Multimedia Tools and Applications (2023) 82:19441–19461
https://doi.org/10.1007/s11042-022-13835-7

Clustering paper shreds of different sizes

Alia Madain[1]

Received: 16 March 2021 / Revised: 31 March 2022 / Accepted: 6 September 2022 /
Published online: 28 November 2022

Abstract
Although paper shredding is widely used to prevent confidential papers from being misused, still it cannot be considered a convenient process. The size of paper shreds became smaller and smaller, as new methods of shredded paper reassembly and reconstruction are evolving. This paper focuses on clustering, which is a possible phase in the assembly process. This work considers real strip-cut shreds, in addition to images shredded by a simulator in one direction to make strip-cut shreds of different sizes, from wide to narrow shreds, and images shredded in two directions, possibly reflecting cross-cut and micro-cut shreds. K-means is used to cluster shreds, the features tested are gray-level ranges, and the well-known gray-level co-occurrence matrix, invariant moments, segmentation-based fractal texture analysis algorithm, and color moments. The number of shreds grouped in the same cluster with originally adjacent neighbors is used to indicate clustering effectiveness, in addition to the overall accuracy of strip-cut shreds clustering. When the number of clusters is 5, and the k-means experiments run 100 times for 38 images, the overall accuracy of gray-level ranges in simulated strip-cut shreds is 84.87, 89.27, and 93.5 percent in the three different sizes tested, also in cross-cut and micro-cut shreds, gray-level ranges achieve a relatively high number of shreds with 3 and 4 originally adjacent neighbors found in the same cluster.

Keywords Clustering · Shredded paper reconstruction · Strip-cut shreds · Cross-cut shreds · Micro-cut shreds · GLCM · Invariant moment · SFTA algorithm · Color moments · K-means

1 Introduction

Almost all types of businesses, small, medium, or large use paper shredders. Many homes own paper shredders as well. Paper shredders are mechanical machines used to cut papers into strips or smaller particles. The main purpose of using shredders is to protect information from unauthorized access. Access to information is significantly harder when the paper

✉ Alia Madain
asmadain@just.edu.jo

[1] The Department of Computer Science, Jordan University of Science and Technology, Irbid 22110, Jordan

is shredded into small pieces, and it is usually considered impractical to reconstruct the document.

Securely damaging confidential documents is not easy, paper shredders are known of being noisy, require frequent cleaning, and do not shred big bulks of paper at the same time, they might require time to cool down, and the more secure the machine, the more expensive it is.

As in the case of any product, the choice of which paper shredder to buy depends on the user's requirements. Focusing on the main purpose of shredding machines, which is keeping documents secure and confidential, and if we assume that the user requires the highest level of security and the budget allows, then the choice will be the machine that produces the smallest size of paper particles or shreds.

Many shredders are available at the market and can be classified into strip-cut, cross-cut and micro-cut shredders. Shredders can also differ in the number of papers the machine is capable of shredding at a time and the capability to shred other media such as CDs, DVDs, and credit cards. The main difference between cross-cut shredders and micro-cut shredders is the size of the shreds that results from the shredding process. This feature is quite important as the size of the shred determines the machine's level of security, which is the main purpose of buying it.

Some standards specify the levels of security and sizes of shredded paper particles. DIN 66399 (Deutsches Institut für Normung), NSA/CSS Specification 02-01 for High-Security Crosscut Paper Shredders, and ISO/IEC 21964 Information technology — Destruction of data carriers are a few examples. The classification into strip-cut, cross-cut and micro-cut gives a broad range of possibilities, so it is important to investigate the actual size of strips or paper particles produced by the machine to be able to map the product to a certain security level. Reassembling or reconstructing shredded paper is supposed to be a hard problem for humans and machines for shredders to be effective. Nonetheless, the main reason for manufacturing machines that cut smaller and smaller particles is the fear of the possibility of reassembling the documents.

This research tests clustering paper shreds, which is a possible phase in assembling shredded papers. As in the case of solving a puzzle, the ability to arrange the pieces of the puzzle into meaningful groups lowers the number of possibilities and makes finding the right position of any piece an easier task. The effectiveness of grouping depends on the actual number of originally adjacent neighbors found in the same group and the possibility to find more groups within the group.

In this paper, clustering is applied in multiple experiments, where different shred sizes are considered. Shredded paper in one direction, vertical or horizontal, also called strip-cut shreds can be wide or narrow. This paper reports the results of using gray-level ranges, Gray-Level Co-occurrence Matrix (GLCM), Hu moments, Segmentation-based Fractal Texture Analysis algorithm (SFTA), and color moments as feature vectors to cluster shreds of different sizes. The use of k-means clustering over each feature vector separately shows that the effectiveness of grouping the shreds into clusters differs between wide and narrow shreds. In all cases, the use of gray-level ranges to group shreds achieved the best overall accuracy.

Shredded paper in vertical and horizontal directions may have different sizes as well; the shreds can be relatively big or quite small. The experimental results give the average number of shreds with one, two, three, four, or no originally adjacent shreds in the same cluster. As with the shredded paper in one direction, the features used in clustering are gray-level ranges, GLCM, HU moments, SFTA algorithm, and color moments. Finally, the gray-level ranges are used to cluster instances from a real data set of strip-cut shredded papers.

This paper is organized as follows: Section 2 gives a brief literature review. Section 3 describes the methods deployed in this research for extracting features. Sections 4 and 5 provides details on clustering and the evaluation of the clustering effectiveness. Section 6 summarizes the data set, results and discussion. Section 7 gives a descriptive comparison of available literature. Finally, Section 8 concludes the work done and provides direction for future research.

