OTCYMIST: Otsu-Canny Minimal Spanning Tree for Born-Digital Images

Deepak Kumar and A G Ramakrishnan
Medical Intelligence and Language Engineering Laboratory
Department of Electrical Engineering, Indian Institute of Science
Bangalore - 560012, India
deepak,ramkiag@ee.iisc.ernet.in

***Abstract*—Text segmentation and localization algorithms are proposed for the born-digital image dataset. Binarization and edge detection are separately carried out on the three colour planes of the image. Connected components (CC's) obtained from the binarized image are thresholded based on their area and aspect ratio. CC's which contain sufficient edge pixels are retained. A novel approach is presented, where the text components are represented as nodes of a graph. Nodes correspond to the centroids of the individual CC's. Long edges are broken from the minimum spanning tree of the graph. Pairwise height ratio is also used to remove likely non-text components. A new minimum spanning tree is created from the remaining nodes. Horizontal grouping is performed on the CC's to generate bounding boxes of text strings. Overlapping bounding boxes are removed using an overlap area threshold. Non-overlapping and minimally overlapping bounding boxes are used for text segmentation. Vertical splitting is applied to generate bounding boxes at the word level. The proposed method is applied on all the images of the test dataset and values of precision, recall and H-mean are obtained using different approaches.**

***Keywords*-Binarization; Edge detection; Minimum spanning tree; Text segmentation; Text localization;**

I. Introduction

Text embedded in landmarks, milestones, display boards, hoardings and building names acts as one of the visual aids. Text, if present in an image, contributes to tagging information. Simon et.al initiated a robust reading competition in International Conference for Document Analysis and Recognition (ICDAR) in 2003 for scenic images [1] and now there are few publicly available datasets [2]. Karatzas et.al organized a competition on reading text from Web and Email images [3]. An image with text superimposed by a software is known as born-digital image. Born-digital images are used in web pages and e-mail as name, logo or ads. Figure 1 shows two sample images from Born-Digital image dataset. The resolution of the text present in the image and anti-aliasing of text and the background form the major differences between scenic and born-digital images. Unlike born-digital images, a large amount of variation in the background may be present in scenic images due to illumination changes. Algorithms that extract text from scenic and/or born-digital images act as points of reference for tagging text in videos. In this paper, we propose an algorithm for text segmentation and localization for born-digital images. This Otsu-Canny Minimum spanning tree (OTCYMIST) algorithm assumes prior information about the presence of text in the image.

Figure 1: Sample images from Born-Digtial image competition dataset [3].

The proposed algorithm is the winner of Text Segmentation Task in the Competition organized during ICDAR 2011 [3]. It consists of binarization, edge detection, minimum spanning tree formation and horizontal text grouping. Each process is independent and acts as a module in the chain of the OTCYMIST algorithm.

II. Segmentation

Initially k-means clustering was used to form clusters. In training dataset, due to low resolution characters, small width characters were merged into the background cluster. Therefore, a thresholding algorithm was used in segmentation task. Each of the R, G and B colour planes are separately segmented using 'Otsu' global thresholding technique [4]. Figure 2 shows the block diagram of the initial part of the proposed algorithm. In Otsu's method, the threshold calculation is posed as an optimization problem and it is maximized. This thresholding scheme is used in binarization of document images, and works well for images of good quality. This threshold will split the histogram into two parts. The threshold value is calculated as shown below:

$$\sigma_B^2(k*) = \max \frac{[\mu_T\omega(k) - \mu(k)]^2}{\omega(k)[1-\omega(k)]} \tag{1}$$

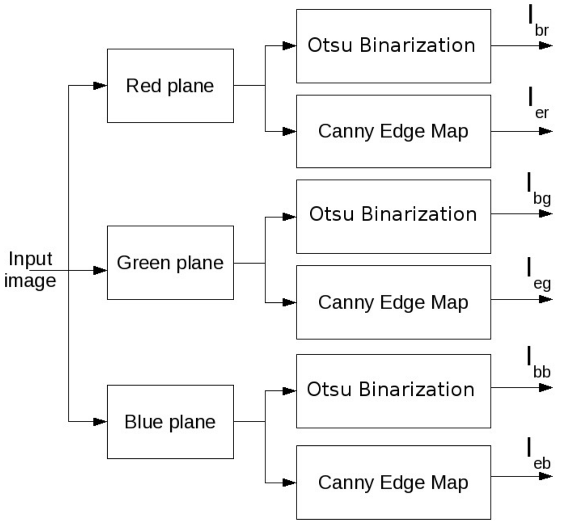


Figure 2: The first part of the proposed OTCYMIST method. This comprises the binarization and edge map modules for the individual colour channels.

