Research on Character Mosaic and Paper Scrap Recovery based on Simulated Annealing Algorithm

PAN DaSheng[1, a]

[1]Department of Physics and Telecom Engineering of Baise University, Baise 533000,China

[a] bspandsh@126.com

Keywords: Image Recovery; Character Mosaic; Paper Scrap; Simulated Annealing Algorithm.

Abstract. With the rapid development of machine learning and document analysis techniques, the research on the character mosaic and paper scrap recovery based on simulated annealing algorithm is urgent needed. In this paper, we approach is proposed based on pieces of text line features, form line semi-auto stitching method, mainly introduced the fragments of text line direction, form line direction acquisition algorithm, and the debris semi-auto stitching algorithm based on boundary intersection distance, pieces of the algorithm does not depend on the geometric features, implementation is simple and reliability is better. Also analyses the algorithm of computing workload, analysis shows that for the majority of less than 200 of the total number of pieces of actual scraps of paper, joining together the computing workload within the scope of the permit. The experiment simulation proves the effectiveness and feasibility of our proposed method.

Introduction

Conventional document scraps of paper pieces of computer stitching method generally used on the edge of the sharp point features, Angle, area features such as geometry, search and match of the adjacent scraps of paper and splicing, the splicing method based on boundary geometric feature does not apply to edge shape similar to a scrap of paper. The popular feature selection and analysis methods could be found in the literature reviews of [1-5]. But people were torn shredding, habit to save time, I always will overlap a scrap of paper, and then tear, then a scrap of paper overlap, continue to tear again, again and again like this, and until it is satisfied with the size a scrap of paper. This tear process will produce a lot of shape is very similar to a scrap of paper, if using fragments of boundary feature when stitching, stitching effect is not ideal. For this type of scraps of paper similar to the edge of the stitching, the ideal computer stitching process should be similar to artificial splicing process, i.e. splicing not only should consider to be the edge stitching a scrap of paper matches, but also judge debris inside the handwriting of bolt or debris inside the text content of matches, but due to the limitation of theory and technology, let computer has the identification within the pieces the handwriting on the edge of the disconnection, and understand the meaning of the text image intelligent almost impossible. But using the existing technology which can retrieve text pieces the same geometry characteristic information, such as the text of a line with high, text information such as line spacing, stitching together pieces such as the use of the information, the splicing efficiency is better than simply using the method of boundary geometry features. With powerful image processing tools, such as Photoshop, Microsoft paint, image editing, modifying, much easier. Digital image security appraisal in recent years become emerging in the field of information security and extremely important research topic, also is the key technology of image content security. Research institutions and scholars at home and abroad for the passive security appraisal of the research work focus on two aspects: on the one hand, is based on single feature in the tamper with the changes before and after testing the method advantage is don't need a photo gallery of training classifier, directly on the image authenticity verification, but has its limitations, the detection accuracy is low [6-8]. On the other hand is based on image feature elements more analysis of the detection, identification testing methods of the class is mainly by extracting the image to be detected a variety of statistical features and combination, the final decision is obtained by classifier to classify the method need gallery training classifier, increase the complexity of computation, but its detection accuracy is higher.

Because most of the text of the document text line parallel to the direction and single row direction and form, if fragments of text line or form on the edge of the fragment fracture, so with its adjacent a scrap of paper in edge must have the same height and the distance between the same text line or forms, this feature can be easily selected out from the shape similar to that of many fragments adjacent pieces. For text line or the height of the form line features, spacing of handwriting recognition than offline identification and the understanding of the text image a bit easier to implement and use of debris within the text line features or table splicing shape similar to a scrap of paper in theory is feasible. On the other hand, due to the lack of computer digital image analysis, let the computer to completely in the sense of debris automatic stitching is almost impossible, to ensure the accuracy of stitching, need to join in the process of joining together of artificial interference process. Generally splicing pieces when using a computer search first not to match the target fragments stitching pieces, and according to match in order to choose a fragment, the operator according to the human brain further analysis results give up or receive selected pieces. This semi-auto stitching method to comprehensive utilization of the computer high speed calculation ability and character image recognition and understanding ability, joining together the efficiency is higher than pure manual joining together the accuracy is better than pure computer stitching.

