MRF Based Text Binarization in Complex Images using Stroke Feature

Yanna Wang Cunzhao Shi Baihua Xiao Chunheng Wang
The State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China
Email:{wangyanna2013, cunzhao.shi, baihua.xiao, chunheng.wang}@ia.ac.cn

Abstract—**This paper presents a novel binarization technique for text images based on Markov Random Field (MRF) framework. We regard stroke as an obvious feature of text to produce clustering result, which will be optimized by MRF model combining color, texture, context features to get the final binarization. The main innovations of our method are: (1) the integrated image is split into sub-images on which we can automatically acquire seed pixels of foreground and background using stroke feature; and (2) diverse weights are attached to seed pixels according to their location information, then highly confident cluster centers of sub-image can be acquired by gathering weighted seeds. The experimental results show that our method is robust and accurate on both video and scene images.**

Index Terms—**binarization, text image, stroke, sub-image, weight, MRF**

I. Introduction

With the rapid development of Internet, the number of images and videos is intensely growing. Text plays an important role in images and videos, brings us significative semantic information, and provides pivotal clues for image indexing and retrieval. In the meanwhile, because of complex condition such as uneven light, diverse size, varied orientation, changeable color and low resolution, recognizing text accurately is challenging. Generally the overall process of text recognition includes: text detection, text localization, text binarization and text recognition [1]. Among the four phases, text binarization is an intermediate connecting detection, localization and recognition, so that binarization is particularly crucial in text recognition. In this paper, we mainly focus on binarization for text with complex background.

The binarization methods proposed in the past could be divided into three categories as follows: the first class is threshold method. This method chooses a suitable threshold to discriminate foreground pixels from background. The traditional threshold methods include Otsu [2], Niblack [3], kittler [4], Sato [5], Sauvola [6] and so on. Global threshold approaches [2][4] select a validly static threshold for the image, while the local threshold approach [3] uses a dynamic threshold based on a window across the image. Threshold methods are effective for images with simple background and high contrast, but fail on condition of high complexity. The second class is based on clustering [7][8][9][10]. These methods utilize color information (RGB, LAB, HIS etc.) to cluster for binarization. Liu [11] used Gaussian mixture models to model the feature vector of three neighboring characters, then text extraction is completed through labeling each connected component in the binary image as character or non-character according to its neighbors. But clustering methods require users to set initial cluster centers and number of clusters. Recently, the third class based on energy minimization has been widely used . These methods construct models, including MRF [12], CRF [13], Graph Cut [14], Grab Cut [15], to extract text from an image. Here we call them the graph-cut methods. These methods have achieved promising results in terms of image binarization.

The aim of binarization is to distinguish pixels of foreground from background in the image. In this paper, we settle the binarization problem in the framework of MRF due to its effectiveness in image segmentation [12] [16]. In the matter of initializing highly confident foreground and background seed pixels which will be used for producing cluster centers, to avoid human interactions (like Graph cut [14] and Grab cut [15]), we put forward an approach to seek seed pixels of foreground and background automatically with stroke feature on edge image. In order to alleviate the influence of different strokes between adjacent characters and get the stroke reliably, we analyze edges connected components (CCs) to locate text characters and split the whole image into several edge sub-images. Afterward, based on the fact that most of texts are located in the central area of the text bounding box, we allocate higher weight for foreground pixels around the center of bounding box. With the diverse weights according to the location of pixels, highly confident initial cluster centers of foreground and background in sub-images are obtained. By merging the centers of sub-images, we acquire the cluster centers of the integrated image. Finally, combing features of color, texture and context [17], we construct the MRF model to smooth the results of the previous step to get the binarization.

The remainder of this paper is organized as follows. In the section II, we introduce the binarization method proposed in this paper in detail. The concrete experiment is presented in the section III, and the section IV is conclusion.

