Handwritten/Machine-printed and Arabic/Latin Mathematical Formula Discrimination and Recognition

Kawther Khazri Ayeb, Afef Kacem Echi

Handwritten/Machine-printed and Arabic/Latin Mathematical Formula Discrimination and Recognition

Kawther Khazri Ayeb
Université de Tunis
LaTICE, Tunis, Tunisia
Email: kawther.khazri@yahoo.fr

Afef Kacem Echi
Université de Tunis
LaTICE, Tunis, Tunisia
Email: afef.kacem@ensit.rnu.tn

Abdel Belaïd
Université of Lorraine
LORIA, Nancy, France
Email: abdel.belaid@loria.fr

***Abstract*—In this paper we mainly introduce a method for mathematical formula script and type identification based on handcrafted features. Arabic/Latin scripts are discriminated by detecting specific symbols based on their pixel density and distribution. Once the formula script identified, we proposed to separate between machine-printed and handwritten formulas. For that, we analyzed the spaces between some specific symbols and their neighbors and the pixel density of some other symbols presenting discriminative differences in their form when they are handwritten or machine-printed. The use of such structural characteristics makes the complexity of the method considerably lower than learning based methods. To finally recognize formulas, two main steps are followed: symbol recognition and formula structure analysis. For the first step, we used a combination of statistical features and an instance-based classifier. For the second step, we proceeded by top-down and bottom-up parsing scheme based on operator dominance. A set of replacement rules is defined by a coordinate grammar. Formula parsing consists of applying, from the dominant operator and its context, the appropriate rule to divide the formula into sub-formulas which will be recursively analyzed by the same way. Carried experiments on various mathematical formulas, show the efficiency of both script and type identification and recognition proposed methods.**

I. Introduction

Research on script and type identification aims to create systems able to discriminate automatically between the different forms in which a document is presented, including the language and the way it is written in (machine-printed or handwritten), in order to select the appropriate recognition system to a given document. The state of the art on the script identification shows that no work deals with mathematical formulas. Existing works treat this problem for text. Also, few systems are interested at the same time in Arabic/Latin and Printed/handwritten script identification. In this context, we propose a method to automatically discriminate between Arabic and Latin, handwritten and machine-printed mathematical formulas. This work comes as a part of our research on off-line recognition of mathematical formulas [1], [2] and [3].

The rest of the paper is organized as follows. In section II, we start with a synthesis of the existing systems for script identification. In section III, we present the proposed method for mathematical formula script and type identification. Section IV is dedicated to confusion analysis and resolution. Experimental results are reported in section V. A description of the formula recognition method can be found in section VI where the focus is on symbol recognition and formula structure analysis steps. Finally, conclusion and future works are drawn in section VII.

II. Related Works

Most research works in the field of script identification focus principally in text documents. To the best of our knowledge, no work treats this problem for mathematical formulas. In this section we will state some principal works related to this field.

From our study of some systems presented in the literature, we notice that there are mainly two approaches to identify the script: global and local approaches. The global approach needs training and test data sets, since it uses classifier based on feature vector. We can quote in this category, systems based on the horizontal and vertical projection profiles analysis like the one proposed by Wood and al [4]. we can also quote the work done by Saïdani and al. [5] where authors used histogram of oriented gradients HOG in combination with Bayes-based classifier for Arabic/Latin script and type identification. As we mentioned before global approaches, needs learning step which can increase the complexity of the method. Authors in [6] used bi-level co-occurrence technique for machine-printed and handwritten text identification in nosy documents images. A co-occurrence count is the number of times a given pair of pixels appears at a fixed distance and orientation. As the most of the information in the image is represented by the black-black pairs, authors have only considered them to extract related features.

