Geometric Cognition Method of Biomimetic Pattern Recognition and Its Application in the Digitalization of Modern Guangdong Customs Archives

CHEN YONGSHENG 1, CHEN LILAN 1

1School of Information Management, Sun Yat-sen University, Guangzhou, China.(email: Chen Yongsheng: cyszdl111@163.com; Chen lilan: chenllan@mail2.sysu.edu.cn)

Corresponding author: Chen Lilan

ABSTRACT This paper puts forward a digitalization model of Modern Guangdong Customs archives based on biomimetic pattern recognition. By introducing biomimetic pattern recognition, the digitalization of Modern Guangdong Customs archives becomes a social production of information network. With this model, we can not only make full use of the wisdom of biomimetic pattern recognition, but also digitalize the Modern Guangdong Customs archives with high efficiency and low cost, thus fully exploring the internal value of these archives. In this paper, firstly, we propose Neuron Model of Digital Biomimetics, and investigate Geometric Sequential Learning Ahead Mask in digitalizing Modern Guangdong Customs archives. Then, we analyze the application of digital biomimetics to the digitalization of Modern Guangdong Customs archives. Finally, we analyze the of comparison of the work cost among Taobao, Professional Institutions, Crowdsourcing and Digital Biomimetics, which proves that the geometric neuron model is reasonable and effective.

INDEX TERMS Biomimetic Pattern Recognition; Digitalization; Geometric Sequential Learning Ahead Mask; Modern Guangdong Customs Archives

I. INTRODUCTION Digitalization of archives has changed the carrier of traditional archives, but it still faces problems such as huge quantity and variety of archives in the archival digitalization process. Procedures in processing archives such as digital photography, information acquisition, image processing and computer-aided design require a lot of human resources, time and economic costs. To digitalize such a variety of archives, if we only rely on a small number of professionals in the archival organizations, we can achieve half the result with great cost of human, financial and material resources. In view of the difficulties existing in the digitalization of archives, how can we find a fast, efficient, cost-effective and time-saving way to solve the problem?

With the production requirements and the development of science and technology, people have realized that biomimetic pattern recognition system has been one of the main ways to develop new technologies since the 1950s, and have consciously regarded biology as the source of various technological ideas, design principles and inventions [1]. Chemistry, physics, mathematics and technological models have been used to carry out in-depth research on biological systems, which has promoted the development of biology greatly and made rapid progress in the study of the functional mechanism of organisms. From then on, imitating organisms was no longer fascinating fantasy, but could be achieved. Biologists and engineers were actively working together to start applying information and knowledge gained from the biology to improve the old, out-of-date technical engineering equipment or create new one. Biology began to be applied to technological innovation and revolution in all walks of life, and first achieved success in military departments such as automatic control, aviation and navigation. Hence biology and information engineering technology are combined and mutually infiltrated, bringing in a new science – biomimetic pattern recognition [2]–[4]. In brief, biomimetic pattern recognition is a science that imitates biological characteristics. To be exact, biomimetic pattern recognition is a comprehensive
science that studies the structure, characteristics, functions, energy conversion, information control and other excellent characteristics of biological information systems, and applies them to technical systems, improving existing technical engineering equipment and creating new information systems for technological processes. From the perspective of biology, biomimetic pattern recognition belongs to a branch of applied biology. From the perspective of engineering technology, biomimetic pattern recognition provides new principles, methods and ways for designing and building new technological equipment based on the research of biological systems, such as deep learning, multi-modal learning and so on [5]–[8].