II. The Binarization Method

A. Method Overview

Fig. 1 illustrates the proposed method. Given a text image, we apply CCs analysis technology in edge image to split the whole image into several edge sub-images. We acquire highly confident seed pixels of foreground and background using stroke feature on edge image, and then obtain cluster

978-1-4799-1805-8/15/$31.00 ©2015 IEEE

centers by gathering the seeds which have been endowed with diverse weights. Further, the cluster centers of integrated image are acquired using k-means taking advantage of the cluster centers of sub-images. Finally, combined with features of color, texture and context, MRF model optimizes the preliminary result with min-cut/max-flow algorithms [18] to acquire the final binarization.

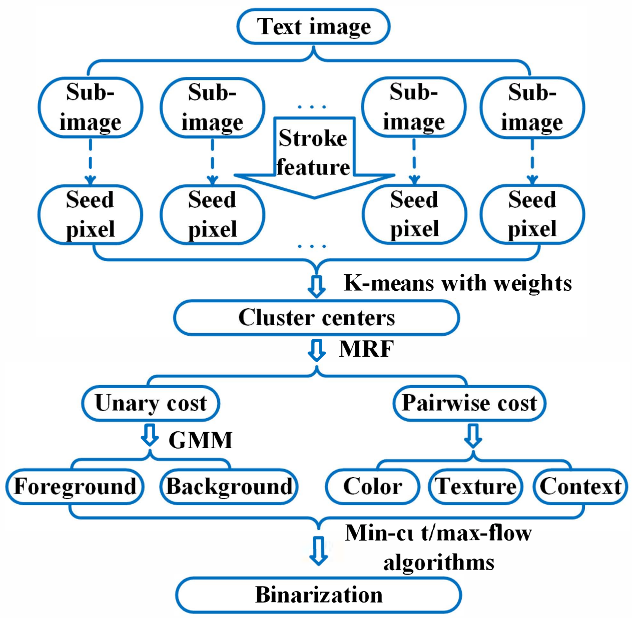


Fig. 1. Flowchart of proposed method

B. Splitting Images

In order to gain reliable stroke feature, and avoid mutual interference between different characters, we split entire image into edge sub-images. Edges are obvious disparate between text and background, meanwhile, edges are robust to the size, the color and uneven light of text, so that precise detection of the image edges is crucial for splitting text image. In term of edge detector, we choose the Canny operator owing to its high consistency, accuracy and attribution of reserving more edge pixels of text.

Splitting images consists of three phases. Firstly, extracting edges by Canny operator. Secondly, CCs analysis is used to get the candidate sub-images. Finally, fusing CCs to obtain ultimate edge sub-images result.

Fig. 2 shows the process of splitting. (a) is the original image, (b) are the character sub- image, (c) are the edge sub-images.

C. Acquiring Seed Pixels

For the purpose of performing the binarization process automatically, we propose an algorithm based on stroke feature to acquire highly confident seed pixels of foreground and background.

Our algorithm is based on the observation that the diversity between foreground and background is that foreground pixels locate inside closed Canny edges, while background outside. We start from every pixel in the edge sub-image to count the number of pixels which has the same binary value of the original pixel in its four directions (up, down, left and right) until the first different binary value pixel appear respectively. We represent the count in the horizontal (directions of left and right) as $stroke_{horizontal}$ and vertical (directions of up and down) as $stroke_{vertical}$. At the same time we count the number of directions (up, down, left and right) in which we can find a different binary value pixel from original pixel called $edge_{number}$. Here $width_{sub}$ denotes sub-image width and $height_{sub}$ denotes sub-image height.

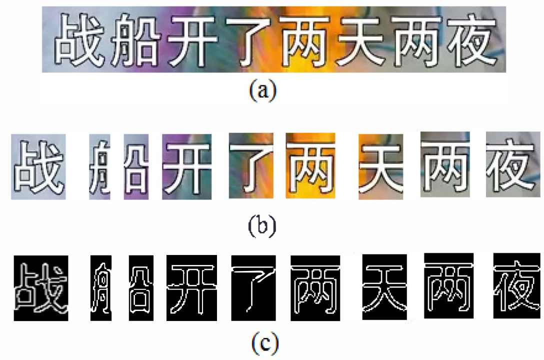


Fig. 2. An illustration of splitting.

