Historical Document Binarization Combining Semantic Labeling and Graph Cuts

Kalyan Ram Ayyalasomayajula, Anders Brun

Centre for Image Analysis, Dept. of Information Technology
Uppsala University
Sweden
{kalyan.ram,anders.brun}@it.uu.se

Abstract. Most data mining applications on collections of historical documents require binarization of the digitized images as a pre-processing step. Historical documents are often subjected to degradations such as parchment aging, smudges and bleed through from the other side. The text is sometimes printed, but more often handwritten. Mathematical modeling of the appearance of the text, as well as the background and all kinds of degradations, is challenging. In the current work we try to tackle binarization as pixel classification problem. We first apply semantic segmentation, using fully convolutional neural networks. In order to improve the sharpness of the result, we then apply a graph cut algorithm. The labels from the semantic segmentation are used as source and sink estimates, with the probability map used for pruning the edges in the graph cut. The results obtained show significant improvement over the state of the art approach.

Keywords: binarization, semantic labeling, deep learning, graph cut, zero shot learning

1 Introduction

In historical document image analysis, binarized images record each pixel as background (parchment/paper) or foreground (text/ink) by preserving most of the relevant visual information in the image. A high-quality binarization significantly simplifies further tasks to be performed on the document image, such as word spotting and transcription. The challenges commonly faced in this area are similar to the problem of uneven illumination for thresholding algorithms. However, historical documents may in addition have other artefacts such as; bleed through; fading or paling of the ink in some areas; smudges, stains and blots covering the text; text on textured background and handwritten documents with heavy-feeble pen strokes for cursive or calligraphic effects to name a few. In general, this makes the task of historical document binarization very challenging as shown in Fig.1. The task is often subjective and garners interest in the field, which has led to the document image binarization content (DIBCO) [12] for automatic methods with minimum parameter setting.

The field of research into binarization in general has led many important methods. A classical approach is the thresholding approach from Otsu [11], which tries to maximize the gray level separation between foreground (FG) and back ground (BG) classes. In an ideal scenario it often suffices to fall back on such a global threshold method. However, local intensity variation and other artifacts introduced in imaging capturing procedure have led to a more *locally adaptive* techniques, such as the methods from Niblack [10], Sauvola and Pietikainen [13]. The techniques discussed in all these methods are generic and applicable to any image in general, however winning entries of DIBCO in the past have developed methods intended to improve binarization through modeling properties of FG/BG in documents images specifically. Lu et al. [8] have for instance modeled background using polynomial smoothing followed by local thresholding on detected text strokes, Bar-Yousef et al. [2] iteratively grow FG and BG within a 7×7 window.

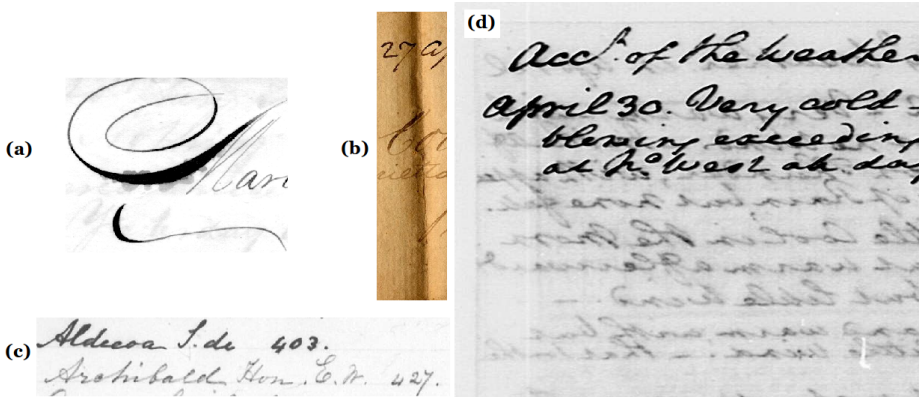

Fig. 1. Examples of typical image degradations are show in the figure (a) smudging of text (b) degradation of paper from aging (c) unevenness of ink in writing (d) bleed through from the ink on the other side of the document

The basic algorithm proposed in this paper draws motivations from other ideas that employed use of Markov random fields (MRF) for binarization such as Howe[4], Mishra et al. [9].These methods take both the global as well as the local aspects of the image into consideration in order to do the pixel labeling. The former uses the Laplacian of the image to obtain invariance in BG intensity, followed by a graph-cut with suitable source-sink (seed points for FG-BG respectively) and edge estimates required to build an image graph. Although the fundamental idea governing Howe's method have been explored previously as separate methods, combining them into an energy function proved particularly

effective. Further improvement of Howe's approach was proposed by Ayyalasomayajula and Brun [1] through detection of seeds for the source and sink estimate with effective detection of edges by exploiting the inherent topology by defining a binarization space.