Therefore, we conduct research on character mosaic and paper scrap recovery based on simulated annealing algorithm. In this paper, we firstly introduce the basic concepts of simulated annealing algorithm (SAA) with theoretical analysis, and then the joint of it with paper recovery is conducted. The experimental result indicates the effectiveness of the proposed approach. Detailed discussion will be down in the following sections.

Simulated Annealing Algorithm (SAA) and the Modification

The General Overview. The automatic joint issue belonged to the computer vision and pattern recognition area, which was accomplished by the computer processing to obtain the information of the paper scrap, like shape and color. Then we analyze shredding automatic or semi-automatic to recover. At present, most of the shredding joint work done by hand. Although some researchers have done abroad, few research achievements can be found specific application background of scraps auto repair technology. Information entropy to measure digital image contains information digital image is composed of many pixels, different pixel combination show different information. From a statistical point of view, and can make use of image information entropy to represent the image pixel distribution. Joining together will change the image pixel distribution and therefore we can the information entropy to identify whether the image has been tampered with. In the formula 1, we give the mathematical definition of information entropy. The formula 2 provides the approximation.

$$H\left(f\right)=-\sum_{m=1}^{M}\sum_{n=1}^{N}p_{mm}lbp_{mm} \tag{1}$$

$$H\left(f\right)\approx-\sum_{m=1}^{M}\sum_{n=1}^{N}p_{mm}(p_{mm}-1) \tag{2}$$

Considering the edge of the matching condition, according to each strip ripped up the left and the right column of the image information, calculate it with other strips scraps left and the right column of the image correlation, based on the principle of maximum correlation, stitching together will each strip shredding, the image is to restore to the original text in the form of information.

Formulation of SAA. Standard SAA algorithm begins with a high temperature, the use of mutation probability characteristics of sampling strategy, in the solution space of random search, along with the temperature falling, repeat the sampling process, and finally get the global optimal solution of the problem. But the SA algorithm don't know much about the situation of the whole search space, can't make the search process into the most promising search area, easy to fall into local optimum. Algorithm is introduced into flexible storage structure and the corresponding rule of taboo search to avoid duplication. But at the same time there are the following two questions: (1) algorithm is easy to fall into local optimal solution; (2) the algorithm convergence speed is slow. There is much less than traditional SA and TS algorithm, the corresponding improved algorithms have been proposed, these methods are mainly generated by the new steps to add random disturbance or expand local search. But in the improved algorithm step length adjustment of control parameters in the absence of reliable basis, meet more complex functions, easy to fall into local optimum. In order to solve this problem, this paper puts forward the definition of the concept of function complexity. Function complexity refers to the function value of a function and the wave trough the intensity of change, with the function values across size and function to represent the peaks and troughs change frequency, the definition is in the formula 3.

$$\rho = \log_{10}\left(f - g \right) + \left(m + n \right) \tag{3}$$

SA and TS algorithm application in function optimization generally choose small step length control parameter and the purpose is to calculate the global optimal solution of more accurate. But at the same time it is easy to fall into local optimum or convergence in advance. By a large number of experiment and found the function itself complicated situation related parameters of algorithm, especially the selection of step length control parameters have great influence. On this basis, put forward based on the function complexity method for adaptive adjustment of step length control parameters, in order to avoid convergence algorithm falls into local optimum or ahead. According to the function complexity adaptive adjustment of step length control parameters can make the algorithm to jump out of local optimal solution, and avoid premature convergence, but with the increase of step length control parameters, get the function of the accuracy of the optimal value. The optimization expression is shown in the formula 4.

$$\sum_{p=1}^{P}\left(\frac{f_p\left(s\right) - I_p}{U_p - I_p} \right)^e + \sum_{q=1}^{Q}\sum_{k=1}^{K}\left(\frac{g_{qk} - I_{qk}}{U_{qk} - I_{qk}} \right)^e \quad s.t.\begin{cases} \alpha_k \le R_k \le \beta_k \\ Area\left(C_{kj}\right) \ge \theta_k \end{cases} \tag{4}$$