2 Literature review

The literature on reconstructing or assembling paper shreds can be organized based on the following factors: damage type, shred size, document type, document content, degree of human involvement in assembling shreds, the existence of missing pieces, number of papers considered, the use of real shredded papers vs. simulation, pre or post-processing requirement, and protocol, method, or algorithm used.

The number of shreds and the size of each shred are the main determinants of the time required for reconstruction. According to [27], there are 3 main categories, namely, strip-cut, cross-cut and others (hand shredded or torn). In this section, the literature on shredded paper reconstruction is divided into strip-cut, cross-cut, and hand torn. Then the literature on clustering and the use of clustering in reconstruction is given.

Some of the techniques used in reconstructing strip-cut shreds are linear scoring [28], image-based techniques [18, 19], MPEG-7 standard descriptors [35], Hungarians algorithm [1], compatibility functions [23–26], and word-path metric and greedy composition optimal matching solver [17].

Some techniques of the strip-cut reconstruction are applied to different types of content such as handwritten documents [33] and documents written in certain languages [38, 39].

Cross-cut shreds reconstruction is considered more challenging than the reconstruction of strip-cut shreds, some strategies deployed are: semi-automatic assembly where feature extraction and matching are done before a human user continues the reconstruction [9] dual combination and divide-and-conquer strategies [6], constrained seed K-means algorithm and ant colony algorithm [7], visual analytic approach [4], lineation algorithm and k-means [5], memetic algorithm [31, 36], similarities of stroke and typesetting features [37, 41], and information quantity [42].

Additionally, the reconstruction of hand-torn pages was studied and different approaches were used such as feature matching [16], contour maps [3], and reconstruction using a matching graph and a spanning tree [20].

The focus of this study is the grouping of paper shreds as a step in the reconstruction process of shredded papers. Clustering was studied alone in clustering strip-cut Chinese homologous pieces [40] and strip-cut shredded documents clustering followed by human reconstruction [34].

The use of clustering in reconstructing shredded papers was also studied in the literature before, for example, clustering as preprocessing [2], row clustering [11], and the use of clustering as a part of the reconstruction process itself [32].

Classifying and reconstructing damaged or shredded papers in the literature focus on hand-torn, strip-cut, or cross-cut shreds. To my knowledge there are no published papers on reconstructing micro-cut shreds, this may be because it is unexpected to see recognizable objects in the shreds, it is very challenging for humans to reconstruct the papers and the amount of time and effort required to collect and scan the shreds is huge using the current technology.

Theoretically, the sizes of the shreds used in the shredding simulation tested may reflect the three classes of shredders, strip-cut, cross-cut, and micro-cut. The features used to describe the shreds are the image gray levels arranged in ranges, invariant moments, GLCM, SFTA, and color moments. After representing each shred by a vector, k-means analysis is used to cluster the shreds in groups.

3 Features extraction

To cluster the shreds, it is required to decide which parameters identify and describe a paper shred and which method is used for clustering. Each shredded paper particle is represented in a vector or an array of parameters. The features used are the gray-level ranges, GLCM, invariant moments, SFTA, and color moments. Every set of parameters describes certain aspects of the shred. Then the feature vectors become the input of the k-means clustering approach.

3.1 Gray-level co-occurrence matrix (GLCM)

GLCM or Haralick descriptors [12] are used in many fields and found effective in different applications, for example, GLCM was proposed as an effective method to extract and analyze information regarding human skin texture, which can potentially be applied in evaluating effects of medical and cosmetic treatments [22].

In this subsection, the following definitions are used: Let G be a matrix whose element g_{ij} is found by calculating how often a pixel with the intensity (gray-level) value i occurs in a specific spatial relationship to a pixel with the intensity value j. Note that intensity values i and j are greater than or equal to 1 and less than or equal to the number of possible intensity levels. G is the co-occurrence matrix. Also, let the probability p_{ij} be the ij-th element of G/ng, where ng is equal to the sum of the elements of G, and let k be the row (or column) dimension of the square matrix G. A normalized G is formed by dividing each of its terms by ng [10].

Three statistical properties derived from the co-occurrence matrix are used to describe the shredded paper particle, the first one is the contrast, also known as the variance or inertia. The contrast is given in (1). It returns a measure of the intensity contrast between a pixel and its neighbor over the whole image.

$$contrast = \sum_{i=1}^{k}\sum_{j=1}^{k}(i-j)^2 p_{ij} \tag{1}$$

where the contrast cannot be a negative number and ranges from 0 to $(size(G,1)-1)^2$, note that the contrast is equal to zero for a constant image.

The second property used is energy, also known as uniformity, which returns the sum of squared elements in (G). The energy is given by (2).

$$energy = \sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}^2 \tag{2}$$

The output range is [0 1], where the energy is 1 for a constant image.

The third and final property used is homogeneity, which returns a value that measures the closeness of the distribution of elements in (G) to the diagonal of (G). The homogeneity equation is as follows:

$$homogeneity = \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{p_{ij}}{1+|i-j|} \tag{3}$$

The output range is [0 1], where homogeneity is 1 for a diagonal (G). GLCM is considered expensive to compute, the work in [13] and [21] proposed faster and more efficient implementations of GLCM. If n x m corresponds to the dimensions of the image, then the calculation of the co-occurrence matrix requires reading the n x m gray-level values. In our application the shred sizes are small, for each problem, the number of pixels processed is the same as the whole original image, and only 3 features are used.

