DOCUMENT LAYOUT ANALYSIS USING INVERSE SUPPORT VECTOR MACHINE (I-SVM) FOR HINDI NEWS PAPER IN IMAGE PROCESSING

Rimmy Kathuria

M .Tech Student of MRIU, Faridabad.

Abstract: Optical character recognition (OCR) is the translation of handwritten, typed or impressed word images into a form that the computer can manipulate. There are various steps in OCR. One of them is classification. To do the Classification, we must have to compare a training data with many feature vectors. A classifier is needed to compare the feature vector of input and the feature vector of training data using inverse support vector machine. This paper proposed classification techniques, which are used to recognize character or word, and about some related work which has been done. It will also present a novel learning based framework to extract articles from newspaper images using bounding box method. The input to the system comprises blocks of text and graphics, obtained using standard image processing techniques. The fixed point model uses contextual information and features of each block to learn the layout of newspaper images and attains a contraction mapping to assign a unique label to every block. Experimental results show the applicability of our algorithm in document newspaper layout labeling and article extraction.

Keywords: ***OCR, I-SVM, CANNY, SOBEL, Hindi newspaper script.***

I. INTRODUCTION

This Research gives an overview about document layout analysis. It presents in detail one method which is part of the pre-processing phase, called Morphological operator. In doing so it introduces a special technique of how determining the bounding box and inverse support vector machine. Furthermore

it presents numerous ways of approaches to speed up the process of document layout.The second part deals with page segmentation which is one of the huge parts in document image analysis. It illustrates a means of partitioning using bounding boxes of different entities. Reasonable success has been achieved at developing mono lingual OCR systems in Indian scripts. Scientists, optimistically, have started to look beyond. Development of bilingual OCR systems and OCR systems with capability to identify the text areas are some of the pointers to future activities in Indian scenario. The separation of text and non-text regions before considering the document layout for OCR is an important task.

In this paper, we present a biologically inspired, multichannel filtering scheme for page layout analysis. The same scheme has been used for script recognition as well. Parameter tuning is mostly done heuristically. It has also been seen to be computationally viable for commercial OCR system development.

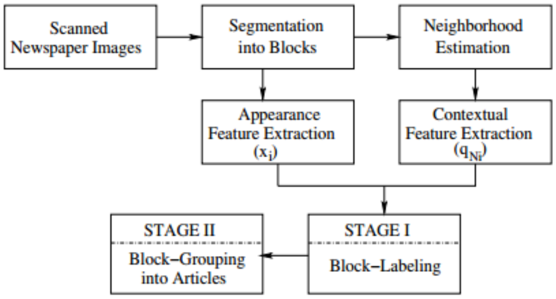
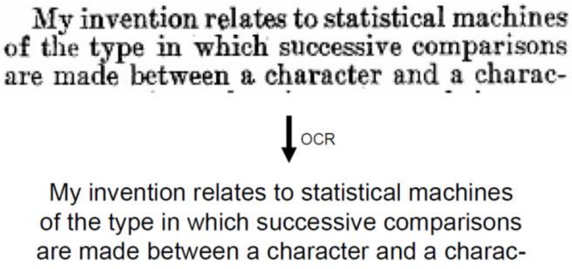

Scanned Newspaper Images
Segmentation into Blocks
Neighborhood Estimation
Appearance Feature Extraction (x_i)
Contextual Feature Extraction (q_{Ni})
STAGE II
Block–Grouping into Articles
STAGE I
Block–Labeling

Figure 1: Block diagram of the OCR

II. OCR

OCR Stands for Optical Character Recognition. It is Extracts the text from a given image. It is invented by Gustav Tauschek. Tauschek obtained a patent on OCR 1929 in Germany and 1935 in USA.


