Document Image Binarization using LSTM: A Sequence Learning Approach

Muhammad Zeshan Afzal[1,4], Joan Pastor-Pellicer[2], Faisal Shafait[3], Thomas M. Breuel[1], Andreas Dengel[4], Marcus Liwicki[1,4]

afzal@iupr.com, jpastor@dsic.upv.es, faisal.shafait@seecs.edu.pk, tmb@iupr.com, Andreas.Dengel@dfki.de, marcus.eichenberger-liwicki@unifr.ch

[1]Technical University of Kaiserslautern, Germany.
[2]Departamento de Sistemas Informáticos y Computación Universitat Politècnica de València València, Spain
[3]School of Electrical and Computer Science, NUST, Islamabad, Pakistan
[4]German Research Center for Artificial Intelligence (DFKI), Germany.

ABSTRACT

We propose to address the problem of Document Image Binarization (DIB) using Long Short-Term Memory (LSTM) which is specialized in processing very long sequences. Thus, the image is considered as a 2D sequence of pixels and in accordance to this a 2D LSTM is employed for the classification of each pixel as text or background. The proposed approach processes the information using local context and then propagates the information globally in order to achieve better visual coherence. The method is robust against most of the document artifacts. We show that with a very simple network without any feature extraction and with limited amount of data the proposed approach works reasonably well for the DIBCO 2013 dataset. Furthermore a synthetic dataset is considered to measure the performance of the proposed approach with both binarization and OCR groundtruth. The proposed approach significantly outperforms standard binarization approaches both for F-Measure and OCR accuracy with the availability of enough training samples.

Keywords

Document Image Binarization, Long Short Term Memory, Neural Network, Optical Character Recognition

1. INTRODUCTION

DIB is the process of segregating text from images where text consists of the pixels of interest in the image (typically, the ink for handwritten text). Binarization of document images is often very important pre-processing step for the Document Image Processing (DIP) pipeline. Binarization of the document images has long been established as a tool [12] to facilitate the processes of text recognition, layout analysis, etc. There exist many methods which use local and global statistics (mean, variance) or tools from image processing such as filtering, inpainting, morphology, etc. Either the statistical operations have been used alone or in combination with the image processing techniques to achieve better performance [14].

Document images may contain different types of degradations and noise. Sample images are shown in Figure 1. They contain different amounts of bleedthrough text. Figure 1a contains uneven background and the major artifact is the stamp in the middle of the page. According to the provided

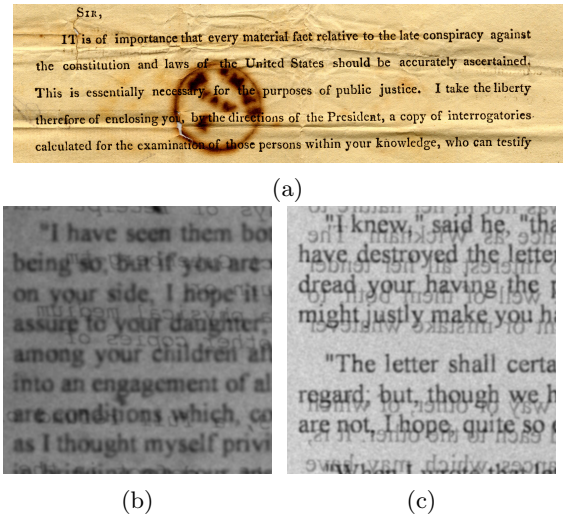


Figure 1: Sample image of degraded documents. (a) image from the DIBCO 2013 dataset. While the page itself has artifacts, the most prominent one is the stamp in the middle of the page (b) Image with bleedthrough visible between the text lines. (c) Image with varying background, text is also has varying grey levels and bleedthrough degrading the original text.

