[Prasad*, 5(3): March, 2016] ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

RECOVERY OF DOCUMENT TEXT FROM TORN FRAGMENTS USING IMAGE PROCESSING

C.Prasad*, Dr.Mahesh, Dr.S.A.K. Jilani
* Department of ECE, Madanapalle Institute of Technology and Science,
Madanapalle,Andhrapradesh,India

DOI: 10.5281/zenodo.46999

ABSTRACT

Recovery of document from its torn or damaged fragments play an important role in the field of forensics and archival study. Reconstruction of the torn papers manually with the help of glue and tapes etc., is tedious, time consuming and not satisfactory. For torn images reconstruction we go for image mosaicing, where we reconstruct the image using features (corners) and RANSAC with homography.But for the torn fragments there is no such similarity portion between fragments. Hence we propose a new process to recover the original document form its torn pieces by using the Binary image processing techniques with region properties of the torn pieces. Our mehodology for recovery of torn pieces can be solved in three simple stages. Initially the torn pieces of the document are acquired as input. The torn pieces are straightening to axis using HORIZON function and they are concatenated. The torn fragments are segmented based on the regionpropertiethen concatenated the segmented images. Finally by creating mask the concatenated images are going to combined.

KEYWORDS: RANSAC,Homography,Binary image processing, horizon, regionproperties,imagemosaicing

INTRODUCTION

A Document is a bonded physical or digital representation of information designed with the capacity to communicate. In prototypical usage, a Document is understood as a paper artifact containing information in the form of ink marks. Questioned Document Examination (QDE) is a sub-field of forensic sciences and it is related to the federal, civil, law enforcement, and justice areas. The task of document examination is to compare a questioned document, using a Scientific method to a series of known standards, i.e., signature Verification, handwriting identification, etc. In order to perform a reliable analysis, forensic document examiner must count on well preserved documents. However, very often questioned documents suffer damages at several levels, such as, torn edges, moisture, obliteration, charring, and shredding.

In the latter case, shredding can be performed by a machine or by hand. In both cases, documents need to be recovered back so that forensic document examiners can analyze them. The amount of time necessary to reconstruct a document depends on the size and the number of fragments, and it can be measured in days or even weeks. Sometimes some fragments of the document can be missing, and for this reason, the document can be only partially reconstructed. Even then, the manual effort of the forensic which is tedious and laborious can be alleviated. One problem faced when reconstructing documents by hand lies in its manipulation. The physical reconstruction of a document modifies some aspects of the original document because products like glue and adhesive tape are added into it. This type of manipulation is known as destructive analysis. In this paper we focus on automatic software solution for recovery of documents shredded by hand using binary image processing techniques with region properties of the torn pieces. We are using mat lab software with image tool box for our approach for recovery.

RELATED WORK

A.Image Mosaicing

Image mosaic is a process of stitching of multiple images having some similarities between them. Image mosaicing can be performed by using the feature based method [1,2, 3,4].Feature based methods consider low level features of the images like edges or corners and reduces the computational complexity and time. The Process of image mosaicing starts with the acquisition of images. Then to match the images we will go for low level feature detection by using the techniques called corner detection algorithms and then we will match the features found by considered algorithms by using the nearest neighbor method. After matching is performed, using the RANSAC method we will eliminate any outliers and will try to estimate HOMOGRAPHY between the pair of images. Then finally to align the images in the warp and blend images step and output misaiced image is obtained.

B. Literature Survey

Document stitching or recovery of document from its torn pieces is absolutely different from Image mosaicing techniques. Image mosaicing required similarities between the images, but you shouldn't find the similarities between the torn fragments. We are introducing a new approach for recovery of document text form the torn fragments using binary image processing techniques. Here I am giving a brief discussion about binary image processing image segmentation, which are useful for build our methodology.

B.1. Binary Image Processing

Binary image processing is also referred as Morphological image processing. Morphology is a branch of biology deals with the form of structure of the plants and animals. Mathematical morphology is a collection of non-linear processes applied to images to remove details smaller than a certain reference shape, called s*tructuring element*. Mathematical morphology is a tool for extracting the image components which are important for representation and description. The technique was developed by Matheron and Serra at Ecole des Mines in Paris. Mathematical morphology is totally based on the set theory. Mathematical morphology is used as the basic for the image segmentation procedure with a broad range of applications and it also plays a major role in procedures for image description. Structuring element in morphological place an important role with different shape and size and these are to be defined by 0s and 1s.