The glorious mission of biomimetic pattern recognition is to provide the most reliable, flexible, efficient and cost-effective information technology system analogous to the biological system for the benefit of mankind. The research objects of this study are the Modern Guangdong Customs archives, focusing on the problems in the digitalization of these archives. In the past, the main body of the digitalization work was the archival organization itself. But the main body was restrained in the support of information technology, blocking the digitalization of archives to a certain extent, which is not conducive to the sharing of archival resources. There are many kinds of archival documents, which contain abundant archival knowledge. To some extent, this isolated processing model hinders the diffusion and dissemination of knowledge. By introducing the biomimetic pattern recognition, the photocopies of the archives could be released on the biomimetic pattern recognition platform after certain processing, and then the archives are textualized, translated and proofread according to the acquisition theory of biology and the information acquisition methods. In this way, archive users can better understand the content of Modern Guangdong Customs archives while digitalizing them, which is conducive to improving the public’s awareness of the protection of and participation in Customs archives. As far as archives are concerned, having a textual, translational and proofreading users bioinformatics chain can enhance the dissemination and diffusion of academic resources. For further research and discussion, there are still volumes to be consulted and theories to rely on. Therefore, the archival digitization based on the biomimetic pattern recognition effectively lessens the isolation problem in archival digitization and promotes the efficiency of archival digitalization. After introducing the biomimetic pattern recognition, archives and other archival organizations can break through the barriers and find a large number of human resources outside the institutions. Through the biomimetic pattern recognition, archival organizations can release the task requirements on the network. Thus a large number of Internet users then can get informed of the work requirements through the biomimetic pattern recognition platform, and the willing and capable talents become digital biomimetic users through the Internet platform. By this means, archival organizations can obtain more professional talents for the archival digitalization, optimizing the distribution of resources. Digital biomimetics users select appropriate bio-information packages for their professional areas of interest, and begin to refine their work and digitalize the archives. In this way, the problem of excessive material cost and time cost faced by professionals in archival organizations when digitalizing Customs archives is effectively solved, and the method of biomimetic pattern recognition can excavate the core value of the Modern Guangdong Customs archives. The digitalization of Modern Guangdong Customs archives by digital biomimetics users can help more archives to be sorted out more smoothly and quickly. The classification and gathering of archival resources by professionals outside the archival organizations can improve the efficiency of archival digitalization, forming a systematic and scientific disciplinary system more quickly, which is conducive to the profound research and comprehensive utilization of archives. Besides, many archives, such as ethnic genealogy, archives of late Qing Dynasty and ancient documents, have extremely important scientific research value. It is of far-reaching significance to bring in biomimetic pattern recognition to sort out archives in detail, systematically and thoroughly, so as to excavate the core value of archives more fully.

II. METHOD

A. DIGITALIZATION OF MODERN GUANGDONG CUSTOMS ARCHIVES BASED ON BIOMIMETIC PATTERN RECOGNITION

1) Neuron of Digital Biomimetics Users

Biomimetic pattern network has considerable neurons and infinite potential, thus making full use of biomimetic pattern network can solve the difficult problems in archival digitalization. Although the activation of neurons in biomimetic pattern recognition has been deeply discussed by scholars, it has not been mentioned in the biomimetic pattern recognition of Modern Guangdong Customs archival digitalization. This paper puts forward for the first time the research on the application of biomimetic pattern recognition to the digitalization of Modern Guangdong Customs archives. But in the digital biomimetics, how to attract and activate digital biomimetics users’ neurons is a question.

Internal Activation of Neurons

When doing the same thing, some people are happy to do it with good results, some people need incentives to do it, with the result being good; still some have to be compelled to do it with poor effect. Why? The essential reason is that people with genuine interest in something will try their utmost to do it well. Many of the participants in the Modern Guangdong Customs archival digitalization based on the biomimetic pattern recognition are digital biomimetics users who have sincere interest in Modern Guangdong Customs archives, and some of them are researchers in the academic field of Modern Guangdong Customs archives. For amateurs, they not only fully contacted with the Modern Guangdong Customs archival resources, which is conducive to the successful completion of the Modern Guangdong
A mathematical description of a user’s neuron is:

\[ y = f(\Phi(x_1, x_2, x_3, \ldots, x_n) - \theta) \]  

(1)

Where \( y \) is the output of the neuron, \( f \) is the excitation function of the neuron, \( \Phi \) is the function, and \( \theta \) is the threshold of the neuron. In the pattern recognition application, \( f \) is a geometric activity function, called geometric neuron. That is:

\[ f(\Phi(x_1, x_2, x_3, \ldots, x_n) - \theta) = \begin{cases} 
1, & \Phi(x_1, x_2, x_3, \ldots, x_n) - \theta \geq 0 \\
0, & \text{otherwise}
\end{cases} \]  

(2)

In this way, a geometric neuron is an activity functionally equivalent to a hyper-geometry volume in high-dimensional space, as shown in Formula 2, forming an archival cognitive geometric space.

Because each neuron has its corresponding geometric volume, for a geometric feedforward network, its hidden layer corresponds to many geometric volumes, and its output layer is the operation of finding the geometric volume. In this way, we can use geometric volume operation to analyze and study artificial neural network, and design novel neural network models with different geometric volumes and algorithms.