where,

$$\omega(k) = \sum_{i=1}^{k} p_i \tag{2}$$

$$\mu(k) = \sum_{i=1}^{k} ip_i \tag{3}$$

$$\mu_T = \sum_{i=1}^{L} ip_i \tag{4}$$

Here, 'L' is the total number of gray levels and 'p_i' is the normalized probability distribution from histogram of the image.

The connected components (CC's) in the binarized image are labeled. The complement version of binarized image is also considered for possible text components, to account for the presence of inverse text.

Edges are detected for individual colour planes using Canny's [5] edge detection algorithm. The edge pixels are termed as edgels. These edgels are used to select the CC's from the binarized image.

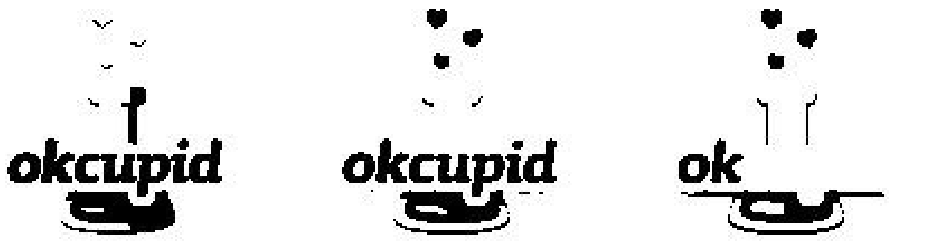
The proposed algorithm is split into sections for ease of description. The processing involved in some of the sections are required to be performed repeatedly on individual colour planes. Figure 3 shows the binarized result for each of the colour channels, for one of the sample images shown in Figure 1.

III. Pruning of connected components

Figure 4 shows the second section of the OTCYMIST algorithm. This part of the processing is carried out on each of the binarized planes and their complements. Complement form is used to counter the polarity of the text existing in an image. The area, bounding box and pixel list are obtained for each CC and are thresholded to remove possible non-text components. During the training set validation period, we observed a lot of stray pixels resulting from thresholding. The CC's with more than 5 pixels were only retained, thereby removing the stray pixels. Aspect ratio is the ratio of height to width. Aspect ratio is calculated for each CC using the dimensions of the bounding box of CC. The range of aspect ratio is fixed between 0.01 to 20. The CC's which have pixels on the image boundary are removed; this ensures that broken characters do not appear in the next step of the algorithm. In the case of low resolution text characters, there will be a displacement one pixel in the Canny edge operation. Hence, the CC's are morphologically thickened. The thickened CC's are placed on the edge map of the colour plane and the number of edgels present within the thickened CC is counted. Non-text components do not have edgels within it. The components with edgel count less than 6 are removed. The training dataset had only English characters. The maximum Euler number for English characters is 2. The CC's whose Euler number is greater than 2 are also removed. The CC's preserved in the binarized plane and its complemented version are morphologically thinned and clubbed together to form a single thinned image plane. Figure 5 shows the resulting individual thinned image planes for the first sample image shown in Figure 1.

IV. Minimum spanning tree

Figure 6 shows another section of the algorithm, which is performed separately on each of the thinned images obtained



(a) Binarized red plane (I_{br}) (b) Binarized green plane (I_{bg}) (c) Binarized blue plane (I_{bb})

Figure 3: Results of Otsu's binarization of colour channels for the first image in Fig. 1.

