 **COMSATS University Islamabad**

**Project Report**

**for**

**Document Reconstruction**

(Reconstruction of shredded documents from digital images)

Version 1.0

***By***

**Osama Zafar CU/FA17-BCS-076/ISB**

**Syed Muhammad Shahzaib CU/FA17-BCS-079/ISB**

**Syed Faizan Ahmed CU/FA17-BCS-085/ISB**

***Submitted To*Shehla Saif**

Submission Date: (8-August-2020)

***Bachelor of Science in Computer Science (SP-2020)***

.

* **Introduction**

Repairing of fragments is important in that it can be applied to many areas such as restoration of archeological findings, the obtainment of modern war information, obtaining the judicial exhibit, repairing the relics and acquiring military intelligence. This project will aim to reconstruct shredded documents to full-fledged documents. A direct and valid way is splicing the fragments manually. But when there was a huge quantity of the fragments, splicing manually would consume vast human resources. The better way in this era of technology is to develop an automated system to save time and human resources.

* **Problem Statement**

Getting a full-fledged document from shredded documents is a very useful thing. Document reconstruction is act of arranging shredded puzzle like document pieces into meaningful documents. It allows obtaining critical intelligence, judicial exhibits, reconstruction of ancient manuscripts and relics e.tc.

* **Literature Review**

There are numerous research papers written in field of puzzle solving, document reconstruction using feature matching, character recognition and context prediction via genetic algorithms, word matrix and context comparison etc. Literature reviewed for the purpose of development and understanding of the core concept of style transfer are as follows.

**Table 1: Related Literature**

|  |  |
| --- | --- |
| **Document Name** | **Authors** |
| Reconstruction of Shredded Paper Documents by Feature Matching | Peng Li  Xi Fang  Lianglu Pan  Yi Piao  Mengjun Jiao |

* **Process Methodology**

Pieces of shredded documents will be scanned and stored as digital images. Preprocessing was be done to create dataset by scaling and cropping region of interest (ROI) the specific document images into multiple parts like they were shredded. Then using image stitching clusters of pieces are to be created and weights get assigned to each specific cluster on basis of word matching form the bag of words.

* **Tools and Technologies**

**Table 2: Tools and Technologies for Proposed Project**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tools**  **And**  **Technologies** | **Tools** | **Version** | **Rationale** |
| MatplotLib | V3.2.0 | Python Library |
| OpenCV | 3.4.2 | Python Library |
| Colab Notebook | ---- | Notebook |
| Adobe Photoshop | CSC 6 | Design Work |
| MS Word | 2015 | Documentation |
| Pencil | 2.0.5 | Mockups Creation |
| **Technology** | **Version** | **Rationale** |
| Python | 3.8.0 | Programming language |

* **Code Snippets**
* **Result**
* **For Shallow Reconstruction**

The total number of input document images were 35. After shredding there were 140 parts. According to those images, After reconstrcution we achieved this result.

|  |  |
| --- | --- |
| **Percentage Reconstrution Accuracy** | **Frequency** |
| 25% | 36 Parts Garbage Single Piece Pages |
| 50% | 20 Parts 10 Pages |
| 75% | 12 Parts 4 Pages |
| 100% | 72 Parts 18 Pages |

Average percentage accuracy = 76.5% Parts were correctly matched and 51% pages were reconstructed completely

|  |  |
| --- | --- |
| **Category** | **Number** |
| True positive | 107 Parts |
| False positive | 31 Parts |
| Negative | 3 Parts |

* **For Deep Reconstruction**

Total 10 document images shredded to 80 pieces were fed for deep reconstruction.

|  |  |
| --- | --- |
| **Percentage accuracy** | **Numbers** |
| 12.5% | 21 Parts |
| 25% | 2 Parts |
| 37.5% | 6 Parts |
| 50% | 4 Parts |
| 62.5% | 5 Parts |
| 75% | 6 Parts |
| 87.5% | 28 Parts |
| 100% | 8 Parts |

Average percentage accuracy = 73.25% Parts were correctly matched and 10% pages were reconstructed completely

|  |  |
| --- | --- |
| **Category** | **Number** |
| True positive | 59 |
| False positive | 18 |
| Negative | 3 |

* **Conclusion**

Through use of optical character recognition (OCR) algorithm for word construction and clustering algorithms using weights based on word matrix matching and context analysis for reconstruction of document shredded pieces will be stitched together to reconstruct the actual image of document.

* **References**
* Reconstruction of Shredded Paper Documents by Feature Matching
* [A Solution to Reconstruct Cross-Cut Shredded Text Documents Based on Character Recognition and Genetic Algorithm](https://www.hindawi.com/journals/aaa/2014/829602/)
* [Optical Character Recognision Using OpenCV](https://www.learnopencv.com/tag/ocr/)
* [Data Analysis with Python](https://cloudxlab.com/blog/numpy-pandas-introduction/)
* **Links**

Github Link for Notebooks and Dataset

<https://github.com/osamazafar980/Document-Reconstruction>

Youtube Link for Video

<https://youtu.be/B6gI_7Pb0S0>