My invention relates to statistical machines
of the type in which successive comparisons
are made between a character and a charac-

My invention relates to statistical machines
of the type in which successive comparisons
are made between a character and a charac-

Figure 2: Example of OCR

III. Data for work

Data in a newspaper document analysis are usually captured by optical scanning and stored in a file of picture elements, called pixels that are sampled in a grid pattern throughout the document. These pixels may have values: OFF (0) or ON (1) for binary images, 0–255 for gray-scale images and 3 channels of 0–255 colour values for colour images.
At a typical sampling resolution of 120 pixels per centimeter, a 20 x 30 cm page would yield an image of 2400x3600 pixels. When the document is on a different medium such as microfilm, palm leaves, or fabric, photographic methods are often used to capture images. In any case, it is important to understand that the image of the document contains only raw data that must be further analyzed to collect the information.

IV. PROPOSED IMPLEMENTATION

We have used scanned images of Hindi newspapers for our experimentation. Each document provides for a number of blocks, typically an average of multiple blocks exists in a single newspaper image. We use Bounding box to obtain these blocks.

This is the most important stage of document layout OCR system. Relevant information from the selected data has been extracted for classification. Different shapes of the character parts have been selected from feature selection algorithm. They might be curves or points or linear shapes. But most of them are open curve shaped. In feature extraction, two techniques are proposed here. One is support vector based sub line direction and the other one is bounding box based shape detection. Here, directions are extracted by sub line direction and bounding box procedures from selected portion of character image.

A. Preprocessing Of newspaper data

The gray scale image is first binaries using the method described in our earlier work [2]. Horizontal and vertical lines are then removed on the basis of aspect ratio of connected components.

B. Inverse-SVM classification

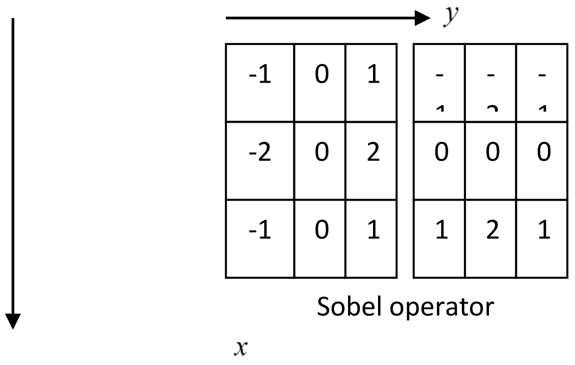
The classification strategy we have adopted combines both deterministic and probabilistic decision making. First, using training samples, we try to learn the various conflicts among classes and equivalent representations for every class. . From here, we use probabilistic classification to assign a label to the test pattern. The important point to observe here is, the switch to conflict resolution can happen at any level (from 1 to 4), since conflicts can be defined at various levels. Usually this will be dictated by certain trade-offs between accuracy and precision-recall, as we show later. The classifier we use to resolve conflicts is the Inverse Support Vector Machine (I-SVM). Though it is not mandatory to use SVMs, their utility in character recognition has been well documented. We have chosen RBF kernel, and the SVM formulation "support vector classification" for multi-class classification.

C. SOBEL OPERATOR

The most popular approximation of equation (1) but using a 3×3 mask is the following.

$$\nabla f \cong \left|(z_7+2z_8+z_9)-(z_1+2z_2+z_3)\right|+\left|(z_3+2z_6+z_9)-(z_1+2z_4+z_7)\right| \quad (1)$$

This approximation is known as the **Sobel** operator.



