Text Fragment Recovery Method Based on Imperial Competition Algorithm

Dongyu Guo [a)], Zeyuan Tao [b)]

A College of Automation, Wuhan University of Technology, Wuhan 430070, China

[a)] Corresponding author: guodongyu1209@163.com
[b)] tzy@whut.edu.cn

Abstract. In this paper, a text fragment splicing method based on K-means and Empire competition algorithm is proposed. First, use K-means clustering to find out a piece of paper that belongs to one line; Then, define the edge distance of the paper, and establish a combined optimization model; using the imperial competition algorithm to optimize the solution. The experimental results show that the speed and accuracy of this algorithm are improved compared with the literature algorithm and have certain application value.

Key words: Imperial Competition Algorithm; K-MEANS clustering; text fragment splicing.

INTRODUCTION

Text fragment splicing technology has a strong applicability in the field of historical document restoration and military intelligence restoration. Images and recovery is a typical application in the field of computer vision. In the computer field, the existing text fragment splicing method has simulated annealing algorithm [1]; Genetic algorithm [2]; Ant Colony Algorithm [3], etc. But when there are too many text fragments, the classic optimization algorithm is not enough to get the optimal solution. This paper presents a method of text fragment restoration based on imperialist competitive algorithm.

The Imperial Competition Algorithm (ICA) was proposed by Atashpaz-Gargari and Lucas in 2007. The Imperial Competition Algorithm is applied to solving TSP [1], power system design [2]et.

This paper based on the existing research, proposes a discrete ICA algorithm to solve the problem of text segment splicing. First, import the text fragment image information into the computer. Then, the image is binarized. transform the binary image into a matrix for clustering. Finally, according to the discrete ICA algorithm, the overall matching degree is the highest means the restored image is obtained. After many simulation experiments, the results are compared with other algorithms, and the feasibility and effectiveness of the algorithm in the splicing of fragments are proved.

RESTORATION OF TEXT FRAGMENTS BASED ON ICA

Convert Image Information into a Matrix.

The text fragments used in this paper have the same size and shape, and they are all $p \times q$ pixels. First, the text image is imported and stored in the computer. In order to facilitate the further processing, the text fragment image is transformed into grayscale image, and the binary image of black word white background is further obtained. In the actual situation, the image of text fragment image binarization is shown in figure 1.

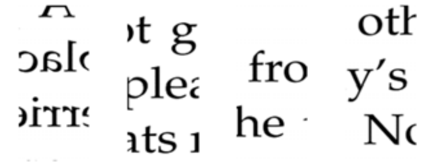


FIGURE 1. Text fragments example

The binary image information is stored in the computer as p × q size,0-1 matrix S,

$$S^{i}=\begin{pmatrix} s^{i}_{1,1} & \cdots & s^{i}_{1,q} \\ \vdots & \ddots & \vdots \\ s^{i}_{1,p} & \cdots & s^{i}_{p,q} \end{pmatrix} \tag{1}$$

Picture K- mean Clustering

Since the original text segment image is cut in two directions on the two-dimensional plane, the same row can be clustered first using k-means clustering to simplify the subsequent calculation. The steps are as follows:

Step1: Select m text segments as cluster centers$Z_1(1), Z_2(1), \cdots, Z_m(1)$, and the parenthetical numbers in the parentheses represent the number of iterations of the cluster center.

Step2: Use the shortest distance principle to assign the sample to one of the m cluster centers, if

$$\min\{\|X\text{-}Z_i(k)\|, i=1,2,\cdots,m\}=\|X\text{-}Z_j(k)\|=D_j(k) \tag{2}$$

Then $X \in S_j(k)$.In the formula, k is the sub ordinal number of the iterative operation.

Step3: Calculate the new vector value of each cluster center.$Z_j(k+1), j=1,2,\cdots,m$.

$$Z_j(k+1)=\frac{1}{N_j}\sum_{X\in S_j(k)} X, j=1,2,\cdots,m \tag{3}$$

Then use the average vector as a new clustering center. In this step, we need to calculate the sample mean vectors of M clusters separately.