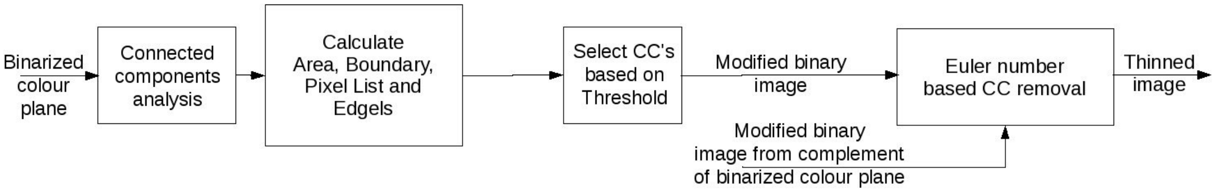
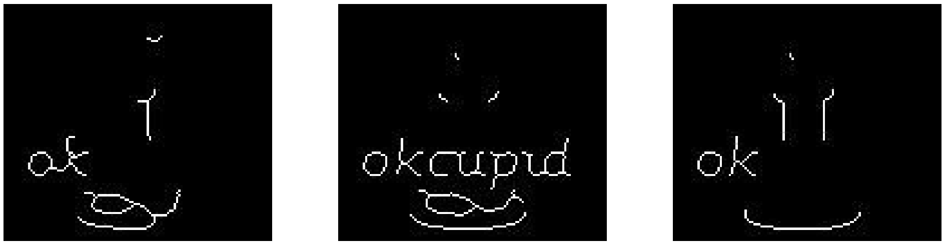

Binarized colour plane
Connected components analysis
Calculate Area, Boundary, Pixel List and Edgels
Select CC's based on Threshold
Modified binary image
Euler number based CC removal
Thinned image
Modified binary image from complement of binarized colour plane

Figure 4: Second section of OTCYMIST algorithm separately processes each binarized colour plane (I$_{br}$, I$_{bg}$ and I$_{bb}$) and its complement to form a thinned image.



(a) Thinned red plane (I$_{tr}$) (b) Thinned green plane (I$_{tg}$) (c) Thinned blue plane (I$_{tb}$)

Figure 5: Thinned planes obtained after the merging operation in second section, where the text and non-text components are easily identified.

from the three colour channels. The connected components are labeled in each of these images. The centroid of these CC's are used as the nodes of a graph. In a text string, the characters are closer when compared to non-text components. Minimum spanning tree is generated for the graph using tree algorithm proposed by Prim [6]. Prim's algorithm uses shortest distance between the nodes and bridges the nodes to generate spanning tree for the graph. Isolated nodes with long edges are broken using a length threshold 2.5 times the mean value of the edges present in the spanning tree. These isolated nodes are presumed to be non-text components appearing in the thinned image.

A text string has characters of similar height. The preserved nodes of spanning tree are subjected to mutual height ratio test. The height of each CC is compared with other CC's. If more than two CC's have height ratio in the range 0.5 to 2, then that CC is retained. The retained CC's are converted into nodes and another pass of minimum spanning tree is carried out. The mean value of the edges from minimum spanning tree is reduced if non-text CC's have been removed. Some of the non-text components with heights similar to that of text components may not be removed by the above check. These non-text components are usually far away from the string of characters. Second pass of minimum spanning tree is to ensure the removal of those non-text components also.

The result of this section is used to retrieve the actual connected components from the binarized image. The connected components from all the three binarized images are grouped into a single plane for text segmentation and localization.

V. Horizontal grouping of text components

Figure 7 shows the final section of the algorithm. Since most of the training images had only horizontal text, we horizontally group appropriate CC's into words. The CC's from the combined image plane are sorted in ascending order, based on the earliest occurring top row of the CC. The bounding box and centroid value of each CC are obtained. The CC's with centroids falling within the vertical range of the bounding box of a candidate CC, are grouped together. This process is repeated until all CC's have been grouped. Figure 8 shows CC's before and after horizontal

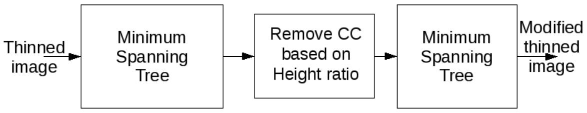


Figure 6: Minimum spanning tree and height ratio check in the proposed OTCYMIST method. This section operates on each of the three thinned planes (I$_{tr}$, I$_{tg}$ and I$_{tb}$) obtained from the previous section.

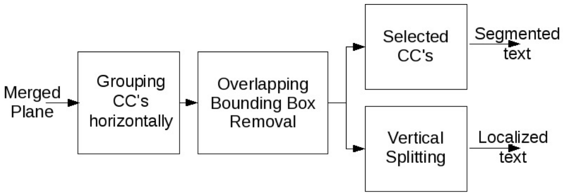


Figure 7: The final section of the OTCYMIST method which segments and localizes the text. Vertical splitting block locates the individual words from a group of words.