The Proposed Methodology

The General Feature Selection. Before joining together fragments of debris within the image binarization processing, generally using Sobel gradient operator or other gradient operator of debris image processing, to get a word boundary, thus obtain fragments within text line direction, text line features such as height, spacing, the gradient is greater than the given threshold point take red, or white. In order to improve the accuracy of the analysis, hypothesis has not flown line of text direction along the horizontal direction, the text of Chinese characters, with space between Chinese characters and Chinese characters, Chinese characters are the width and height ratio of 1/3 ~ 3. This means that each text image with other words there is gap between image and text image width and the height of the ratio between 1/3 ~ 3, if the pieces with words, English words should be split into image class Chinese character image, image segmentation is English words into the width and height of approximate class Chinese character image, if the same Chinese characters image by very closely, so that the industry of Chinese characters between red dot adjacent to each other, also need to remove too much red dot between Chinese character image. Due to the randomness fragments are placed, pieces of text line direction is usually arbitrary direction, to confirm the purpose of the text line direction is to place any direction of the fragment level adjust according to the text line direction, to facilitate the joining together of the computer. In general most, according to the text image number line height and minimum on the basis of the selected direction is the actual text line direction.

Registration method based on the characteristics of not using the image pixel values directly, but by the characteristics of the pixel image derived, and then on the basis of image features, the corresponding characteristics of image overlap areas to search matching, this kind of stitching algorithm has higher robustness and robustness. Registration method based on the characteristics of the two processes: feature extraction and feature matching. First we extract from two image grayscale change obvious features such as dot, line, area, form feature set. Then the two images corresponding centralized feature matching algorithm is used as much as possible there will be corresponding relationship of characteristics of selected. A series of image segmentation techniques have been used on feature extraction and edge detection. Space of the extracted features is closed boundary, boundary, cross the line, and other characteristics. Feature matching algorithms are: cross correlation, distance transformation, dynamic programming and structure matching, such as chain code related algorithm. The selector is defined as the formula 5.

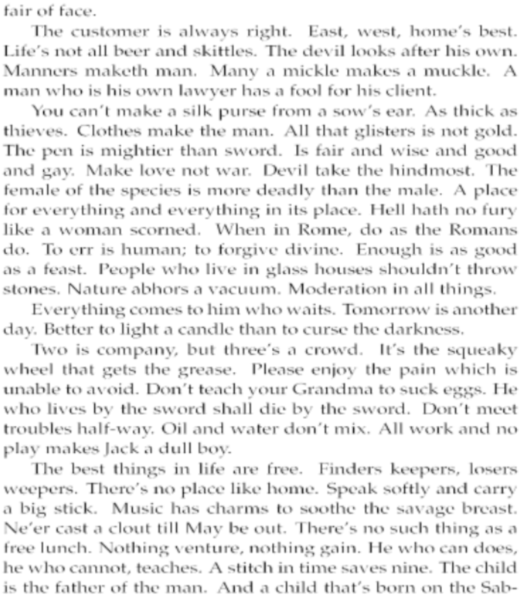
$$f(s) = \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ik} x_{ik} a_i \tag{5}$$

The Detailed Steps. Debris within the text image after the pretreatment, to adopt the following scan algorithm can obtain fragments of text line direction. (1) To cut pieces within a certain point as the origin of coordinates, horizontal direction as the X axis direction and vertical direction as Y-axis direction, in isometric within 60 direction. (2) the direction of each selected as the new coordinate system for the X axis, and the original pixel in the integer coordinates, under the new coordinate system, if the new coordinates after coordinate transformation is not the integer, then according to the value of 4 and 5 into the method of point in the new system change the coordinate values, only the color properties remain unchanged. (3) the new coordinates a scrap of paper on the edge of the highs and lows coordinates, starting from the lowest to each row (with the same Y coordinates) number starting from 1, calculate the number of white, red dot number and width of each line (the number of pixels). (4) according to the number of each row of red dot and white dot number and debris width direction of the total number of character image as well as text line high sum. (5) The total number of words in each direction image, text line height sum according to the order from big into small. (6) The number of selected text images, text lines most high and the smallest in that direction as fragments of text line direction. Sure good pieces of text line direction or table line direction, line the pieces according to the text line direction or form as a horizontal direction reposition the debris and correct direction of debris within the text direction may be up, also may be down. With the direction of the text together scraps, stitching is bordered by the debris around the boundary of two pieces of different text direction splicing, fragments of joining together either as the left border, either as the right boundary. Need special attention is, if the text line or line form is missing, even alignment, good location, a fragment one form text line or in a separate pieces on line or text line, can't find the corresponding form when calculating distance equal to the number of consecutive points should be missing the form line and matching form text line or ignored. Also because of the influence of the calculation error, the position of the two corresponding point height should be equal in theory, but in fact, the position of the two heights will have deviation, also continuously match although in theory, the distance between points are equal, but in fact is not absolutely equal, but there is a certain deviation. If adjacent two pieces are on both sides of text line or table line, then the alignment location and corresponding text line or line form and on both sides of the distance between the boundary nodes value equal to the number of consecutive points up, together. So the actual splicing pieces, can be calculated first target fragments connect with positive and negative all not stitching pieces in all possible stitching location distance equal to the number of consecutive points, and then the and descending order, and at the top pieces with the target of the adjacent the possibility of the highest. In many situations, as two paper scraps were both considered as the suitable one to be jointed with the objective paper scrap by the computer, the computer would automatically chose one of them which was considered the better to joint without the logical analysis. At this time, it was easy to produce a large-scale and extensive joint error. Hence the artificial interference was of significance as the solution error of the computer happened.