Step4: If $Z_j(k+1) \neq Z_j(k), j=1,2,\cdots,m$,then back to step 2, reclassify the pattern samples one by one, and iterate over the calculation. If $Z_j(k+1) = Z_j(k), j=1,2,\cdots,m$, algorithm convergence, clustering finished.

The Establishment of the Text Fragments Splicing Model.

After k-means clustering, each line of the fragment is analyzed and it is found that the proximity index can be set as the matching degree of the image edge pixels. The reference [3] provides a solution model for the splicing of scrap paper. The goal of stitching is to minimize the value of the edge gap of each sub image in the combination of images based on stitching order. The concrete mathematical expression of this model is as follows:

$$\min C=\sum_{i=1}^{n}\sum_{j=1}^{n} dis_{ij}x_{ij} \tag{4}$$

$$st.\sum_{j=1}^{n} x_{ij}=1, i=1,2,\cdots,n \tag{5}$$

$$\sum_{i=1}^{n} x_{ij} = 1, j=1,2,\cdots,n \tag{6}$$

$$\sum_{i,j\in s} x_{ij} \leq |s|-1,\ 2\leq |s| \leq n-1 \tag{7}$$

In $x_{ij} \in \{0,1\}, i,j = 1,2,\cdots n, i \neq j$,$x_{ij}$ is 0-1variable, when it is equal to 1, it means that the paper i is splice into the front of the paper j, otherwise it does not splice. The last three main constraints said each fragment will only to a piece of paper, each piece will only spell ahead in front of a piece of paper, in addition to the fore and aft two pieces, other paper does not constitute a loop. Through the analysis of the above mathematical models, we can find that the mathematical model of the problem is consistent with the traveling salesman problem model and belongs to the asymmetric traveling salesman problem.

ICA for Solving the Segmentation of Text Fragments

Coding

First, encode the original Empire. The coding methods of TSP include path representation, sequential representation and neighbor representation. In this paper, the neighborhood representation is chosen to encode. Each initial empire can be regarded as a solution to this problem. Taking ten texts as an example, and the text fragments in turn are 5-8-1-9-2-10-6-3-7-4.

Initialization of the Empire

In the period of the initialization of the Empire, random generation of K initial states, calculated the normalized cost value C_i of each country, the normalized cost value of each country is calculated by the maximum image edge gap value of all solutions minus image edge gap value of this current solution. That is C_i=max $\{c_n\}$-c_i; c_i is the cost value of each country, that is the image edge gap of the current solution. According to the national standardized cost value, all countries are arranged in descending order. Selecting the top Nimp countries as colonial countries, the number of colonies Ncol=Npop $-N$imp.

Then calculate the power of each colonial country. The number of colonies allocated by each colonial country is related to the power of the colonial power.

$$P_i = c_i / \sum_{i=1}^{N_{imp}} c_i \tag{8}$$

The number of colonies assigned to it NC_i=round$\{P_i \times N_{col}\}$, and Pi is the power value of the colonials' power. Round function is rounded to the integral function.

Colonial assimilation

Assimilation is a process of internal colonial empire gradually tends to colonial countries, this paper takes the following four steps to generate new assimilation colonies.

Step1: The probability of randomly generating between 0 and 1 in each urban coding position in the colony.

Step2: In the colonial location where all probability is greater than or equal to ρ, the colonial city code of that location is used as assimilate colony number.

Step3: For the location of the colony with smaller probability that ρ, if the city number at the location does not appear at the rest of the colonies, the city number of the colonies in this position is used as the number of the assimilation colonies. Step2 and Step3 are collectively referred to as the replacement process.

