DATABASES FOR RECOGNITION OF HANDWRITTEN ARABIC CHEQUES

Alohali, Y.; Cheriet, M.; Suen, C.Y.

DATABASES FOR RECOGNITION OF HANDWRITTEN ARABIC CHEQUES

YOUSEF AL-OHALI[1], MOHAMED CHERIET[1, 2] AND CHING SUEN[1]

1. CENPARMI, ConcordiaUniversity GM-606, 1455 de Maisonneuve W., Montreal, Quebec H3G 1M8, Canada

2. Imagery, Vision and Artificial Intelligence Laboratory, École de Technologie Supérieure, University of Québec, 1100, Notre-Dame West, Montréal, Québec H3C 1K3, Canada

E-mail: {yousef, suen@cenparmi.concordia.ca}, cheriet@gpa.etsmtl.ca

This paper describes an effort toward building Arabic cheque databases for research in recognition of handwritten Arabic cheques. Databases of Arabic legal amounts, Arabic sub-words, courtesy amounts, Indian digits, and Arabic cheques are provided. This paper highlights the characteristics of the Arabic language and presents the various steps that have been completed to achieve this goal including segmentation, binarization, tagging and validation.

1 Introduction

Cheque Processing involves all the tasks a bank officer may have to do to process an incoming cheque for a client. This includes: accessing account numbers, verifying names and signatures on the cheque, verifying the date of the cheque, matching the legal amount with the courtesy amount and verifying the paying ability of the issuer. While extensive efforts have already been devoted to Latin and oriental cheque processing systems [3,4,5], to the best of our knowledge, no attempts have been made towards an Arabic cheque processing system. This was partially due to the lack of a supporting infrastructure required to conduct, develop and compare such systems. A major effort to open this area is to provide a real world database that can be used for such purposes. This work provides databases for Arabic legal amounts and courtesy amounts (written in Indian digit). It is interesting to note that Indian digits are more popular than Arabic numerals in some parts of the Arabian world.

The rest of this paper is organized as follows. After the introduction, we give a brief description of the Arabic written language focusing on the current application, and mentioning the major characteristics that may affect a pattern recognition system. Data collection is covered next, followed by pre-processing section. Tagging comes next, followed by validation, and databases that came as a result of this project. We finish this paper with some concluding remarks.

In: L.R.B. Schomaker and L.G. Vuurpijl (Eds.), Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, September 11-13 2000, Amsterdam, ISBN 90-76942-01-3, Nijmegen: International Unipen Foundation, pp 601-606

2 Brief description of the Arabic written language

Arabic is semi-cursive in nature. Out of the 28 basic Arabic letters, 22 are cursive letters while 6 are non-cursive. Within one word, a cursive letter should be connected to the succeeding letter, while non-cursive letter can not be connected to any succeeding letter. Thus, an Arabic word may be decomposed into more than one sub-word, each represents one or more connected letters with their corresponding secondary components. In addition, Arabic defines two types of secondary components: dot and Hamzah (a zigzag-like shape). The number and position of secondary components play a factor in identifying different letters. Due to connectivity, an Arabic letter may change significantly depending on its position within a sub-word, identity of neighbouring letters, the writing font, and the way the writer connects successive letters [1,2]. Arabic handwritten letters differ in height and width. Moreover, Arabic allows the presence of diacritics that control the pronunciation of words and possibly their meanings. However, such diacritics are only used in formal documents or in cases of contextual ambiguity. Unlike Latin, Arabic is written from right to left.

The vocabulary of Arabic legal amounts is larger than those found in Latin languages. This is due to three major factors. First, Arabic has three different forms: singular, double, and plural (figure 1).

Second, double and plural nouns have up to four different forms according to their grammatical positions (figure 2). Third, two forms are defined for feminine and masculine countable things (figure 3).

