An Imaging Technique for Retrieval of Lost Content in Damaged Documents

Neelam Bhardwaj
Research Scholar, Computer Science and Engineering Department
Motilal Nehru National Institute of Technology
Allahabad, India.
Member, Computer Society of India

Suneeta Agarwal, PhD
Head, Computer Science and Engineering Department
Motilal Nehru National Institute of Technology,
Allahabad, India,
Member, IEEE

ABSTRACT

It is very common that some useful contents of documents are lost or hidden intentionally or accidently due to several reasons e.g. Whitener, pasting of paper, ink spreading, fading, dirt etc. The importance of these damaged documents may in terms of some research result, historic event, pacts or any important piece of information. The retrieval of such lost contents is a latent research area. The available approaches only guess for the lost contents by exploring the remaining intact information. But no approach is found yet to say that the retrieval is exactly of original ones. We have proposed a new approach and developed an experimental setup which involves imaging by sensing the light after passing through the damaged document instead reflection and then applying OCR on acquired images for retrieval of lost or hidden contents. Experiments are carried out on various test documents. Good results are obtained. The applicability of this scheme is limited for physically available documents only, but ensures the originality of retrieved contents.

General Terms

Imaging, Computer application, contents retrieval in damaged documents.

Keywords

Imaging, damaged document, content retrieval, OCR.

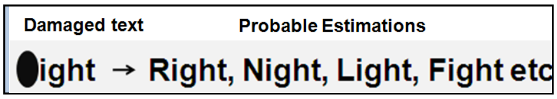
1. INTRODUCTION

Preservation of document is very important for any type of advancement in life. It is very common that some useful contents of documents may be lost or hidden intentionally or accidently due to several reasons e.g. using whitener, pasting of paper, ink spreading, fading, dirt etc. The importance of these damaged documents may in terms of some research result, historic event, pacts or any important piece of information. The retrieval of such lost contents is a latent research area.

Some important work has been done for content retrieval from degraded documents. A. Antonacopoulos and D. Karatzas [1] presented digital historical document lifecycle based approach where the expert knowledge of the historian/archivist user is required at different stages. Later A. Antonacopoulos and D. Karatzas [2] again developed an unsupervised machine base to extract the relation between the objects at symbolic level. In this approach, first an appropriate training algorithm is developed and then used. F. Drira [3] presented their work towards document restoration. A. Antonacopoulos and C. C. Castilla [4] assessed that the condition and individual nature of characters in degraded documents necessitate a departure from existing thresholding approaches and designed a flexible approach. To overcome the difficulties presented by such documents this approach works by flexibly analyzing at each character level and cautiously repairing it. S. Pletschacher et. al. [5] presented a new semi-supervised clustering framework to the recognition of heavily degraded characters in historical typewritten documents. Lots of character classes with similar shapes are clustered and are indexed by pseudo code for easy synthesize of degraded document images. Xia Yong et. al. [6] presented a new way to improve the performance of retrieval by fuzzy coding strategy.

All the above discussed techniques work for degraded documents. Till now, for damaged documents where retrieval of completely lost and invisible contents are required, only linguistic, reading comprehension, vocabulary skills and prior knowledge of semantics are utilized to estimation the probable contents. Very little literature is available in this domain of content estimation. Mohamed Cheriet et. al. [7] worked on a guide for semantic knowledge in word completion.

No approach is found yet to retrieve the actual hidden information. The estimations always have the probability of inaccuracy and some time may be correct also. But how the meaning may be different because of wrong predictions is shown in figure 1 and figure 2.



Damaged text	Probable Estimations
●ight →	Right, Night, Light, Fight etc

Fig 1: Example for estimation about damaged word

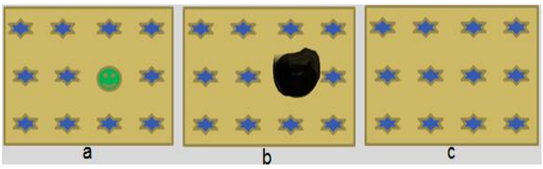


Fig 2: Example for estimation about hidden object

In this paper, we have attempted to achieve the same goal of text retrieval from damaged documents by an entirely different approach than in literature. We have developed an experimental setup which involves imaging by sensing the light after passing through the damaged document instead reflection and then applying OCR on acquired images for retrieval of lost or hidden contents. We also experimented with proposed technique to recover the missing words in the damaged documents effectively and with a great ease.

