|  |
| --- |
| Close-up image showing the leaf-sides of two oversized books side-by-side on a bookshelf, with additional books in soft focus background |
| Document Layout Analysis  Mini report |
| |  |  |  | | --- | --- | --- | |  | 7/1/16 | Digital Image Processing | |

**Sania Irfan-B12101110**

**Tooba Aziz-B12101110**

**Objective:**

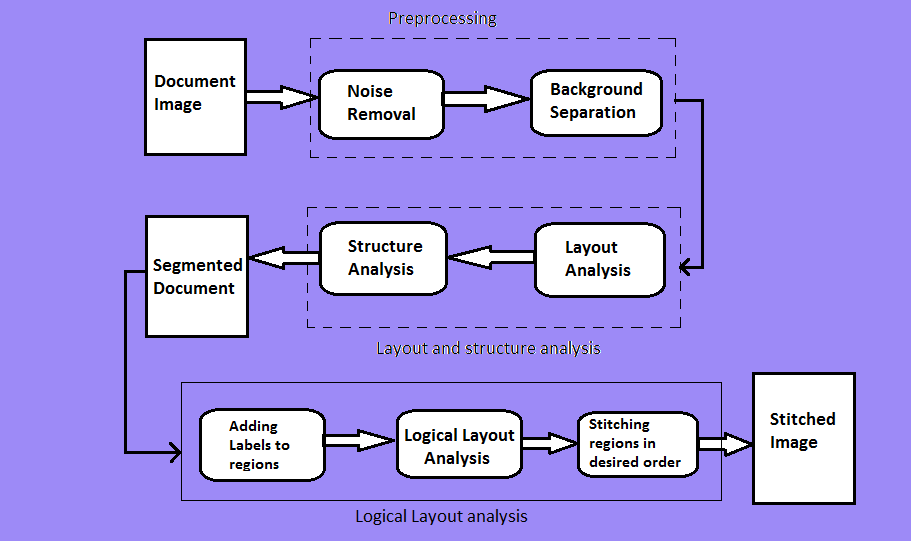
The process of document structure and layout analysis tries to decompose a given document image into its component regions and understand their functional roles and relationships.

**Description:**

A document image is composed of a variety of physical entities or regions such as text blocks, lines, words, ﬁgures, tables, and background. We could also assign functional or logical labels such as sentences, titles, captions, author names, and addresses to some of these regions. The process of document structure and layout analysis tries to decompose a given document image into its component regions and understand their functional roles and relationships. The processing is carried out in multiple steps, such as preprocessing, page decomposition, structure understanding, etc

Many documents, such as newspapers, magazines and brochures, contain very complex layout due to the placement of ﬁgures, titles, and captions, complex backgrounds, artistic text formatting, etc. A human reader uses a variety of additional cues such as context, conventions and information about language/script, along with a complex reasoning process to decipher the contents of a document. Automatic analysis of an arbitrary document with complex layout is an extremely difficult task and is beyond the capabilities of the state-of-the-art document structure and layout analysis systems. This is interesting since documents are designed to be effective and clear to human interpretation unlike natural images.   
As mentioned before, we distinguish between the physical layout of a document and its logical structure.

**Flow of work:**



**DOCUMENT PROCESSING:**

**Pre-processing:**

* Image resizing



* Noise removal



* Foreground detection (Thresholding)



**Region Segmentation:**

* Dilation

**Region Classification:**

* Column vs head (“Bounding box”)
* Column vs Image (OCR)

**Cropped image Placement:**

* Using function **cat(cell{ })**

**Coding Description:**

**Step 1:**

* Background separation was done by thresholding and inverting the image.

I = rgb2gray(img);

level = graythresh(I);

BW = im2bw(I,level);

BW=~BW;

**Step 2:**

* Structural layout analysis is done by morphological operations.

