

Cuneiform Symbols Recognition Based on K-Means and Neural Network

Naktal M. Edan

College of Computer Sciences and Mathematics
University of Mosul, Mosul, Iraq

Received on: 21/10/2012

Accepted on: 30/01/2013

ABSTRACT

Cuneiform is the ancient writing system in the world. But, there is no clear interest recognition cuneiform symbol, despite its importance. This research interested in building an algorithm for cuneiform symbol recognition. Firstly, the Sumerian texts were entered through the scanner and make some initial preprocessing operations such as segmentation for the purpose of cutting the text and getting a cuneiform symbol. Then, features were extracted for each symbol by using vertical and horizontal projections, centre of gravity, and connected component. Because of the large number of cuneiform symbols, the similar symbols are clustered by K-means algorithm, then multilayer neural networks are used to classify the symbol within the same cluster. The proposed algorithm gave good results.

Keywords: Symbols Recognition, K-Means algorithm, and Neural Network.

تمييز العلامات المسمارية بناءً على K-Mean والشبكة العصبية

نكتل مؤيد عيدان

كلية علوم الحاسوب والرياضيات

جامعة الموصل، الموصل، العراق

تاريخ قبول البحث: 2013/01/30

تاريخ استلام البحث: 2012/10/21



المخلص

تعتبر العلامات المسمارية أول نظام للكتابة عرفه العالم ولكن ليس هناك اهتمام واضح بتمييز العلامات المسمارية على الرغم من أهميتها. يهتم البحث ببناء خوارزمية لتمييز العلامات المسمارية. يتم أولاً إدخال النصوص السومرية عن طريق الماسح الضوئي وإجراء بعض عمليات المعالجة الأولية عليها مثل التقطيع لغرض تقطيع النص والحصول على العلامات المسمارية. ثم يتم استخراج الصفات من صور العلامات باستخدام الإسقاط العمودي والأفقي، ومركز الثقل، والعناصر المتصلة. وبسبب ارتفاع عدد العلامات المسمارية يتم أولاً عنقدة العلامات المتشابهة باستخدام خوارزمية K-means ثم يتم استخدام شبكة عصبية اصطناعية متعددة الطبقات لتمييز العلامات داخل العنقود الواحد وقد أعطت الخوارزمية المقترحة نتائج جيدة. الكلمات المفتاحية: العلامات المسمارية، خوارزمية K-mean، الشبكات العصبية.


1. Introduction:


Sumerians created the ancient writing system about 3000 B.C.E which is Cuneiform writing system. Cuneiform was used in commerce by many ancient civilizations and was valued for its practical use. The Sumerian writing system was a very important and powerful factor in the early eastern world. Cuneiform, which is the earliest known system of writing, was the Sumerians most important and useful contribution. Cuneiform, which is the earliest known system of writing, began before the first written history. [1]

For more than 35 centuries, through several stages of development, the cuneiform writing system was in use, The original Sumerian script was adapted for the writing of the Akkadian, Eblaite, Elamite, Hittite, Luwian, Hattic, Hurrian, and Urartian languages, and it inspired the Ugaritic and Old Persian alphabets. During the Neo-Assyrian Empire, Cuneiform writing was gradually replaced by the Phoenician alphabet. [2]

Cuneiform (from Latin meaning 'wedge-shaped') is composed of a series of short straight wedge-shaped strokes made with a stylus into a tablet of soft clay. The strokes are thickest at the top, like , on harder material, it is more looking like . At first, symbols were written from top to bottom; later, they were turned onto their sides and written from left to right. Harder materials were also used in later periods. Five basic orientations are applied: horizontal, two diagonals, a hook and a vertical stroke:



The up-diagonal stroke has limited use. These five components occur in two different sizes. A small hook often not being distinguishable from a short diagonal. The two diagonal strokes are not used as an individual sign, but the other types are. Reverse orientations (e.g. with the head of the wedge at the bottom) are hardly attested. The inverse vertical  is rare on old tablets (sometimes the vertical in the sign for 'hand',

which normally is . [2]

2. Related Work:

Reading cuneiform symbols is an important subject for understanding cuneiform tablets contents. Where there is a great number of cuneiform tablets which is copied as hand written. [3]

Yousif H. and et al (2006) proposed a method that used the intensity profile curves for selected pixels in the hand written images of Cuneiform text to differentiate between them. [3]

Al-Saif studied Cuneiform symbols photographs and suggested an alphabetic method depended on the geometric shape, where the symbol code generated form number of horizontal, vertical, oblique pin and the number of cross section pin. To get images with isolated pin high preprocess method, where performed to measure the exact threshold value to isolated the object from the background which is used for image slicing, each slice contains one pin. [4]

3. Proposed Cuneiform Symbols Recognition Algorithm

The steps of proposed algorithm are illustrated in flowchart which is shown in Figure(1).