Generally, documents are damaged by occurrence of defects due to addition of layer of substance on their surfaces. This addition of layer results in increasing the thickness effectively at places of hidden contents or objects. In fact the increased thickness becomes almost double as compared to the remaining undamaged portion of document for passage of light through it. This keen observation led us to develop an experimental set up which involves imaging or scanning by sensing the light after passing through the damaged document instead of the usual way of scanning i.e. by sensing the reflected light from their surface and then applying OCR software on the acquired images for content retrieval. This technique is not applicable for lost contents because of scratches.

At present, the applicability of our scheme is limited for physically available documents only, but works for restoration towards originality.

The paper is organized as follows: after an introduction and related work in section 1, Section 2 introduces our proposed imaging technique and experimental set-up developed for the purpose, Section 3 describes its applicability in content retrieval, and Section 4 gives the framework of the overall content retrieval approach. The different types of input documents that have been prepared for experiments and the corresponding results are discussed in Section 5. Finally we conclude with result analysis, limitations and future scope in Section 6.

2. PROPOSED IMAGING AND EXPERIMENTAL SETUP

Imaging involved scanning and image acquisition. Conventionally documents are scanned by sensing the light reflected from their surface. In proposed technique, the images are acquired by sensing the light after passing through the damaged document as shown in figure 3. Here the spatial luminous intensity distributions of acquired images are direct proportional with intensity distribution of light passed through subject documents and exposing on the focal plan array of image acquisition device. For better results, the imaging should be performed in dark room to minimize the effect of light reflection during image acquisition.

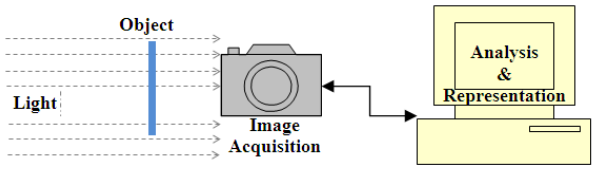


Fig 3: Proposed imaging and setup

An experimental setup for imaging is prepared by us as shown in figure 4. In terms of hardware used, wooden box, white 5w CFL as light source, 4 mm thick transparent glass as a base for document and 8 mega pixel cameras were used. The vertical distance between light source and transparent glass/document is 300 mm and size of wooden box is 150x150 mm2. The image acquisition device was kept at a vertical height of 600 mm from damaged document.