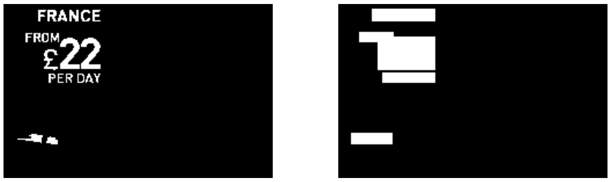


Figure 8: CC's before horizontal grouping process. After grouping, each group is replaced by a white mask, five distinctive groups can be found.

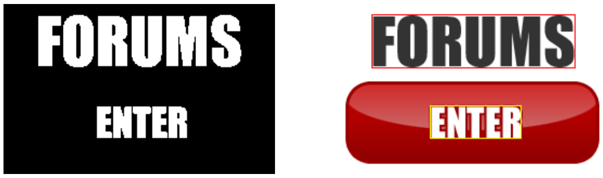


Figure 10: Segmentation and Localization result of OTCYMIST algorithm on the second image in Fig. 1, where text has both polarities.

Table I: Text segmentation results of ICDAR 2011 Robust Reading Competition: Challenge-1 [3] evaluated using [7].

Method	Well Segmented	Merged	Lost	Recall	Precision	H-mean
Anthimopoulos	72.98	12.18	14.83	86.15	71.19	77.96
OTCYMIST	64.14	15.69	20.15	80.62	72.06	76.10
Textorter	58.12	9.50	32.37	65.23	63.63	64.42
SASA	41.58	10.97	47.43	71.68	55.44	62.52

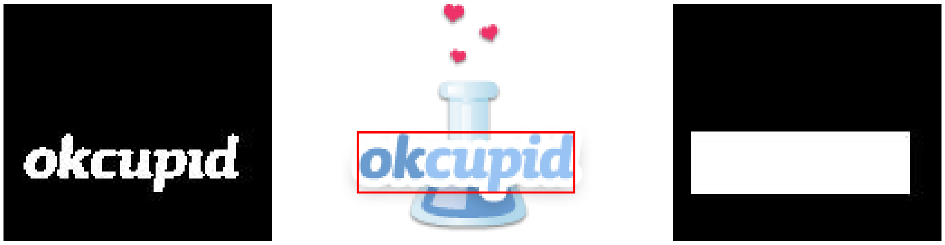
grouping process. Through validation, we noted that non-text components which pass through the minimum spanning tree module had large overlap with other components. Hence, the bounding box values of grouped CC's are tested for overlap condition, and those with high overlap are removed. We found that a bounding box should be removed if the area of its overlap with another BB is more than 20 percent of its own area. Bounding boxes containing a single CC are also removed, since we are interested only in bounding boxes containing words.

The CC's from the selected bounding boxes are preserved as segmented text at the level of pixels. Bounding boxes containing more than 4 CC's are tested for the presence of multiple words. If the maximum value of gap between two neighboring CC's is higher than the mean value of the inter-component gaps by more than 3 pixels, then a threshold of (max value − 2) pixels is used to split the bounding box into two. If the height of the bounding box of the horizontally grouped CC's is less than 11 pixels, a fixed threshold of 3 pixels is used to split the bounding box. This splitting into individual words is termed as vertical splitting. After this, the bounding box values are calculated and the text localization details are created at the output stage.

This final section of the OTCYMIST algorithm gives two outputs. Selected CC's form the segmented text at the pixel level. Vertical splitting block outputs the final bounding boxes. The bounding box list provides the text localization information. Figure 9 shows the results of segmentation and localization of the text for the first example image shown in Figure 1. Figure 10 shows the results of text and inverse text obtained using our algorithm for the second example image shown in Figure 1 .