HIP '15 August 22, 2015, Nancy, France

DOI: http://dx.doi.org/10.1145/2809544.2809561

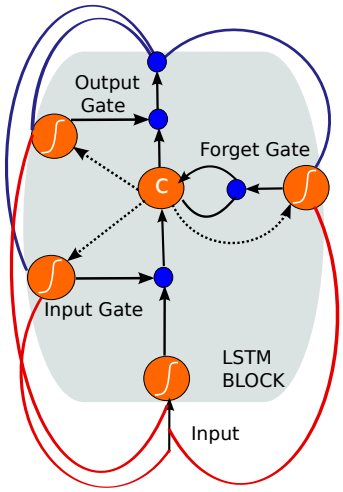


Figure 2: Sample LSTM cell. The internal state is controlled by three multiplicative gates i.e. the input, forget and output gates. The connections from input to multiplicative gates are depicted by red color. Due to the recurrent nature of the network each cell receives the input from itself and as well as from the other units from the previous time step.

groundtruth, some of the pixels belonging to stamp are considered as background and others as text due to an overlap over a relatively large region. Figure 1b and 1c show the images with dark and light background. The images also contain varying amount of bleedthrough which exist at different locations. These sample images give an idea of the difficulty for DIB. While existing binarization methods prove to be effective for document image binarization, they are heuristic in nature and rely on the parameter tuning to fit the needs of the specific dataset under consideration. The quantity of parameters of a binarization method is dependent on the image processing and statistical techniques used as intermediate steps. Some approaches demonstrate that no parameter tuning is required [14]. Essentially the default parameters are used which may require adjustment if the dataset under consideration changes.

Recently, LSTM, a variant of recurrent neural networks, has been shown to outperform state-of-the-art methods in different applications which require sequence learning, such as time series prediction, human action recognition and handwriting recognition in [7–9, 23]. A typical architecture of the LSTM memory cell is depicted in Figure 2. The multiplicative gates maintain the internal state in order to model long term dependencies.

The major contribution of this paper is modeling the image as a 2D sequence and proposing the image binarization as a sequence learning problem. We explore the potentials of LSTM for document image binarization. The viability of the approach has been evaluated on DIBCO 2013 dataset yielding comparable results with the current state-of-the-art. Furthermore to evaluate the OCR accuracy, we use a dataset of degraded documents with known binarization and OCR ground truth and show that our approach produces significantly better results in comparison to standard binarization methods.

This papers is structured as follows: Section 2 introduces some related work on the DIB task, Section 3 explains how Bidirectional LSTM are applied for DIB, and Section 4 describes the performed experiments on the two databases. In the last section, we present our conclusions.

2. RELATED WORK

There are many criteria for categorising binarization methods based on the type of documents. One such categorization could be devised on the basis of number of colour channels. While colour provides more information for the binarization as described in [3, 15, 26] for example, the work in DIB is focused more towards the grey level document image which is the focus of this paper.

Binarization methods can be coarsely divided into two broad classes. The first class of methods is based on heuristics and the second one uses machine learning. Furthermore, heuristic binarization methods can be categorized into the three categories global, local, and hybrid methods. As implied by the name, global methods consider a single threshold value for an image to be binarized. The threshold is generally determined in global methods by gray levels histogram analysis and finding a local minimum. For global methods, among many others, one can refer to methods based on discriminant analysis [16], and histogram separation [28]. The obvious advantage of global methods is being extremely fast. However, a major drawback with a global threshold is that they compromise over the spatial relationships of the pixels and sometimes completely ignore them, which turns out to be even more problematic in the presence of noise. In general, global methods for binarization suffer a great deal when the degradations in the image are purely local.

The alternative category of binarization methods uses local threshold value. It considers a different threshold value for each pixel by taking into account the gray values of the neighbouring pixels. Local threshold methods cover to some extent the shortcomings of the global methods in dealing with brightness variations in documents. Examples of local methods are methods based on the concept of Niblack [13,25] which is further refined by Sauvola [22]. A threshold is determined based on the local window statistics around the central pixel.

There are also hybrid methods of binarization that employ both local and global techniques for binarization. For instance [27] proposes two-step binarization method by first applying a global threshold, followed by a local threshold that is adjusted according to the document spatial information. There are also works that apply global thresholding to correct the result of a previous local thresholding (such as [24]). A combined approach which works very well for handwritten document images and uses background estimation as well as the local and global image statistics is presented in [14]. All these methods either require lot of parameters to tune or they are hard coded into the steps of binarization to suit the requirements of a certain set of documents.