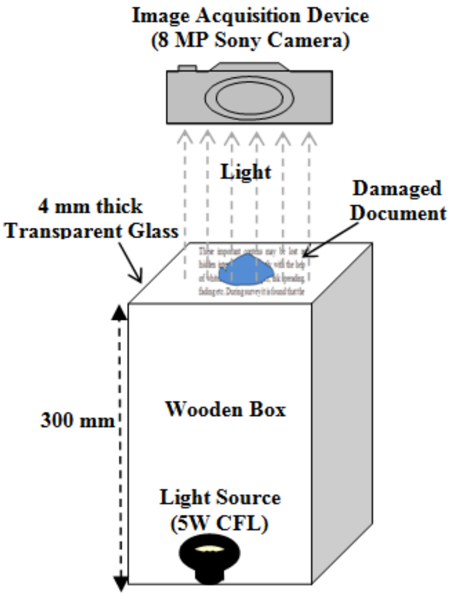


Fig 4: Experimental imaging setup

3. APPLICABILITY PROPOSED IMAGING FOR RETREIVAL OF HIDDEN CONTENT

In the damaged documents, the spatial locations where the information is hidden beneath substances, causing damages, have almost double thickness as compared to the remaining part for passage of light through them. Hence according to Beer-Lambert law, the intensity of light exposing on the focal plan array through these relatively thick damaged regions is approximately less by a factor of ‘$10^{2\alpha}$’ as compared to the remaining portion. Where, α = absorption coefficient of color material. The multiplication of 2 to α is taken assuming that in case of damage by addition of substances, the surface thickness at damage location of document becomes double. This causes in their relatively less and comparable intensity in the luminous domain and the hidden contents are bring out with good contrast in the image, acquired with proposed technique. After analyzing these regions, the hidden contents may be retrieved toward their originality.

4. CONTENT RETREIVAL APPROACH

The objective is to retrieve the hidden or lost text information from input damaged document. The proposed approach is divided in three stages a) Imaging b) Image analysis c) Recognition by OCR. The imaging includes scanning and image acquisition and is performed using set up shown in figure 4. During image acquisition, the light source was switched ON while the Camera flash & all room lights were kept OFF. Camera position i.e. distance from document was adjusted for acquiring different images depending on font size of the document. The image processing (noise removal, contrast enhancement and morphological processing) is done on acquired image using MATLAB 7.0 image processing tool. Then this processed image is passed as input to OCR ABBYY Fine Reader 11 to get the outcomes in MS WORD format. The complete process is illustrated in figure 5.

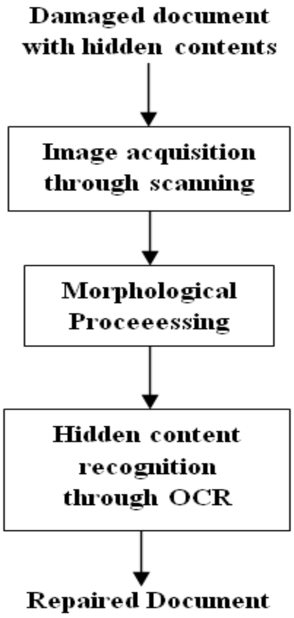

Damaged document with hidden contents
Image acquisition through scanning
Morphological Proceeessing
Hidden content recognition through OCR
Repaired Document

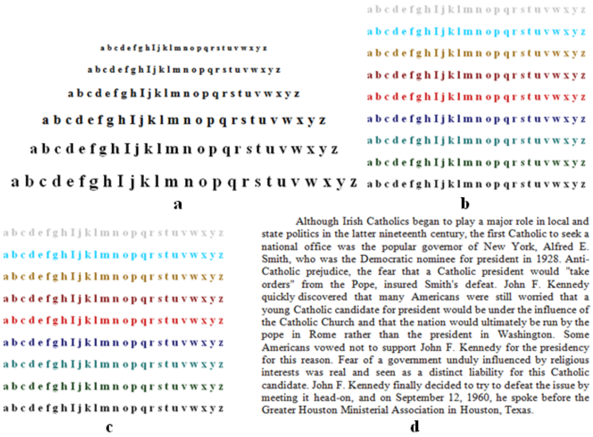
Fig 5: Content retrieval approach

5. EXPERIMENTS AND RESULTS

5.1. Input documents

The input damaged documents are prepared by pasting of white sheet and by spreading of whitener over the original documents of figure 6(a) to 6(d). In Fig 6(a), the document is having alphabets with times roman font and sizes varying from 8 to 18. Figure 6(b) and figure 6(c) have same alphabets series with same font size but with different colors and figure 6(d) is a paragraph of historic speech.

The input documents of figure 6(a) to figure 6(d) are damaged purposefully by pasting of paper and by applying whitener over them as shown sequentially in figure 7(a) to figure 7(d). The prediction of any character/text beneath the damage is almost impossible in these cases of damaged inputs neither visually nor by Optical Character Recognizer (OCR).



