Building A Handwritten Cuneiform Character Imageset

Kenji Yamauchi[1], **Hajime Yamamoto**[2], **Wakaha Mori**[3]
[1]TIS Inc., Japan, yamauchi.kenji@tis.co.jp
[2]Kyoto University, Japan
[3]The Institute for Cultural Studies of Ancient Iraq.

Abstract

Digitization of cuneiform documents is important to boost the research activity on ancient Middle East and some projects have been launched in around 2,000. However, the digitization process is laborious due to the huge scale of the documents and no trustful (semi-)automatic method has established. In this paper, we focused on a cuneiform document digitization task, realization of Optical Character Recognition (OCR) method from the handwritten copies of original materials. Currently, as the first step toward development of such methods, we are constructing a handwritten cuneiform character imageset named with professional assistance. This imageset contains typical stroke patterns for handwriting each frequently appearing cuneiform character and will be able to support the development of handwritten cuneiform OCR system.

Keywords: Cuneiform, OCR, Dataset

1. Motivation

Cuneiform characters (signes) were used to write several languages (e.g. Sumerian, Akkadian) for about 3,000 years in ancient Middle East (so-colled Mesopotamia). Each character superficially contains wedges (called *cuneus* in Latin) which you can see as triangular shape. Usually, the ancient scribes adopted the clay tablets as a material on which they drawn the signs. The characters were similar to Old Japanese in terms of the orthography (both phonogramatically and ideogramically used) and the number of character classes (almost 600).

Because of the diversity of category and written languages, cuneiform documents are linguistically and historically important. Since the amount of the documents is huge (at least 300,000 documents have been published, and many remain unpublished), digitization of cuneiform documents are indispensable to conduct academic research efficiently.

Some digitization projects of cuneiform documents have been launched in around 2000. For example, some studies tried to 3D scanning of original clay tablets (e.g. Digital Hammurabi Project (Watkins and Snyder, 2003)) and others treated transcription and translation (e.g. Cuneiform Digital Library initiative (CDLI) [1], Open Richly Annotated Cuneiform Corpus (ORACC)[2]).

To facilitate those digitization projects, (semi-)automatic digitization methods are required. Unfortunately, current digitization projects have been conducted manually. Few studies tried to detection of a character class from handwritten copies of original tablets (Massa et al., 2016; Rothacker et al., 2015), grammatical analysis (Homburg and Chiarcos, 2016) and automatic machine translation (Pagé-Perron et al., 2017). However, reliable (semi-)automatic digitization methods have been not established yet.

In this paper, we introduce our ongoing project, a construction of an imageset named Handwritten Cuneiform Character Collection (HCCC)[3]. HCCC is intended to be an image collection of frequently used cuneiform characters written with some stroke patterns and to be used for Optical Character Recognition (OCR) from handwritten copies of original clay tablets.

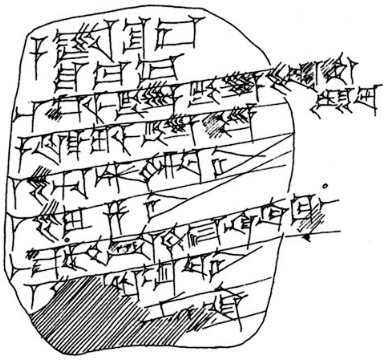


Figure 1: An example of handwritten copy of cuneiform tablet

Although our final goal is to develop an end-to-end method to directly produce transliterated scripts from original tablets, some limitations such as the condition of materials of the tablets prevents us from accessing the tablets. Fortunately, we can obtain handwritten copies of original tablets depicted by the scholars such as Figure 1[4] and many of such copies have not been transcripted yet. For example, there are about 70,000 digitally untranscribed hand copies out of all approximately 300,000 documents registered in CDLI.

Therefore, we set first goal to achieve OCR from the handwritten copies of original tablets and are constructing the imageset as explained above. This image set is the first attempt to collect handwritten cuneiform characters as far as we surveyed despite of the small scale of the dataset (at most approximately 200 images per chararcter class) like Omniglot dataset (Lake et al., 2011) for now.

[1] `http://cdli.ucla.edu/`

[2] `http://oracc.museum.upenn.edu/`

[3] Will be available on `https://github.com/yustoris/hccc`

[4] Excerpted from `http://cdli.ucla.edu/search/archival_view.php?ObjectID=P101022`

2. Imageset Construction Process

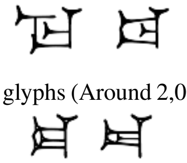
In this section, we describe the ongoing construction process of HCCC.