As mentioned above, some binarization methods make use of machine learning techniques [2, 10, 11]. The machine learning techniques are applied in different ways. There are some approaches which try to find appropriate binarization method to sub-regions of a document, such as [5] in which an input image is partitioned into $n \times n$ regions, and uses support vector machines to decide what kind of binarization action to take. With a similar approach, [4] tries to choose

the optimal binarization method (among a set of candidate approaches) using a neural network model. The images are then binarized using the threshold that is obtained from the network according to the input histograms. The application of machine learning techniques for binarization purposes is further extended by works such as [21] that use Multilayer Perceptrons (MLPs) to classify a pixel as black, according to the pixel's gray value and those of the neighbouring pixels. Also Markov Random Fields [10, 11] have been used for the image binarization.

We propose a learning method for DIB which treats images as 2D sequences. Our image is considered as 2D sequence in order to model complex spatial dependencies. We use LSTM which can process very long sequences. It does not require any handcrafted feature extraction or parameter in contrast with other heuristic and machine learning based binarization methods.

3. BINARIZATION USING LONG SHORT-TERM MEMORY

For binarization methods to work effectively for a wide range of documents, it is required to understand the semantics of the image in order to be able to distinguish between the text and the artifacts which are present in the image. To learn the underlying semantics we propose to view image binarization as a sequence learning problem. The local information has been processed within a neighborhood and then it is propagated through a recurrent neural network, LSTM in our case.

The structure of the document image has complex spatial dependencies. The spatial relationships are learned using 2D BLSTM (Bidirectional Long-Short Term Memory). Each pixel is considered as one time step. The same learning procedure which is used in temporal domain to learn temporal dependencies is applied to learn the spatial dependencies in the 2D image.

The standard LSTM architecture by formulation is one dimensional. In each cell there is one recurrent connection which is connected to one forget gate as it has been shown in Figure. 2. The one dimensional architecture is suitable for the sequence labelling tasks which are one dimensional in nature such as OCR where the text line is the 1D sequence or the speech recognition where the features are inherently one dimensional in nature. In order to optimally predict the value of a single pixel, contextual information from both dimensions is needed.

While the standard LSTM is one dimensional because it contains only one recurrent connection and one forget gate, it could easily be extended to multidimensional by introducing more self connections and forget gates according to the dimensions of the data. In the case of binarization we need 2D BLSTM which will have two recurrent connections with two forget gates.

The simplistic overview of the learning procedure with a single hidden layer has been shown in Figure. 3. The images are divided into patches of size $n \times n$. These patches are further propagated to the LSTMs. Although the architecture has one hidden layer, but it is divided into four independent LSTM for each direction of the x and y dimensions. As each dimension can be traversed from two directions, each LSTM independently processes the patches from all the four possible directions.

After processing this information for the whole patch, the values of each LSTM are further provided to the output layer. Thus, the output now have context of the image at a current pixel from 4 different directions which is essential for the correct classification of current patch. The size of the output is the same as the size of the input where each output value indicates the probability of each pixel to be either text or background.

The feature maps learned by each LSTM (one for each direction) for DIB are shown in the Figure. 4. It is essential to analyze the feature maps for the better understanding of the learning procedure. The input to the 2D LSTM network is a patch from the image. The same patch is provided to all of the four independent LSTM. In the proposed sample, each LSTM consists of 5 memory cell. The figure shows 20 feature maps where each 5 feature maps belong to a different LSTM. The red arrows illustrate the direction in which the image has been processed by each LSTM. The feature maps show that each memory cell learns different types of features from the input image. The proposed system learns different features from the training data. When a test image is given to the network, the learnt weights extract several features maps which are provided to the output layer. The output layer combines all this information and produces the output final image.

4. EXPERIMENTS AND RESULTS

4.1 Datasets

We consider two types of datasets for the evaluation of the proposed approach. The first dataset consists of the standard datasets used in DIBCO contests. We consider the images from the competitions DIBCO09 [6], HDIBCO10 [17], DIBCO11 [18], and HDIBCO12 [19] as our training set. We take DIBCO13 [20] as our test set. The DIBCO2013 dataset contains both printed and handwritten documents as the training set and is presumably the most complicated among all of the DIBCO datasets. All datasets contain both color and grey level images, as mentioned previously, color images are converted to grey level for processing. Additionally, we perform local contrast normalization followed by the histogram equalization.

It is not possible to check the OCR performance of the binarization in the DIBCO datasets since a transcription ground truth is not provided. We generate a synthetical dataset for this purpose. We mimic the artifacts of historical documents and then we applied them to real images of printed books for which a transcription ground truth is available. This dataset consists of 100 degraded images each of size 2550×3300, and half of the images were used for the training and the rest for testing.

