their attribute sets *M(E1)* and *M(E2)is :*

$$sim(E1, E2) = \left\| \frac{M(E1) \cap M(E2)}{M(E1) \cup M(E2)} \right\|$$

When our goal is to evaluate the semantic similarity, not only the number of similar components, we choose to evaluate manually different type of test queries by different level of similarity:

i. Identical similarity, like: $\sqrt{x+y}$ and $\sqrt{a-b}$

ii. Sub expression similarity: $\frac{2}{\sqrt{x+y}}$ and $\sqrt{x+y}$

iii. Categorical similarity:
$\int x^2\,dx$ and $\int (x+x^3)^2\,dx$ are both Integrals with polynomial functions.

We decide to give a score of tree points for identical similarity, two points for sub expression similarity and categorical similarity and zero for non-relevant expressions.

Table I, shows the performance results of the proposed system using Euclidean distance, Minkowski distance, and Correlation distance compared to the statistical approach. The score is based on the top 10 results of 10 test queries. A perfect score is 120 points.

TABLE I. Performance Results

Approach	PS (Euclidian)	PS (Minkowski)	PS (Correlation)	Statistical
Score/120	100	113	102	81
Score (%)	0.83	0.94	0.85	0.68

The experiments show that results obtained using the proposed system outperforms the statistical approach. Using Minkowski distance our system become more efficient (a score of 94%) Table II and III show test queries with their relevant expressions in the dataset using simultaneously the proposed system with Minkowski distance and the statistical approach.

TABLE II. Relevant Expressions of each one of the Test Queries using the Proposed System with Minkowski Distance

Query	Relevant expressions
$\int x^2 + 1\,dx$	$\int x^2 + 1\,dx \int \rho^2 + 2\,d\rho$, $\int (x+3)^2 dx$, $\int x^3 + 2x^2 + 2\,dx$, $\int x^3 + 2x^2 + x + 1 dx$
$\frac{-b+\sqrt{b^2-4ac}}{2a}$	$\frac{-b+\sqrt{b^2-4ac}}{2a}, \frac{-b-\sqrt{b^2-4ac}}{2a}, \frac{x+\sqrt{x^2-2}}{x-2}, \frac{\sqrt{س^2-س}}{س}, \frac{1-x^2}{\sqrt{x^2-2x}}$
$\sqrt{x+1}$	$\sqrt{x+1}$, $\sqrt{س+2}$, $2\sqrt{a+b}$, $\sqrt{x+x^2+1}$, $\sqrt{\sqrt{x+1}+1}$
$\sin x\, cos^2 x$	$\sin x\, cos^2 x, \frac{\sin x}{cos^2 x} \sin x + cos\, x, \int \sin x\, cos^2 x\, dx$, $cos^2 x + 1$,
$x^3 + 2x^2 + 3$	$x^3 + 2x^2 + 3$, $2س^3 + س^2 + 1$, $x^3 + x^2 + x + 1$, $(x+1)^3 + x$, $y^4 + y^3 - y^2 + 1$

TABLE III. RELEVANT EXPRESSIONS OF EACH ONE OF THE TEST QUERIES USING STATESTICAL APPROACH

Query	Relevant expressions
$\int x^2+1\,dx$	$\int x^2+1\,dx \int \rho^2+2\,d\rho$,$\int \frac{1}{y^2}\,dy$,$\int (x+3)^2 dx$, $\int cos(x^2+1)$,
$\frac{-b+\sqrt{b^2-4ac}}{2a}$	$\frac{-b+\sqrt{b^2-4ac}}{2a}$, $\frac{-b-\sqrt{b^2-4ac}}{2a}$,$\frac{x+\sqrt{x^2-2}}{x-2}$, $\frac{\sqrt{س^2-س}}{س}$, $x^2-2xy+\sqrt{\frac{x}{2y}}$
$\sqrt{x+1}$	$\sqrt{x+1}$, $\sqrt{س+2}$,,$x+1+\sqrt{2}$,$\sqrt{x}+2$,$\sqrt{x+1}+1$
$\sin x\, cos^2 x$	$\sin x\, cos^2 x$,$\frac{sin\, x}{cos^2 x}$, $sin^2 x+cos\, x$,$\int \sin x\, cos^2 x\, dx$, $x^2\, sin\, x$
x^3+2x^2+3	x^3+2x^2+3,$2س^3+س^2+1$,x^3+x^2+x+1, $\frac{x^3}{x^2+1}$, $y^4+y^3-y^2+1$

Our system with Minkowski distance allows a better detection of categorical similarity. We notice that the most of relevant expressions returned by the proposed system exactly match the query. For the statistical approach the absence of a semantic input can generate wrong outputs like relevance between $\sin x\, cos^2 x$ with $\frac{sin\, x}{cos^2 x}$ and $\sqrt{x+1}$ with $x+1+\sqrt{2}$.