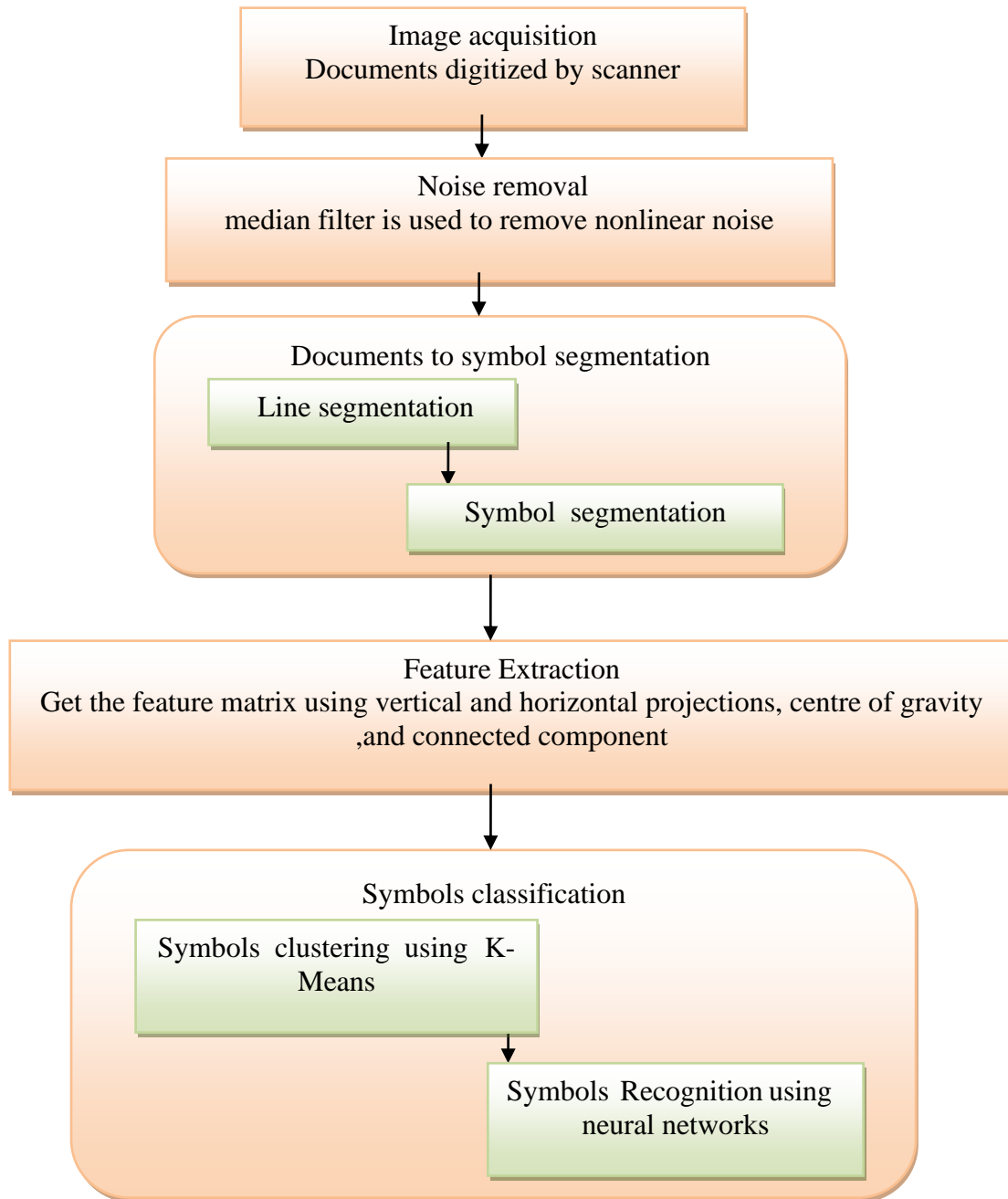


Figure (1). The Steps of the Proposed Algorithm

3.1 Data Acquisition and Preprocessing:

The cuneiform symbols are scanned by a scanner device with a resolution of 300 dpi and it is stored as a portable gray map (PGM) format file. Then, performed preprocessing steps which generally include smoothing, segmentation, ... etc. Noise errors caused by the data acquisition system, need to be eliminated from the scanned document, median filter is used to remove non-linear noise, where A 3*3 window is used to examine each pixel.