Although Irish Catholics began to play a major role in local and state politics in the latter nineteenth century, the first Catholic to seek a national office was the popular governor of New York, Alfred E. Smith, who was the Democratic nominee for president in 1928. Anti-Catholic prejudice, the fear that a Catholic president would "take orders" from the Pope, insured Smith's defeat. John F. Kennedy quickly discovered that many Americans were still worried that a young Catholic candidate for president would be under the influence of the Catholic Church and that the nation would ultimately be run by the pope in Rome rather than the president in Washington. Some Americans vowed not to support John F. Kennedy for the presidency for this reason. Fear of a government unduly influenced by religious interests was real and seen as a distinct liability for this Catholic candidate. John F. Kennedy finally decided to try to defeat the issue by meeting it head-on, and on September 12, 1960, he spoke before the Greater Houston Ministerial Association in Houston, Texas.

Fig 6: Original documents.

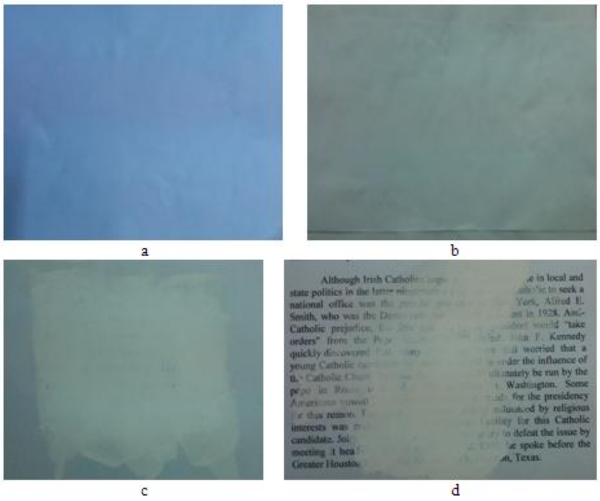


Fig 7: Damaged document. a) Damage by pasting a double layered white paper. b) Damage by pasting a double layered white paper. c) Damage by whitener. d) Historical speech document damaged by whitener.

5.2. Images acquired with proposed imaging

The damaged input documents of figure 7 are then passed through the proposed imaging using above discussed experimental setup. In the resulting images, hidden characters now become visible. These images are further processed for noise removal, smoothening and contrast enhancement. The resulting images are shown in figure 8.

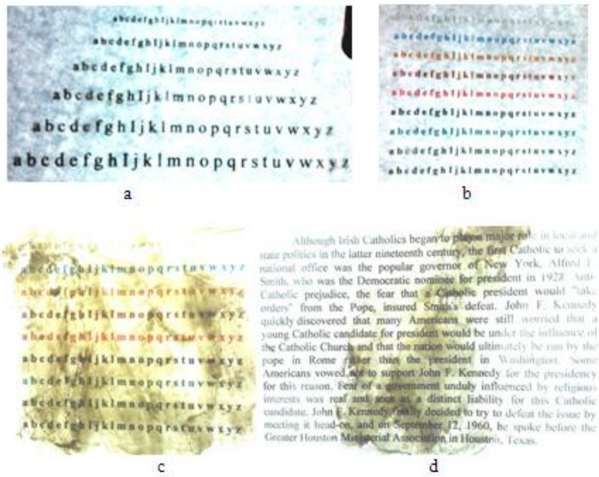
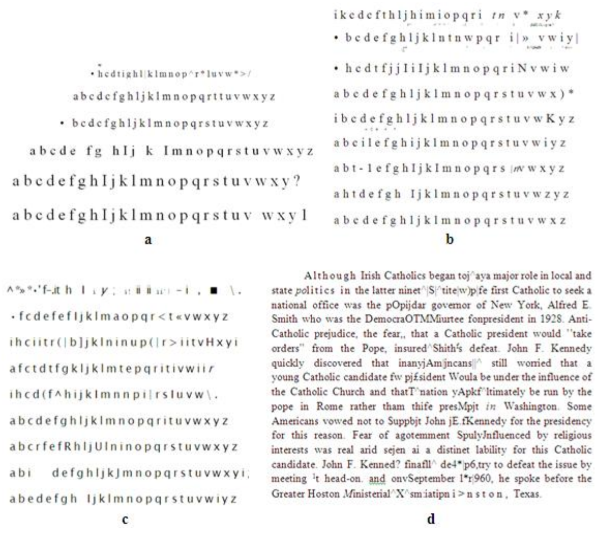


Fig 8: Images acquired after Imaging of damaged documents of figure 7.