3.2 Invariant moments

A set of 2-D moment invariants that are insensitive to translation, scale change, mirroring (to within a minus sign), and rotation could be flexible enough to learn almost any set of patterns. In this subsection, we limit the equations to the first two moment invariants as they are the only ones used in the experiments. The following description and equations of the moments come from [10]. Further discussions on the invariant moments' concepts and derivations could be found in [15].

The 2-D moment of order (p+q) of a digital image f(x,y) of size $m \times n$ is defined as:

$$m_{pq} = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} x^p y^q f(x,y) \tag{4}$$

where p = 0 , 1 , 2 , . . . and q = 0 , 1, 2 , . . . are integers. The corresponding central moment of order (p + q) is defined as:

$$\mu_{pq} = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1}(x-\overline{x})^p(y-\overline{y})^q f(x,y) \tag{5}$$

where

$$\overline{x} = \frac{m_{10}}{m_{00}} \quad and \quad \overline{y} = \frac{m_{01}}{m_{00}} \tag{6}$$

The central moment of order (p + q) after normalization, denoted η_{pq}, is defined as:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \tag{7}$$

where

$$\gamma = \frac{p+q}{2} + 1 \tag{8}$$

for p + q = 2, 3, ... The first two moment invariants are:

$$\phi_1 = \eta_{20} + \eta_{02} \tag{9}$$

and

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{10}$$

The time required for calculating moments increases with the moment order [14]. We use the first two moments. The time required for each shred is based on the dimension of the shred.

3.3 Gray-level ranges

This parameter reflects the number of occurrences of each gray-level range in the shredded paper particle. This set of parameters requires 16 numbers. Each number refers to 16 gray levels or a range of gray levels, where the 256 gray levels are divided into 16 ranges as shown in Table 1.

The first number in the vector is the probability of gray levels in the range [1-16], the second parameter represents the range [17-32], and the last parameter, or the 16th parameter represents the range [241-256]. The time required to calculate the gray-level ranges for each shred is based on the dimensions of the shred.

3.4 Segmentation-based fractal texture analysis (SFTA)

SFTA is an algorithm used to describe segmented texture patterns given by [8]. The method computes the fractal dimensions of regions border from a set of binary images extracted from the original grayscale image. The algorithm receives a grayscale image and the number of thresholds as input and produces a feature vector as output.

At first, the algorithm deals with converting the input image to corresponding binary images, the algorithm utilizes the multi-level Otsu algorithm given in [29] to compute the set of thresholds. The region boundaries are calculated from binary images pixel by pixel, where a given pixel is considered a boundary pixel when it is white and any of its eight neighbors is black. The resulting feature vector is a concatenation of the binary images' size, mean gray level, and boundaries' fractal dimension.

3.5 Color moments

Color moments are used to describe color distribution over an image. Moments can be computed for all color models, and they cover mean, standard deviation, skewness, and

Table 1 Ranges of Gray-Level Values

Parameters/ Gray-level	Range Beginning	Range End
1	1	16
2	17	32
3	33	48
4	49	64
5	65	80
6	81	96
7	97	112
8	113	128
9	129	144
10	145	160
11	161	176
12	177	192
13	193	208
14	209	224
15	225	240
16	241	256

kurtosis, in addition to higher-order color moments. In the experiments, the input image is an RGB image, where the three channels are processed separately to compute the mean and the standard deviation and the results are combined in a feature vector. The moments are rotation invariant and will result in the same feature vector even when shreds are not in the right orientation. The time required for each shred is based on the dimensions of the shred.

4 Clustering

k-means clustering, an unsupervised method of clustering, was used. The k-means function used divides the n-by-p data matrix containing the feature vectors of all paper particles or shreds (n), into k clusters, and returns an n-by-1 vector containing cluster indices of each shredded paper particle. Where p is the feature vector, while n is the total number of particles. The function depends on the squared Euclidean distance metric and the k-means++ algorithm for cluster center initialization.

5 Clustering evaluation

The assessment of the real shredded paper assembly is not straightforward for documents shredded in two directions. The work in [30] studied cross-cut real shredded papers data set. The work proposes the use of precision, recall, and the number of edit operations necessary to convert the assessed assembly to a perfect one.

This paper uses two evaluation methods, both of which depend on the fact that the information regarding the position of each shred in the original media before shredding is available and it is possible to check which shreds are in the same cluster as their originally adjacent neighbors.

The first evaluation method is used in all experiments, the clustering is evaluated by the number of shreds found in a cluster with one, two, three, four, or no originally adjacent shreds. The second evaluation method is used in the clustering of strip-cut shredded papers, where the shredding is done in one direction (vertical or horizontal). The experiments of strip-cut shredding are done using a computer-based simulation and real shredded paper.

Calculating the overall accuracy of the clustering requires the number of originally adjacent neighbors found in the same cluster (d). If the edges are counted twice the number is divided by 2. The accuracy also requires the possible number of adjacent neighbors to be found in the same cluster as this is determined by the fact that originally not all shreds do have adjacent neighbors (those on the borders and the corners of the original paper) and the clustering itself lowers the possible number of adjacent neighbors found in the same cluster.

To find the number of possible adjacent neighbors (m), we can assume two neighbors for each shred in strip-cut shredding ($2 \times n$), where n is the number of shreds, afterwards, the outer borders of the first and the last shred are subtracted ($2 \times n$) $- 2$, and finally, the result is divided by 2, so that the same border is not counted twice. This is equivalent to the number of shreds minus one (n-1).