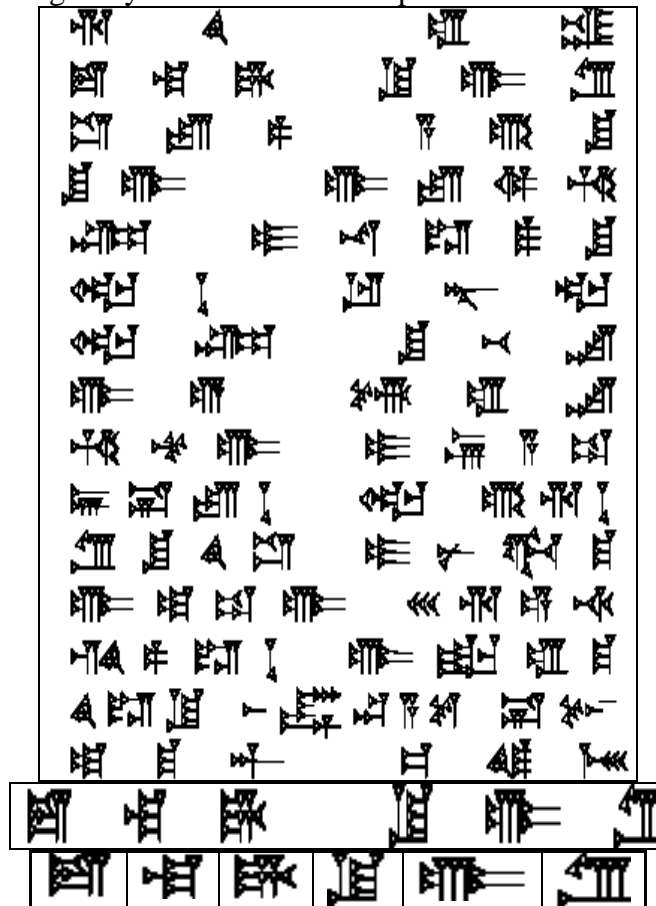
3.2 Segmentation:

Segmentation is a necessary step in order to isolate the cuneiform symbols which will be passed to the recognition stage for recognizing the symbol correctly, the scanned image must be segmented to set of images which only contain one symbol which will be recognized. Usually, line separation followed by separates the text line into cuneiform symbols.

This is accomplished by examining the horizontal histogram profile. It focuses on identifying physical gaps using only the components.[5] (see Figure (2)). Then, horizontal and vertical projection of cuneiform symbols image are used to find the outer rectangle of the cuneiform symbol image, where outer rectangle is a rectangle with the least size that all pixels of symbol are in it.

3.3 Feature Extraction:

Feature extraction is critical in any recognition system. The classifier performance depends directly on the features which have been extracted [6]. In this research, both structural and statistical features of an Cuneiform Symbols have been used to recognize each symbol as a separate object where vertical and horizontal projections, center of gravity and connected-component are the features used here.



Figuer (2). sample of Codex Hammurabi Symbols Ssegmentation

3.3.1 Vertical and Horizontal Projections Features:

Binary image refers to a two-dimensional binary function $f(x,y)$, whose value is either 0 or 1, where x and y denote vertical and horizontal coordinates of the pixel respectively, and where the coordinates origin is located at the top left corner of the

image. The maximum dimensions of the image are x_{max} and y_{max} , where x_{max} and y_{max} are the number of rows and the number of columns of a binary image, respectively.

Vertical Projection of a symbol image is applied as a graphical representation, showing a visual impression of distribution pixels in the symbol body. Mathematically, the Vertical Projection of an image can be computed by the following equation: [6].

The horizontal projection is defined as:

$$h(i) = \sum p(i, j) \quad \dots(1)$$

and the vertical projection as:

$$v(j) = \sum p(i, j) \quad \dots(2)$$

where i is the row number and j is the column number, P is the pixel value.[7].