Morphological operations are defined by moving the structuring element over the binary image to be modified; in such a way that it is centered over every image pixel at some point and a logical operation is performed on the pixels covered by structuring element and yielding binary output. Morphological operations are like convolution process. As similar to convolution kernel, the structuring element can be of any size, and can contain complements of 1s and 0s. At each pixel position, a specified logical operation is performed between the structuring element and the binary image that to be modified. The outcome of morphological operation is depends upon size and the content in the structuring element and the nature of the logical operation. Dilation and erosion are the basic operations in morphological image processing. Dilation is an operation which 'grows' or 'thickens' objects in the images. The specific manner and the extent of the thickness or growing of the object are controlled by the *structuring element.* The mathematical representation of the dilation process of A and B is given below.

$$A \oplus B = \left\{ z \left| \left(\hat{B} \right)_z \bigcap A \neq \phi \right. \right\}$$

It says that the dilation of A and B is the set consisting of al structuring element origin locations where the reflected and translated B overlaps at least some portion of A. The mat lab function for the structuring element is *strel*,*strel* Create morphological structuring element and we can give the shape and size for the *strel* function as diamond, disk, and line, octagon, pair, periodic line, and rectangle. The syntax for structuring element is

se= strel('line',LEN,DEG)

se = strel('line',5,90) ; Here structuring element shape is line with length 5 and angle 90 degrees. It creates a flat linear structuring element that is symmetric with respect to the neighborhood center. DEG specifies the angle (in degrees) of the line as measured in counterclockwise direction from the horizontal axis. LEN is approximately the distance between the centers of the structuring element members at opposite ends of the line.

B.2 Image Segmentation

Image segmentation is the process of partitioning the images into meaningful regions. Image segmentation are based on threshold, edge based segmentation, Region based, clustering techniques and Matching. For our proposed algorithm we use region based segmentation.

PREVIOUS WORK

Reconstruction of fragmented pieces of an object in order to get back the original one is a critical and time consuming task, which is dealt and maneuvered in several departments for example archaeology (reassembling mixed remnants obtained from excavation of ancient rooms), forensics and investigation science (recovering torn or shredded documents). Many researches were done on the document recovery from its torn pieces or fragments. Mosaicking of torn document images [5] by Nagaraj B. Patil is done by scan the two torn pieces of shredded document such that the non-uniform sides face each other. He considered the non-uniform boundary values and if they match the translation process done and then the images are combined if not the matching process of non-uniform boundary vales is continued.Reconstruction of torn page using corner and segment matching [6] had done using semiautomatic techniques. Torn pieces joined by comparing edge length and angles. If the image features were matched then the two torn pieces were club together. ArindamBiswas, ParthaBhowmick had done the Reconstruction of Torn Documents Using Contour Maps proposed a [7] technique for reconstruction of hand-torn documents. And Reconstructing shredded documents through feature matching [8] by Edson Justino by polygonal approximation for reducing complexity for boundaries and then extracts relevant features of the polygon to carry out the local reconstruction. The ambiguities resulting from the local reconstruction are resolved and the pieces are merged together as we search for a global solution.Ankush Roy had given an approach of “A Probabilistic Model for Reconstruction of Torn Forensic Documents” [9] is presents a probabilistic approach to reconstruct a document. By iteratively calculates the probability of an arrangement subject to some constraints and attempts to produce the best possible configuration (or solution) using low level image statistics.

PROPOSED METHOD

Our Methodology for recovery of document using the binary image processing techniques involves three main steps they are

1. Image Acquisition
2. Pre-Processing
   (i) Image Straightening
   (ii) Concatenation
   (iii) Binarisation
3. Joining Process
   (i) Region Segmentation
   (ii) Horizontal and Vertical Correction

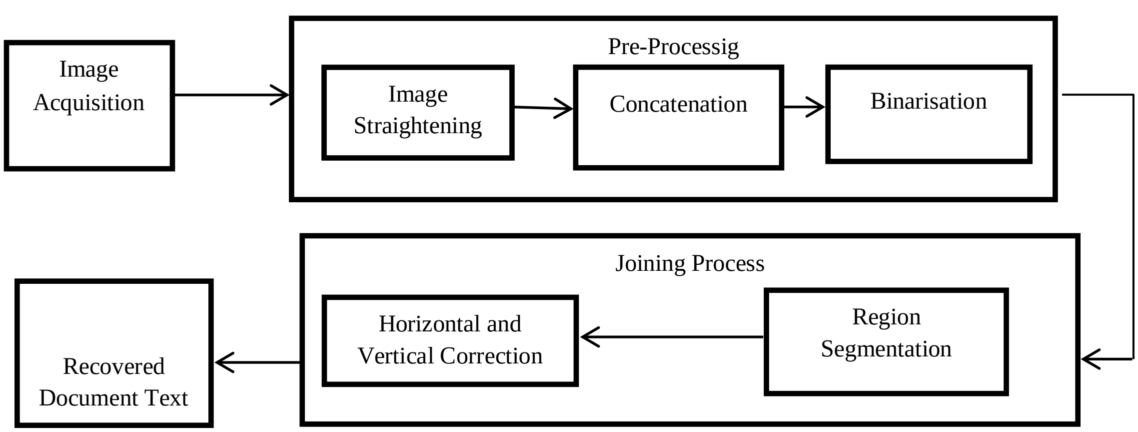


Fig 1.Recovery of Document text From torn Fragments

A software based solution for the recovery of document will give better results than the manmade stitching and also we can save the time much more than manual process. Fig1. Shows the step by step procedure in performing Recovery of document text from torn pieces.