4.2 Evaluation

For computing the output of each pixel it is needed to compute first the features map of each of 4 LSTM. Processing long sequences could be very slow. For that, dividing the images in paches $512x512$ give us a tradeoff between performance and efficency. The smaller images are padded to make them of the proposed size. These patches are not very large but essentially contain enough information to provide both local and global context for the recurrent neural network.

We report F-Measure for DIBCO datasets and both OCR

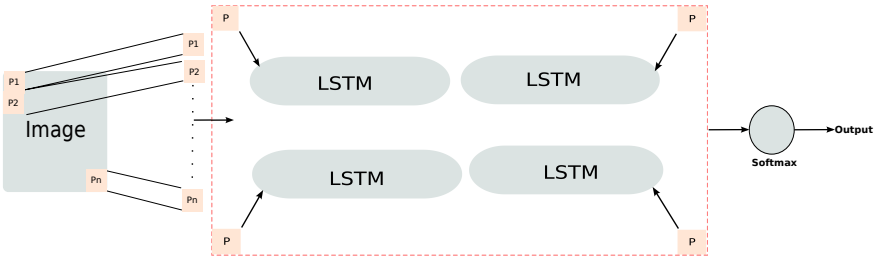


Figure 3: Sample 2D LSTM for the proposed architecture. Four individual LSTMs with 5 memory cells scan the image from different directions. The result of each individual component is gathered at the output layer hence providing the context from all the dimensions for the final classification.

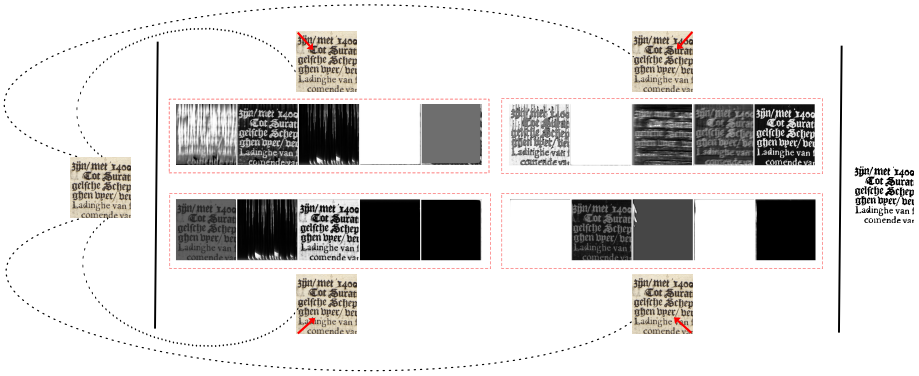
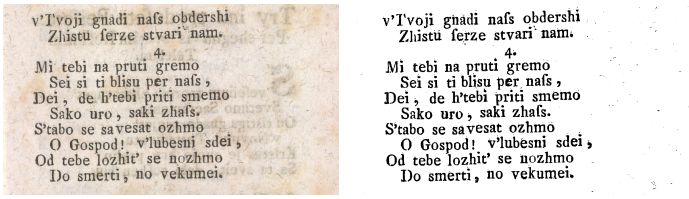
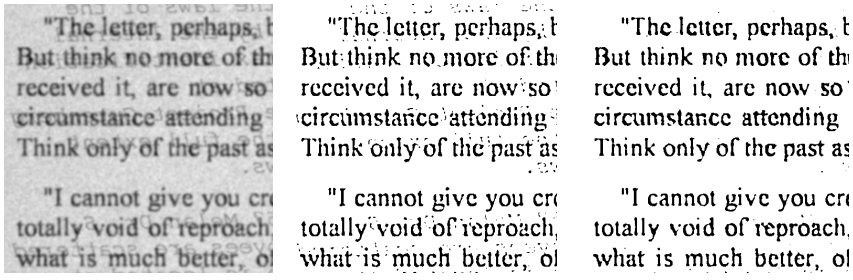


Figure 4: A cropped part of an image is used to display the input, output and the learned representation of the hidden layer. The hidden layer consists of four LSTM each one referring to a different dimension and direction. These directions are depicted with red arrows on the input image for each corresponding layer.