Still grouping the shreds into clusters will lower the possible maximum number of adjacent neighbors. So (c-1) will be subtracted from the maximum, where c is the number of clusters. So, the number of possible adjacent neighbors is given by:

$$m = (n - 1) - (c - 1) \tag{11}$$

The accuracy is a percentage of shreds that are adjacent in the original paper and found in the same cluster to the maximum possible number of correct clustering.

$$accuracy = \frac{d}{m} \times 100 \tag{12}$$

6 Results and discussion

The experiments can be divided based on the number of directions shredding is made, so shredding in one direction is given in Section 6.2 and results when the images are shredded in two directions is given in Section 6.3. The data sets used in the experiments are presented in Section 6.1.

In all experiments, Matlab R2016a was used. The machine's operating system is 64-bit, Windows 10. The processor is Intel Core i7-8550U with 12 GB RAM. The images were converted whenever the features required a different format. For example, the RGB images are converted to have a bit depth of 8 bits when the gray levels are counted and saved into a vector representing the ranges.

The experiments were made for every feature separately. Notice that the lengths of the vectors describing the paper shreds vary as the length depends on the number of parameters or the features chosen. The implementation of invariant moments from Gonzalez [10] was used after normalization, and k-means and GLCM built-in functions were used.

6.1 Data sets

Two data sets were used in the experiments, the first one consists of 38 RGB images, used to make ideal shreds done by a shredder simulator. Samples from the first data set content are given in Fig. 1. The second data set contains 3 samples of real paper shreds, the content shredded is simply images printed from the first data set.

The whole image orientation can be portrait or landscape if it is part of a document. Also, the possible shredding orientation of an image depends on the paper size and the shredder size.

All ideal shreds and real paper shreds are in their right orientation. Most feature extraction methods used in ideal shreds experiments are invariant to rotation and the one used in real papers is invariant as well.

6.1.1 Computer generated shreds (ideal shreds)

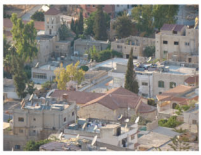
It is more convenient and practical to use a computer-shredded image. The shreds are made by a computer program and not actual scanned paper, where the scanned papers may have borders and possible white borders at each paper shred that may affect the result. Also, it is assumed that the images do not contain white borders or blank shreds. The 38 images of the first data set have the same width of 864 pixels and the same height of 648 pixels. The ideal shreds made by shredding the images in one dimension, vertically or horizontally, are made in three sizes, namely, wide, medium, and narrow fragments. The ideal shreds made by shredding the images in two directions, horizontally and vertically, are made in two sizes. All ideal shreds are identical in terms of size except the shreds of the last row/column. Starting from the left, the image is shredded to particles of the specified size, the last column and/or the last row in the image might be shredded to parts smaller than the specified size.


(a) Jordan University-1


(b) Jordan University-2

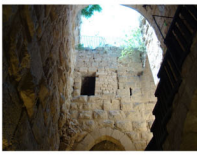

(c) Jordan University-3


(d) Fuheis-1


(e) Fuheis-2


(f) Umm Qais-1


(g) Umm Qais-2


(h) Ajloun Castle


(i) Gerash-1


(j) Gerash-2


(k) Al-Kharanah-1


(l) Al-Kharanah-2


(m) Al-Kharanah-3


(n) Al-Azraq-1


(o) Al-Azraq-2


(p) Al-Azraq-3

Fig. 1 Samples from the Data set

6.1.2 Real paper shreds

Three images from the first data set were printed, shredded, and scanned. The images were printed on A4 papers so that each image is part of the page and does not cover the whole page. The shredding was done using Rexel Alpha strip-cut shredder labeled as P1 security. The papers were shredded into 27 shreds. A mask was built for each shred using edge detection techniques.

6.2 Shredding in one direction (vertical or horizontal cuts)

Shredding paper in one direction or strip-cut shredding may refer to cutting paper into wide shreds or thin shreds depending on the shredder. Three sizes are used in clustering experiments to reflect the differences when wider or thinner shreds are assembled. If one unfolded paper is to be shredded using a strip-cut machine, the direction in which the shredding is done may depend on the size of the paper, if the paper is small enough both directions are possible. In the following experiments, the same direction is assumed for 38 images.

The appropriate number of clusters in shredding in one direction can be determined partially by the number of paper shreds available to be reassembled. Knowing the shredder information and the size of the paper, A4 for example, indicates an approximation of the total number of shreds from a single paper. The number of clusters should be reasonably less than the number of shreds to make the clustering meaningful. The determination of the best number of clusters should consider a more efficient merging process as well.

The first step is to shred the paper, then to get the features vector of each shredded paper particle, and to cluster the vectors into groups, and finally, the accuracy of clustering is calculated as a percentage, as given in equation (12).

6.2.1 Wide fragments - real paper data set

The clustering was applied to a real data set, Table 2 shows the results. The gray-level vector was found for each shred and k-means clustering was applied 100 and 10K times to group shreds in 5 clusters. The table shows the average number of shreds where no neighbors were found, the average number where exactly one neighbor was found, and finally, the average number of shreds with two neighbors found. The overall accuracy ranges from 76.13 to 88.72 percent.

6.2.2 Strip-cut wide fragments

As the images are of the same size, the total number of wide shreds is 26 in each image of the 38 tested images. Table 3 shows the average of shredding 38 images grouped into 5 clusters. As running the k-means results vary based on the initial random point chosen, all the results given in the table show the average of running k-means 100 and 10K times. The averages reported in the table of running k-means 100 or 10K times are quite close with minor differences.