2.1. Target Classes and Glyphs

We began by determining target classes and glyphs for the construction. As denoted previously in Section 1., the number of cuneiform character classes are up to hundreds. However, most of them are ligatures of basic characters or less frequently appeared. For example, 𒅥 (GU_7, *to eat*) is composed of 𒅗 (KA, *mouth*) and 𒃻 (GAR, *bread*). Therefore we'll focus on most frequently appeared and probably most basic classes. More precisely, we selected most frequently appeared 50 character classes based on the result of counting occurence of the character classes on the available cuneiform translitelized corpora, Electronic Text Corpus of Sumerian Literature (ETCSL) [5]. The number of characters classified into the chosen 50 classes accounts for roughly 70% of the total character appearances.

Glyphs of cuneiform characters are also various depending on ages in which the characters were used, because the characters' shape had been gradually simplified during thousands years. In this work, we limited to most oldest (by Ur III, about 2,000 BC) ones. Since the oldest glyphs have complex shapes on each character class, we can distinguish classes more easily compared to newer glyphs as shown in Figure 2. This discernibility make us possible to collect images with less ambiguity between different classes.

However, there is one exception, i.e. we cannot distinguish E_2 𒂍 and KID 𒆤 clearly without any context. Thus, we classify those two characters into a same class.



(a) Old glyphs (Around 2,000 BC)

(b) New glyphs (Around 600 BC)

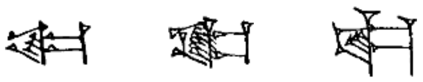
Figure 2: Comparison on glyphs of MA (left) and BA (right)

2.2. Resource

The original resource from which we obtained the candidates is openly accessible one, academic handcopy collections published in the 19th century and handcopies available on CDLI [6]. Most of the candidates have been collected from the former, and supplementary employ the latter when we cannot find characters which belong to the target classes.

We considered stroke variation depending on authors of handcopies. This stroke variations mainly causes by difference of the way how to write a cuneus of the character. For instance, Figure 3 shows three examples of such stroke variations for a same character class KA 𒅗.

When we tackle to OCR from the handcopies, we have to treat that variations. Therefore, we decided to include several stroke variations in each character class as possible as we can.



(a) Variation 1 (b) Variation 2 (c) Variation 3

Figure 3: Stroke variations in same character class

2.3. Generation of Candidates

We firstly derived *candidate character images* from given resource. Each candidate is expected to be an image which contains exact one cuneiform character. They were heuristically and automatically clipped from original handcopied tablet images.

2.3.1. Generation Method

The adopted heuristic method consists of following three procedures.

Firstly, we derived regions which are probably drawn tablets from given resource in which each page has several handcopied tablets. We converted original images into greyscale ones and clipped regions enclosed by contours detected by an algorithm described in Suzuki and Abe (1985). From those regions, we filtered out regions whose areas are under a fixed value.

After extracting tablet regions, we erased extra lines separating section as shown in Figure 4 by detecting lines whose length are longer than fixed rate of width of the tablet images. This erasing process is needed because those separation lines were commonly used by cuneiform scribes and are noises for extracting characters.

Finally, we extracted character candidates from the tablet images. We applied Gausian smoothing and detecting regions by enclosed contours with the same method when we used for extracting the tablet images. After filtered out areas whose area are smaller than a heuristically defined value and consequently the rest candidates were resized to 64x64. As a result of those procedures, we derived 147,010 character candidate images with 64x64 and grayscale.

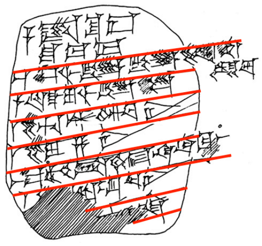


Figure 4: Figure 1 colored separation lines by red

2.3.2. Simple Filtering

Since these candidates are generated by automatic procedures, we can easily obtain a large number of character images. While we successfully got "complete" candidates, i.e. which can be clearly recognized as a target character, many candidates are "incomplete". That is, (Type-1) just noises, (Type-2) complete one character but contains noises,