3) Geometric Sequential Learning Ahead Mask (GSLAM)

In this paper, we propose a Geometric Sequential Learning Ahead Mask (GSLAM) model based on sequential learning ahead masking model put forward by academician Shoujue Wang [9]–[12], which is a neural network model and algorithm for archival cognitive pattern recognition. This algorithm employs a different network structure from feedforward network [9] (as shown in Figure 1) – General Feed-Forward Network (GFFN). It can be seen from the figure that each neuron in the hidden layer is connected to all inputs of the neuron, and the output of the neurons with smaller numerical order will be connected to the ahead layer neurons with larger numerical order. Since we adopt the geometric ordering in archival recognition, thus the network structure can also be called Archives General Feed-Forward Cognition Network (AGFFN). It can be seen that by setting weights, the AGFFN can form a feedforward network with arbitrary layers (the maximum number of layers is \( M \), and each layer has a series of geometric neurons. Therefore, the traditional multi-layer feedforward network is a special case of the general feedforward network. Let the dimension of the input sample be \( N \) (so that the number of input nodes of the neural network is \( N \)), and the number of geometric elements of the neural network is \( M \).

Suppose the value range for each input node is \([1, 1]\), and each neuron uses the following geometry structure. The for-
cognitive body or volume, which cuts the input space into two parts just like a group of geometries. However, the forward masking ability of the hidden layer unit enables the former part of the geometry not to be affected by subsequent covering, that is to say, the part covered will be removed and will not participate in further division. In this way, only one class of samples can be separated for each cut, and all the samples of different classes can be separated by covering one after another. This is the basic idea of learning algorithm of GSLAM model, which we can call parting algorithm. Figure 2 depicts this process visually. GSLAM higher-order neural network can, without affecting the trained neural network archival classifier, enlarge the geometric coverage of the new archival samples caused by the realization of network size through increasing extra higher-order neurons, thus realizing the archival classifier’s function of continuous expansion.

In the archival training and recognition process, the transfer function of AGFFN network neurons is geometry hard limiter function (Formula 2). Let the AGFFN network neuron input space be N-dimension space $R^n$, the output be continuous value, the network input be $X = (x_1, x_2, \ldots, x_n)$, the connection weight(s) kernel matrix be $W = (w_1, w_2, \ldots, w_N)$, the connection weight $W = (w_1, w_2, \ldots, w_N)$, orientation weight $W' = (w'_1, w'_2, \ldots, w'_N)$, and the threshold of the higher-order neuron $\theta$, and the output of the neuron will be

$$O_j = f(W^* (X_j - W') - \theta)$$

(3)

where $f$ is the constructed corresponding geometric neuron activation function.

The construction of AGFFN neuron parameter

After the training, the fixed archival neurons are listed according to the order of their generation, from the small number to the larger. Let the number of network neurons be $N$, the maximum connecting number of neurons be $N + a$, $a = 1, 2, \ldots$, which are the serial number of network neurons. And the connection of $M$ results from the input node of $M$, and the connection of $a - 1$ comes from the output of neurons whose serial numbers are smaller than $a$, the activation threshold of one neuron is $\theta_i$, which is fixed in the training process.

4) The application theory and method description

The learning algorithm of GSLAM model is divided into five steps:

- The training samples are fed into the network, with each sample as the core by turns, the distance between the core and the other samples is then calculated. Find the number of separable samples of the same kind of each candidate neuron from the sample output list, then choose the neuron with the largest number of separable samples of the same kind from the candidate sample group as the network neuron, and remove the isolated samples from the training set, feeding the removed samples into the network for further training until the whole sample set being divided completely. Now, the whole network construction is completed.

- Make geometries with all archive database samples randomly, and calculate the optimal threshold for each geometry, so as to extract as many samples of the same kind from the geometry as possible. Set all archive database samples A, threshold T, Geometry G, the kinds of archive k

- Further adjust geometry, and threshold $T = T + \Delta T$ according to certain rules, making it possible to divide more samples of the same kind.

- Record a geometric weighted, with the largest number of samples, threshold T and the cut-out category number k (archive), and remove the cut samples from the archive sample set;

- Neural geometry volume about archive samples is added to the hidden layer, and the W and T recorded in the previous step are taken as their weights and thresholds, and the connection weights between the neuron and the first k output archive neural geometry volume is supposed to be 1, and the connection weight between the neuron and other output neurons is 0.