Step4: The process of reconstruction. For the city number that is not appearing in the assimilation colony, the increase distance to all possible locations in the assimilative colonial city sequence is calculated in turn, and the city number is inserted at the minimum increase distance. The increase distance N of inserting the city numbered i into the city code number a and b adjacent position is N=dis(a,i)+dis(i,b)-dis(a,b).

dis (a, i) represents the distance between a cities and i cities. After the assimilation process is completed, the original colony is replaced by the newly generated assimilation colony.

Step5: Calculate the power value of the new colony, if it is superior to the colonial country it belongs to, will replace the solution of the colonial country. The specific process is shown in Figure 2. ρ is taken 0.5.

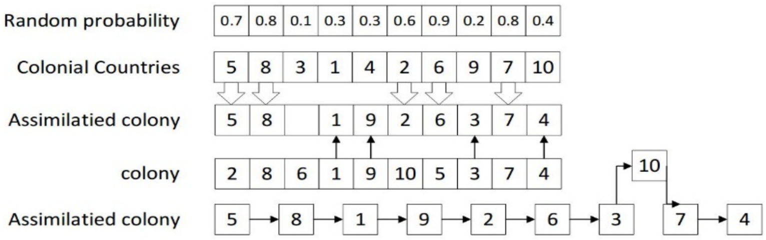


FIGURE 2. Assimilation process using ten text fragments as an example

Colonial revolution

In this paper, the colonial revolution was set up as exchange the number of two random locations. This can avoid invalidation of random search while giving full play to the mechanism of revolution to maintain the function of population diversity. Establishing a formula of adjustment from the distance between colonial and colonial countries to the probability of revolution $\mathrm{PT_i}$.

$$\mathrm{PT_i}=\eta\times\mathrm{dist(i)}/\sum_{i=1}^{\mathrm{NC_i}}\mathrm{dist(i)}+(1\text{-}\eta)\times\mathrm{rand}(0,1) \tag{9}$$

dist.(i) is the distance between colony and colonial countries. ci represents the cost of the i colony; ck represents the cost of the k colony; dist.(i) is defined as $\mathrm{dist(i)} = \mathrm{C_i} - C_k$.When the probability of revolution $\mathrm{PT_i}$ is greater than the given threshold μ, revolute the colony. According to the requirement of revolution, when the empire is far away from the colony, that is, when the solution of the colony is relatively poor, the probability of its revolution will increase. The specific process of the colonial revolution is as follows:

Step1: Random generation of α, β,And satisfy $\alpha, \beta \in (1, \mathrm{m} \times \mathrm{n})$.

Step2: Interchanging two coded numbers of α, β position in the colony, create a new colony

Step3: Calculating the power value of the new colony, if it is superior to the colonial country to which it belongs, will replace the solution of the colonial state.

Imperial competition and parameter selection in the algorithm

Imperial competition is the key step of algorithm convergence, which originated from the social behavior of colonial redistribution. Specific operation steps are shown in the reference [1], [4].

ICA has three important parameters: the total number of countries, the number of colonial countries, and the threshold of colonial revolution probability. A large number of case studies show that the algorithm performs best when the number of colonial countries accounts for 10%~20% of the state. Based on a large number of experiments, the best parameters of ICA are $\mathrm{N} = 100, \mathrm{N_{im}} = 15, \mu = 0.3$. When using these parameters, ICA has the best performance.

Algorithm Verification and Result Analysis

The characteristics of noise sensitivity, time consuming and effectiveness are very important in the problem of text stitching. In order to verify the characteristics of the algorithm, MATLAB programming is used to implement the above algorithm, and verified and compared by reference [3].

The examples of the application of the test picture in reference [3] are 10*15 and 10*12. The test is runs on the 4.0GB RAM 2.20GHz computer, programmed by Matlab2015b.

In order to analyze and compare the running effect of ICA algorithm on the problem of text fragment splicing, it runs independently 20 times. Calculate each run time and take 20 running time averages.