(a) (b)

Figure 5: The input from DIBCO13 dataset is shown in (a) and the corresponding binarization result using LSTM is shown in (b)



(a) (b) (c)

Figure 6: The input image from synthetic dataset shown in (a) and the corresponding binarization results of Sauvola and LSTM are shown in (b) and (c) respectively

accuracy and F-Measure for the synthetically generated dataset. The F-Measure is used commonly for evaluate DIB performance. It is the harmonic mean of *precision* and *recall* where precision (also called positive predictive value) is the fraction of well classified foreground predicted pixels, and recall (also known as sensitivity) is the fraction of ground-truth foreground correctly.

Several configurations have been tried for our models, but we got competitive results using a simple 2D BLSTM network with 5 memory cell for each LSTM , that makes 20 cells (5×4) before the output layer. A different neural network is trained for each of the two datasets used in this paper.

We used $1e^{-6}$ and 0.9 as the values of learning rate and the momentum, we have seen that this parameters have worked fine for both LSTMs. All the experiments are performed using RNNLIB[1].

It is important to mention that the images of all DIBCO datasets are essentially different from one another. Documents with more complicated artifacts are introduced progressively every year to check the performance of the current state of the art. An example of the binarization on an image of DIBCO is shown in Figure 5, where 5a is the input image and 5b the corresponding output. The Figure 6 shows the binarization results of Sauvola and the proposed approach for a patch from the synthetical dataset. As we can observe in both cases, the proposed method is able to learn the bleedthrough artifacts.

While learning methods require a large amount of training data and also similar type of images, we show that with the limited amount of data and with varying artifacts which are different in every image, the proposed approach shows comparable results with the state of the art. The proposed approach achieves an F-Measure of 87.91 on DIBCO13 dataset in comparison to the state-of-the-art [20] achieves 92.12.

We compare the results of our approach with standard Sauvola binarization for syntheticaly generated dataset. For finding the best parameters for Sauvola binarization we used F-Measure for error evaluation. The method of Sauvola binarization has two parameters k and w. A grid search over an interval $[0.1, 0.35]$ with step size of 0.1 and $[3, 53]$ with step size of 2 respectively for k and w has been performed. The best values are 0.17 and 15 for k *and* w respectively. A sample binarized image is shown in Figure. 6. The proposed approach outperforms both Sauvola and Percentile Filter [1] as depicted in Table. I. It achieves an accuracy of 94.53 and 81.73 for F-Measure and OCR accuracy respectively. We used the latest version[2] of Ocropus for computing the OCR accuracy.

Method	F-Measure	OCR accuracy(edit distance)
Sauvola	92.07	31.19
Percentile Filter	89.39	69.6
LSTM (proposed)	94.53	81.73

Table 1: Comparison of proposed method with Sauvola and Percentile Filter.

5. CONCLUSION AND FUTURE WORK

We have proposed a novel technique for image binarization which works well for the degraded document images and outperforms standard methods of document image binarization both for pixel based (F-Measure) and OCR based measures. The proposed method requires no parameter tuning and works well without any feature extraction. It also does not require any pre- or post-processing steps. The proposed method is able to learn the background artifacts present in the image and is able to binarize without severely affecting the quality of the image. The proposed work is rather unique in comparison to other learning based approaches for binarization. It is due to its ability to explain the kind of features which have been automatically extracted from the images. The features can be visually inspected. Therefore, these features provide the insight of the learning process and can help to modify the architecture based on visual information. We plan further to improve the binarization process using deep recurrent neural network architectures.

[1]http://sourceforge.net/projects/rnnl/

[2]https://github.com/tmbdev/ocropy

Acknowledgment

This work was partially funded by the BMBF project Kallimachos (01UG1415C).

6. REFERENCES

[1] M. Z. Afzal, M. Krämer, S. S. Bukhari, M. R. Yousefi, F. Shafait, and T. M. Breuel. Robust binarization of stereo and monocular document images using percentile filter. In *C CBDAR 2013, Washington, DC, USA, August 23, 2013,*, pages 139–149, 2013.

[2] N. Babaguchi and K. Yamada. Connectionist model binarization. In *Pattern Recognition*, pages 51–56, 1990.