5.3. OCR results

The images of figure 8 are then passed as input to OCR ABBYY Fine Reader 11. The outcomes of OCR are in MS Word format and are shown in figure 9.These results are further quantized to know the degree of accuracy.



•hcdtighl|klmnop^r*luvw*>/
abcdcfghljklmnopqrttuvwxyz
• bcdcfghljklmnopqrstuvwxyz
abcde fg hIj k Imnopqrstuvwxyz
abcdefghIjklmnopqrstuvwxy?
abcdefghIjklmnopqrstuv wxy1

a

ikcdcfthljhimiopqri tn v* xyk
• bcdefghljklntnwpqr i|» vwiy|
• hcdtfjjIiIjklmnopqriNvwiw
abcdefghljklmnopqrstuvwx)*
ibcdefghljklmnopqrstuvwKyz
abcilefghijklmnopqrstuvwiyz
abt-1efghIjkImnopqrs |nv wxyz
ahtdefgh Ijklmnopqrstuvwzyz
abcdefghljklmnopqrstuvwxz

b

^*»*•'f-.it h I ι y ; ıı ii ii ıı ı -i , ■ \.
•fcdefefIjkImaopqr<t«vwxyz
ihciitr(|b]jkIninup(|r>iitvHxyi
afctdtfgkIjkImtepqritivwiir
ihcd(f^hijkImnnpi|rsIuvw\.
abcdefghIjkImnopqrituvwxyz
abcrfefRhIjUIninopqrstuvwxyz
abi defghIjkJmnopqrstuvwxyi;
abedefgh IjkImnopqrstuvwiyz

c

Although Irish Catholics began toj^aya major role in local and state politics in the latter ninet^|S|^tite|w)p|fe first Catholic to seek a national office was the pOpijdar governor of New York, Alfred E. Smith who was the DemocraOTMMiurtee fonpresident in 1928. Anti-Catholic prejudice, the fear,, that a Catholic president would "take orders" from the Pope, insured^Shith's defeat. John F. Kennedy quickly discovered that inanyjAm|jncans||^ still worried that a young Catholic candidate fw pj£sident Woula be under the influence of the Catholic Church and thatT^nation yApkf^ltimately be run by the pope in Rome rather tham thife presMpjt *in* Washington. Some Americans vowed not to Suppbjt John jE.fKennedy for the presidency for this reason. Fear of agotemment SpulyJnfluenced by religious interests was real arid sejen ai a distinet lability for this Catholic candidate. John F. Kenned? finafll^ de4*|p6,try to defeat the issue by meeting ¹t head-on. and onvSeptember 1*r|960, he spoke before the Greater Hoston Ministerial^X^sm:iatipn i>nston, Texas.

d

Fig 9: OCR outcomes for images of figure 8

5.4. OCR results analysis

The results obtained above are analyzed in-terms of the number of correctly detected alphabets with respect to damaged alphabets. The percentage of errors are calculated using the formula-

$$\textbf{Error (\%)} = (\alpha-\beta)*100/\alpha$$

Where, α =Total number of hidden characters beneath the damages, β= Number of correctly retrieved characters. The results are shown in the plots of error with respect to input parameters like font sizes and colors. The results are further discussed for all above mentioned four input cases.

5.4.1. *Case 1: Document having alphabets with varying font size (figure 6(a) and damaged by pasting of double layer of white paper sheet over it (figure 7(a)).*

In figure 10, error is plotted with respect to different font sizes. It is observed that errors are less in retrieving large characters in comparison to errors in small characters. To a small font size it is still retrieving characters up to a certain level of accuracy, which was not possible otherwise. In case of font size 8 character "a" is detected as "." while taking font size "18" of same character, it is detected correctly.