The second strategy's type is based on the appearance of the script to identify and it searches for the existence or the absence of intrinsic features of each script. The detection of these features is rely on observation on the morphology of the script. In this context, authors in [8] use Gaussian Mixture Models, GMM, to deal with the Arabic/French identification between multi-font printed text documents and multi-script handwritten texts at the world level. They divide each word image in fixed length analysis window sliding and represent each one with 35 features including affine moments invariants, X-Y position of the top and bottom extrema, cumulated

projection values, etc. these features are used with the GMM to estimate the script category likelihoods. Using a 20000 word images database, they obtain 99.1% identification rate. The system, they present, encountered problems with printed words due to diversity of used fonts. In the same context, in [9], an identification system at word level, based on steerable pyramid transform was proposed. The features are extracted from pyramid sub bands and served to classify the scripts on only one script among the scripts to identify. Hybrid systems exploit all available information in text block, text line or word, and connected component level. Among the relevant research in this context, we cite the work presented in [7], which is a contribution for Arabic and Latin script identification, in printed and handwritten documents. Authors, in this work, use morphological features at the text bloc level and geometrical features at the text line and connected component level. Using a KNN classifier, they achieve a correct identification rate of 88% on a data set of 400 big size text block and 92% on a data set of 335 little size text images. However, their proposed method suffers from confusion between printed and handwritten text and between Arabic and handwritten Latin. Inspired from some related works, we propose here a method to automatically separate between Arabic and Latin mathematical formulas of handwritten or machine-printed types.

III. Formula Script and Type Identification

A. Arabic/Latin Formula Script Discrimination

The proposed method for formulas script identification is based on its structural characteristics. Considering the visual differences between Arabic and Latin script, we look for the presence of some distinctive symbols in specific position to distinguish each script as explained below.

1) Detection of Arabic mathematical function: Arabic mathematical formulas uses Arabic words to identify certain name of function like ظَا for the sinus function, ظتَا for the cosine function, جَا for the tangent and جتَا for the cotangent function (see Figure 1). The extraction of these name of

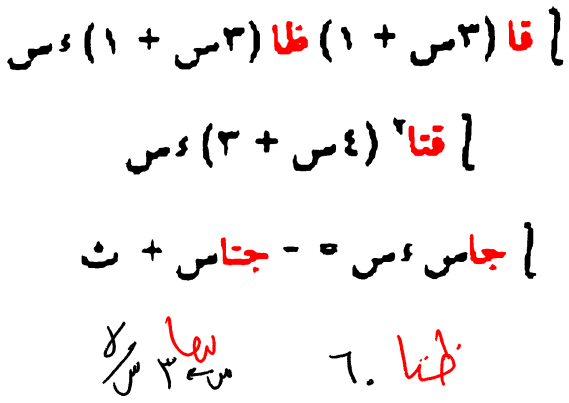

Fig. 1. Arabic name of functions.

functions is based, first, on pixel density analysis and second, on their specific morphologies characterized by the existence of a left pole in their connected component. A pole is all form with maximum pixel density above the upper.

2) Detection of Discriminant Mathematical Symbols: Some Latin mathematical symbols have their vertical mirror images in Arabic script, like the integral, the summation and the root symbols (see Figure 2). Detection of these symbols is

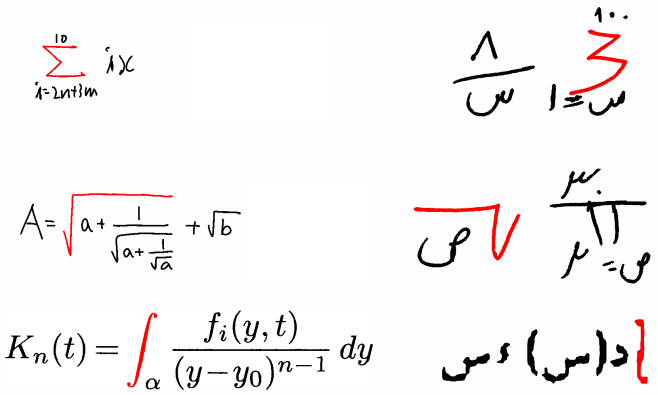

Fig. 2. Latin mathematical symbol and their mirrored one in Arabic context.

also based on their morphological characteristics. Integral, for example, is vertically elongated. Summation has a square form. the root symbol has an elongated horizontal line in the top of its connected component. Finding these symbols in a formula and analyzing their pixel density help as to identify the script's language of the formula.