1. Image Acquisition

The definition of Image Acquisition is the process of capturing the images from the real world and store them into processing device i.e. Computer. For our approach we need to get the torn pieces into processing device. It can be done by capturing the torn pieces by a digital camera having minimum 8MP resolution for the better visibility of the text in the fragments. The figures below shows he input images i.e. torn pieces for the recovering the original document.

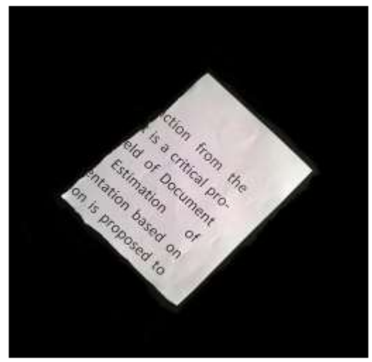


Fig 2.a.Input image1 ***2.b.Input Image 2***

The images are of jpg, jpeg, bmp, tiff type and the input images are with different size and with different rotational angle. We use the Mat lab software for propose method, so we need to save input images into Mat lab folder.

2. Pre-Processing

The acquired input images are then given to Pre-processing stage where the Image Straightening, Concatenation and Binarisation processes are performed. Image straightening is the process of making the image parallel to the any of the axis. Concatenation is defined as the process of placing the images one after another i.e. in a single image. Binarisation process converts the gray scale image to binary image.

2.1.image Straightening

The input images or torn pieces that we captured are with different rotational angle. We need to straighten the image such that the image must be parallel to any of the axis and make sure that the two image has to be parallel either X or Y Axis.

angle=horizon(A, PRECISION, METHOD, DISKSIZE) aligns the image ‘A ’with the specified method. The methods used for image straightening are

- “Fast Fourier Transform” which is a default method.
- “Hough transform” which finds lines in the image.
- “blot"’ Finds blots and estimates their's orientation. Blot method allows additional parameter DISKSIZE that defines the filter size of morphological transformation

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

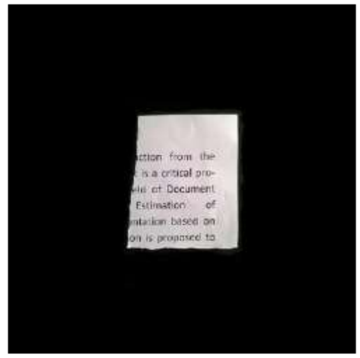


Fig 3.a. Straighten Image1 *3.b. Straighten Image2*

The rotational property of the Fast Fourier Transform (fft) states that if a function or image is rotated by the angle, then its Fourier transform also rotated by an equal amount of angle.

$$f(m,n) \longrightarrow f(r\cos\theta, r\sin\theta)$$

$$DFT\left[f(r\cos\theta, r\sin\theta)\right] \longrightarrow F\left[R\cos\phi, R\sin\phi\right]$$

$$DFT\left[f(r\cos\theta+\theta_0, r\sin\theta+\theta_0)\right] \longrightarrow F\left[R\cos\varphi+\varphi_0, R\sin\varphi+\varphi_0\right]$$

From the above equations it is clear that if an image or a function is rotated by an angle θ_0 then the spectrum of that image or function will also rotated. So horizon (fft) gives the rotational angle of the image and by rotates the image exactly opposite to that angle we get the straightened image. Syntax for rotation of the image for getting straight image is

im_straight=imrotate(A, -angle,'bicubic')

A is the input image and angle which is estimated by horizon and bicubic is interpolation method.

2.1.Image Concatenation

Concatenation of the images is the process of placed the images one after another because they are related. To perform this task we need to make the sizes of the images equal. A pre-defined function ***imresize*** help us for making the sizes equal.

Mat lab function cat, with the syntax as cat ***(DIM,A,B)*** concatenates the arrays A and B along the dimension DIM. Dimension may be first, second or third. If we use first dimension the images are place by up and down manner. And if the dimension is second then the images are placed side by side.