The vertical and horizontal projection of a cuneiform symbol image after obtained, are divided into four equal vertical and horizontal strips. Then, the maximum and minimum values of each strip are extracted as the second type of features, thus, this type of features consists of sixteen elements in the feature vector .

3.3.2 The Center of Gravity Features:

The gravity of the symbol image is another part of features vector. The vertical and horizontal centers of gravity are determined by the following Equation (Eq. 3 & Eq. 4):

$$C_x = \frac{M(1,0)}{M(0,0)} \quad \dots(3)$$

$$C_y = \frac{M(0,1)}{M(0,0)} \quad \dots(4)$$

where, C_x and C_y are the horizontal and vertical centers of gravity, respectively, M_{pq} is the geometrical moments of rank $p+q$. [Al Tameemi]

3.3.3 The Connected-Component Features

Connected-component labeling is used in computer vision to detect connected regions in binary digital images, although color images and data with higher-dimensionality can also be processed. When integrated into an image recognition system or human-computer interaction interface, connected component labeling can operate on a variety of information [8] and [9].

4. Classification:

Because of the large number of Cuneiform symbols, then recognition in traditional methods is usually not practical, since there is a large number of training patterns and the dimension of the input and output space is fairly large.

Therefore, it is usual and practical to first cluster the training patterns to a reasonable number of groups by using a clustering algorithm such as K-means and, then build private neural network for every cluster to train and test symbols in that cluster.

A clustering algorithm is a kind of an unsupervised learning algorithm and is used when the class of each training pattern is not known. But, a neural network is a supervised learning network. At least, the class of each training pattern is known [10]. So, it is better to take advantage of the information of these class memberships when training patterns are clustered. Namely, the training patterns are clustered into classes, then neural networks are built to training and testing Cuneiform symbols for every cluster.

4.1. K-means Clustering Algorithm:

One of the most widely used algorithms for clustering is the k-means algorithm [11]. It is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem [12]. It is a local search algorithm and follows a simple and an easy way to classify n data points into k clusters. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. The better choice is to place them as much as possible far away from each other. [12].

The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point, we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function:

$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2 \quad \dots(5)$$

Where, $\|x_i^{(j)} - c_j\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j , is an indicator of the distance of the n data points from their respective cluster centers. [12] and [11].

4.2. Multilayer Neural Network (MLP):

The multilayer neural network (MLP) basic structure is comprised as input layer, one or more hidden layer(s), and an output layer. Each processing element (node) in a particular layer is fully connected to every processing element in the succeeding layer. The number of nodes, in the input layer, is set according to the actual number of features used. The number of output nodes is set according to the number of pattern classes [13].

A multilayer neural network (MLP) trained by a backpropagation of errors algorithm. It is simply a gradient descent method to minimize the total squared error of the output computed by the network. The training of a network by backpropagation involves three phases. These phases are feed forward of the input training pattern, backpropagation of the error of this pattern, and updating the weights. After training, the network is applied by using the feed forward phase of the training algorithm Figure(3) shows multilayer neural network architecture [14].

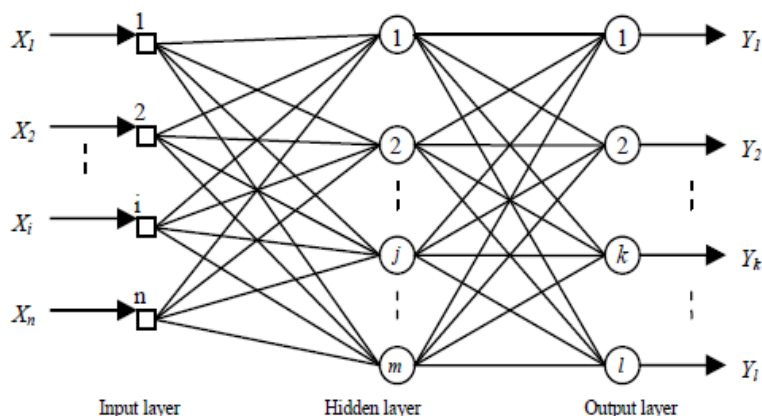


Figure (3). Multilayer Neural Network (MLP)