If it is assumed that each paper was shredded into 26 wide shreds, the use of gray-level features to cluster the shreds into 10 clusters will result in an average overall accuracy of 80.38 percent for 100 tests.

The time to find the features vector for all 26 shreds of a single image is as follows: gray-level ranges: 0.15 seconds, GLCM: 0.084 seconds, all HU moments: 0.14 seconds, SFTA: 16.07 seconds, and color moments: 0.07 seconds. The use of k-means in one test takes time between 0.002 and 0.008 seconds.

Table 2 Overall accuracy and average number of neighbors found in the real paper data set

Images/ results	Average number of shreds without neighbors		Average number of shreds with exactly one neighbor		Average number of shreds with 2 neighbors		Overall Accuracy	
Number of tests	100	10K	100	10K	100	10K	100	10K
Image 1	1.83	1.74	11.56	11.49	13.61	13.77	88.14	88.72
Image 2	3.51	3.51	13.46	13.49	10.03	10.01	76.18	76.13
Image 3	2.11	2.17	12.82	12.71	12.07	12.12	84	83.97

Table 3 Overall accuracy and average number of neighbors found in strip-cut wide fragments

Features/ results	Average number of shreds without neighbors		Average number of shreds with exactly one neighbor		Average number of shreds with 2 neighbors		Overall Accuracy	
Number of tests	100	10K	100	10K	100	10K	100	10K
Gray-level ranges	2.60	2.62	11.15	11.09	12.25	12.30	84.87	84.96
GLCM	6.47	6.42	12.10	12.10	7.43	7.48	64.18	64.42
Invariant moments	5.87	5.87	11.02	11.04	9.10	9.09	69.59	69.59
SFTA	6.07	6.07	11.41	11.41	8.52	8.52	67.73	67.75
Color moments	3.94	3.96	12.08	12.12	9.98	9.93	76.30	76.11

Table 4 Overall accuracy and average number of neighbors found in strip-cut medium fragments

Features/ results	Average number of shreds without neighbors		Average number of shreds with exactly one neighbor		Average number of shreds with 2 neighbors		Overall Accuracy	
Number of tests	100	10K	100	10K	100	10K	100	10K
Gray-level ranges	2.14	2.15	15.59	15.57	33.27	33.28	89.27	89.28
GLCM	9.95	9.94	21.03	20.92	20.02	20.14	66.38	66.52
Invariant moments	9.21	9.23	19.81	19.90	21.98	21.87	69.31	69.17
SFTA	7.68	7.62	20.55	20.54	22.77	22.84	71.84	71.98
Color moments	4.93	4.94	17.99	17.99	28.08	28.08	80.59	80.58

Table 5 Overall accuracy and average number of neighbors found in strip-cut narrow fragments

Features/ results	Average number of shreds without neighbors		Average number of shreds with exactly one neighbor		Average number of shreds with 2 neighbors		Overall Accuracy	
Number of tests	100	10K	100	10K	100	10K	100	10K
Gray-level ranges	3.42	3.40	25.01	25.02	144.57	144.58	93.50	93.51
GLCM	34.70	34.52	58.71	58.76	79.59	79.73	64.85	64.94
Invariant moments	13.28	13.26	50.56	50.58	109.16	109.16	80.03	80.03
SFTA algorithm	20.01	20.10	52.05	51.95	100.94	100.96	75.58	75.55
Color moments	6.77	6.74	36.31	36.25	129.93	130.01	88.14	88.18

Table 6 Average number of neighbors found in medium fragments

Features/ results	Average number of shreds without neighbors		Average number of shreds with exactly one neighbor		Average number of shreds with 2 neighbors		Average number of shreds with 3 neighbors		Average number of shreds with 4 neighbors	
Number of tests	100	10K	100	10K	100	10K	100	10K	100	10K
Gray-level ranges	11.69	16.57	48.31	58.47	143.2	141.55	174.92	172.25	80.88	70.15
GLCM	37.57	43.4	93.78	97.75	155.58	147.37	129.2	126.01	42.88	44.4
Invariant moments	22.34	20	71.89	77.7	157.31	170.01	146.17	139.95	61.28	51.34
SFTA	63.37	53.73	103.95	116.41	153.21	155.74	104.53	111.73	33.94	21.39
Color moments	15.84	19.34	59.23	71.93	135.16	145.82	174.86	160.01	73.92	61.91

Table 7 Average number of neighbors found when 10 clusters are used

Shred size/ results	Average number of shreds without neighbors	Average number of shreds with exactly one neighbor	Average number of shreds with 2 neighbors	Average number of shreds with 3 neighbors	Average number of shreds with 4 neighbors
Medium fragments	32.42	96.66	187.32	113.13	29.48
Small fragments	146.19	425.86	904.64	888.62	402.69

6.2.3 Strip-cut medium fragments

In the medium fragments, the shreds are smaller than the wide shreds. Using the same 38 images, each image has a total number of 51 shreds. Table 4 shows the average overall accuracy in 38 images in every feature set when 5 clusters are assumed. Dividing 51 shreds into 10 clusters using the gray-level ranges as the features set results in an overall accuracy of 84.24 percent for 100 tests. The clusters may contain an unequal number of shreds, in other words, if 50 shreds are grouped into 10 clusters, some clusters may contain 5 shreds, others 3, and so on.

Comparing wide and medium fragments in terms of the overall accuracy from different feature sets, it can be noticed that the order is similar, where the gray-level ranges overall accuracy is highest.