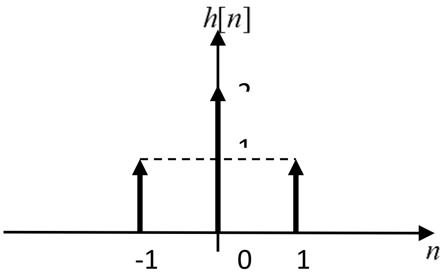
-1	0	1
-2	0	2
-1	0	1

-1	-2	-1
0	0	0
1	2	1

Sobel operator

If we consider the left of mask the Sobel operator, this causes differentiation along the y direction. If we isolate the following part of the mask and treat it as a one dimensional mask, we are interested in finding the effects of that mask. We will therefore, treat this mask as a one dimensional impulse response $h[n]$ of the form

$$h[n]=\begin{cases}1 & n=-1\\ 2 & n=0\\ 1 & n=1\\ 0 & \text{otherwise}\end{cases}$$



The above response applied to a signal $x[n]$ yields a signal $y[n]=x[n-1]+2x[n]+x[n+1]$ or in z-transform domain

$$Y(z)=(z^{-1}+2+z)X(z)\Rightarrow Y(j\omega)=2(\cos\omega+1)X(j\omega). \quad (2)$$

Therefore, $h[n]$ is the impulse response of a system with transfer function

$H(j\omega)=2(\cos\omega+1)=|H(j\omega)|$ Shown in the figure below for $[0,\pi]$. This is a low pass filter type of response. Therefore, we can claim that the Sobel operator has a differentiation effect along one of the two directions and a smoothing effect along the other direction.

D. Finding text lines

Precise identification of text lines is an important part for most OCR systems. It is also very useful for document layout analysis. Numerous methods have been proposed for text line and baseline finding. Some methods attempt to find just text lines, e.g. by using Hough transforms, projection profiles, and Radon transforms. Others find text lines as part of more general document layout analysis tasks, e.g. XY cuts, whitespace segmentation, Voronoi diagrams, and distance-based grouping. Usually, such methods start by performing a rough analysis of the layout, often based on the proximity of bounding boxes of connected components or connected components themselves. More precise base- and text-line models are employed in subsequent steps, as required. The text line fitting algorithm correctly identified all lines present in the documents when the extracted bounding boxes allowed it to do so. No spurious text lines were detected, and some short one-word lines were ignored. One particularly interesting property of this algorithm is that it allows variations in text line orientations. This permits the algorithm to be applied to document images captured by photo or video cameras rather than by scanners.

E. MORPHOLOGICAL OPERATION

A morphological operation is used to remove the noise and smooth the shape of the candidate text areas. The element used here is also cross-shaped with size 11x45. The size of the elements for the morphological operations and the geometrical constraints give to the algorithm the ability to detect text in a specific range of character sizes (12-48 pixels).

F. Canny edge detection

We use canny edge detector applied in grey scale images. Canny uses Sobel masks in order to find the edge magnitude of the image, in gray scale, and then uses non-Maxima suppression and hysteresis thresholding. With these two post-processing operations canny edge detector manage to remove no maxima pixels, preserving the connectivity of the contours. After computing the Canny edge map, dilation by an element 5x21 is performed to connect the character contours of every text line. Dilation by a cross-shaped element 5x21 is performed to connect the character contours of every text line.

G. Pseudo code of proposed method

```
1. Initialize input image from dataset
2. If input image in RGB format then convert in gray image
3. If no then go to next step
4. Now do the preprocessing step using Bounding box
5. Also apply the morphology operation
6. Then extract the feature
7. Testing and classification using Inverse support vector machine
8. Find the F-measure, Accuracy and Precision Recall value.
9. Finally get the output result
10. End
```

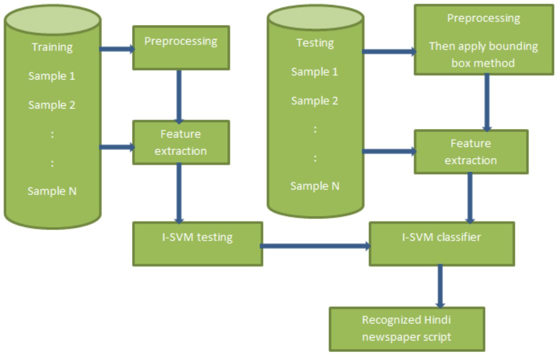
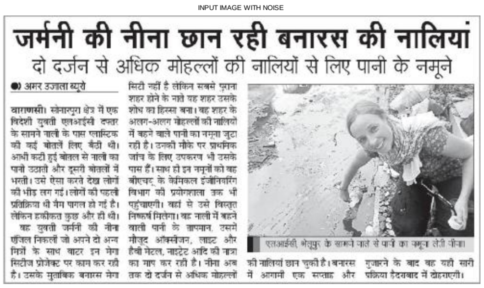