TABLE 1. The time-consuming comparison results

	Lingo [3]	**ACO [3]**	**ICA**
The Maximum time (S)	98.503	5.981	5.029
The minimum time (S)	93.617	5.268	4.422
The average time (S)	95.545	5.691	4.781

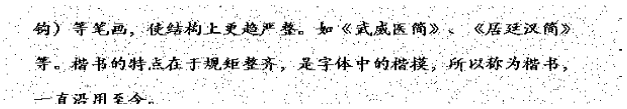
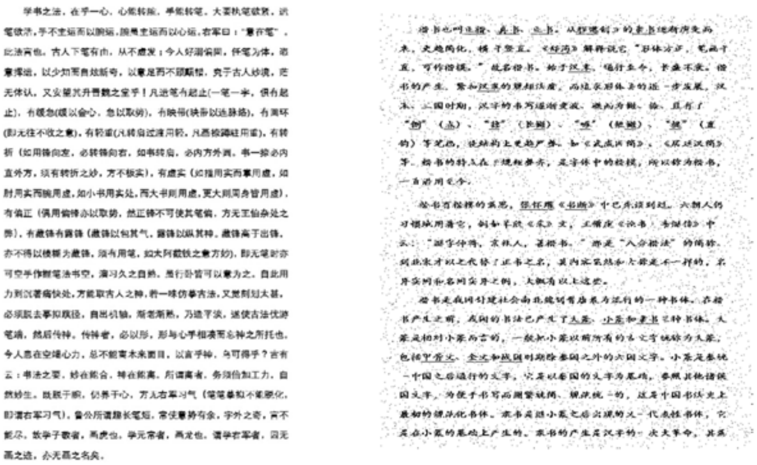

钧）等笔画，使结构上更趋严整。如《武威医简》、《居延汉简》
等。楷书的特点在于规矩整齐，是字体中的楷模，所以称为楷书，
一直沿用至今。

FIGURE 3. The row images after K-means clustering


学书之法，在乎一心，心能转腕，手能转笔，大要执笔欲紧，运笔欲活，手不主运而以腕运，腕虽主运而以心运。右军曰："意在笔"，此法言也。古人下笔有由，从不虚发；今人好溺偏固，任笔为体，恣意挥运，以少知而自炫新奇，以意足而不顾颠错，究于古人妙境，茫无体认，又安望其升晋魏之堂乎！凡运笔有起止(一笔一字，俱有起止)，有缓急(缓以会心，急以取势)，有映带(映带以连脉络)，有回环(即无往不收之意)，有轻重(凡转肩过渡用轻，凡画捺蹲驻用重)，有转折（如用锋向左，必转锋向右，如书转肩，必内方外圆，书一捺必内直外方，须有转折之妙，方不板实），有虚实（如指用实而掌用虚，如肘用实而腕用虚，如小书用实处，而大书则用虚，更大则周身皆用虚），有偏正（偶用偏锋亦以取势，然正锋不可使其笔偏，方无王伯杂处之弊），有藏锋有露锋（藏锋以包其气，露锋以纵其神，藏锋高于出锋，亦不得以模糊为藏锋，须有用笔，如太阿截铁之意方妙)，即无笔时亦可空手作握笔法书空，演习久之自熟，虽行卧皆可以意为之。自此用力到沉著痛快处，方能取古人之神，若一味仿摹古法，又觉刻划太甚，必须脱去摹拟蹊径，自出机轴，渐老渐熟，乃造平淡，遂使古法优游笔端，然后传神，传神者，必以形，形与心手相凑而忘神之所托也。今人患在空竭心力，总不能离本来面目，以言乎神，乌可得乎？古有云：书法之要，妙在能合，神在能离，所谓离者，务须倍加工力，自然妙生，既脱于腕，仍养于心，方无右军习气（笔笔摹拟不能脱化，即谓右军习气），鲁公所谓趋长笔短，常使意势有余，字外之奇，言不能尽，故学子敬者，画虎也，学元常者，画龙也，谓学右军者，因无画之迹，亦无画之名矣。