الف الفين آلاف

Figure1: Singular, double and plural forms for the word "thousand"

الفين الفي الفان الفا

Figure2: Four grammatical forms of the same word

ثلاث ثلاثة

Figure3: Feminine and masculine forms of the word "three"

In addition, a few common spelling mistakes and/or colloquials occur in writing some Arabic numbers (figures 4,5). These factors affect the identity of letters and the number of sub-words in a word. We found more different words than sub-words in the lexicon. That was one of the reasons to consider sub-word as the basic unit of Arabic legal amounts. However, this does not prevent others from using words as their basic units.

In principle, Arabic allows legal amounts to be written in any order, i.e. starting from the most significant digit, from the least significant digit or from the middle. However, eloquence measurements and people habit excluded most permutations.

3 Data collection

The first step toward building a database is to find suitable sources of data. Finding a real world source of data becomes a problem when dealing with applications that carry sensitive or private information like bank cheques. Through collaboration with Al Rajhi Banking and Investment Corp. (the biggest banking corporation in Saudi Arabia), we were able to collect about 7000 real world grey-level cheque images. The gathering process involved scanning the real cheques at the bank's centre, and removing all personal (private) information including names, account numbers, and signatures. The cheques were scanned in grey level at 300 dpi.

We were also able to collect about 100,000 real world binary images, which will be processed in a future work.

Figure4: Two common forms for the word "hundred"

ثلاثه
ثلاثة

Figure5: Secondary components of the last letter have been ignored, a common mistake.

4 Pre-processing

The next step was to segment cheque images to extract filled-in fields. We concentrated on the legal and courtesy amounts, leaving the date field for future work. That was achieved by localizing the target fields on all kinds of cheque forms. We knew in Saudi Arabia, there were only two types of cheques, which share the same format (structure) but have different sizes.

The next pre-processing step was to binarize and remove noises from the segmented fields. The noises include lines, borders, and pre-printed text that may appear along with the extracted fields. This step has been successfully achieved by adapting the tools available in CENPARMI (which were designed for Canadian cheques) [6,7]. Figures 6 and 7 show a sample Arabic cheque and its corresponding segmented legal amount.

5 Tagging

Tagging intends to label each object in the cleaned legal and courtesy amounts. To do so, we have adopted a tagging tool designed to tag Latin dates, and made it reusable to tag Arabic sub-words and numerals.

The tagging person is required to click at a point of each connected component of the target object and then select a tag from a pre-defined vocabulary. The tool stores the coordinates of each point, adds a delimiter and stores the tag. It allows tagging of touching objects and permits reverse action (undo).

While defining our vocabulary, we have accounted for most differences, even small ones. For instance, two different tags were used to label objects that differ only in their secondary components (dots). This gives more choices for future analysis since it costs nothing to merge similar classes (if such discrimination is not useful for a particular method or application).

This tool produced four sets of tagged objects:

1. Courtesy amount: contains a sequence of coordinates and tags of objects. Objects may include Indian digits, delimiters, commas, decimal points or noise. Coordinates provide unambiguous pointer to the object intended by each tag.
2. Indian digit: contains a reference to the original courtesy amount that produced the current object, followed by the tag of the current object.
3. Legal amount: contains a sequence of coordinates and tags of objects. Objects may include sub-words, or noise. Coordinates provide unambiguous pointer to the object intended by each tag.
4. Arabic sub-word: contains a reference to the original legal amount that produced the current object, followed by the tag of the current object.

Tagging of legal and courtesy amounts were done independently to avoid chances of complex errors that happen in both the legal and courtesy amounts of the same cheque.

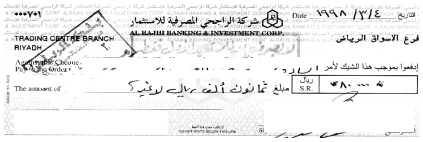

شركة الراجحي المصرفية للاستثمار
AL RAJHI BANKING & INVESTMENT CORP.
TRADING CENTRE BRANCH
RIYADH
فرع الاسواق الرياض
Date ١٩٩٨/٣/٤ التاريخ
ادفعوا بموجب هذا الشيك لأمر
The amount of
مبلغ ثمانون ألف ريال لاغير
S.R. ريال
٨٠٠٠٠

Figure6: A sample of the Arabic cheque database.