Figure 3: Proposed flow chart

V.RESULT

The present growth of document analysis of books and manuscripts demands an immediate solution to access them electronically. This requires research in the area of document image understanding, specifically in the area of document layout analysis. There is an immense scope for such a feature extraction system for a digital Document Images. This paper presents an efficient Inverse support vector machine system for a Hindi newspaper document image collection. A recognition-free approach is followed because recognition based approach is inefficient in terms of performance. The data is pre-processed and segmented for faster matching and extraction. An efficient search technique - Correlation method is used to search in large collection of document images. Performance evaluation using different datasets of documents shows the effectiveness of the approach.


INPUT IMAGE WITH NOISE

जर्मनी की नीना छान रही बनारस की नालियां

दो दर्जन से अधिक मोहल्लों की नालियों से लिए पानी के नमूने

● अमर उजाला ब्यूरो

वाराणसी। सोनारपुरा क्षेत्र में एक विदेशी युवती एलआईसी दफ्तर के सामने नाली के पास प्लास्टिक की कई बोतलें लिए बैठी थी। आधी कटी हुई बोतल से नाली का पानी उठाती और दूसरी बोतलों में भरती। उसे ऐसा करते देख लोगों की भीड़ लग गई। लोगों की पहली प्रतिक्रिया थी मैम पागल हो गई है। लेकिन हकीकत कुछ और ही थी।

वह युवती जर्मनी की नीना एंजिल निकलीं जो अपने दो अन्य मित्रों के साथ वाटर इन मेगा सिटीज प्रोजेक्ट पर काम कर रही है। उसके मुताबिक बनारस मेगा सिटी नहीं है लेकिन सबसे पुराना शहर होने के नाते यह शहर उसके शोध का हिस्सा बना। वह शहर के अलग-अलग मोहल्लों की नालियों में बहने वाले पानी का नमूना जुटा रही है। उनकी मौके पर प्राथमिक जांच के लिए उपकरण भी उसके पास हैं। साथ ही इन नमूनों को वह बीएचयू के केमिकल इंजीनियरिंग विभाग की प्रयोगशाला तक भी पहुंचाएगी। वहां से उसे विस्तृत निष्कर्ष मिलेगा। वह नाली में बहने वाली पानी के तापमान, उसमें मौजूद ऑक्सीजन, लाइट और हैवी मेटल, नाइट्रेट आदि की मात्रा का माप कर रही है। नीना अब तक दो दर्जन से अधिक मोहल्लों की नालियां छान चुकी है। बनारस में आगामी एक सप्ताह और गुजारने के बाद वह यही सारी प्रक्रिया हैदराबाद में दोहराएगी।

एलआईसी भेलूपुर के सामने नाले से पानी का नमूना लेती नीना।

Figure 5.1: Input Image with Noise

For the next test scenario we used for training only the features corresponding to Hindi newspaper OCR letters. The image used for testing contained data, and the construction of the training set, which consisted of images containing examples of each Hindi OCR letter in the Hindi alphabet. We used gray scale for image without noise.