Highly confident pixels of foreground and background acquired are given below:

(1) seed pixels of foreground: if $edge_{number} = 4$ and $stroke_{horizontal} \leqslant 0.5 * width_{sub}$ or $stroke_{vertical} \leqslant 0.5 * height_{sub}$.

(2) seed pixels of background:

case1, for pixels in edge sub-images: $stroke_{horizontal} > 0.5 * width_{sub}$ or $stroke_{vertical} > 0.5 * height_{sub}$. Owing to the harmful effect of illumination and low-resolution, some foreground pixels may appear outside of the closed Canny edges in case of these edges rupture, so that we set $edge_{number} <= 1$ to reduce the number or erroneous judgement of foreground pixels.

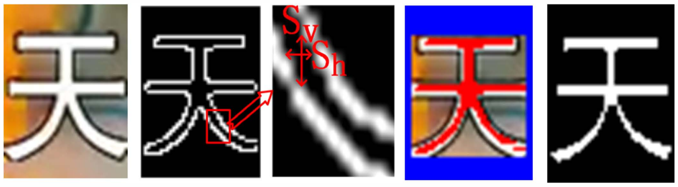
case2, for pixels on boundary of the integrated image: text is usually located in the center of text bounding box after localization, so the pixels on the top, bottom, left, right boundaries of image and not on Canny edges belong to background.

case3, for pixels between adjacent edge sub-images: according to edge density, we estimate it belongs to background or not.

Fig. 3 shows an example of acquiring the foreground and background seeds. In (b), S_h and S_v represent $stroke_{horizontal}$ and $stroke_{vertical}$. In (c), blue and red area represent background and foreground.

D. Segmenting images

The binarization problem is equivalent to labeling every pixel as foreground or background. We assign every seed



(a) (b) (c) (d) (e)

Fig. 3. Illustration of the seeds seeking process: (a) Original sub-image. (b) Edge sub-image.(c)$stroke_{horizontal}$ and $stroke_{vertical}$ of a pixel (d)Initial seeds. (e) Label image after clustering.

pixel of foreground variable weight depending on the pixel's position in the image, then we calculate the cluster centers of background and foreground in each sub-images with these weighted seeds. By merging the centers of sub-images, we acquire the cluster centers of the whole image. Finally, we take advantage of MRF model to get the final binarization.

1) K-means Clustering with Weights

We define a pixel as a variables x_i labeled with y_i (0 or 1) which means the pixel belongs to background or foreground. Clustering is an effective method for image binarization. Here we adopt k-means to generate cluster centers of the integrated image. But k-means clustering often has the following three challenges [19]: (a) how to choose K (number of cluster centers)? (b) how to select initial cluster centers? (c) which clusters belong to foreground?

Our work depends on the observation that foreground pixels usually have the uniform color and most of text are located in the middle of the image in vertical direction after localization. In addition, in consideration of Canny operator unlikely removing every noise, in order to reduce the effect of these noise as far as possible, we assign diverse weights to pixels at different positions of image. We divide the edge sub-image into nine parts along horizontal direction, and the weight of foreground seed pixel depends on its corresponding part's distance to the middle of sub-image, namely, the closer the larger.

The K in k-means is 4 because we analyze background from three aspects and foreground from one when we acquired seed pixels. For a given image, we firstly gather information of seed pixels' RGB color, location, texture information as features to calculate the four cluster centers in each sub-image, then we can get the cluster centers of the integrated image by merging the centers of sub-images contained in this given image.

2) Markov Random Field Model Based Binarization

Through k-means clustering above, the cluster centers in the whole image are obtained. Afterward, we remove noises and smooth binary image with MRF framework. We define each pixel as a node in MRF and get unary and pairwise cost to construct the energy function. The energy function is expressed as follows:

$$E(Y,X,\theta) = E_{unary}(Y,X,\theta) + \lambda E_{smoothness}(Y,X). \quad (1)$$

Here, λ is a trade-off coefficient between unary and pairwise cost. $X = \{x_1, x_2, x_3 ...\}$ denotes the features of nodes. $Y = \{y_1, y_2, y_3...\}$ is a vector of labels. θ is parameters related to the model and data. In the above equation, $E_{unary}(Y,X,\theta)$ is data term measuring the inconsistency of inferred label and real data label. $E_{smoothness}(Y,X)$ is smoothness term representing the cost between y_m and y_n to adjacent pixels.