[5] http://etcsl.orinst.ox.ac.uk/

[6] Fully listed in https://github.com/yustoris/hccc

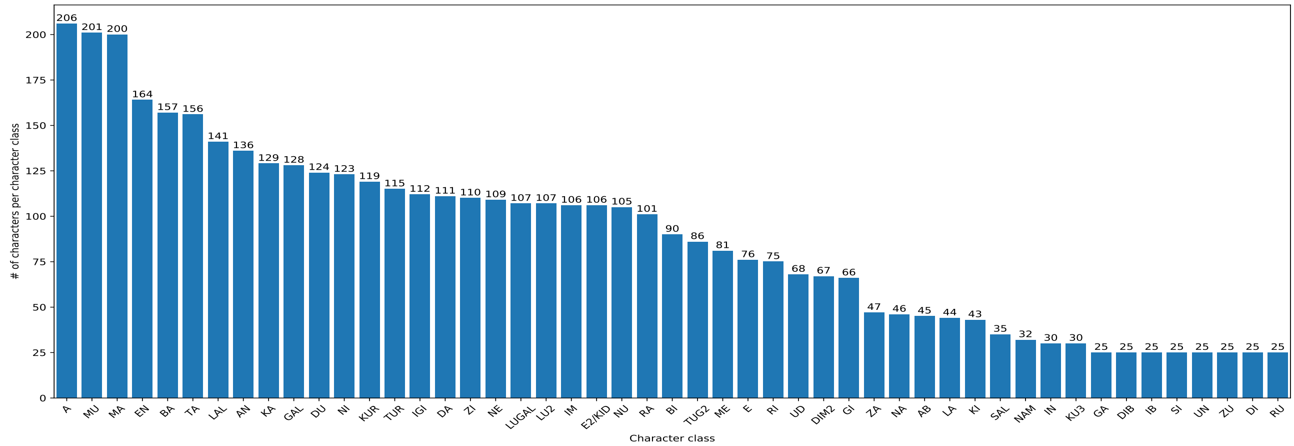


Figure 5: Distribution of currently collected imageset

(Type-3) only part of one character and (Type-4) include multi characters. In this work, we decided to filter out all of those incomplete candidates except (Type-2) because we can denoise and don't have to adjust aspect of the images again. For the same reason, the rotated images also are not included in the "incomplete" images.

From derived candidates images, we manually and roughly filter out (Type-1) images after clustered all the images into classes whose size is temporary fixed. This fixed cluseter size is determined considering the noise images denoted above. The clusering was conducted by converting the images into Bag of Visual Words (BoVW) in which the features are extracted by AKAZE (Alcantarilla et al., 2013) and unsupervisedly classified the converted images into 120 groups with with K Nearest Neighbors. Many of (Type-1) images were clustered into same classes, so we could easily recognize the images to be exclulded. We found 31 classes in which almost all candidate images were (Type-1) ones.

Consequently, we obtained 118,801 roughly filtered candidate images.

2.4. Labelling and Cleaning

After deriving the candidates filtered by the simple method explained above, we label and clean them manually. Because of a large number of incomplete candidates and limited workers, we set a first goal to collect confirmed images at small scale, i.e. at most approximately 200 images per target character class. In future, we'll conquer this scantiness by (semi-)automatic collection methods.

We annotate a character class on complete and (Type-2) incomplete candidates. After labelling, we denoise (Type-2) candidates and fix the direction of the candidates which are rotated above 90 degrees manually.

Futhermore, when we can not collect enough images for a character class from the handcopy collections, we will manually clip the target character from handcopies available on CDLI and also convert it to 64x64 gray scale format following to the auto generation method described in Section 2.3.1..

3. Current Status

Although all collection procedures written above has not been fully completed yet, we have already obtained 4,358 annotated images as of Sep. 9, 2017. More than 100 images are collected for 26 out of 50 character classes, maximum number of images for a class is 206 and minimum is 25. Figure 5 describes more detailed statistics of current number of collected images for each character class. Each bar describes the number of collected images for each character class and the bars are sorted in descending order.

4. Tasks to Enhance Imageset

As described in Section 3., the size of currently built dataset is limited. To scale up the dataset, we have to overcome some tasks.

4.1. Detection of Touched and "Complex" Characters

The heuristic method used in this paper can not detect one character from touched multi characters. This weak point resulted out (Type-4) incomplete candidates which probably were generated because of connecting the two or more characters by Gausian smoothing.