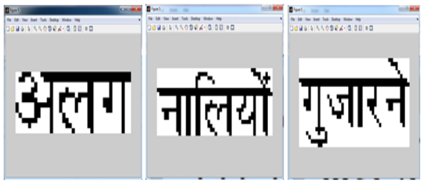
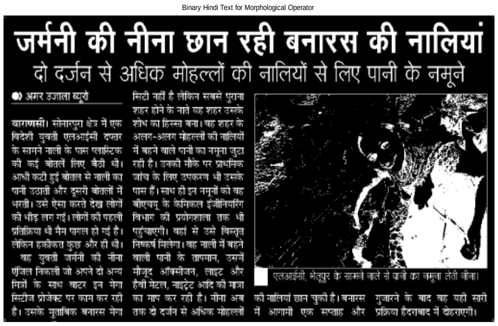


Figure 5.2: Extracted character from input image


Binary Hindi Text for Morphological Operator

जर्मनी की नीना छान रही बनारस की नालियां

दो दर्जन से अधिक मोहल्लों की नालियों से लिए पानी के नमूने

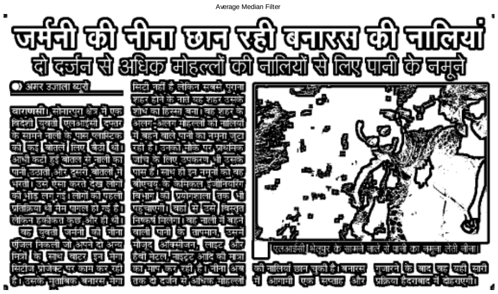
● अमर उजाला ब्यूरो

वाराणसी। सोनारपुरा क्षेत्र में एक विदेशी युवती एलआईसी दफ्तर के सामने नाली के पास प्लास्टिक की कई बोतलें लिए बैठी थी। आधी कटी हुई बोतल से नाली का पानी उठाती और दूसरी बोतलों में भरती। उसे ऐसा करते देख लोगों की भीड़ लग गई। लोगों की पहली प्रतिक्रिया थी मैम पागल हो गई है। लेकिन हकीकत कुछ और ही थी।

वह युवती जर्मनी की नीना एंजिल निकलीं जो अपने दो अन्य मित्रों के साथ वाटर इन मेगा सिटीज प्रोजेक्ट पर काम कर रही है। उसके मुताबिक बनारस मेगा सिटी नहीं है लेकिन सबसे पुराना शहर होने के नाते यह शहर उसके शोध का हिस्सा बना। वह शहर के अलग-अलग मोहल्लों की नालियों में बहने वाले पानी का नमूना जुटा रही है। उनकी मौके पर प्राथमिक जांच के लिए उपकरण भी उसके पास हैं। साथ ही इन नमूनों को वह बीएचयू के केमिकल इंजीनियरिंग विभाग की प्रयोगशाला तक भी पहुंचाएगी। वहां से उसे विस्तृत निष्कर्ष मिलेगा। वह नाली में बहने वाली पानी के तापमान, उसमें मौजूद ऑक्सीजन, लाइट और हैवी मेटल, नाइट्रेट आदि की मात्रा का माप कर रही है। नीना अब तक दो दर्जन से अधिक मोहल्लों की नालियां छान चुकी है। बनारस में आगामी एक सप्ताह और गुजारने के बाद वह यही सारी प्रक्रिया हैदराबाद में दोहराएगी।