Experiment and Simulation Result

Initialization and Experiment Set Up. From the basic knowledge of the image processing, it was known that the pixel value range of each pixel point on the paper scraps to be jointed was 0-255, according to which, the figure could be analyzed. The figure was firstly broken up into the point collection, and the pixel matrix of the point collection of the figure was established for a random figure. The formula 6 shows the corresponding expression. The figure 1 shows the sample image.

$$P_{ij} = \begin{bmatrix} p_{1ij} \\ p_{2ij} \\ \vdots \\ p_{hij} \end{bmatrix}, 0 \le i \le 72, 0 \le j \le 19 \tag{6}$$



fair of face.

The customer is always right. East, west, home's best. Life's not all beer and skittles. The devil looks after his own. Manners maketh man. Many a mickle makes a muckle. A man who is his own lawyer has a fool for his client.

You can't make a silk purse from a sow's ear. As thick as thieves. Clothes make the man. All that glisters is not gold. The pen is mightier than sword. Is fair and wise and good and gay. Make love not war. Devil take the hindmost. The female of the species is more deadly than the male. A place for everything and everything in its place. Hell hath no fury like a woman scorned. When in Rome, do as the Romans do. To err is human; to forgive divine. Enough is as good as a feast. People who live in glass houses shouldn't throw stones. Nature abhors a vacuum. Moderation in all things.

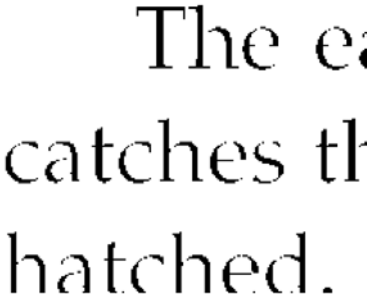
Everything comes to him who waits. Tomorrow is another day. Better to light a candle than to curse the darkness.

Two is company, but three's a crowd. It's the squeaky wheel that gets the grease. Please enjoy the pain which is unable to avoid. Don't teach your Grandma to suck eggs. He who lives by the sword shall die by the sword. Don't meet troubles half-way. Oil and water don't mix. All work and no play makes Jack a dull boy.

The best things in life are free. Finders keepers, losers weepers. There's no place like home. Speak softly and carry a big stick. Music has charms to soothe the savage breast. Ne'er cast a clout till May be out. There's no such thing as a free lunch. Nothing venture, nothing gain. He who can does, he who cannot, teaches. A stitch in time saves nine. The child is the father of the man. And a child that's born on the Sab-

Figure 1.Sample Image for the Experiment

Experiment Result and Analysis. As shown in Figure 2 and 3, obviously the two figures could be totally matched with the objective figure, and in this situation, if the wrong figure was picked by the computer (as shown in Figure 2), the other paper scrap would be jointed with other paper scraps wrongly either, which would result in the domino effect. Meanwhile, if the artificial interference was used with the logical thinking and analysis of human, it could be realized that the word "catens" in the figure to be jointed in Figure 3 didn't actually exist, hence the joint error would be found and corrected, which would lead to a high accuracy and right restoration of the literatures.