- Repeat the above steps until only one class of samples remains in the sample set of the archive database.

With the GSLAM theory, the subsequent samples can increase new learning samples through adding new neurons without impacting the trained network, and can also learn all samples again, just like the reviewing and summarizing process, in which the formerly acquired knowledge and the new knowledge is classified and summarized, forming better-organized knowledge set, which refers to the network structure(Figure 3).

B. DATABASE DESIGN AND THE APPLICATION

ANALYSIS OF ARCHIVAL DIGITALIZATION SYSTEM

BASED ON BIOMIMETIC PATTERN RECOGNITION

1) Database Design

The archival database based on biomimetic pattern recognition is a further gathering and summarization of data resources. Constructing a reasonable and effective archival data
structure not only can improve the development efficiency of the system, but also plays a vital role in the maintenance of the system.

There are six tables in the archival database, which are release table, document table, document category table, user table, document status table and user category table.

The release table is used to store the textualized archives and information release in the later stages, including seven fields. The ID of the textualized archive is the primary key, and the other fields are the document ID, the title of the archive, the author, the keywords, the input date and the full text of the archive.

2) Application for Modern Guangdong Customs Archives

Taking the Modern Guangdong Customs archives as the data source [16], this paper tries to realize Modern Guangdong Customs archival digitalization using the biomimetic pattern recognition. According to statistics, there are 321 fonds of archives in Guangdong Provincial Archives. Among them, the Modern Guangdong Customs archives include 16115 volumes recorded from 1861 to 1949, involving the operation of the Customs, trade agreements, social conditions and public opinions, and most of them are in English. In this biomimetic pattern recognition, a small number of Modern Guangdong Customs archives have been selected as the data resources, which are in the form of microfilms, and we manually sort out the photocopied pictures of the microfilms according to chapters and then integrate the pictures into PDF-format archives. Then we upload and input these archives into the database of the system as task geometric learning and analysis to be distributed. According to the analysis and design of the following modules, the digital biomimetic system of these archives will be further realized.

Therefore, the method in this paper is of great benefit to the digitization of archives, and the establishment of an archival digitization system based on biomimetic pattern recognition is a powerful support for the archival digitization. With such a platform, we can extend the biomimetic pattern recognition to practice, combining the theory with practice, thus further enhancing the scientific research value of this paper. This paper first establishes an overall framework for the analysis in many aspects, and then completes the construction and analysis of the module of biomimetic pattern recognition.

3) Constructed Model of Digital Biomimetics Users

After digital biomimetics is realized in archival users, several WeChat groups are selected as the subjects of digital biomimetics trial operation to test the effectiveness and usability of biomimetic information system. Firstly, 100 complete Modern Guangdong Customs English archives are extracted and classified into 100 task data. Each task data contain necessary task descriptions and operation procedures. The release of digital biomimetics information through the above channel activates the tasks externally, recruiting data partitioning personnel. Since undergraduates have acquired...
certain English, the recruitment process is very democratic, and generally all the applicants can pass the qualification examination. Of these 100 archives, the number of pictures in each archive is by and large fixed, and the clarity of the pictures is different. Initially, the salary paid for completing the task of a data package was 20 RMB, and the salary would be transferred directly to the WeChat wallet of the digital biomimetics users after the examiner submits the reviewed salary list. During a certain period of trial operation of the digital biomimetics recognition, 100 series of task data were distributed, and the digitization of these 100 archives in the data task packages was completed. All the selected archives are in English, with some having a larger professional English vocabulary. In fact, those series of task data with clear pictures, less workload and smaller professional English vocabulary were distributed faster and completed in a short time, with task rejection rarely taking place. On the contrary, archives with unclear pictures, long English sentences and relatively more professional expressions were rarely successfully distributed, with some task packages remaining unavailable for a long time. So we increased the salary to 2030 RMB for the series of task data with blurred pictures or larger professional English vocabulary needing a longer time to complete. After the salary change, the tasks were successfully distributed and completed, but the completion effect of 100 archival documents is not the same.