3) Detection of Arabic Indic Numbers: Some used digits in Arabic formulas can be confused easily with other symbols in Latin script. The Arabic digit zero (٠)for example can be confused with a dot. The Arabic digit one (١) can be confused with the absolute value line. The Arabic digit six (٦) can be confused with the Latin digit seven (7), etc. (see Figure 3).

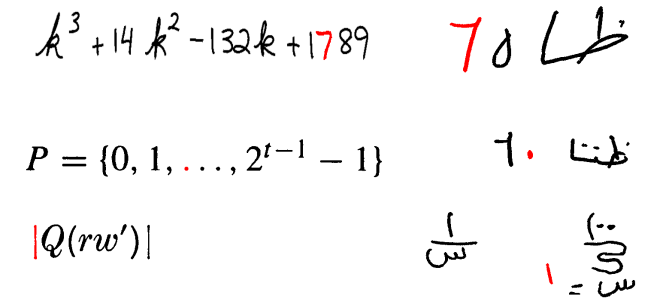

Fig. 3. Ambiguous symbols.

Other Arabic digits have distinctive morphology and can be used to identify Arabic script. In our method, for example, Arabic digit three (٣) is used to identify Arabic formulas.

4) Detection of Some Arabic Letters Curvature: Some Arabic letters are characterized by their large bottom curvature

like the Arabic letter Seen (س), letter Sad (ص), letter Noun (ن), etc. extraction of these letters help as to distinguish Arabic formulas.

5) Detection of Some Latin Letters: Latin mathematical formulas use Latin letters. Some of these letters can be easily confused with other symbols in Arabic script like the letter q and the Arabic digit nine (٩). Other letters have distinctive morphology and can be used to identify Latin script like for example the letter b and d, characterized by their vertically elongated form, composed of a bottom blob and a vertical upper pole. The same for the letter p, characterized by an upper blob and a vertical lower pole. These letters are used by our system to identify Latin formulas.

B. Handwritten/Machine-printed Formula Type Discrimination

To distinguish between handwritten and printed formulas, we analyzed the spaces between some symbols and their neighbors.We also studied pixel distribution of some symbols presenting clear difference between their forms in machine printed and handwritten context.The main idea is to benefit from space irregularities between formulas symbols and from pixel distribution irregularities inside the symbol itself.

1) Detection of Irregularities in Pixel Distribution: From our observation in mathematical formulas in both language, we notice the frequent presence of straight lines whether vertical or horizontal in many symbols like the summation symbol, the horizontal fraction bar, the root symbol, etc. An other pertinent characteristic that marks the mathematical script is the present of many symbols presenting symmetry in their pixel distribution, whether vertical symmetry like in the comparison symbols greater than of less than, or horizontal symmetry like in the Arabic digits seven and height (resp. ٧ and ٨), or central symmetry like in the integral sign or the addition and multiplication sign. In a handwritten context, lots of irregularities in the writing are easily identified and respecting writing rules is not evident. Detecting irregularities in pixel distribution in some symbols can lead us to determine the type of the writing, machine printed or handwritten. Figure 4 shows examples of these irregularities. The horizontal fraction bar, for instance, is generally not well aligned in handwritten script, unlike in printed script, it is generally perfectly straight (Figure 4 (a)). The same case for the horizontal line in the root symbol (Figure 4 (b)) or in the summation symbol(Figure 4 (c)). Other Symbols like the integral sign, the addition sign, the comparison symbols greater than of less than, the Arabic digits seven and height (resp. ٧ and ٨), etc. present perfect symmetry in machine printed context which is generally not found in handwritten (see Figure 4 (d) and (e)).