The time to find the features vector for all 51 shreds for a single image is as follows: gray-level ranges: 0.19 seconds, GLCM: 0.15 seconds, all HU moments: 0.13 seconds, SFTA: 36.26 seconds, and color moments: 0.12 seconds. The use of k-means in one test takes time between 0.002 and 0.003 seconds.

6.2.4 Strip-cut narrow fragments

In this experiment, the same 38 images are shredded in one direction for 173 shreds. Table 5 summarizes the overall accuracy achieved using different feature vectors and k-means to cluster shreds into 5 clusters. Testing k-means clustering to divide the shreds into 10 clusters using gray-level ranges results in an overall accuracy of 89.34 percent.

The time to find the features vector for all 173 shreds for a single image is as follows: gray-level ranges: 0.45 seconds, GLCM: 0.41 seconds, all HU moments: 0.39 seconds, SFTA: 123.08 seconds, and color moments: 0.28 seconds. The use of k-means in one test takes time between 0.003 and 0.006 seconds.

6.3 Shredding in two directions (vertically and horizontally)

Cross-cut shredders and micro-cut shredders cut the paper in two directions. The clustering experiments are done using ideal shreds or in other words, a program shreds the images.

Table 8 Average number of neighbors found in small fragments

Features/ results	Average number of shreds without neighbors		Average number of shreds with exactly one neighbor		Average number of shreds with 2 neighbors		Average number of shreds with 3 neighbors		Average number of shreds with 4 neighbors	
Number of tests	100	10K	100	10K	100	10K	100	10K	100	10K
Gray-level ranges	55.30	69.18	200.76	263.99	580.68	660.65	1054.71	1039.99	876.54	734.18
GLCM	211.57	212.75	474.55	495.2	749.3	769.18	831.3	816.67	501.28	474.12
Invariant moments	63.43	67.77	252.35	290.74	735.11	819.43	947.84	946	769.27	644.06
SFTA	349.73	208.63	337.26	398.14	760.92	846.91	835.62	868.79	484.47	445.53
Color moments	56.01	72.22	238.19	297.45	621.92	747.99	995.64	963.84	856.24	686.5

Table 9 Comparison

Factors/ Reference	document type, document content	damage type, shred size	number of papers considered in a single experiment	real shredded papers vs. simulation	Method used
This paper	images	One direction, or two directions, strip-cut, cross-cut, and micro-cut	Single page	Simulation for all sizes and additional real data set for strip-cut shreds	Features: Gray-level ranges, GLCM, Invariant moments, SFTA algorithm, Color moments. Clustering: k-means
Ukovich and Ramponi [34]	Printed pages with or without ink notes	Strip-cut (Orthogonal to the text direction)	Mixed	Real-world dataset and virtual data set	Features: line spacing, document layout, presence of a marker, text edge energy, paper/ink color, squared paper. Clustering: Agglomerative hierarchical approach with stepwise optimization
Xing et al. [40]	Chinese printed pages	Strip-cut (vertically)	Multiple pages	Real-world data set	Three steps: computing the cluster number, identifying the starting point of clusters, and achieving clustering based on the regional division.
Yang and Wang [41]	Chinese printed pages	Cross-cut	Single page	computer-simulated cutting	Characteristics of character and typesetting based on vertical and horizontal projections

6.3.1 Medium fragments

The same 38 images used in strip-cut shreds are used here. The number of shreds is 459 in each image. The image is cut vertically and horizontally. The appropriate number of clusters in this context is harder to find and the method used in calculating the overall accuracy in strip-cut shreds does not apply directly as it requires the number of edges affected by clustering.

The accuracy is calculated based on the number of edges found and the maximum possible number of edges that could be found. As the maximum possible number depends on the number of edges effected by the clustering, the exact overall accuracy is not calculated here as it can only be approximated. Table 6 shows the average results of 100 and 10K tests when 5 clusters are assumed.

The use of gray-level ranges to group the shreds into 10 clusters, will result on average in 32.42 shreds in clusters with no originally adjacent neighbors, 96.66 shreds with exactly one originally adjacent neighbor, 187.32 shreds with 2 originally adjacent neighbors, 113.13 with 3, and 29.48 with 4, as shown in Table 7.

6.3.2 Small fragments

The number of shreds is 2768 in each image. The image is cut vertically and horizontally. Table 8 shows the average results of 100 and 10K tests when 5 clusters are assumed. The use of gray-level ranges to group the shreds into 10 clusters, will result on average in 146.19 shreds in clusters with no originally adjacent neighbors, 425.86 shreds with exactly one originally adjacent neighbor, 904.64 shreds with 2 originally adjacent neighbors, 888.62 with 3, and 402.69 with 4, as shown in Table 7.

7 Comparison

Table 9 describes literature tackling the clustering of paper shreds in terms of: document type and content, damage type and shred size, number of papers considered, real shredded papers vs. simulation, and methods used. None of the methods given in the table assumes the existence of missing pieces in the clustering experiments.

8 Conclusions and future work

This paper suggests the possibility of working on the hard problem of reassembling shredded paper, even for small sizes of paper shreds. The work focuses on clustering, which is a possible phase in shredded paper assembly. Vertically shredded images in addition to images shredded vertically and horizontally were considered. A set of features were used in the experiments to describe the shreds followed by k-means clustering. The results evaluation depends on the available information regarding the position of each shred and the neighborhood of each shred in the original image. Knowing the adjacent shreds of each shred makes it easier to check if the clustering was effective. The experimental results for all sizes and features were reported. The results show that there is potential for the clustering approach to be used in reconstructing the shredded paper.