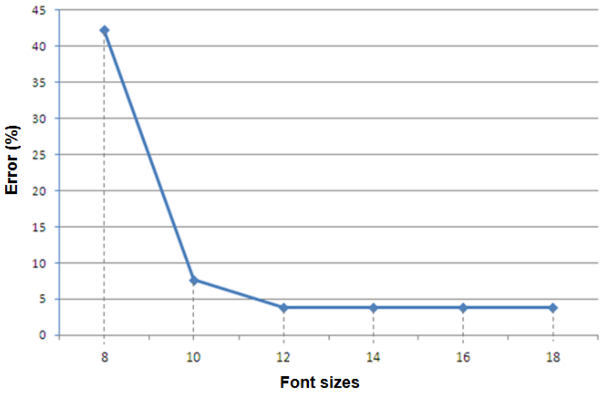


Fig 10: Error Vs Font size for document damaged by pasting of sheet

5.4.2. *Case 2: Input document having different colour alphabets (figure 6(b)), damaged by pasting of single layer of white paper sheet over it (figure 7(b)).*

It is observed in figure 11 that dark colored characters are retrieved more accurately in comparison to the light colored text characters of similar font size constant. The black colored text is completely retrieved. Still light colored text is recovered with certain level of accuracy which was not possible otherwise.

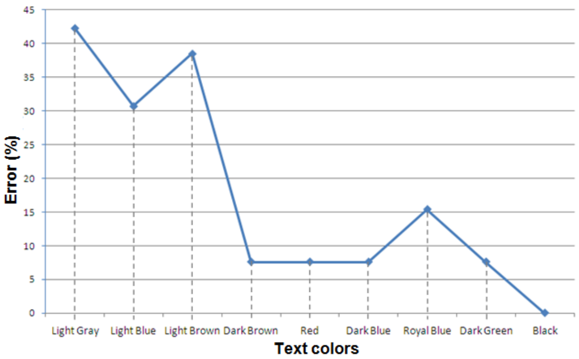


Fig 11: Error Vs color for document damaged by pasting of sheet

5.4.3. *Case 3: Input document having different colourr alphabets [figure 6(c)], damaged by applying whitener over it as shown in figure 7(c).*

It is observed in figure 12 that the variation in character recovery is due to difference in colors of the document. It is again clear from figure 12 that dark colors are detected with more accuracy in comparison to light colored text.

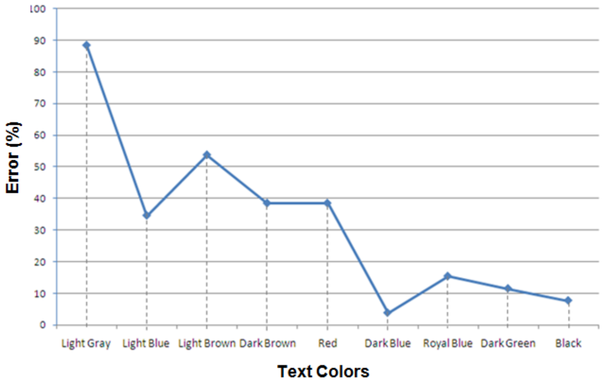


Fig 12: Error Vs color for document damaged by whitener

5.4.4. *Case 4: Input document of figure 6(d) damaged by whitener shown in figure 7(d).*

In this case the input document was originally having an historic speech section, which was damaged by whitener, shown in figure 8(d). The resulting document after passing though the proposed experimental set-up is recovered as shown in figure 9(d). Most of the damaged characters are recovered by the proposed scheme and the missed characters in the words can be corrected easily with the basic linguistic and vocabulary skill e.g. in the sentence "Although Irish Catholics began to jlay a major role in local and state politics in the later ninent//n|h century" words "jlay" is "play" and "ninent//n|h" is "nineteenth" can be easily corrected with available linguistic and vocabulary knowledge.