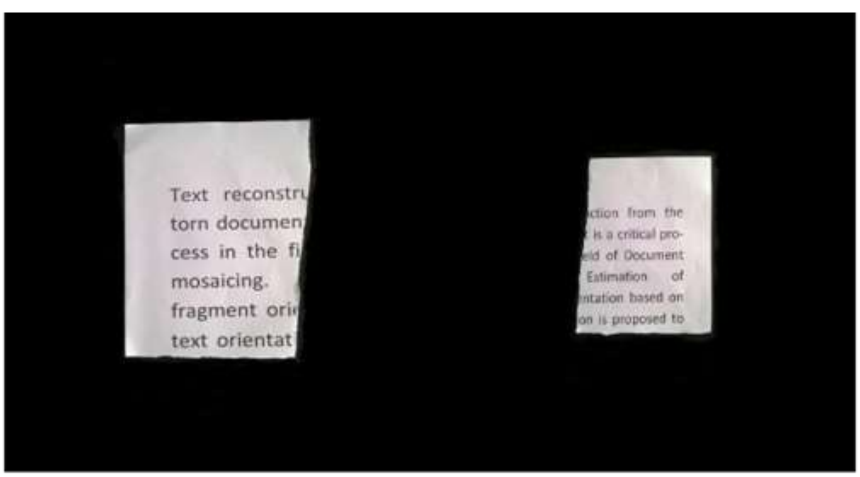


Fig .4.Concatenated Images

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

2.2.Binarisation

We are going to perform the recovery of torn fragments based on the binary image processing techniques. So it is required to convert the image to binary form. ***im2bw***is a mat lab function that converts the gray scale. Syntax for conversion of image to binary is

imbinary=im2bw(gray scale image)

Here *im2bw* produces binary images from indexed, intensity, or RGB images. To do this, it converts the given image to gray scale format (if it is not already an intensity image), and then converts this gray scale image to binary by thresholding.

The binary image contain small unnecessary components, these are effect the process while performing segmentation. So, we need to remove the details smaller than a certain reference. For this we are using image dilation process with line shaped structuring element. The dilated images are outlined for the further process.

The outlined image is consisting of text but we don't require the text in the binary image. We want the regions of the image. So fill the image region, for that ***imfill,*** whichis a mat lab function, help us to fill regions and holes. The mat lab syntax for imfill is given by

BW_filled=imfill (imbinary,'holes')

This function fills holes in the input image im1 and gives the outcome images as BW. Here a hole is a set of background pixels that can't be reached by filling in from edge of the image.

Fig.5. Images filled With Holes

3.Joining Process

Joining Process is the final stage for recovering of the document. This stage includes the region based segmentation, horizontal and vertical corrections of the image using mask creation.

3.1.Region Segmentation

The filled images are required to segment along bounding box of each image. For document to be recovered we need only the portion of the torn piece. So we need to divide the filled image using region based segmentation. So by using the region properties partition them along the boundary box of the images. The mat lab function ***regionprops*** help us to find out the region of the images.

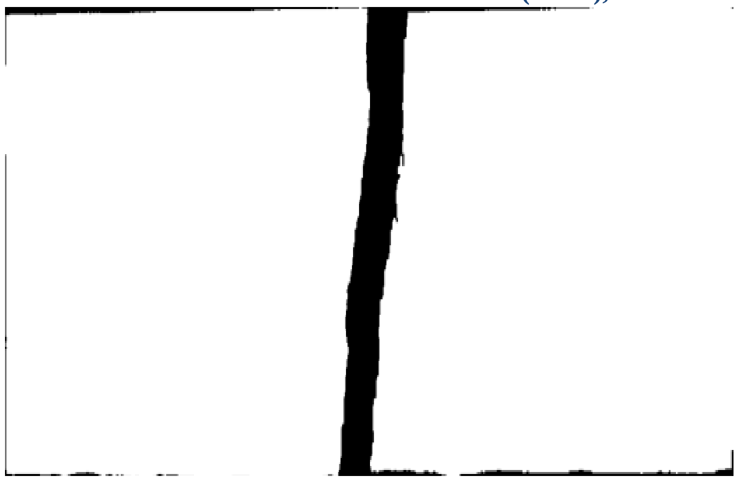


Fig 6.Image Concatenation After Region Segmentation

Mat lab syntax for finding the regional properties of the mage is given by

im_props = regionprops(image, 'all')

The mat lab function **regionprops** measures a set of properties for each connected component (object) in the binary image. By using the shape measurements Area, Bounding box we segment the images. And make the image sizes equal. The segmented images are not with the same size so by considering the first image make remaining same size equal to first image and then concatenate them.

3.1.Horizontal and Vertical Correction

Consider the image concatenated after region segmentation, it consisting of torn pieces which are placed one after another and in between them there you found zeros (black space).If we remove that zeros, we can recover the original document.For this we are going to create a mask which is of the size of the image and which consisting zeros at the top and bottom of the image, which is horizontal mask. With the size of the image left and right side of the image is going to fill with ones, which is vertical mask. Horizontal mask is given by

img (1: bodc,:) = zeros; For upper part of the image
img (lz-bodc : lz, :) = zeros; For bottom part of the size

By the applied horizontal mask we remove the area which is exactly equal size of the horizontal mask, which is called Horizontal correction. After horizontal correction the upper part and bottom part of the image shouldn't consist of zeros.

Fig 7. Image with Horizontal Mask

Now the image doesn't have any zeros at the upper and bottom part, but the left and right side of the image consisting zeros. So we need to remove them by using the vertical correction. For this create a mask with row and column size as the horizontal correction image.