[3] E. Badekas, N. A. Nikolaou, and N. Papamarkos. Text localization and binarization in complex color documents. In *MLDM Posters*, pages 1–15, 2007.

[4] R. Chamchong and C. Fung. Optimal selection of binarization techniques for the processing of ancient palm leaf manuscripts. *Systems Man and Cybernetics (SMC)*, pages 3796–3800, 2010.

[5] C.-H. Chou, W.-H. Lin, and F. Chang. A binarization method with learning-built rules for document images produced by cameras. *Pattern Recognition*, 43(4):1518–1530, Apr. 2010.

[6] B. Gatos, K. Ntirogiannis, and I. Pratikakis. DIBCO 2009: document image binarization contest. *IJDAR*, 14(1):35–44, 2011.

[7] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):855–68, May 2009.

[8] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks : the official journal of the International Neural Network Society*, 18(5-6):602–10, 2005.

[9] A. Graves and J. Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 545–552, 2009.

[10] J. Kuk, N. Cho, and K. Lee. MAP-MRF approach for binarization of degraded document image. *Image Processing, 2008. ICIP 2008*, pages 2612–2615, 2008.

[11] T. Lelore and F. Bouchara. Document Image Binarisation Using Markov Field Model. In *2009 10th International Conference on Document Analysis and Recognition*, number 2, pages 551–555. Ieee, 2009.

[12] G. Nagy. Twenty years of document image analysis in pami. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(1):38–62, Jan. 2000.

[13] W. Niblack. *An Introduction to digital image processing*. Prentice-Hall, Englewood Cliffs, NJ, 1986.

[14] K. Ntirogiannis, B. Gatos, and I. Pratikakis. A combined approach for the binarization of handwritten document images. *Pattern Recognition Letters*, 35(0):3 – 15, 2014. Frontiers in Handwriting Processing.

[15] H. Orii, H. Kawano, H. Maeda, and N. Ikoma. Text-color-independent binarization for degraded document image based on map-mrf approach. *IEICE Transactions*, 94-A(11):2342–2349, 2011.

[16] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, C(1):62–66, 1975.

[17] I. Pratikakis, B. Gatos, and K. Ntirogiannis. H-DIBCO 2010 - handwritten document image binarization competition. In *ICFHR 2010, Kolkata, India, 16-18 November 2010*, pages 727–732, 2010.

[18] I. Pratikakis, B. Gatos, and K. Ntirogiannis. Icdar 2011 document image binarization contest (dibco 2011). In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 1506 –1510, sept. 2011.

[19] I. Pratikakis, B. Gatos, and K. Ntirogiannis. ICFHR 2012 competition on handwritten document image binarization (h-dibco 2012). In *ICFHR*, pages 817–822, 2012.

[20] I. Pratikakis, B. Gatos, and K. Ntirogiannis. ICDAR 2013 document image binarization contest (dibco 2013). In *ICDAR*, pages 1471–1476, 2013.

[21] T. Sari, A. Kefali, and H. Bahi. An MLP for binarizing images of old manuscripts. *Frontiers in Handwriting Recognition*, pages 247–251, 2012.

[22] J. Sauvola and M. Pietikäinen. Adaptive Document Image Binarization. *Pattern Recognition*, 33:225–236, 2000.

[23] J. Schmidhuber, D. Wierstra, and F. J. Gomez. Evolino : Hybrid Neuroevolution / Optimal Linear Search for Sequence Learning Recurrent Neural Network. In *19th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 853–858, 2005.

[24] H. Tanaka. Threshold Correction of Document Image Binarization for Ruled-line Extraction. *2009 10th International Conference on Document Analysis and Recognition*, pages 541–545, 2009.

[25] O. Trier and A. Jain. Goal-directed evaluation of binarization methods. *Pattern Analysis and Machine Intelligence, ...*, 17(12):1191–1201, 1995.

[26] C.-M. Tsai and H.-J. Lee. Binarization of color document images via luminance and saturation color features. *IEEE Transactions on Image Processing*, 11(4):434–451, 2002.

[27] S. Wu and A. Amin. Automatic thresholding of gray-level using multistage approach. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, pages 493 – 497 vol.1, 2003.

[28] F. Yang, Z. Ma, and M. Xie. A novel binarization approach for license plate. *2006 1ST IEEE Conference on Industrial Electronics and Applications*, pages 1–4, May 2006.