Here we construct unary energy function making use of Gaussian mixture models. We define the unary item as:

$$E_{unary}(Y,X,\theta) = -\sum_i \log p(y_i|x_i). \quad (2)$$

Here, $p(y_i|x_i) = p(x_i|y_i)p(y_i)/p(x_i)$. We ignore $p(x_i)$ and assume $p(y_i = 0) = p(y_i = 1)$, so we think there is no difference between $p(y_i|x_i)$ and $p(x_i|y_i)$.

the smoothness item is:

$$\begin{aligned} E_{smoothness}(Y,X) = \quad & \lambda_1 e^{-\frac{1}{2\delta_1^2}(x^i_{color}-x^j_{color})^2} \\ & +\lambda_2 e^{-\frac{1}{2\delta_2^2}(x^i_{texture}-x^j_{texture})^2} \\ & +\lambda_3 e^{-\frac{1}{2\delta_3^2}(x^i_{context}-x^j_{context})^2}. \end{aligned} \quad (3)$$

Here, $(i,j) \subset N$, $i \neq j$, N denotes the eight neighborhood system. x_{color}, $x_{texture}$, $x_{context}$ represent the color, texture, and context feature, respectively. x_{color} denotes RGB feature, $x_{texture}$ denotes the quantity of gradient at a pixel in RGB three channels. $x_{context}$ is the feature which equals to the probabilities of other pixels belonging to foreground in N neighborhood. $\lambda_1, \lambda_2, \lambda_3$ are the weights assigning to above-mentioned three features, $\delta_1, \delta_2, \delta_3$ are the normalized coefficients. These six parameters are learned from the image automatically. We minimize the function using min-cut/max-flow algorithms [18].

We feed clustering results into Gaussian mixture model to represent $p(x_i|y_i)$.

Considering foreground is generally consistent in color and texture, we use a Single Gaussian Model to calculate probabilities of foreground.

$$p(x_i|y_i,\theta) = \mathcal{N}(x_i|\theta). \quad (4)$$

Here $\mathcal{N}$ represents Gaussian distribution.

The background is complex and changeable, so we use the Gaussian Mixture Models here.

$$p(x_i|y_i,\theta) = \prod_{j=1}^{n} \frac{\pi_j}{\sqrt{\det(\sum_j)}} \exp(-\frac{1}{2}(x_i-\mu_j)^T {\Sigma_j}^{-1}(x_i-\mu_j)). \quad (5)$$

Here n denotes the number of Gaussian distribution models. π_j denotes the Gaussian mixture weighting coefficient learned from the image. μ and Σ denote mean and covariance.

Finally, the result of MRF is further filtered by CCs analysis technology.

III. Experimental Results and Discussion

Since there are no public dataset for video images, we collect video images from movies, TV, lectures and news for our experiment. The dataset contains 756 images including Chinese, Japanese, English, numbers etc. The colors, sizes, fonts of these image are diverse and background are highly

Fig. 4. some examples for our experiment

complex. Additionally, we also apply our binarization algorithm on ICDAR 2003 Robust Word Recognition dataset [20] consisting of 1110 word images. The proposed method is evaluated by character and pixel level. Some examples are shown in Fig. 4.

We compare the proposed method with other binarization algorithms: Otsu [2], Niblack [3], Kittler [4], K-means clustering method [16], Howe [21]. In our experiment, parameters for other methods were chosen by selecting the best among several runs with different parameters. In [16], k-means clustering without weights is a preprocessing stage before energy-based binarization. Howe [21] presents a binarization algorithm based on Laplacian energy with Canny edge information and shows a good performance on his dataset. For fair comparison, we convert all the binary results into white text on black background. We analysis the performance of proposed method qualitatively and quantitatively.