Fig 8.Image After Horizontal Correction

img1(:, 1 : bodc) = ones; for left part of the image
img1(: ,bz - bodc : bz) = ones; for right part of the image
Here img1 is horizontal corrected image. And we are specifying mask width for the image

The image with vertical mask shown in the Fig. 9 and it has zeros only in between images. Removing of the black space in between the image is done by the process of vertical correction. Find the location of the zeros in vertical mask applied image and shift and then find the length of the zeros. By circular shifting the images with the length of the zeros the second image image gets combined with the host image or first image.

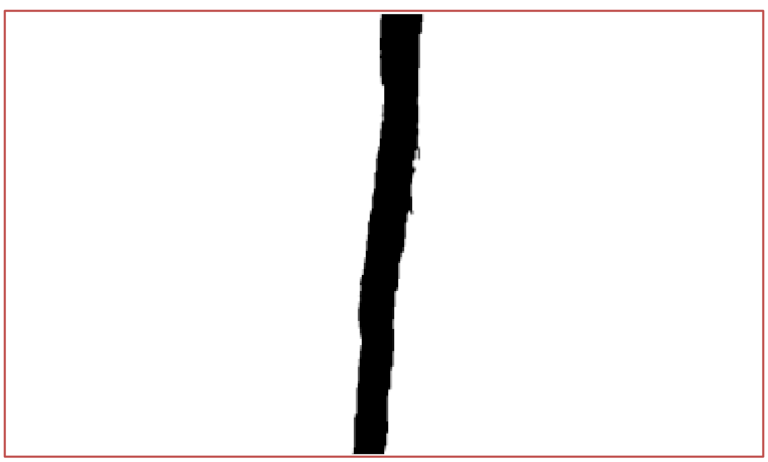


Fig. 9. Image with Vertical Mask

By applying the same operations (region properties base segmentation and horizontal and vertical correction as same as binary image) what we did for the binary image we get the recovered paper. Fig 10 shows the combine (recovered binary image).

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

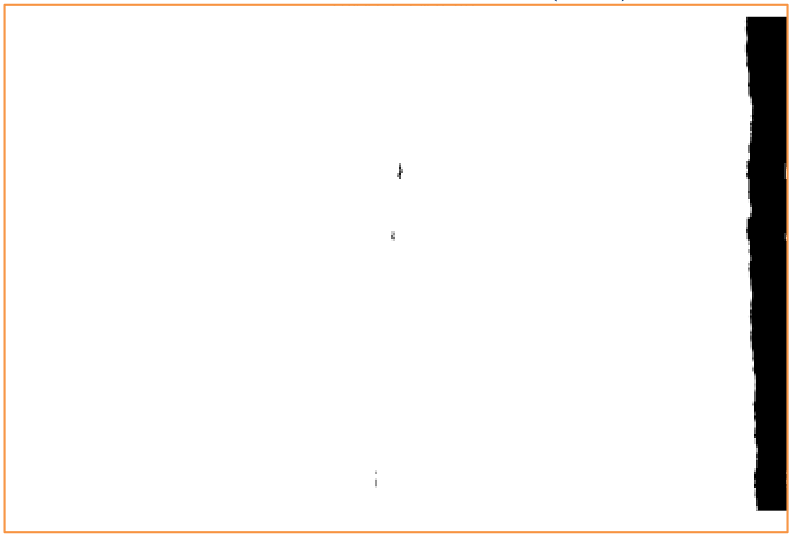


Fig 10.Recovered BinaryDocument

While processing image through mat lab software we observe that for different images the proceeding time or elapsed time of CPU varies. The same operations like region props and horizontal and vertical mask and correction of vertical and horizontal masks we get the recovered Document and text.

EXPERIMENTAL RESULTS

Fig 10.a. Input Image1

10.b.Input Image2

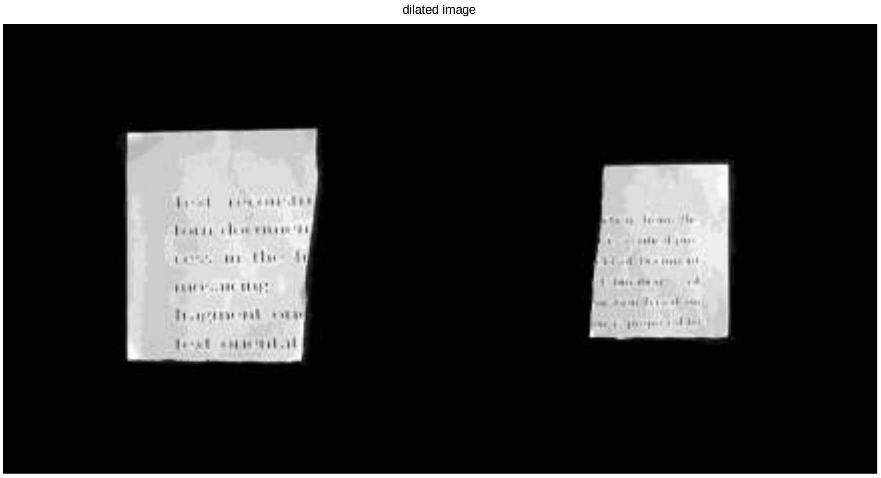
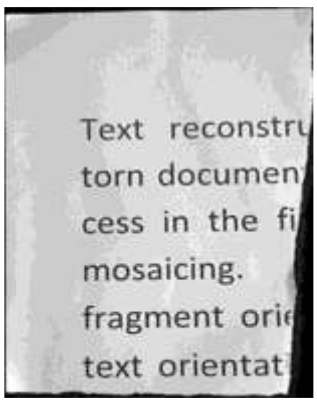


