Reconstruction of shredded documents

Author - Joe Dinn
Student number - 1724858
Project supervisor - Dr Paul L Rosin
Module number - CM3203
Module name - One Semester Individual Project

Project description

Despite the wide availability of computers, use of paper documents is still common. If a paper document contains sensitive data and is no longer needed, it is common practise to destroy it using a shredder. Reconstructing a shredded document is a very relevant problem in areas such as forensics where it can be used to recreate evidence. It is possible for a human to reassemble a shredded document, but it is both time consuming and laborious. The aim of this project is to design an automatic system for this.

Some research into the design of a system for reconstructing a shredded document has already been done. Solutions typically treat reconstruction as a combinatorial optimisation problem: order the fragments in such a way that maximises an objective function. This objective function is usually based on matching features between adjacent fragments. An example of this technique is by matching the shape of adjacent fragments [1]. To reduce the domain of the problem a common technique used is clustering similar fragments together [2].

There are a few complicating factors involved with designing a document reconstruction system:

- Shredding of documents be done in various ways e.g. tearing by hand, strip shredding and cross shredding, each producing different shapes and consistency of fragments. If the fragments are inconsistently shaped, then a jigsaw-solving algorithm can be used, however this will not be effective if they are very regular. This means that it is also important to consider the content of the fragments.

- The content of the documents may vary. For text documents optical character recognition might be a viable strategy but there is likely to be little variation in colour. On the other hand, with a document containing an image, variation in colour may be the easiest feature to match.

- A collection of fragments may be incomplete or may contain fragments from multiple documents.

- There will likely be a large number of fragments. Trying every combination of fragments would take exponential time with respect to the number of fragments, so this is clearly impractical. A heuristic based approach is therefore necessary.

Initially the project will focus on a solution with various restrictions for simplicity, outlined in the objectives section.

Objectives

The primary objective of this project is to implement a program that achieves the following:

- Take as input an image containing scanned fragments from a document.
- Segment the image to isolate the fragments.
- Rearrange the fragments to make a best approximation of the original document.
- Output a single image of the recreated document.

This solution must be able to reconstruct documents that are:

- Text based
- Black and white
- Strip shredded
- Made of a complete collection of fragments (the collection has no missing or excess fragments)

Additionally the solution may be extended to accommodate documents with:

- Non-text content such as images and coloured content
- Arbitrarily shaped fragments
- Missing or excess fragments

The secondary objective of the project is for the implementation to perform comparably or better than other contemporary solutions. To achieve this I will compare the performance of the solution with a number of other solutions, on the same dataset.

Work plan

3/2/20 - 16/2/20

- Research into contemporary solutions for reconstruction of shredded images.
- Identify 3 leading solutions for comparison.
- Find or create a dataset of shredded documents and ground-truths for comparison.
- Identify common metrics for analysing performance and select which will be used in the evaluation of the solution.
  Deliverable: dataset of shredded documents and ground-truths

17/2/20 - 23/2/20

- Implementation of image segmentation for isolating fragments.

24/2/20 - 8/3/20

- Implement an initial solution. This should satisfy the primary objective.
  Deliverable: Initial solution

9/3/20 - 5/4/20

- Experiment with different techniques to improve upon the initial solution. This should satisfy the secondary objective. If possible the solution can also be extended as outlined in the objectives.
  Deliverable: improved solution

6/4/20 - 12/4/20

- Evaluation of contemporary solutions compared to final improved solution.

13/4/20 - 19/4/20

- Visualisation of evaluation and write report, focus on evaluation of solution

20/4/20 - 6/5/20

- Finish report, focus on discussion of overall project
  Deliverable: report

7/5/20

- Final deadline for hand in of report

I plan to meet my supervisor every week to discuss the projects progress. The meetings are scheduled for 2:00 every Monday. In addition to these there will be two progress report meetings,

which will replace the standard meetings in the respective weeks. These will be held after developing an initial solution and after developing an improved final solution. Assuming the work plan goes as expected this will be Monday 8th of March and Monday 6th of April. The time dedicated to report writing leaves a small buffer if earlier stages overrun the allotted time.

I intend to write a brief first draft for each section of the report after finishing the relevant part of the work plan. This will ensure that detail is not omitted in the final report.

Ethics

The project does not require ethical approval.

Refrences

[1] Justino, Oliveira, Freitas “Reconstructing shredded documents through feature matching” Forensic Science International 160 (2006) 140–147 [https://s3.amazonaws.com/academia.edu.documents/48445564/Reconstructing_shredded_documents_throug20160830-6269-1qi6fn7.pdf?response-content-disposition=inline%3B%20filename%3DReconstructing_shredded_documents_throug.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20200202%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20200202T180407Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=69a52adf2cf2cd4bb0c797ab7543fe4bc601bb4f9fec3c15f28c905155a5a72d]

[2] Ukovich, Zacchigna, Ramponi, Schoier “Using Clustering for Document Reconstruction” [https://www.researchgate.net/profile/Giovanni_Gianni_Ramponi/publication/253426240_Using_clustering_for_document_reconstruction_-_art_no_60640J/links/0a85e53bc22f7a43bb000000.pdf]