एलआईसी भेलूपुर के सामने नाले से पानी का नमूना लेती नीना।

Figure 5.3: Binary image


Average Median Filter

जर्मनी की नीना छान रही बनारस की नालियां

दो दर्जन से अधिक मोहल्लों की नालियों से लिए पानी के नमूने

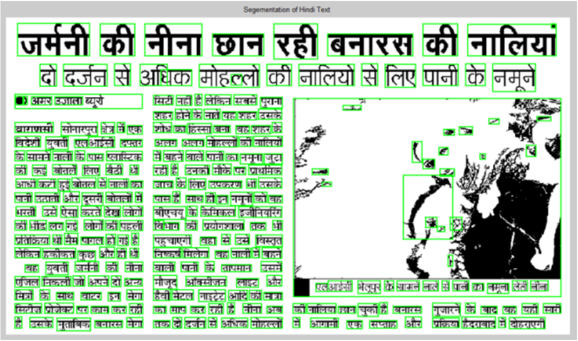
● अमर उजाला ब्यूरो

वाराणसी। सोनारपुरा क्षेत्र में एक विदेशी युवती एलआईसी दफ्तर के सामने नाली के पास प्लास्टिक की कई बोतलें लिए बैठी थी। आधी कटी हुई बोतल से नाली का पानी उठाती और दूसरी बोतलों में भरती। उसे ऐसा करते देख लोगों की भीड़ लग गई। लोगों की पहली प्रतिक्रिया थी मैम पागल हो गई है। लेकिन हकीकत कुछ और ही थी।

वह युवती जर्मनी की नीना एंजिल निकलीं जो अपने दो अन्य मित्रों के साथ वाटर इन मेगा सिटीज प्रोजेक्ट पर काम कर रही है। उसके मुताबिक बनारस मेगा सिटी नहीं है लेकिन सबसे पुराना शहर होने के नाते यह शहर उसके शोध का हिस्सा बना। वह शहर के अलग-अलग मोहल्लों की नालियों में बहने वाले पानी का नमूना जुटा रही है। उनकी मौके पर प्राथमिक जांच के लिए उपकरण भी उसके पास हैं। साथ ही इन नमूनों को वह बीएचयू के केमिकल इंजीनियरिंग विभाग की प्रयोगशाला तक भी पहुंचाएगी। वहां से उसे विस्तृत निष्कर्ष मिलेगा। वह नाली में बहने वाली पानी के तापमान, उसमें मौजूद ऑक्सीजन, लाइट और हैवी मेटल, नाइट्रेट आदि की मात्रा का माप कर रही है। नीना अब तक दो दर्जन से अधिक मोहल्लों की नालियां छान चुकी है। बनारस में आगामी एक सप्ताह और गुजारने के बाद वह यही सारी प्रक्रिया हैदराबाद में दोहराएगी।

एलआईसी भेलूपुर के सामने नाले से पानी का नमूना लेती नीना।

Figure 5.4: Image after applying Median filter on binary image.


Segmentation of Hindi Text

जर्मनी की नीना छान रही बनारस की नालियां

दो दर्जन से अधिक मोहल्लों की नालियों से लिए पानी के नमूने

● अमर उजाला ब्यूरो

वाराणसी। सोनारपुरा क्षेत्र में एक विदेशी युवती एलआईसी दफ्तर के सामने नाली के पास प्लास्टिक की कई बोतलें लिए बैठी थी। आधी कटी हुई बोतल से नाली का पानी उठाती और दूसरी बोतलों में भरती। उसे ऐसा करते देख लोगों की भीड़ लग गई। लोगों की पहली प्रतिक्रिया थी मैम पागल हो गई है। लेकिन हकीकत कुछ और ही थी।

वह युवती जर्मनी की नीना एंजिल निकलीं जो अपने दो अन्य मित्रों के साथ वाटर इन मेगा सिटीज प्रोजेक्ट पर काम कर रही है। उसके मुताबिक बनारस मेगा सिटी नहीं है लेकिन सबसे पुराना शहर होने के नाते यह शहर उसके शोध का हिस्सा बना। वह शहर के अलग-अलग मोहल्लों की नालियों में बहने वाले पानी का नमूना जुटा रही है। उनकी मौके पर प्राथमिक जांच के लिए उपकरण भी उसके पास हैं। साथ ही इन नमूनों को वह बीएचयू के केमिकल इंजीनियरिंग विभाग की प्रयोगशाला तक भी पहुंचाएगी। वहां से उसे विस्तृत निष्कर्ष मिलेगा। वह नाली में बहने वाली पानी के तापमान, उसमें मौजूद ऑक्सीजन, लाइट और हैवी मेटल, नाइट्रेट आदि की मात्रा का माप कर रही है। नीना अब तक दो दर्जन से अधिक मोहल्लों की नालियां छान चुकी है। बनारस में आगामी एक सप्ताह और गुजारने के बाद वह यही सारी प्रक्रिया हैदराबाद में दोहराएगी।