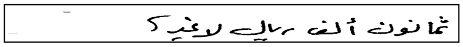


Figure7: Segmented legal amount (from figure 6).

6 Validation

Although the tagging tool was adjusted to prevent or warn for possible errors in the tagging process, yet there are still some traces of mistakes. This is particularly true when dealing with large amounts of data. Therefore, there must be some procedure to verify the truthfulness of the tagging process.

We took an early advantage of the redundancy available in the cheque form by comparing the numerical values of the tags of legal and courtesy amounts of each cheque. We approve the tagged legal and courtesy amounts (and all objects obtained from them) if the two values match. Otherwise, further steps need to be taken to validate or correct the tagged legal and/or courtesy amounts.

Comparing the two amounts requires translation and interpretation of each sequence of tags into its numerical value. While this looks trivial for courtesy amounts, it is not the case for legal amounts. First, each tag should be translated into the appropriate sub-word. Second, each sequence of sub-words needs to be translated into a correct word. This was achieved by means of a context sensitive grammar developed for this purpose. Third, the sequence of words should be interpreted into numerical value. Again, this requires special manipulation since there are various orders to write an amount in Arabic (e.g. from high order to low order).

The validation process approved about 60% of the tagged cheques, which provided about 23,325 sub-words and 9,865 digits. Table 1 shows the distribution of the validated sub-word classes (excluding touching sub-words). Some classes are very rare, though they do exist in the lexicon of handwritten Arabic legal amounts. We decided to keep such classes in the lexicon although they are not very well represented**.** Further processing for the remaining tagged data is scheduled in our future work. It is important to note that this automatic validation process guarantees the correctness of the tagged legal/courtesy amounts, and all Indian digits. For Arabic sub-words, it is possible to have different sub-words that were interchanged without affecting the legal amount. This note should not have any effect if we update our lexicon to merge sub-words having the same value contribution together.

There are various reasons for not approving a tagged cheque. Some of them are listed below:

1. The legal amount (or the courtesy amount) may have been cut by the extraction tool, providing incomplete or incorrect data to the tagging tool.
2. The legal amount may have contained unexpected spelling mistake that left the relevant sub-word untagged (tagged as OTHERs symbol), leaving a gap in the legal amount.
3. There may be missing sub-words (mainly letters) in the original legal amount.
4. An error can be produced by the human tagger.

7 Databases

This research effort has produced a number of databases that can help researchers in various fields. These databases include Arabic legal-amounts database (1,547 legal amounts), Courtesy amounts database (1,547 courtesy amounts written in Indian digits), Arabic sub-words database (23,325 sub-words), and Indian digits database (9,865). Each database mentioned above is divided into training and testing sets. The training set includes 2/3 to 3/4 of the available data. That is true for legal amounts, courtesy amounts, Indian digit classes and most sub-word classes. In a few sub-word classes, this condition could not be satisfied. These databases will be available to researchers upon validating the rest of the tagged data mentioned in section 6.

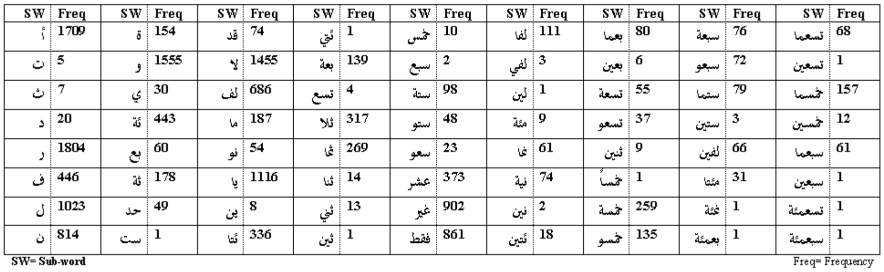
In addition, this work produced a database of complete (original) grey level cheques, which can be used for other research purposes. This database will be available after the approval of the source of the data (Al Rajhi bank).