Fig 11. Dilated Image



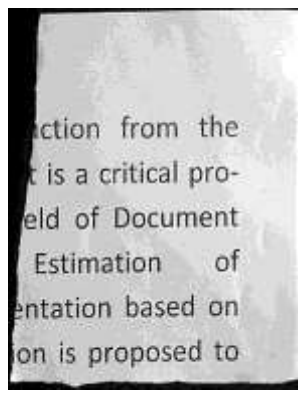


Fig 12.a.Segmemeted Gray Image1 ***12.b.Segmemeted Gray Image1***

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

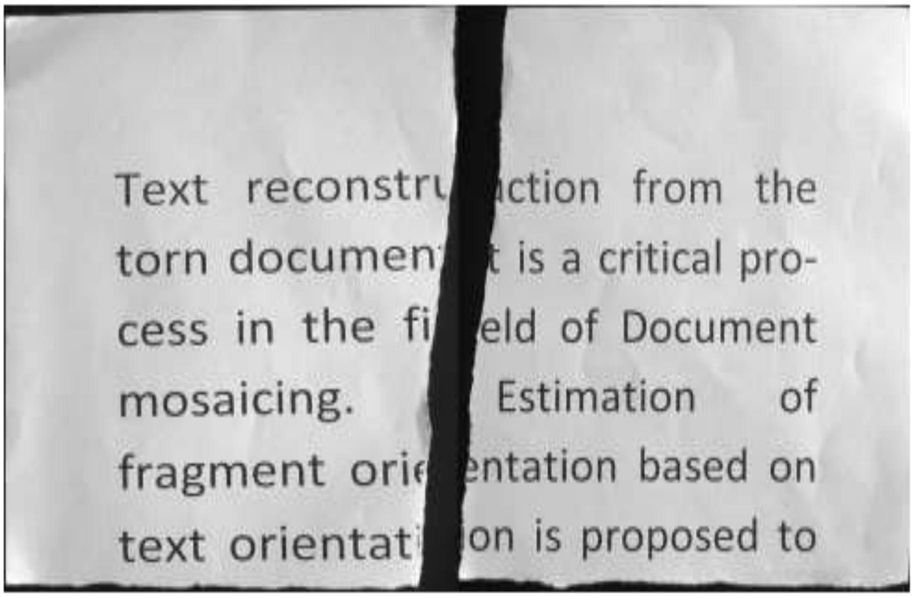


Fig.13. Gray scale Image Concatenation after Region props applied

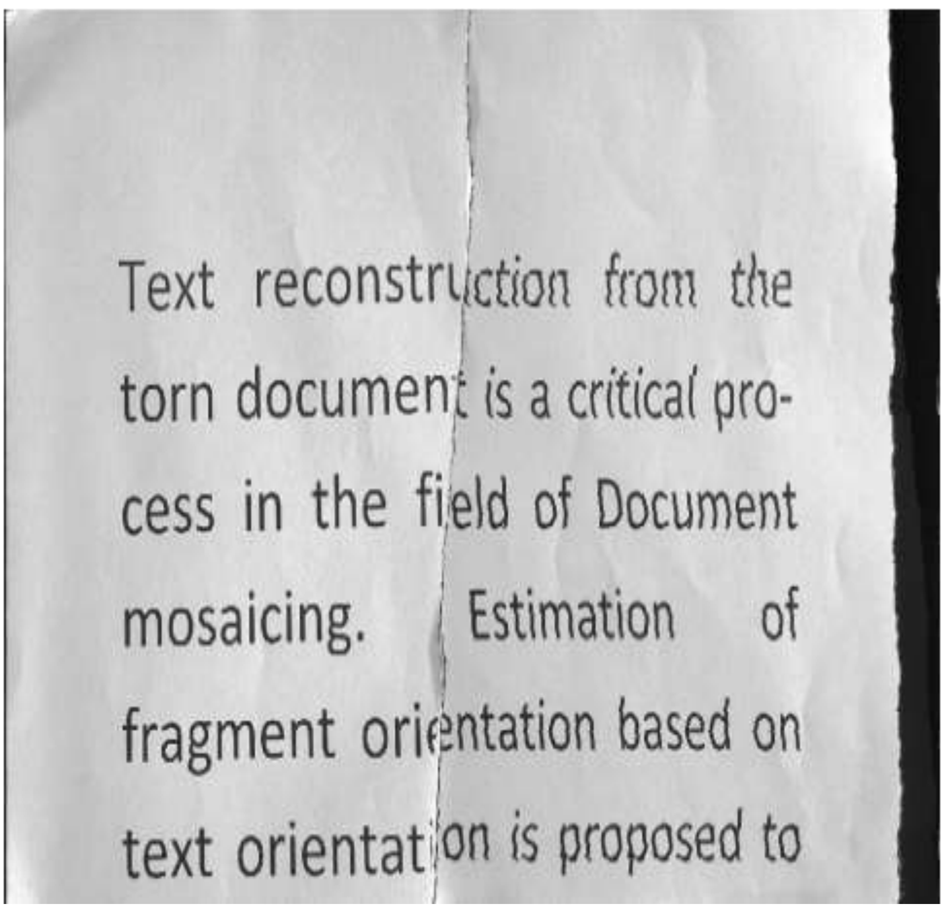


Fig. 14.Recovered Document Text

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

STATISTICAL ANALYSIS

The Table 1 Shows how the CPU time vary with the size of the images. We consider different images with different size. It is clear from the table that if the size of image increases then the CPU time or elapsed time is also increases. In the Table 1, we consider the nine image sets. Eah image set is consisting of two images. The two images in image set are with diffarent size and different rotational angle.

Table 1. Image size Vs. Elapsed time (For Different Images)

S.No	Image Size		Rotational Angle		Elapsed Time
	I1	I2	Angle1	Angle2	
1	412*320	512*512	16	-44.4000	7.1150
2	512*512	512*512	0	0.8000	7.3744
3	512*512	724*724	0	-43.900	7.4643
4	1632*920	1632*920	-3.2000	-0.2000	8.4953
5	628*779	3328*1872	43.5000	0.2000	10.5359
6	512*512	3264*1840	42.3000	-2.2000	11.3115
7	3264*1840	1028*792	-34.4000	17.3000	12.2205
8	3264*1840	920*920	2.7000	-29.3000	12.7833
9	881*1024	1840*3264	-38	-22.7000	13.8060