On the other hand, some characters such as IN 𒅔 consist of multi parts which tend to be separated by simple Gausian somoothing based character detection. Therefore, it is also needed to detect these "complex" characters as one character precisely.

To tackle to those touched and "complex" characters, it may be effective to apply some existing character recognition methods such as Rothacker et al. (2015) with training features for each target character class from current dataset.

4.2. Efficient Classification

Other task is to classify derived candidates into the target character classes more effectively. The candidates are too many to conduct the collection of characters by hand. Therefore, it is needed to combine automatic method and manual correction to achive more efficient and precise image collection.

However, regular supervised classification method such as ResNet (He et al., 2015) is not suitable to apply the ongoing dataset building because these methods require plenty of training data and we have only limited scale dataset now. To resolve this problem, we will refer some studies tries to develop classifiers from tiny training data. In particular, Koch et al. (2015) applied siamnese convolutional network to treat One Shot Lerning, which is a task creating classifier from only one sample for each class as we human beings are often able to learn object shape from only one instance of the target object.

We will apply those One Shot Lerning methods to develop reliable classifier from current small dataset and to conduct compilation of character images more efficiently.

4.3. Collection of Other Ages Glyphs

As denoted in Section 2.1., we limited target age for collection. However, for constructing more practical dataset, we have to collect other glyphs of other ages.

The important and typical glyphs are ones used in Neo-Assyrian and Hittite in addition to the collected glyphs in this paper. Thus, the future collection targets are those two glyphs. Since those glyphs are relatively simplified and the total number is smaller compared to oldest ones, we will have to tackle to the ambiguity between different character classes are higher as previously described in Figure 2b.

5. Conclusion and Future Work

We conducted a first attempt to build handwritten cuneiform character imagest, HCCC. The scale of built imageset HCCC is currently small compared to several large handwritten character dataset, such as CASIA dataset (a large handwritten Chinese character dataset (Liu et al., 2011)) or MNIST for now. However, the way to enrich the imageset has been already suggseted as explained in Section 4..

After finishing imageset construction, we will try to achive cuneiform character recognition from given handwritten source by image processing using HCCC and by considering linguistic context. The reason why we is to tackle to diversity of cuneiform languages and ambiguity of character glyph. In particular, the latter problem is needed to treat both graphically and linguistically. Each cuneiform character can be used both phonogramatically and ideogramically, and some cuneiform characters are too similar to distinguish. Furthermore, many excavated documents often lacks some characters due to poor material condition. Those graphical ambiguity and incompleteness would be solved by taking into account linguistic context such as proposed in Dȟondt et al. (2016).

6. Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Number 17K13549 (Grant-in-Aid for Young Scientists (B)) and 16K02642 (Grant-in-Aid for Scientific Research (C)).

7. References

Alcantarilla, P. F., Nuevo, J., and Bartoli, A. (2013). Fast explicit diffusion for acceleratedfeatures in nonlinear scale spaces. In *Proc. of the 2013 BMVC*.

Dȟondt, E., Grouin, C., and Grau, B. (2016). Low-resource ocr error detection and correction in french clinical texts. In *In Proc. of 7th LOUHI*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. In *Proc. of CVPR*.

Homburg, T. and Chiarcos, C. (2016). Akkadian word segmentation. In *Proc. of LREC 2016*.

Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *Proc. of Deep Learning Workshop, ICML'15*.

Lake, B. M., Salakhutdinov, R., Gross, J., and Tenenbaum, J. B. (2011). One shot learning of simple visual concepts. In *Proc. of the 33rd ACCSS*.

Massa, J., Bogacz, B., Krömker, S., and Mara, H. (2016). Cuneiform detection in vectorized raster images. In *21th CVWW*.

Pagé-Perron, E., Sukhareva, M., Khait, I., and Chiarcos, C. (2017). Machine translation and automated analysis of the sumerian language. In *Proc. of the 2017 LaTeCH-CLfL*.

Rothacker, L., Fisseler, D., Müller, G. G. W., Weichert, F., and Fink, G. A. (2015). Retrieving cuneiform structures in a segmentation-free word spotting framework. In *Proc. of the 3rd IWHDIP*.

Suzuki, S. and Abe, K. (1985). Topological structural analysis of digitized binary images by border following. *CVGIP*.

Watkins, L. and Snyder, D. (2003). The digital hammurabi project. In *Proc. of MW*.