VI. Experimentation

The total number of training images given is 420 [3]. The ground truth of training dataset was in pixel format for text segmentation and bounding box format for text localization. In the training images, the location and the number of words and characters vary randomly. This ground truth of training samples were used for tuning parameters in the cross validation process. During testing phase, the competition organizers provided 102 images. Text localization performance is evaluated as described in [8]. ABBYY OCR software was also used in Text Localization as Baseline Method [9]. The performance of the baseline method is comparable to the best performing algorithm. Text segmentation performance, as described in [7], is used as primary evaluation for benchmarking algorithms and precision, recall values at pixel level were used as secondary indices. Columns titled 'Well Segmented', 'Merged' and 'Lost' in Table I indicate recognition performance on the segmented image. Our OTCYMIST algorithm is the winner of text segmentation task in the competition and is ranked



(a) Text Segmentation (b) Text Localization (c) Text Localization mask

Figure 9: Result of the proposed OTCYMIST algorithm for the born-digital image in Fig. 1.

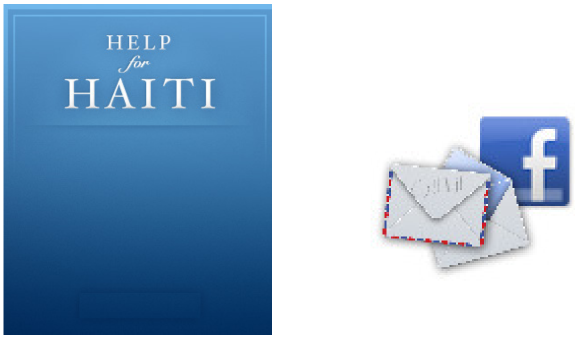


Figure 11: Images from the born-digtial competition dataset, where OTCYMIST method fails. The first image has low contrast. The second image has a single character, which gets eliminated in the proposed method.

Table II: Text localization results of ICDAR 2011 Robust Reading Competition: Challenge-1 [3] evaluated using [8].

Method	Recall	Precision	H-mean
Anthimopoulos	81.88	87.35	84.53
TDM IACAS	69.70	85.83	76.93
TH-TextLoc	73.06	80.39	76.55
Textorter	69.08	85.54	76.43
Baseline Method	69.94	83.92	76.30
OTCYMIST	75.65	63.85	69.25
SASA	64.91	67.38	66.12
TextHunter	58.43	75.52	65.88

fourth in text localization task. Table I lists the results of text segmentation task and Table II gives the results of text localization task [3].

Figure 11 shows some sample test images, on which our method fails at the level of binarization or thresholding the CC's.

VII. Conclusion

We have proposed a new, effective method for the segmentation and localization of text in born-digital images. Minimum spanning tree is used in our method. We have approached image binarization by splitting the different colour channels of an image, thus achieving higher recall rate than other algorithms. In our future work, we propose to incorporate a classifier for text components to improve the precision of our algorithm.

References

[1] S. M. Lucas et.al, *ICDAR 2003 Robust Reading Competitions: Entries, Results, and Future Directions*, International Journal on Document Analysis and Recognition, vol. 7, no. 2, pp. 105-122, June 2005.

[2] IAPR TC11 Reading Systems-Datasets List, http://www.iapr-tc11/mediawiki/index.php/Datasets

[3] D. Karatzas, S. Robles Mestre, J. Mas, F. Nourbakhsh and P. Pratim Roy, *ICDAR 2011 Robust Reading Competition - Challenge 1: Reading Text in Born-Digital Images (Web and Email)*, In Proc. 11th International Conference of Document Analysis and Recognition, 2011 , September 2011, http://www.cv.uab.es/icdar2011competition/

[4] N. Otsu, *A Thresholding Selection Method from Gray-level Histogram*, IEEE Transanctions on Systems, Man and Cybernetics, vol. 9, pp. 62-66, March 1979.

[5] J. Canny, *A Computational Approach to Edge Detection*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 8, no. 6, pp. 679-698, November 1986.

[6] R. Prim, *Shortest connection networks and some generalizations*, Bell System Technical Journal, 38:1389-1401, 1957.

[7] A. Clavelli, D. Karatzas and J. Llados, *A Framework for the Assessment of Text Extraction Algorithms on Complex Colour Images*, In Proc. 9th IARP International Workshop on Document Analysis Systems, ACM Press, pp. 19-28, Boston, MA, USA, 2010.

[8] C. Wolf and J. M. Jolin, *Object count/area graphs for the evaluation of object detection and segmentation algorithms*, International Journal of Document Analysis, 8(4), pp.280-296, 2006.

[9] Online Source: http://finereader.abbyy.com/, 2011.