A lot can be done for future research, as the results of the clustering approach used in the experiments depend heavily on the features chosen, other feature extraction methods and

clustering algorithms can be tested. Also, the use of other document types and documents with backgrounds is left for future work.

Acknowledgements The author would like to thank the anonymous reviewers for their valuable comments.

Declarations

Conflict of Interests The author declares that there is no conflict of interest.

References

1. Alhaj F, Sharieh A, Sleit A (2019) Reconstructing colored strip-shredded documents based on the hungarians algorithm. In: 2019 2nd international conference on new trends in computing sciences (ICTCS), pp 1–6. https://doi.org/10.1109/ICTCS.2019.8923048
2. Atallah AS, Emary E, El-Mahallawy MS (2015) A step toward speeding up cross-cut shredded document reconstruction. In: 2015 Fifth international conference on communication systems and network technologies, pp 345–349. https://doi.org/10.1109/CSNT.2015.69
3. Biswas A, Bhowmick P, Bhattacharya BB (2005) Reconstruction of torn documents using contour maps. In: IEEE International conference on image processing 2005, vol 3, pp III–517–20. https://doi.org/10.1109/ICIP.2005.1530442
4. Butler P, Chakraborty P, Ramakrishan N (2012) The deshredder: a visual analytic approach to reconstructing shredded documents. In: 2012 IEEE Conference on visual analytics science and technology (VAST), pp 113–122. https://doi.org/10.1109/VAST.2012.6400560
5. Chen G, Wu J, Jia C, Zhang Y (2017) A pipeline for reconstructing cross-shredded english document. In: 2017 2nd international conference on image, vision and computing (ICIVC), pp 1034–1039. https://doi.org/10.1109/ICIVC.2017.7984711
6. Chen J, Ke D, Wang Z, Liu Y (2017) A high splicing accuracy solution to reconstruction of cross-cut shredded text document problem. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-017-5389-z
7. Chen J, Tian M, Qi X, Wang W, Liu Y (2019) A solution to reconstruct cross-cut shredded text documents based on constrained seed k-means algorithm and ant colony algorithm. Expert Syst Appl 127:35–46. https://doi.org/10.1016/j.eswa.2019.02.039
8. Costa AF, Humpire-Mamani G, Traina AJM (2012) An efficient algorithm for fractal analysis of textures. In: SIBGRAPI 2012 (XXV Conference on graphics, patterns and images), pp 39–46. https://doi.org/10.1109/SIBGRAPI.2012.15
9. Deever A, Gallagher A (2012) Semi-automatic assembly of real cross-cut shredded documents. In: 2012 19Th IEEE international conference on image processing, pp 233–236. https://doi.org/10.1109/ICIP.2012.6466838
10. Gonzalez RC, Woods RE, Eddins SL (2009) Digital image processing using MATLAB, 2nd edn Gatesmark Publishing. Printed in the United States of America
11. Guo S, Lao S, Guo J, Xiang H (2015) A semi-automatic solution archive for cross-cut shredded text documents reconstruction. In: Zhang YJ (ed) Image and graphics. Springer International Publishing, Cham, pp 447–461. https://doi.org/10.1007/978-3-319-21978-3_39
12. Haralick RM, Shanmugam K, Dinstein I (1973) Textural features for image classification. IEEE Trans Systems Man Cybern SMC-3(6):610–621. https://doi.org/10.1109/TSMC.1973.4309314
13. Hong H, Zheng L, Pan S (2018) Computation of gray level co-occurrence matrix based on cuda and optimization for medical computer vision application. IEEE Access 6:67762–67770. https://doi.org/10.1109/ACCESS.2018.2877697
14. Htet ZW, Koldaev VD, Teplova YO, Kremer EA, Fedorov PA (2018) The evaluation of computational complexity of moment invariants in image processing. In: 2018 IEEE Conference of russian young researchers in electrical and electronic engineering (EIConrus), pp 1844–1848. https://doi.org/10.1109/EIConRus.2018.8317466
15. Hu MK (1962) Visual pattern recognition by moment invariants. IRE Trans Inform Theory 8(2):179–187. https://doi.org/10.1109/TIT.1962.1057692
16. Justino E, Oliveira LS, Freitas C (2006) Reconstructing shredded documents through feature matching. Forensic Sci Int 160(2):140–147. https://doi.org/10.1016/j.forsciint.2005.09.001