2) Detection of Irregularities in Relative Placement of Symbols: Symbol placement is very important for the understanding of mathematical formulas. In fact, placing symbols is guided by precise rules that reveal the type of relationship between them. For example the operands of the fraction bar are generally centered at the top and the bottom (Figure 5

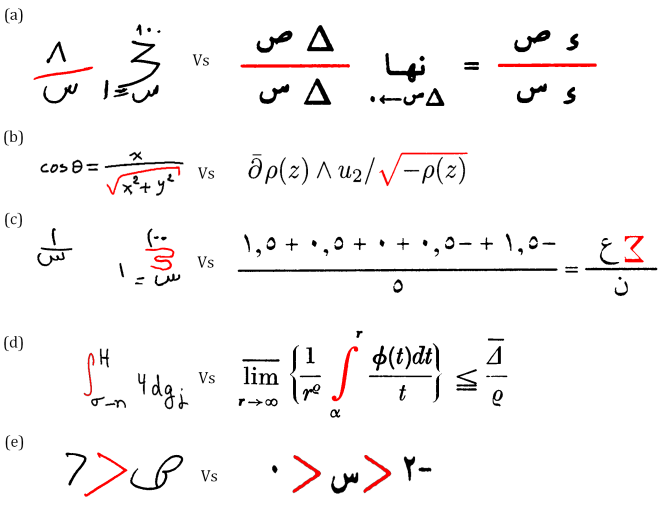

Fig. 4. Irregularities in symbols.

(a)). Closing parenthesis and closing square bracket must be in the same line and with the same height as their counterparts (Figure 5 (b)). the operands of binary mathematical operators have the same distance that separates them from the operator (Figure 5 (c)). Unlike in machine printed scripts, respecting these rules in handwritten is relative. Taking benefits form this characteristic, we start with detecting some symbols like the horizontal fraction bar, the closing parenthesis, the closing square brackets and the addition sign. Than, we calculate distances between them and the symbols they are in relation with to decide about the type of the formula.

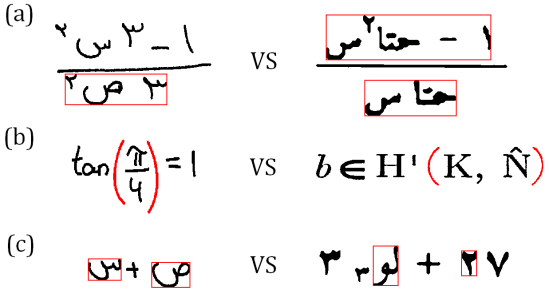

Fig. 5. Irregularities in symbol distribution.

IV. Errors Analysis

Identifying the language and the type of a mathematical formula is an initial primordial step that precedes the formula recognition in advanced steps. Formula's language and type identification errors lead in the majority of cases to formula's recognition errors. That is why we focus in this part on the analysis of confusion cases and proposal to treat them.

The proposed system is based in the detection of specific symbols in the formula. Tow types of confusion can be encountered during this step.

- Confusions due to the similarity in the morphology of some symbols: There symbols with the same morphology and belongs to different script language like, for example, the Arabic digit three (٣) in printed formulas and the Latin latter. Also, the Arabic summation can be confused with the Latin digit three.
- Confusions du to touching symbols: For example the Arabic digit seven (٧) when attached with and horizontal fraction bar can be easily confused with a Latin root.

To enhance the script identification, we propose to reorder the detection of symbols from the less ambiguous to the most ambiguous one.

V. Experimental Results

To evaluate performance of the proposed formula script and type identification method, we carried primary experiments on our database which consists of 250 machine-printed Arabic mathematical formulas, extracted from different mathematical books of several Arabic countries, and 250 different handwritten Arabic mathematical formulas written by 5 different writers. We also used for test a data base of 250 Latin machine-printed formulas extracted from InftyMDB-1, and 250 handwritten Latin mathematical formulas extracted from the CROHME database.

Table I, show our obtained results for script and type identification.