5.4.5. Imaging at alphabet level

In the above shown OCR results, maximum errors occurs to distinguish the alphabets having closed boundaries e.g. ‘a’, 'o’, 'z’ etc and small font sizes. Then to improve the results, we performed imaging of damaged document at alphabet level. In figure13, the results are shown. This way, the error may be minimized up to a great extent but the time consumption for recovery of characters will be more.

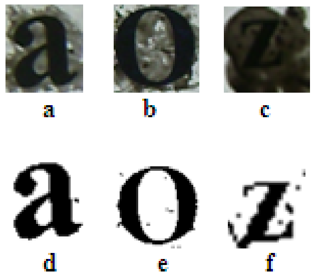


Fig 13: a, b & c are images acquired by zooming on particular alphabet during imaging and c, d & e are converted binary images.

5.4.6. Character on similar background

There may be the cases where document seems to be undamaged but some of the character or contents are morphed with same color background. To test the applicability of proposed scheme, we have taken the one black colored case as shown in figure 14 (a). The outcomes obtained after imaging this input with proposed set-up is shown in figure 14(b). Finally the recognized text is shown in figure 14(c) which was not possible otherwise through conventional scanning technique.

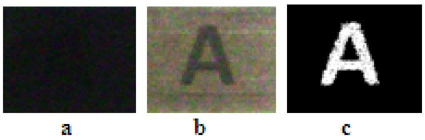


Fig 14: Damaged document. A) alphabet with same background. b) Image with proposed imaging. c) Binary image.

6. CONCLUSION

Obtained results are encouraging. The approach is able to retrieve actual hidden contents effectively. Currently, proposed technique has the limitation that the damaged document must be having some light passing property. But in future, the imaging with IR and x-ray transmission may also be explored to overcome the restriction of proposed methodology as S. Cubero et. al. [8] done recently for quality inspection in fruits and vegetables by using IR transmission (Also the automatic engineering version to detect damaged locations and to perform imaging at alphabet level is future scope.

7. ACKNOWLEDGMENT

Sincere thanks to Mr. Vikash Bhardwaj, Scientist, ADRDE (DRDO) for help in creating the innovative scanning set-up.

8. REFERENCES

[1] A. Antonacopoulos and D. Karatzas, 2004. The Lifecycle of a Digital Historical Document: Structure and Content. Proceedings of the ACM Symposium on Document Engineering (DocEng 2004), pp. 147-154.

[2] A. Antonacopoulos and D. Karatzas, 2005. Semantics-Based Content Extraction in Typewritten Historical Documents. Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR2005), Seoul, South Korea, IEEE, vol. 1, pp. 48-53.

[3] F. Drira, 2006. Towards restoring historic documents degraded over time. Document Image Analysis for Libraries (DIAL '06). IEEE Second International Conference, pp. 357-364.

[4] A. Antonacopoulos and C.C. Castilla, 2006. Flexible Text Recovery from Degraded Typewritten Historical Documents. 18th International Conference on Pattern Recognition (ICPR), IEEE, vol. 2, pp. 1062-1065.

[5] S. Pletschacher1, J. Hu and A. Antonacopoulos, 2009. A New Framework for Recognition of Heavily Degraded Characters in Historical Type written Documents Based on Semi-Supervised Clustering. 10th International Conference on Document Analysis and Recognition, IEEE, pp. 506-510.

[6] Xia Yong, Jia Xu-Hui and Wang Kuan-Quan, 2012. International conference on Systems and Informatics (ICSAI), IEEE xplore, pp. 261 – 264.

[7] Mohamed Cheriet, Nawwaf Karma, Cheng Lin Liu, Chingy Suen, 2007. Character Recognition Systems. A Guide to Students and Practioners, Wiley Publications.

[8] S. Cubero, N. Aleixoa, E. Molto, J. G. Sanchis and J. Blasco, 2011. Advances in machine vision application for automatic inspection and quality evaluation of fruits and vegetables. Food and Bioprocess Technology, Springer, vol. 5, issue 4, pp. 487-504.