%DILATION

%FOR COLUMN and IMAGE

seType1 = strel('line',40,180);

seType2 = strel('line',50,90);

bwSub = imdilate(BW,[seType1 seType2]);

bwSub=imfill(bwSub,'holes');

%FOR HEADING

cropped=imcrop(bwSub,[1 1 3075 380]);

se=strel('square',25);

cropped=imdilate(cropped,se);

Pos=[1 1 3075 370];

r = round(Pos(1));

c = round(Pos(2));

if(r ==0)

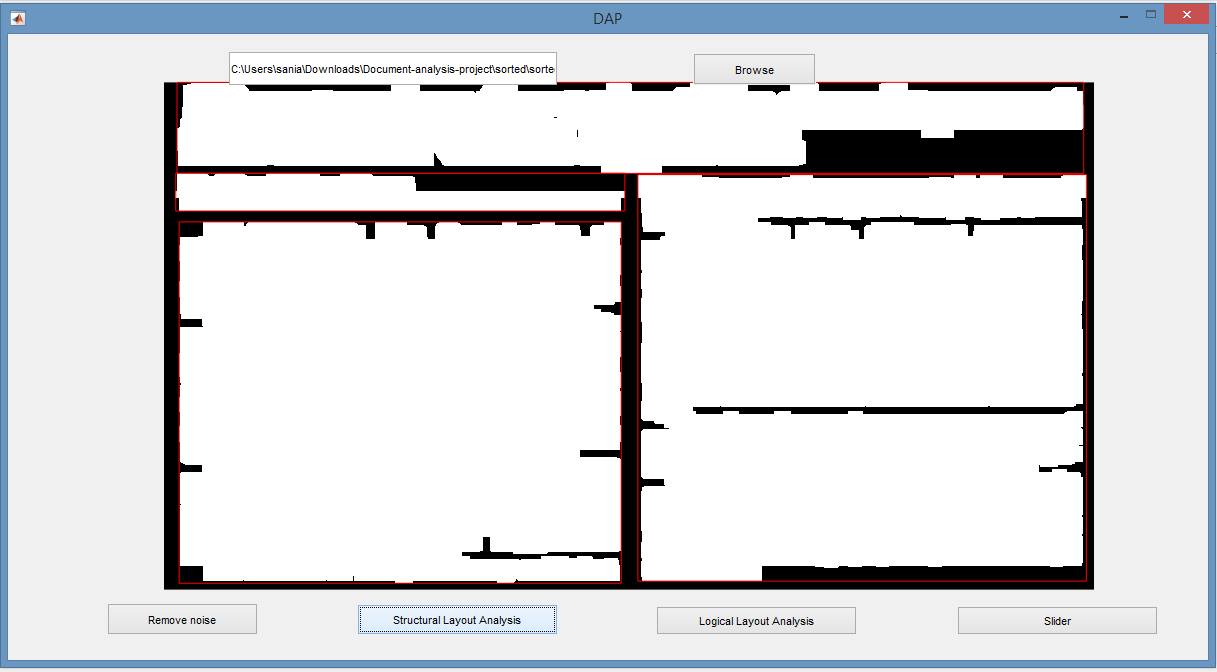
r=1;

end

sizeR = size(cropped,1);

sizeC = size(cropped,2);

bwSub(r:sizeR+r-1,c:sizeC+c-1) = cropped;



**Step 3:**

* Logical layout analysis is done by adding labels to the segmented image regions.

[Ilabel num]=bwlabel(bwSub);

stats = regionprops(Ilabel, 'BoundingBox');

Iprops=regionprops(Ilabel);

Ibox=[Iprops.BoundingBox];

Ibox=reshape(Ibox,[4 num]);

LabelImg=img;

for k=1:length(stats)

c=imcrop(LabelImg,Ibox(:,k));

c=imresize(c,[480 640]);

filename=sprintf('C:\\Users\\sania\\Desktop\\Output\_Images\\%02d.jpg',k);

imwrite(c,filename,'jpg');

results=ocr(c);

d=results.Text;

thisBB = stats(k).BoundingBox ;

if stats(k).BoundingBox(4) > 400

if d~=0

label = 'col';

else

label='image';

end

elseif stats(k).BoundingBox(4) < 400

label = 'head';

end

LabelImg = insertObjectAnnotation(LabelImg,'rectangle',[thisBB(1),thisBB(2),thisBB(3),thisBB(4)]...