TABLE I
Overall accuracy results

	Arabic		Latin	
	Handwritten	Printed	Handwritten	Printed
Formulas	250	250	250	250
Script identification rate (%)	78.4	86.4	77.2	89.2
Total script identification rate (%)	82.8			
Script identification error rate (%)	4	0.4	9.2	5.2
Total script identification error rate (%)	4.7			
Type identification rate (%)	58.8	93.2	52	99.2
Total type identification rate (%)	75.8			
Type identification error rate (%)	41.2	6.8	48	0.8
Total type identification error rate (%)	24.2			

Since our method is based principally on the identification of specific mathematical symbols than the analysis of their structural characteristics. The major reason for errors in script and type detection are due basically to confusions in the detection of these symbols. Table II shows the identification and error rate relative to the extraction of these symbols.

When observing the confusion cases, we found that they are due principally to the similarities in the morphology of certain

TABLE II
Specific symbols extraction result.

Symbol	Formulas containing this symbol	Correctly identified formulas	Identification rate	Error rate
Arabic mathematical name of function	225	223	99%	8%
Arabic integral	26	20	77%	1,4%
Latin integral	99	76	77%	0,4%
Latin summation	82	82	100%	0,2%
Arabic summation	12	9	75%	17,2%
Latin root	92	72	78%	0,4%
Arabic root	58	50	86%	0,2%
Arabic digit 3	152	125	82%	0,4%
Arabic letters	351	297	85%	1,2%
Latin letters	161	144	89%	4,4%
Horizontal fraction bar	314	309	98%	1,5%
Closing parenthesis	455	446	98%	1,9%
Opening parenthesis	443	383	86%	13,5%

symbols. For example, our extraction of the mathematical Arabic name of functions like ظَا for the sinus function, ظتَا for the cosine function, جَا for the tangent and جتَا for the cotangent function, is based on the fact that they are characterized by a big pixel density in the left and in the bottom of the image, thing that is frequent in Latin script with symbols like the letter L. The integral also is frequently confused with the absolute bar especially in handwritten. The Arabic summation is confused with the digit 3. What we propose to reduce the impact of these confusions on the identification results is to start the extraction with symbols presenting less confusion. For errors concerning type's identification, we note that some handwritten formulas are well written and respect the rules of mathematical formulas. For that reason, we prefer to not propose a solution in this stage and see the results of their recognition with our system for machine printed mathematical formulas recognition presented in [1].

VI. Formula Recognition

The recognition method consists of two main steps: symbol recognition and formula structure analysis. For symbol recognition, we used an instance-based classifier: K* [**?**] and extracted 23 statistical features from Hu moments, Run-length histograms, Zernike moments, bi-level co-occurrence and white pixel portion. The proposed symbol recognizer is able to identify 50 symbol classes, including some variable and function names, arithmetic operators, literal and mirrored symbols, Mashrek Arab numbers, etc. Although the symbol recognizer achieved a good accuracy in primary tests, its failure to distinguish certain common symbols would be

bothersome to later steps of our system. In fact, certain distinct symbols are in close resemblance such as the horizontal fraction bar and minus sign, the Arabic digits two and three (٢,٣), the Arabic letter س and ص, or between mirror symbols such as the opening and the closing parenthesis or brackets ([,], (,)), the comparative operators more than and less than ($<$, $>$), the Arabic digits seven and eight (٧, ٨). Observing the event of confusion, we remark that confused symbols have roughly similar morphologies. It is also due to rotational invariance of Hu and Zernike moments. We consider some of the misrecognition cases to be too difficult for any classifier to resolve without considering symbol context. That is why we keep resolving some of these confusion cases for formula structure analysis step which consists of lexical, geometric and syntactic analysis .

In lexical analysis, we assigned a label to each symbol or group of symbols which refers to its syntactic category. For multi-part symbols ($=, \leq, \geq$, etc.), Arabic letters, having diacritic points, (أ, ب, ت, etc.) and function names (نهَا, ظتَا, طتَا, etc.), vertical regroupment is required.