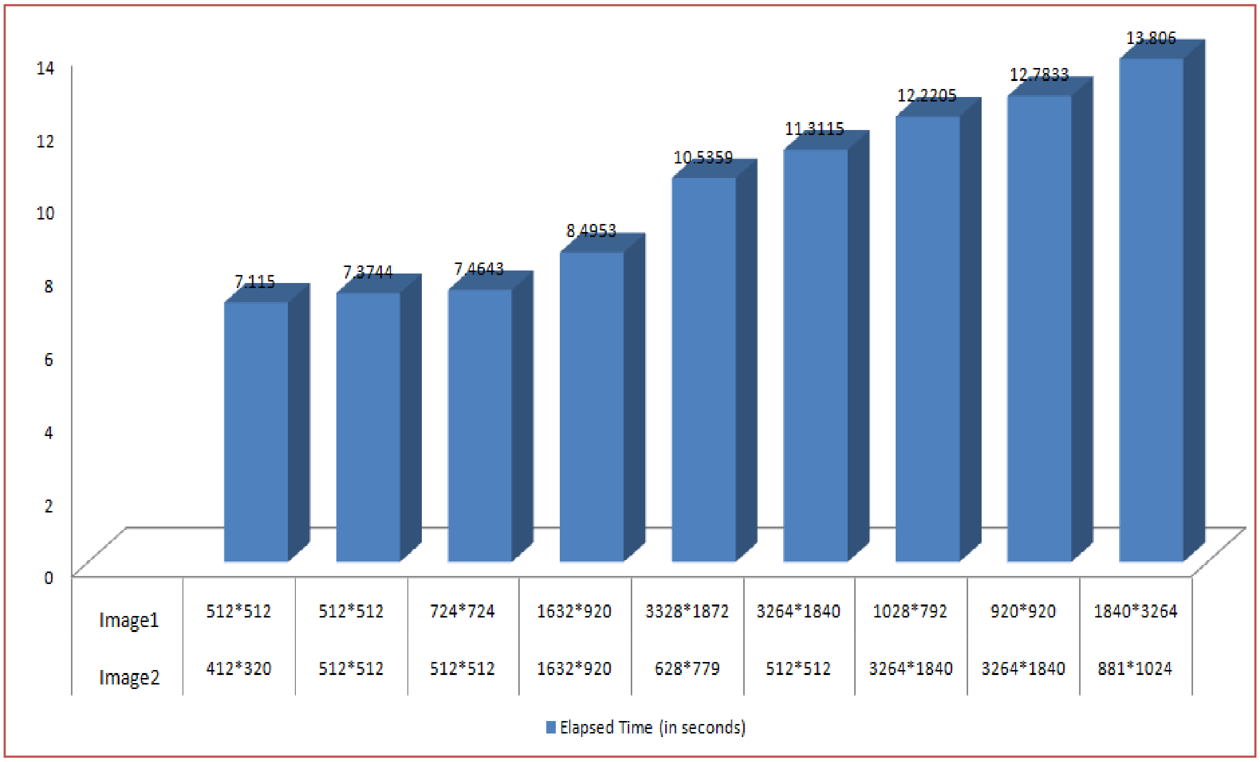


Fig 15.Image size Vs Elapsed Time

Form the above Fig 15, whichgives the performance analysis of our methodology. Here all the images we consider are with different from each other both in sizes well as in shape. Image1, Image2 are two images required for recovery of document text. We are considering nine sets of images; each set is having two images for stitching. From top to bottom the image sizes are creasing. Similarly the Elapsed Time is also increases. The image set with 412*320 and 512*512 takes 7.115 seconds and For image with 1632*920 , 1632*920 image takes 8.4953 seconds. So it is clear from the table the CPU time is directly proportional with the size of images.

CONCLUSION

In our methodology we are going to recover the document text from the torn fragments using image processing. We can recover the document text form the multiple torn fragments also. We can use the same procedure for the recovery of the document text from the color document fragments.

ACKNOWLEDGEMENTS

I extended my sincere thanks to Assoc.ProfessorDr.Mahesh, Professor Dr.S.A.K.Jilani and also my earnest thanks to Dr.A.R.Reddy, head of the department of E.C.E, MITS, Madanapalle for their support to carry out successfully of this project work.

REFERENCES

1. Dr.Mahesh, S.Mahammadinigar "Image Mosaic Using FAST Corner Detection" *International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 1, ISSN: 2278 – 909X, Issue 6*, December 2012.
2. Dr.Mahesh, R.Nagendra and K.Subramayam "SIFT Based Image Mosaic algorithm" International Journal of Communications & Engineering (IJAE) Vol.3, Issue 1, (ISSN: 0988-0382E) PP_9-15, July 2012
3. Dr.Mahesh and Y.Murali "Image Mosaic Using Speeded Up Robust Feature Detection" International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 1, Issue 3, (ISSN: 2278 – 909X), PP-40-45, September 2012
4. Dr.Mahesh, Dr. M.V. Subramanyam, "Feature Based Image Mosaic Using Steerable Filters and Harris Corner Detector" *I.J. Image, Graphics and Signal Processing,(IJIGSP)* Vol.5, No.6, *may* 2013 pp.9-15.(ISSN Online : 2074-9082, ISSN Print : 2074-9074) (*Free journal*): DOI: 10.5815/ijigsp.2013.06.02