,label,'TextBoxOpacity',0.5,'FontSize',70);

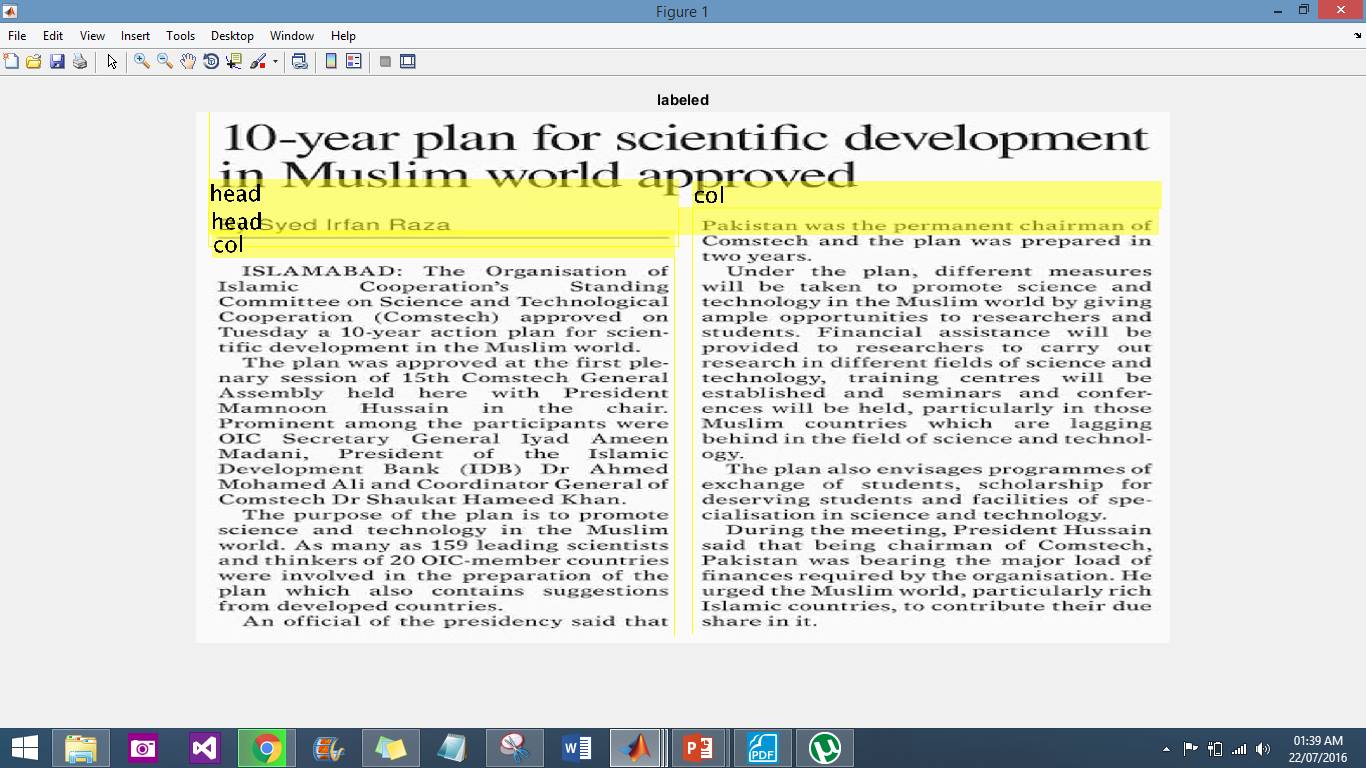
end

imshow(LabelImg);

* Textual and non-textual regions are separated by OCR.

results=ocr(c);

d=results.Text;



**Step 4:**

* Image stitching is then performed on the segments formed in the previous steps and are placed together vertically with a scrollbar on it.

Dir=fullfile('C:\Users\sania\Downloads\Document-analysis-project\Output\_Images');

paper = imageSet(Dir);

for i=1:paper.Count

imgCell{i} = read(paper,i);

end

img=cat(1,imgCell{:});

hFig = figure('Toolbar','none',...

'Menubar','none');

hIm = imshow(img);

SP = imscrollpanel(hFig,hIm);