For geometric analysis, we defined ten spatial structure: Left, Right, Above, Below, Left and Right Superscript, Left subscript, Inside and Delimited. These spatial relations, in conjunction with context, are used to resolve confusions between symbols having similar morphologies. For example, in order for a symbol to be considered as a fraction bar, it should have no empty parts above and below.

To syntactically analyze the formula, we proposed a top-down and bottom-up parser which selects, from the dominant operator, the appropriate rule to apply in order to divide the formula into sub-formulas which will be analyzed by the same way (see Algorithm). More details about the used grammar and formula parsing can be found in our previous work [1].

procedure ANALYSE(Z, Var T)▷ Z is the formula zone and T is its resulting syntactic tree
  $D \leftarrow Dominant\_operator(Z)$
  if $D \neq NULL$ **then**
    $[R, C_i] \leftarrow select\_rule(Z, D)$ ▷ R is the rule where D is in the right side and C_i is the context of D
    $Z_i \leftarrow delimit\_zone(C_i)$
    $ANALYSE(Z_i, T_i)$
    $T \leftarrow R \times (D, T_i)$
  else ▷ the formula is reduced to a letter, an integer or a float.
    $T \leftarrow R \times (Z)$
  end if
end procedure

To evaluate the formula recognition method, we measured the performance of each level in the formula parsing process. We tested the mathematical symbol recognizer on 1018 symbols extracted from 100 test formulas and using 5000 samples for a training step. We achieved a recognition rate of 89.9%. But, when we considered symbol context, the recognition rate was increased to 96.18%. The proposed syntax directed system was tested on a database of 110 mathematical formulas of different types. The overall system has shown its efficiency on a reasonable number of practical mathematical formulas with a recognition rate of 91%.

VII. Conclusion

In this work, the focus was on the problem of mathematical formula script and type identification. Without training, the proposed identification system succeeds to achieve acceptable results in primary tests on a database composed of multi-font printed and multi-script handwritten formulas. We achieved a formula script identification rate of 83% and a formula type identification rate of 76%. Considerable amelioration are expected if we take into account extraction of more specific symbols. As future work we plan to enhance performance of the proposed script and type identification method in addition to the formula recognition method.

References

[1] K. Khazri Ayeb, A. Kacem Echi, and A. Belaïd, *A Syntax Directed System for the Recognition of Printed Arabic Mathematical Formulas*, ICDAR 2015, pp. 186-190

[2] K. Khazri, A. Kacem and A. Belaïd, *Recognition of Machine- Printed Arabic Mathematical Formulas*, ICTIA, 2014.

[3] A. Kacem, K. Khazri and A. Belaïd, *Reconnaissance de formules mathmatiques arabes par une approche dirige syntaxe*, CIFED, 2010.

[4] S. L. Wood, Xiaozhong Yao, K. Krishnamurthi and L. Dang, Language identification for printed text independent of segmentation, ICIP 1995, V. 3, PP. 428 - 431

[5] A. Saidani and A. Kacem, Arabic/Latin and Machineprinted/Handwritten Word Discrimination using HOG-based Shape Descriptor, ELCVIA 2015

[6] Y. Zheng, H. Li, and D. Doermann, Machine Printed Text and Handwriting Identification in Noisy Document Images, , IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 26, n 3, pp. 337 - 353, 2004.

[7] S. Kanoun, I. Moalla, A. Ennaji, and A. Alimi, Script Identification for Arabic and Latin, Printed and Handwritten Documents, DAS 2000, pp. 159-165.

[8] A. Mezghani, F. Slimane, S. Kanoun and V. Margner, Identification of Arabic/French - Handwritten/Printed Words using GMM - Based System, Coria 2014

[9] S. Habboubi, S. S. Maddouri and H. Amiri, Discrimination between Arabic and Latin from bilingual documents, CCCA 2011