VI. CONCLUSION

In this paper, we proposed a semantic approach to retrieve mathematical expressions. Based on MathML, we extract a multilevel representation for each mathematical expression. We used KNN with different types of distances to measure similarity between each representation level. In the light of our experiments we can conclude that the results are encouraging. Our system outperforms significantly the statistical approach and the implementation of Minkowski distance allows a better detection of categorical similarity. In our future work we have two goals, first to take into consideration more types of math expressions and second to evaluate our system on the web scale level.

REFERENCES

[1] R. Munavalli, R. Miner. “Mathfind: A math-aware search engine”. 29th annual international ACM SIGIR conference on research and development in information retrieval. New York, NY, USA ACM. p735, (2006).

[2] P. Libbrecht, E. Melis. “Methods to access and retrieve mathematical content and activemath” The 2nd international congress on mathematical software. ICMS’06 Castro Urdiales, SPAIN, 1-3 September (2006).

[3] T. T. Nguyen, S. H. Cheung, K. Chang. “A lattice-based approach for mathematical search using Formal Concept Analysis”. In Expert Systems with Applications 39. 5820-5828. (2012)

[4] D. Waud. “What is next?” Available at: http://www.tex.uk/cgi-bin/texfaq2html.(2003)

[5] S. Buswell, O. Caprotti, D. P. Carlisle, M. C. Dewar, M. Gaetano, and M. Kohlhase, “The open math standard” version 2.0.(2004)

[6] P. Jipsen, “Translating ASCII math notation to mathml and graphics”. Available at: http://www.chapman.edu/jipsen/mathml/asciimath.html.(2007)

[7] R. Ausbrooks, S. Buswell, S. Dalmas, S. Devitt, A. Diaz, R. Hunter, B. Smith, N. Soiffer, R. Sutor, S. Watt. “Mathematical markup language” (mathml)version 2.0. (2000).

[8] M. Nghiem, G. Yoko, Y. Matsubayashi, A. Aizawa. “Automatic Approach to understanding mathematical expressions using Mathml Parallel Markup corpora”. The 26th Anuual Conference of the Japanese Society for Artificial Intelligence, (JSAI2012) Yamaguchi, June (2012)

[9] B. R. Miller, A. Youssef. “Augmenting presentation mathml for search”. The 7th international conference on mathematical knowledge management MKM ’08 ,pp. 536-542, (2008).

[10] K. Yokoi, A Aizawa. “Towards Digital Mathematics Library” pp. 27-35, (2009).

[11] K. S.Cheung, D. Vogel,, “Complexity reduction in lattice-based information retrieval”. Information retrieval, 8, pp 285-299, (2005).

[12] S.Q. Yang, X. D. Tian: “A maintenance algorithm of FDS based mathematical expression index”. International Conference on machine learning and Cybernetics, Lanzhou, 13-16 July, (2014).

[13] S. Q. Yang, X. D. Tian, B. T. Yu, F. Yang. “A matching model of mathematical expressions with FDS based index”. International Journal on machine learning and Cybernetics. 6, pp 993-1004, (2015).

[14] X. lin, L. Gao, X. Hu, Z. Tang, Y. Xiao, X. Liu. “A mathematical retrieval system for formulae in layout presentations”. The 37th international ACM SIGIR conference on research & development in information retrieval (SIGIR 14) ACM, pp 697-706, (2014).

[15] MathSearch, http://wme.lzu.edu.cn/mathsearch/index.html

[16] W. Guo, W. Su, L. Li, N. L. Cui. “MQL : a Mathematical Formula Query Language for Mathematics Search”. The 4th IEEE International conference on Computational Science and Engineering (CSE2011). Dalian, pp 245-250, (2011).

[17] X. Hu, L. Gao, X. Lin, J. B. Baker. “WikiMirs : a mathematical information retrieval system for Wikipedia”. The 13th ACM/IEEE-Cs joint conference on digital libraries ACM, pp 11-20, Indianapolis, IN, USA , 22 – 26 July, (2013).

[18] P. B. Stanley, A. I Karshmer. “ Translating MathML into Nemeth Braille Code”. The 10th International Conference, ICCP 2006, LNCS 4061, LNCS 4061, pp. 1175-1182, Linz, Austria, (2006)

[19] B. V. Dasarathy, “Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques”, Mc Graw-Hill Computer Science Series, IEEE Computer Society Press, Las Alamitos, California, pp. 217-224, (1991).

[20] M. Oujaoura, R. El Ayachi, B. Minaoui, M. Fakir, B. Bouikhalene, O. Bencharef. “Invariant descriptors and classifiers combination for recognition of isolated printed Tifinagh characters.” International Journal of Advanced Computer Science and Applications 3.2: pp22-28, (2013).

[21] O. Bencharef, Y. Chihab, N. Moussaid, M. Oujaoura. “Data set for Tifinagh handwriting character recognition.” Data in brief 4: pp11-13, (2015).