Moreover, it is not difficult to deduce a database of Arabic words. This is achievable using the legal amounts database or the sub-words database. It is also possible to derive a database of Arabic dates from the Arabic cheques database.

8 Conclusion

A substantial amount of effort has been devoted toward building Arabic cheque databases, a very important infrastructure to develop and compare pattern recognition systems for the Arabic based cheque-processing systems. This paper describes the main steps that have been completed to provide such databases. The paper also gives a list of useful databases that have been produced, to be produced, or could be produced from this work.

Table 1: Distribution of the validated sub-word classes



SW	Freq	SW	Freq	SW	Freq	SW	Freq	SW	Freq	SW	Freq	SW	Freq	SW	Freq	SW	Freq
أ	1709	ة	154	قد	74	ئتي	1	خمس	10	لفا	111	بعما	80	سبعة	76	تسعما	68
ت	5	و	1555	لا	1455	بعة	139	سبع	2	لفي	3	بعين	6	سبعو	72	تسعين	1
ث	7	ي	30	لف	686	تسع	4	ستة	98	لين	1	تسعة	55	ستما	79	خمسما	157
د	20	ئة	443	ما	187	ثلا	317	ستو	48	مئة	9	تسعو	37	ستين	3	خمسين	12
ر	1804	بع	60	نو	54	ثما	269	سعو	23	نا	61	ثنين	9	لفين	66	سبعما	61
ف	446	ثة	178	يا	1116	ثنا	14	عشر	373	نية	74	خمساً	1	مئتا	31	سبعين	1
ل	1023	حد	49	ين	8	ثني	13	غير	902	نين	2	خمسة	259	ثمئة	1	تسعمئة	1
ن	814	ست	1	ئتا	336	ثين	1	فقط	861	ئتين	18	خمسو	135	بعمئة	1	سبعمئة	1

SW= Sub-word

Freq= Frequency

9 Acknowledgements

Our special thanks are due to Al Rajhi Banking and Investment Corp. for their understanding of our research goals and for the facilities they have provided to us. The Saudi Cultural Mission in Ottawa has been thankfully very cooperative and supportive of this work too.

Thanks are also due to the Natural Sciences and Engineering Research Council of Canada, and the FCAR Program of the Ministry of Education of Quebec, for their support of this research project.

Our thanks are also extended to Christine P. Nadal, at CENPARMI, for her continues help in this project.

References

1. 1. B. Al-Badr and S. Mahmoud, "Survey and bibliography of Arabic Optical text recognition,"' *Signal Processing*, 41, pp. 49-77, 1995.
2. 2. Adnan Amin, "Off-line Arabic Character Recognition: The State of the Art," *Pattern Recognition*, Vol. 31, No. 5, pp. 517-529, 1998.
3. 3. Michel Gilloux and Manuel Leroux, "Recognition of Cursive Script amounts on Postal Cheques,"' *European Conf. Dedicated to Postal Technologies*, Nantes, France, pp. 705-712, June 1993.
4. 4. D. Guillevic and C. Y. Suen, "Recognition of Legal Amounts on Bank Cheques,"' *Pattern Analysis and Applic.* , 1, pp. 28-41, 1998.
5. C. Suen, L. Lam, D. Guillevic, N. Strathy, M. Cheriet, J. Said and R. Fan, "Bank Check Processing System," *International Journal of Imaging Systems and Technology*, Vol. 7, pp. 392-403, 1996.
6. X. Ye, M. Cheriet, C. Y. Suen and K. Liu, "Extraction of bankcheck items by mathematical morphology," *Intl. J. Document Analysis and Recognition* in press, 2000.
7. X. Ye, M. Cheriet, and C. Y. Suen, "Model-Based Character Extraction from Complex Backgrounds," *Proc. ICDAR99*, pp. 511-514, 1999.