5. Results and Discussions:

The experiments are conducted on cuneiform symbols which are extracted from Hammurabi codex by segmentation. For each symbol, 6 samples are extracted 4 samples are used for training and 2 ones are used for testing. The features are extracted for each symbol by vertical and horizontal projections, centre of gravity, and connected component. k-means clustering algorithm are applied. To find an appropriate number of cuneiform symbols classes, different values of K are considered . The value yielding a lower classification error rate is chosen as the best value. Here, total symbol classes are taken as = 5. For each symbol classes multilayer neural network were built and the number of input, hidden, output nodes in the network depended on the number of symbol in that class. In testing stage firstly the classes of unknown cuneiform symbols must be specified by calculating Euclidean distance between the symbols and centroid of each cluster, then it is fed to the appropriate Multilayer Neural Network. Table (1) shows the final recognition rates.

Table 1. Final Recognition Results

Class No.	No. of symbols in class	Recognition Rate %	FAR %	FRR %
Class 1	100	90.4	6.3	3.3
Class2	67	95.1	2.4	2.5
Class3	187	88.4	1.6	10
Class4	79	91.2	6.1	2.7
Class5	167	83.3	9.4	7.3

6. Conclusion:

Although, there are tens of thousands of clay tablets have been discovered a small amount of work has been done that deals with digitized and recognized cuneiform symbols. This research proposed a new algorithm for cuneiform symbols recognition, where there is no clear interest recognition algorithms.

Due to the large number of cuneiform symbols and to get a high recognition rates the symbols are clustered in depended on their features to classes, for each class there is appropriate multilayer neural network that is used for recognizing the symbol, where the number of neural network used here is equal to the number of classes which is equal to 5. The algorithm proposed in this research is efficient and produces a good recognition rates.

REFERENCES

- [1] T.N., T.D.H., 1988, "The Evolution of Cuneiform, History & Thought of Western Man", Rich East High School, <Http://www.richeast.org/htwm/index.html>.
- [2] Heise John, 1996 , "Akkadian language' on the origin and development of cuneiform". <http://www.sron.nl/~jheise/akkadian/Welcome.html>.
- [3] Yousif H., Rahma A., Alani H., 2006, "Cuneiform symbols recognition Using intensity curves", The International Arab Journal of Information Technology, Vol.3, No. 3, PP:237-241.
- [4] Al-Saif K. , 2002 , "Cuneiform symbols recognition" , A Ph.D. Thesis , College of Computer and Mathematical Sciences University of Mosul , Iraq.
- [5] Romeo-Pakker K., Miled H., and Lecourtier Y.,1995, "A New Approach for Latin/Arabic Character Segmentation," Proc. Int'l Conf. Document Analysis and Recognition, pp. 874-877.
- [6] Al Tameemi A.M., Zheng L., Khalifa M. , 2011, "Off-Line Arabic Words Classification using Multi-Set Features", Information Technology Journal 10 (9): PP:1754-1760.
- [7] Aljuaid H., Mohamad D., Sarfraz M., 2009,"Arabic Handwriting Recognition Using Projection Profile and Genetic Approach", Fifth International Conference on Signal Image Technology and Internet Based Systems , PP:118-125.
- [8] Samet H. , Tamminen M. 1988. "Efficient Component Labeling of Images of Arbitrary Dimension Represented by Linear Bintreees". IEEE Transactions on. Pattern Analysis and Machine Intelligence (TIEEE Trans. Pattern Anal. Mach. Intell.) 10: 579. doi:10.1109/34.3918.
- [9] Michael B. Dillencourt, Hannan Samet , Markku Tamminen, 1992. "A general approach to connected-component labeling for arbitrary image representations". J. ACM. <http://doi.acm.org/10.1145/128749.128750>.
- [10] Musavi, M., Ahmed, W., Chan, K., Faris, K., and Hummels, D. 1992." On the training of radial basis function classifiers". Neural Networks, 5, 595-603.
- [11] Vattani., A., 2011, "k-means requires exponentially many iterations even in the plane". Discrete and Computational Geometry 45 (4): 596–616. doi:10.1007/s00454-011-9340-1. <http://cseweb.ucsd.edu/users/avattani/papers/kmeans-journal.pdf>.
- [12] MacQueen J. B., 1967, "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297.
- [13] Gauri, S. K. 2010, "Control chart pattern recognition using feature-based", Springer, Int J Adv Manuf Technol 48:1061–1073.
- [14] Callan R., 1999, "The Essence of Neural Network", Prentice Hall, New Jersey.