एलआईसी भेलूपुर के सामने नाले से पानी का नमूना लेती नीना।

Figure 5.5: Segmented image

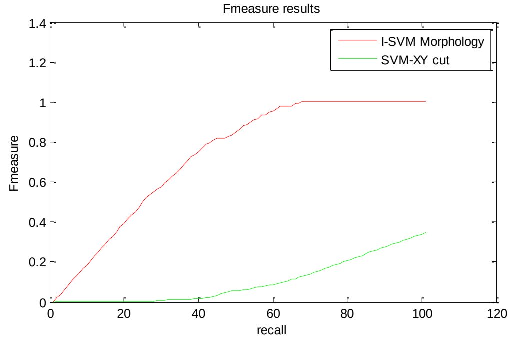


Figure 5.6: comparisons between ISVM-morphology and SVM X-Y cut for f-measure and recall value

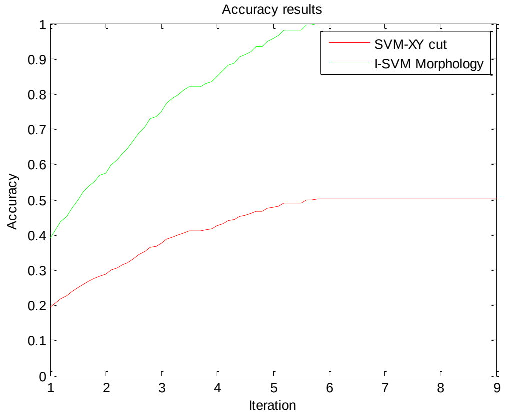


Figure 5.7 comparisons between ISVM-morphology and SVM X-Y cut for Accuracy on iteration value

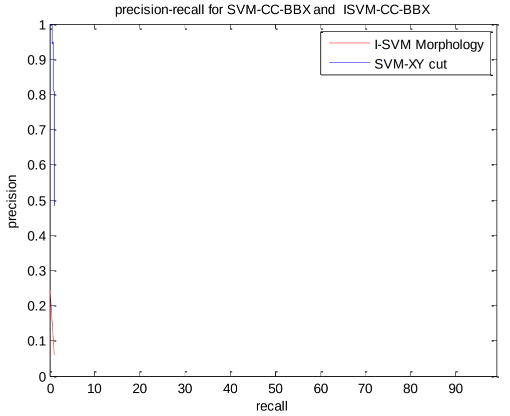


Figure 5.8 comparisons between ISVM-morphology and SVM X-Y cut for precision and recall value

Evaluation strategy

A text line must have influence to the final evaluation measure proportional to the number of containing characters and not to the number of its pixels

The number of characters in a box cannot be defined by the algorithm but it can be approximated by the ratio width/height of the bounding box

$$\text{Recall} := \frac{\sum_{i=1}^{N} \frac{EGD_i}{hg_i^2}}{\sum_{i=1}^{N} \frac{EG_i}{hg_i^2}}$$

$$\text{Precision} := \frac{\sum_{i=1}^{M} \frac{EDG_i}{hd_i^2}}{\sum_{i=1}^{M} \frac{ED_i}{hd_i^2}}$$

$$F = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

hg_i is the height of the ith ground truth bounding box . hd_i is the height of the ith detection bounding box.

EGi is the number of pixel of the ith ground truth bounding box.

EDi is the number of pixel of the ith detection bounding box.

EGD is the number of pixel of the intersection that belongs to ith ground truth bounding box.

EGDi is the number of pixel of the intersection that belongs to ith detection bounding box.