楷书也叫正楷、真书、正书。从程邈创立的隶书逐渐演变而来，更趋简化，横平竖直。《辞海》解释说它"形体方正，笔画平直，可作楷模。"故名楷书。始于汉末，通行至今，长盛不衰。楷书的产生，紧扣汉隶的规矩法度，而追求形体美的进一步发展，汉末、三国时期，汉字的书写逐渐变波、磔而为撇、捺，且有了"侧"（点）、"掠"（长撇）、"啄"（短撇）、"趯"（直钩）等笔画，使结构上更趋严整。如《武威医简》、《居延汉简》等。楷书的特点在于规矩整齐，是字体中的楷模，所以称为楷书，一直沿用至今。

楷书有楷模的意思，张怀瓘《书断》中已有提到过。六朝人仍习惯地用着它，例如羊欣《采》文，王僧虔《论书·书赋传》中云："韦字仲将，京兆人，善楷书。"都是"八分楷法"的简称。到北宋才以之代替了正书之名，其内容虽然和今天是不一样的，名异实同和名同实异之例，大概有以上这些。

楷书是我国封建社会南北朝到隋唐朝为流行的一种书体。在楷书产生之前，我国的书法已产生了大篆、小篆和隶书三种书体。大篆是相对小篆而言的，一般把小篆以前所有的古文字统称为大篆，包括甲骨文、金文和战国时期除秦国之外的六国文字。小篆是秦统一中国之后通行的文字，它是以秦国的文字为基础，参照其他诸侯国文字，为便于书写而删繁就简、推陈统一的，这是中国书法史上最初的规范化书体。隶书是继小篆之后出现的又一代表性书体，它是在小篆的基础上产生的。隶书的产生是汉字的一次大革命，其基

FIGURE 4. Assimilation process using ten text fragments as an example

Fig.3 is an image that is clustered into a row using the k-means method, but the method needs manual correction. Figure 4, Figure 5, successfully restored the image, verifying the effectiveness of the algorithm. The image with salt and pepper noise can also be restored successfully, which shows that the algorithm is robust. We can see that the algorithm is more efficient in comparison of timeliness.

CONCLUSION

In this paper, a new method based on improved discrete ICA algorithm is proposed to solve the problem of text fragments mosaicking. use two value method to extract of fragment image information. Clustering fragments into rows using k-means clustering. Using the matching degree of the edge information of each line fragment as a measure of whether or not the debris is adjacent or not, a restoration model with the best overall matching degree is established. The improved discrete ICA algorithm is used for encoding and solving. Experimental results show that the discrete ICA algorithm used in this paper is a practical and feasible algorithm to solve the text fragment restoration model. The algorithm has good stability, high efficiency and strong robustness.

REFERENCES

1. Nemati K, Shamsuddin S M, Kamarposhti M S. Using Imperial Competitive Algorithm for Solving Traveling Salesman Problem and Comparing the Efficiency of the Proposed Algorithm with Methods in Use[J]. Australian Journal of Basic & Applied Sciences, 2011, 5(10).
2. Gharavi H, Ardehali M M, Ghanbari-Tichi S. Imperial competitive algorithm optimization of fuzzy multi-objective design of a hybrid green power system with considerations for economics, reliability, and environmental emissions[J]. Renewable Energy, 2015, 78:427-437.
3. Zixiao Pan, Mei Wang, "A New Method of Shredded Paper Image Restoration Based on Ant Colony Algorithm," Chinese Automation Congress (CAC). Jinan: Academic, 2017, pp. 5526-5530.
4. Atashpaz-Gargari E, Lucas C. Imperialist competitive algorithm: An algorithm for optimization inspired by imperialistic competition[C]// Evolutionary Computation, 2007. CEC 2007. IEEE Congress on. IEEE, 2008:4661-4667.
5. Pan Z, Wang M. A New Method of Shredded Paper Image Stitching and Restoration[C]// International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration. IEEE Computer Society, 2017:55-58.