5. Mosaicing of torn document images Nagaraj B. Patil, V. M Viswanatha& Sanjay Pande M. B in IJRRAS 7 (1) , April 2011.
6. Reconstruction of torn page using corner and segment matching by Ms.PoonamD.Kesarkar, Ms.Mridula RC. Prasad, Prof.S.L.Tade in International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 2, Issue 6, June 2013.
7. Reconstruction of Torn Documents Using Contour Maps by ArindamBiswas, ParthaBhowmick, Bhargab B. Bhattacharya in 0-7803-9134-9/05/$20.00 ©2005 IEEE.
8. Reconstructing shredded documents through feature matching by Edson Justino ,Luiz S. Oliveira ,CinthiaFreitas in Forensic Science International 160 (2006) 140–147
9. A Probabilistic Model for Reconstruction of Torn Forensic Documents by Ankush Roy, UtpalGarain in 2013 12th International Conference on Document Analysis and Recognition.

AUTHOR BIBLIOGRAPHY

	**C.Prasad** He is Pursuing M.Tech in Madanapalle Institute of Technology & Science (MITS), Department of ECE, Angallu. He received B.Tech Degree from Siddharth Institute of Engineering and Technology, Puttur, Department of Electronics & Communication Engineering (ECE), 2012. His areas of interests are Image processing ,signal processing and communication systems.
	**Dr. Mahesh** Dr. Maheshreceived the B.E. degree from Gulbarga University, Gulbarga, Karnataka, M.Tech degree from JNTU Hyderabad, A.P, and Ph.D. from JNTU Anantapur,A.P India. He is working as an Associate Prof, in the Department of Electronics and Communications Engineering, Madanapalle Institute of Technology, Madanapalle. He has more than 15 years of experience in teaching field.
	**Dr.S.A.K. Jilani** Dr.S.A.K. Jilani is professor in the Dept. of Electronics and Communication Engineering, Madanapalle Institute of Technology and Science, Madanapalle. He has been teaching for the past twelve years. Earlier he was working as an R&D professional in electronics industries. He obtained his Ph.D. degree in 2002 and has published more than 25 papers in various national and international journals. He has guided more than 50 M.Tech, M.Sc., M.C.A and B.Tech projects. Areas of his interest are artificial intelligence, computer visions, digital signal processing and embedded systems.