III. EXPERIMENT AND RESULTS

A. EVALUATION AND IMPROVEMENT OF BIOMIMETIC PATTERN RECOGNITION BASED ON THE EXPERIMENT

Through the experiment with the digitalization and translation of the 100 Modern Guangdong Customs archives based on biomimetic pattern recognition, the GSLAM model designed in this paper is feasible. The experience and enlightenment summarized in the experiment are of great reference value in the actual implementation.

For about 1,000 pictures needing textualization, a total of about 100 people participated and the final completion was made by 500 people, the completion rate being 87.4%. For about 800 archives (including repetitive translation), a total of about 200 people participated and the final completion was made by 40 people, the completion rate being 100%. If there had been no recruitment time limit, the number of participants would be larger. Compared with contracting the task to a specific number of people, the user participation based on digital biomimetics is higher, and if implemented on a larger scale, the effect will be better.

B. SHORTER TIME CONSUMED IN THE DIGITAL BIOMIMETICS RECOGNITION

For the textualization of 100 Modern Guangdong Customs archives, if biomimetic pattern recognition is not adopted, and 10 people are assigned to complete the textualization task at the speed of 3 minutes per picture, considering the difficulty of hard-to-identify handwriting recognition, the completion time required with continuous work (8 hours per day) is about 28 days; while based on biomimetic pattern recognition, it takes about 5 days to complete the recognition and textualization of all the standard pieces and sub-writings.

1) Time Consumption Analysis of the Modern Guangdong Customs Archival Digitalization

Based on the experiment, time consumption mainly concentrates on two major parts. The first part is the preparation time, i.e., the time spent on data collation and task packaging. The second part is the time spent on user recruitment, that is, the time it takes to release task information through network channels and recruit users. Whereas the users’ working time is relatively short. This indicates that the time needed to be measured in the biomimetic pattern recognition is the preparation time before the work starts.

There are 16115 volumes of Modern Guangdong Customs archives in total, and each volume contains 100 to 1000 archives, averaging 4 pages, thus transferring all these archives into scanned forms amounts to about 3223,000 copies in total. And it is expected to take more than 10 years, or even longer period, to outsource the textualization and translation of these archives to professional institutions.

Under the biomimetic pattern recognition, it is necessary to increase the investment in labor costs for the digitalization of
the whole Modern Guangdong Customs archives. Applying GSLAM algorithm to the construction of the biomimetic pattern recognition based on the website, the effect of network promotion will be more obvious.

2) Cost Analysis of the Translation Experiment

In this paper, the work evaluation dimension table for different quality components and different translation tasks was adopted. All the participants filled in the work evaluation dimension table, totaling 312 non-duplicate translation documents, with a 100% information recovery rate. The users’ task completion time and the number of words were recorded, and data analyzed, the analysis of the average showing that users spent 45 minutes completing each document and there were 550 words per document. This paper compares the cost spent in work among Taobao, Professional Institutions, crowdsourcing [17]–[20] and digital biomimetics, as shown in Figure. 7.

![FIGURE 7: Comparison in Work Cost](image)

IV. CONCLUSIONS

Finally, based on the experiment, this paper puts forward optimization suggestions for the two main parts of user incentives and quality management. The biomimetic pattern recognition designed for the digitalization of Modern Guangdong Customs archives has promising application prospects, but it also has drawbacks, listed as follows.

Firstly, the design of the biomimetic pattern recognition is based on digital information and needs to be supported by digital information network system. In this paper, we only describe the prototype and function of feedforward network implemented by GSLAM of the whole biomimetic pattern recognition system, but the construction of biomimetic pattern recognition has not yet been fully realized. How to improve it reasonably is the next task of this paper.

Secondly, the incentive mechanism of biomimetic pattern recognition is mainly of money incentive; the simple structured cognizable geometric covered body may be a little simple. From the analysis of many successful cases, it can be seen that multiple incentives coexist in order to obtain geometries with good cover and achieve the best motivating effect. Due to the limited experimental literature, it is impossible to explore the geometric coverage combined with different motivation patterns. How to construct complex geometry by information fusion with different motivation patterns is also what we need to do next.

Thirdly, although the quality control system of the whole biomimetic pattern recognition has passed the test of the experiment, how to ensure the high quality of large-scale document digitalization is still the focus of future exploration.

V. ACKNOWLEDGEMENT

The authors gratefully acknowledge the generous support from major projects of the National Social Science Fund of China of 2017(project No.: 17ZDA200)

REFERENCES