A. Qualitative Analysis

Fig. 5 shows an example consisting of four steps in our experiment. Clustering method plays a crucial role in the overall performance of the proposed method. Through adding diverse weights, our clustering result in the first step shows a relatively clear text. But it still contains some noise. Color and texture are obvious characteristics to distinguish foreground from background, so when they are added to MRF model, some noise pixels are removed. In the last step, it takes contextual information into smoothness term, and CCs technology is adopted after MRF model, therefore most of the remaining noise in the third step are filtered. The results of different methods are shown in Fig. 6. Distinctly, our approach generates high-quality text images with least noise.

B. Quantitative Analysis

1) OCR accuracy:
In the paper, we evaluate binarization performance by CER and CRR using ABBYY FineReader 11. CER is the character extraction rate, and CRR is the character recognition rate:

$$CER = N_{segment}/N \qquad CRR = N_{recognize}/N. \tag{6}$$

Here, $N_{segment}$ denotes the number of characters extracted from text image without breaking and adhering to background, $N_{recognize}$ denotes the number of characters recognized from text image with ABBYY Recognition Engine, and N denotes the total number of the characters. The experiment results of OCR accuracy on video images and on scene images are shown in Table I.

Fig. 5. Each steps in our method

The three threshold methods, Otsu, Kitter and Niblack, usually have more noise in images. At the same time, ABBYY Recognition Engine is not robust to noise, so their results on CRR are relatively poor. Our clustering result, with weighted seeds, exceeds k-means in [16]. Howe [21] presents a relatively low performance in our dataset, because this method mainly concentrates on texture information and neglects color in energy function, while there are many multicolored images in our dataset. Through clustering, in our method, we acquire initial label, and can not remove all background noise. Taking advantage of the neighborhood information by the MRF, we wipe off much of the remaining noise and smooth the binary result, and as a result, the improvement of CRR is much more than CER. Due to the influence of uneven illumination and low resolution, it is more difficult to extract full text edge information in scene images than in video images. Performances of our method on scene images is slightly lower than video images. CRR of ABBYY on original images is 44.85% on video dataset and 47.9% on ICDAR 2003 dataset, which are much lower than the results of our proposed binarization 78.99% and 71.32%.

TABLE I
OCR ACCURACY EVALUATION ON VIDEO IMAGES AND ICDAR 2003(%)

Method	Video Images		ICDAR 2003	
	CER	CRR	CER	CRR
OTSU [2]	73.33	53.95	77.55	50.60
Niblack [3]	71.98	43.00	67.30	41.70
Kittler [4]	55.50	38.41	75.92	49.59
Howe [21]	76.68	58.36	79.28	55.63
k-means	87.50	63.30	82.80	59.59
Proposed K-means with weights	89.12	73.15	86.33	60.00
+ MRF with color	89.19	76.03	86.43	68.98
+ MRF with color and texture	89.21	76.77	86.43	69.08
+ MRF with color,texture and context	**90.34**	**78.99**	**88.20**	**71.32**

TABLE II
PIXEL LEVEL SEGMENTATION EVALUATION ON ICDAR 2003 (%)

Method	P	R	F
Otsu [2]	87.03	**90.45**	88.71
Niblack [3]	71.36	82.67	76.60
Kittler [4]	74.31	84.89	79.25
Howe [21]	82.56	87.93	85.16
Proposed	**88.40**	90.09	**89.24**

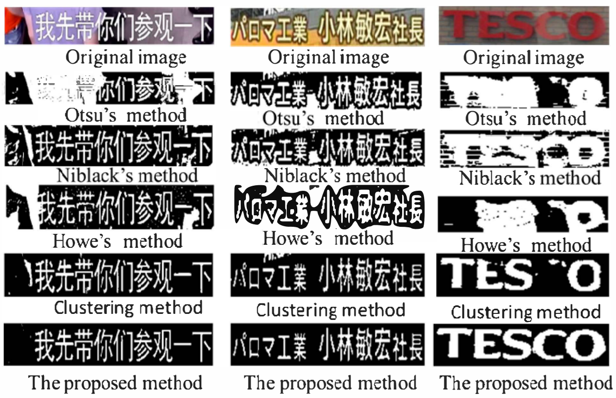


Fig. 6. Comparing Results of Different Methods.