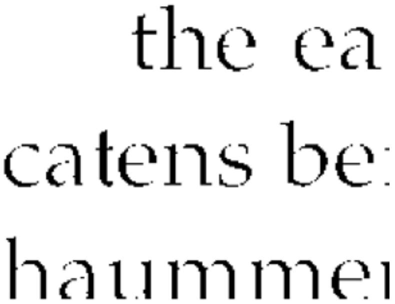


The e

catches tl

hatched.

Figure 2.The Result One



the ea

catens be

haummei

Figure 3. The Result Three

Hence, the artificial interference needed to be added after the initial joint was obtained, and the interference way was: the joint couple of the paper scraps in the running result of the program. With the direction of the text together scraps, stitching is bordered by the debris around the boundary of two pieces of different text direction splicing, fragments of joining together either as the left border, either as the right boundary. Based on the above text direction line, form the characteristics of the debris semi-auto stitching algorithm is relatively rough, but easily into a detailed algorithm is closely related to the computer language, due to space limitations, no longer discuss the practical implementation of the algorithm. The stitching algorithm there are three key steps: (1) the debris boundary for; (2) pieces within the text features, form line access; (3) target fragments and splice pieces in May not stitching location distance equal to the number of consecutive points calculation and sorting. We could see from the corresponding simulation result that our method is robust and accurate.

Conclusion and Summary

The technology of joint and recovery of paper scrap are considered in shape before. In this paper, we conduct research on the character mosaic and paper scrap recovery based on simulated annealing algorithm. It is proposed based on pieces of text line features, form line semi-auto stitching method, mainly introduced the fragments of text line direction, form line direction acquisition algorithm, and the debris semi-auto stitching algorithm based on boundary intersection distance, pieces of the algorithm does not depend on the geometric features, implementation is simple and reliability is better. Also analyses the algorithm of computing workload, analysis shows that for the majority of less than 200 of the total number of pieces of actual scraps of paper, joining together the computing workload within the scope of the permit, if make some improvement of algorithm, joining together can greatly reduce the amount of calculation. The scraps of semi-automatic edge stitching method is suitable for any size, any scraps of paper splicing in the shape of a document. Stitching algorithm is proposed based on the stitching program is developed, and the stitching test is carried out on a real scraps of paper. The experiment shows that our method is efficient.

References

[1] Wen-wen, Y. A. N. G., T. A. O. Jia-qi, Z. H. E. N. G. Lu-tong, S. U. N. Guo-wei, and M. A. I. A-li. "Algorithm for Reassembly of Longitudinally and Transversely Cutting Chinese Character Fragments on a Single-sided Paper." Journal of Yuncheng University 5 (2013): 005.

[2] Dai, Li, Yousai Zhang, Yuanjiang Li, and Haoxiang Wang. "MMW and THz images denoising based on adaptive CBM3D." In Sixth International Conference on Digital Image Processing, pp. 915906-915906. International Society for Optics and Photonics, 2014.

[3] Zheng, Shuai, Ming-Ming Cheng, Jonathan Warrell, Paul Sturgess, Vibhav Vineet, Carsten Rother, and Philip HS Torr. "Dense semantic image segmentation with objects and attributes." In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pp. 3214-3221. IEEE.

[4] Ion, Adrian, João Carreira, and Cristian Sminchisescu. "Probabilistic joint image segmentation and labeling by figure-ground composition." International journal of computer vision 107.1 (2014).

[5] Zhang, Wentai, et al. "FPGA Acceleration for Simultaneous Image Reconstruction and Segmentation based on the Mumford-Shah Regularization." Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. ACM, 2015.

[6] Zorin, D. A. "Estimating the convergence of a simulated annealing algorithm for the problem of constructing multiprocessor schedules." Moscow University Computational Mathematics and Cybernetics 38.2 (2014): 83-90.

[7] Jing, Yiming, et al. "Application of improved simulated annealing optimization algorithms in hardware/software partitioning of the reconfigurable system-on-chip." Parallel Computational Fluid Dynamics. Springer Berlin Heidelberg, 2014.

[8] Plasencia, Manuel, et al. "Geothermal model calibration using a global minimization algorithm based on finding saddle points and minima of the objective function." Computers & Geosciences 65 (2014): 110-117.