17. Liang Y, Li X (2020) Reassembling shredded document stripes using word-path metric and greedy composition optimal matching solver. IEEE Trans Multimed 22(5):1168–1181. https://doi.org/10.1109/TMM.2019.2941777
18. Lin HY, Fan-Chiang WCWada T, Huang F, Lin S (eds) (2009) Image-based techniques for shredded document reconstruction. Springer, Berlin. https://doi.org/10.1007/978-3-540-92957-4_14
19. Lin HY, Fan-Chiang WC (2012) Reconstruction of shredded document based on image feature matching. Expert Syst Appl 39(3):3324–3332. https://doi.org/10.1016/j.eswa.2011.09.019
20. Liu H, Cao S, Yan S (2011) Automated assembly of shredded pieces from multiple photos. IEEE Trans Multimed 13(5):1154–1162. https://doi.org/10.1109/TMM.2011.2160845
21. Muliadi Panggabean T, Elyezer Simaremare M, Siahaan R, Pardede C, Putri Gurning W (2020) Another parallelism technique of glcm implementation using cuda programming. In: 2020 4th International conference on advances in image processing, ICAIP 2020. Association for Computing Machinery, New York, pp 143–151. https://doi.org/10.1145/3441250.3441251
22. Ou X, Pan W, Xiao P (2014) In vivo skin capacitive imaging analysis by using grey level co-occurrence matrix (glcm). Int J Pharmaceut 460(1):28–32. https://doi.org/10.1016/j.ijpharm.2013.10.024
23. Paixao TM, Berriel RF, Boeres MCS, Koerich AL, Badue C, De Souza AF, Oliveira-Santos T (2020) Fast(er) reconstruction of shredded text documents via self-supervised deep asymmetric metric learning. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 14343–14351. https://doi.org/10.1109/CVPR42600.2020.01435
24. Paixão TM, Berriel RF, Boeres MC, Koerich AL, Badue C, De Souza AF, Oliveira-Santos T (2020) Self-supervised deep reconstruction of mixed strip-shredded text documents. Pattern Recognit 107:107535. https://doi.org/10.1016/j.patcog.2020.107535
25. Paixão TM, Berriel RF, Boeres MCS, Badue C, De Souza AF, Oliveira-Santos T (2018) A deep learning-based compatibility score for reconstruction of strip-shredded text documents. In: 2018 31St SIBGRAPI conference on graphics, patterns and images (SIBGRAPI), pp 87–94. https://doi.org/10.1109/SIBGRAPI.2018.00018
26. Paixão TM, Boeres MCS, Freitas COA, Oliveira-Santos T (2019) Exploring character shapes for unsupervised reconstruction of strip-shredded text documents. IEEE Trans Inf Forensics Secur 14(7):1744–1754. https://doi.org/10.1109/TIFS.2018.2885253
27. Patel B, Amin J (2015) Reconstruction of shredded document using image mosaicing technique-a survey. International Journal of Science and Research (IJSR) 4(12):737–740
28. Phienthrakul T, Santitewagun T, Hnoohom N (2015) A linear scoring algorithm for shredded paper reconstruction. In: 2015 11Th international conference on signal-image technology internet-based systems (SITIS), pp 623–627. https://doi.org/10.1109/SITIS.2015.13
29. Ping-sung L, Tse-sheng C, Pau-choo C (2001) A fast algorithm for multilevel thresholding. J Inf Sci Eng 17(5):713–727
30. Saboia P, Goldenstein SBayro-Corrochano E, Hancock E (eds) (2014) Assessing cross-cut shredded document assembly. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-12568-8_34
31. Schauer C, Prandtstetter M, Raidl GRBlesa MJ, Blum C, Raidl G, Roli A, Sampels M (eds) (2010) A memetic algorithm for reconstructing cross-cut shredded text documents. Springer, Berlin. https://doi.org/10.1007/978-3-642-16054-7_8
32. Sleit A, Massad Y, Musaddaq M (2013) An alternative clustering approach for reconstructing cross cut shredded text documents. Telecommun Syst 52(3):1491–1501. https://doi.org/10.1007/s11235-011-9626-x
33. Ukovich A, Ramponi G (2005) Features for the reconstruction of shredded notebook paper. In: IEEE International conference on image processing 2005, vol 3, pp 93–96. https://doi.org/10.1109/ICIP.2005.1530336
34. Ukovich A, Ramponi G (2008) Feature extraction and clustering for the computer-aided reconstruction of strip-cut shredded documents. J Electron Imaging 17:17–17–13. https://doi.org/10.1117/1.2898551
35. Ukovich A, Ramponi G, Doulaverakis H, Kompatsiaris Y, Strintzis MG (2004) Shredded document reconstruction using mpeg-7 standard descriptors. In: Proceedings of the Fourth IEEE international symposium on signal processing and information technology, 2004, pp 334–337. https://doi.org/10.1109/ISSPIT.2004.1433788
36. Wang Y, Ji DC (2014) A two-stage approach for reconstruction of cross-cut shredded text documents. In: 2014 Tenth international conference on computational intelligence and security, pp 12–16. https://doi.org/10.1109/CIS.2014.92
37. Wang Y, Wu B, Gao L, Yang H (2019) Automatic reconstruction of cross-cut chinese document shreds based on the feature of typesetting and strokes. In: 2019 IEEE 4Th international conference on signal and image processing (ICSIP), pp 727–731. https://doi.org/10.1109/SIPROCESS.2019.8868511

38. Xing N, Shi S, Xing Y (2017) Shreds assembly based on character stroke feature. Procedia Computer Science 116:151–157. https://doi.org/10.1016/j.procs.2017.10.060
39. Xing N, Zhang J (2017) Graphical-character-based shredded chinese document reconstruction. Multimed Tools Appl 76:12871–12891. https://doi.org/10.1007/s11042-016-3685-7
40. Xing N, Zhang J, Cao F, Liu P (2017) Practical challenge of shredded documents: Clustering of chinese homologous pieces. Appl Sci 7(9). https://doi.org/10.3390/app7090951
41. Yang H, Wang Y (2021) Automatic splicing of chinese single-sided shreds based on character feature and typesetting characteristics. J Phys Conf Series 1827:012063. https://doi.org/10.1088/1742-6596/1827/1/012063
42. Zhao B, Zhou Y, Zhang Z, Na Y, Ma T (2014) Information quantity based automatic reconstruction of shredded chinese documents. In: 2014 IEEE 26Th international conference on tools with artificial intelligence, pp 1016–1020. https://doi.org/10.1109/ICTAI.2014.154