2) Pixel level accuracy:

We further compare various binarization algorithms on pixel accuracy. Following the experimental settings in [12], we randomly select 200 images from ICDAR 2003 Robust Word Recognition dataset and produce pixel level binarization ground truth for them. In order to guarantee the confidence of pixel level ground truth, we choose images with adequate thick strokes. We present experimental results on pixel level in Table II. Most of images in ICDAR 2003 dataset have simple background, as a result, global threshold methods like Otsu can extract the majority of text pixels which lead to a higher recall than ours, however the precision is lower. Our proposed approach is more robust than other methods getting the best F-score in terms of pixel level evaluation.

IV. CONCLUSION

In this paper, we have introduced a new binarization method for text images. Based on stroke feature as a obvious characteristics, we acquire highly confident seed pixels for k-means clustering automatically. The more reliable cluster centers are obtain using seed pixels attached with diverse weight. We use MRF model and integrate various features including color, texture, context to eliminate noise to improve performance. Experimental results show that our method outperforms other methods above-mentioned on CER, CRR and pixel level on video and scene images. More robust and effective edge extraction method and more confident selecting method for seed pixels of foreground and background could be further studied.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants No. 61172103 and No. 61271429.

REFERENCES

[1] M. R. Lyu, J. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 15, no. 2, pp. 243–255, 2005.

[2] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[3] W. Niblack, *An introduction to digital image processing*. Strandberg Publishing Company, 1985.

[4] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern recognition*, vol. 19, no. 1, pp. 41–47, 1986.

[5] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, "Video ocr for digital news archive," in *Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on*. IEEE, 1998, pp. 52–60.

[6] J. Sauvola, T. Seppanen, S. Haapakoski, and M. Pietikainen, "Adaptive document binarization," in *Document Analysis and Recognition, 1997., Proceedings of the Fourth International Conference on*, vol. 1. IEEE, 1997, pp. 147–152.

[7] K. Wang and J. A. Kangas, "Character location in scene images from digital camera," *Pattern recognition*, vol. 36, no. 10, pp. 2287–2299, 2003.

[8] C. Yi and Y. Tian, "Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification," *Image Processing, IEEE Transactions on*, vol. 21, no. 9, pp. 4256–4268, 2012.

[9] K. Kita and T. Wakahara, "Binarization of color characters in scene images using k-means clustering and support vector machines," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 3183–3186.

[10] L. Fu, W. Wang, and Y. Zhan, "A robust text segmentation approach in complex background based on multiple constraints," in *Advances in Multimedia Information Processing-PCM 2005*. Springer, 2005, pp. 594–605.

[11] X. Liu, H. Fu, and Y. Jia, "Gaussian mixture modeling and learning of neighboring characters for multilingual text extraction in images," *Pattern Recognition*, vol. 41, no. 2, pp. 484–493, 2008.

[12] A. Mishra, K. Alahari, and C. Jawahar, "An mrf model for binarization of natural scene text," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 11–16.

[13] M. S. Cho, J.-H. Seok, S. Lee, and J. H. Kim, "Scene text extraction by superpixel crfs combining multiple character features," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1034–1038.

[14] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 105–112.

[15] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

[16] Z. Zhang and W. Wang, "A novel approach for binarization of overlay text," in *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 4259–4264.

[17] M. Li, M. Bai, C. Wang, and B. Xiao, "Conditional random field for text segmentation from images with complex background," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2295–2308, 2010.

[18] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 9, pp. 1124–1137, 2004.

[19] Y. Song, A. Liu, L. Pang, S. Lin, Y. Zhang, and S. Tang, "A novel image text extraction method based on k-means clustering," in *Computer and Information Science, 2008. ICIS 08. Seventh IEEE/ACIS International Conference on*. IEEE, 2008, pp. 185–190.

[20] "The ICDAR 2003 Robust Reading Datasets," http://algoval.essex.ac.uk/icdar/Dataset.html.

[21] N. R. Howe, "A laplacian energy for document binarization," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 6–10.