MSE = 13.1506

MSE=13.1506

PSNR = 85.0606

VI.CONCLUSION

In this paper, document layout analysis text extraction techniques such as I-SVM, bounding box, sobel operator, morphological method etc. have been discussed. The performance comparison of these methods for document text extraction on the basis of accuracy, precision rate, recall rate, processing time has been done. It is observed that better accuracy is best for document texture Analysis) approach and edge based text extraction techniques. Precision and recall rate is best in case of I-SVM algorithm and segmentation method.

Future work

We plan to exploit the colour homogeneity of text temporal, text detection from frame to frame.Multi-frame integration for image enhancement

REFERENCES

[1] Vijay Singh and Bhupendra Kumar "Document layout analysis for Indian newspapers using contour based symbiotic approach" 2014 International Conference on Computer Communication and Informatics (*ICCCI* -2014), Jan. 03 – 05, 2014, Coimbatore, INDIA.

[2] S. Malakar, S. Halder, R. Sarker, N. Das, S. Basu, M. Nasipuri, *Text line Extraction from Handwritten Document pages using spiral run length smearing algorithm*, International Conference on communications, Devices and Intelligent Systems, Kolkata, Dec. 28-29 (2012) 616-619.

[3] S.J. Ha, B. Jin, N.I. Cho, *Fast Text Line Extraction in Document Images,* 19th IEEE International Conference on Image Processing, Orlando, Sept. 30-Oct 3 (2012) 797-800.

[4] S.V. Seeri, S. Giraddi, Prashant B.M, *A Novel Approach for Kannada Text Extraction*, Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, Tamil Naidu, Mar. 21-23 (2012) 444-448.

[5] Z. Li, J. Luo, *Resolution Enhancement from Document Images for Text Extraction,* 5th International Conference on Multimedia and Ubiquitous Engineering, Loutraki, June 28-30 (2011) 251-256.

[6] D. Zaravi, H. Rostami, A. Malahzaheh, S.S Mortazavi, *Journals Subheadlines Text Extraction Using Wavelet Thresholding and New Projection Profile*, World Academy of Science, Engineering and Technology, 49 (2011) 686-689.

[7] T.V. Hoang, S. Tabbone, *Text Extraction From Graphical Document Images Using Sparse Representation*, International Workshop on Document Analysis Systems, June 9-11 (2010) 143-150.

[8] P. Nagabhushan, S. Nirmala, *Text Extraction in Complex Color Document Images for Enhanced Readability*, Intelligent Information Management, 2 (2010) 120-133.

[9] D. Dunn, W. E. Higgins, and J. Wakeley, "Texture segmentation using 2-D Gabor elementary functions," IEEE transaction on Pattern Analysis and
Machine Intelligence, vol. 16, no. 2, pp. 130-149,
1994.

[l0] W. Chan and G. Coghill, "Text analysis using local energy," Pattern Recognition, vol. 34,
pp. 2523-2532,2001.

[l1] U. Pal and B. B. Chaudhuri, "Script line separation from Indian muliti-script document," in Proceedings of the International Conference on Document Analysis and Recognition, pp. 406-409, 1999.

[I21 D. Dhanya, A. *G.* Ramakrishnan, and P. B. Pati,
"Script identification in printed bilingual docuements," Sadhana, vol. 27, pp. 73-82,2002.

[13] Chih-Wei Hsu, Chih-len Lin. A comparison of methods for multiclass support vector machines. IEEE Trans Neural Networks. 13: 415-425, March (2002).

[14] Vijay singh and Bhupendra kumar "*document layout analysis for Indian newspapers using contour based symbiotic approach",* 2014 International Conference on Computer Communication and Informatics (*ICCCI* - 2014),IEEE.

[15] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[16] F. Liu, Y. Luo, D. Hu, and M. Yoshikawa. A new component based
algorithm for newspaper layout analysis. pages 1176 1180. IEEE Computer Society, 2001.

[17] S. Mao, A. Rosenfeld, and T. Kanungo. Document structure analysis algorithms: a literature survey. volume 5010 of SPIE Proceedings, pages 197–207. SPIE, 2003.