The proposed method improves upon our previous work in Ayyalasomayajula and Brun [1]. The method builds upon three ideas, basic idea is borrowed from the Howe's approach [4] of defining the binarization as labeling problem. The next step is to incorporate the benefits of defining good seed points for foreground and background regions and idea of edge prunning to delineate them as discussed in our previous work [1]. We use the labeled output and class probabilities map from a fully convolution neural network to improve the FG and BG seed points as well as edge estimates. Further details into understanding the key contributions would require an overview of the above two methods. The following section captures the essence of these methods drawing attentions to the aspects that have been improved.

2 Motivation

2.1 Howe's method

This approach defines the target binarization as a pixel labeling problem that minimizes a global energy function. The energy function consists of two parts a data-fidelity term and a smoothness term for continuity. The former part of the energy relies on the Laplacian of the image intensity to estimate foreground/ink and background/parchment. The Laplacian obtained is invariant to intensity variation in background. The smoothness term of the global energy function incorporates continuities along the ink contours allowing finer details in the text to be preserved. For an image I with each pixel indexed *(i,j)* can be labeled $B_{ij} \in \{0, 1\}$ if it belongs to FG or BG, respectively. The energy function, which results in separating the FG and BG can be written as Eq(1)

$$\begin{aligned} E_I(B) = & \sum_{i=0}^{m}\sum_{j=0}^{n}[L_{ij}^0(1-B_{ij}) + L_{ij}^1 B_{ij}] && \textit{(Data fidelity term)} \\ & + \sum_{i=0}^{m}\sum_{j=0}^{n} C_{ij}^h(B_{ij} \neq B_{i+1,j}) && \textit{(Horizontal continuity term)} \\ & + \sum_{i=0}^{m}\sum_{j=0}^{n} C_{ij}^v(B_{ij} \neq B_{i,j+1}) && \textit{(Vertical continuity term)} \end{aligned} \tag{1}$$

where $L_{ij}^0, L_{ij}^1, C_{ij}^h, C_{ij}^v$ are costs associated with labeling a pixel as belonging to FG, BG, smoothness along horizontal and vertical directions respectively at a

pixel indexed (i, j). Labeling costs are governed by Eq(2)

$$\begin{aligned} L^0_{ij} =& \nabla^2 I_{ij} \\ L^1_{ij} =& \begin{cases} -\nabla^2 I_{ij}, & I_{ij} \leq \mu^r_{ij} + 2\sigma^r_{ij} \\ \phi, & I_{ij} > \mu^r_{ij} + 2\sigma^r_{ij} \end{cases} \end{aligned} \tag{2}$$

where $\nabla^2 I$, $\nabla^2 I_{ij}$, are the Laplacian of image I and Laplacian value at the pixel (i, j) respectively. The $\nabla^2 I_{ij}$ values are enough to label the pixel as belonging to FG or BG however this could lead to extremely large values at image borders due to BG class being far large than FG class. In order to avoid such strong label associations a conservative strategy is applied to BG labeling as indicated in the modified form of L^1_{ij} in Eq(2). where pixels with intensities more than two standard deviations $(2\sigma^r_{ij})$ brighter than the local mean μ^r_{ij}, as computed over nearby pixels weighted by a Gaussian of radius r are assigned a constant cost ϕ. This facilitates the possibility of label switching with some penalty but not restricting it all together.

Smoothness cost is governed by Eq(3)

$$\begin{aligned} C^h_{ij} =& \begin{cases} 0, & E_{ij} \wedge (I_{ij} < I_{i+1,j}) \\ 0, & E_{i+1,j} \wedge (I_{ij} \geq I_{i+1,j}) \\ c, & otherwise \end{cases} \\ C^v_{ij} =& \begin{cases} 0, & E_{ij} \wedge (I_{ij} < I_{i,j+1}) \\ 0, & E_{i+1,j} \wedge (I_{ij} \geq I_{i,j+1}) \\ c, & otherwise \end{cases} \end{aligned} \tag{3}$$

which encourages label consistency on either sides of the edges. As per DIBCO requirement Eq(3) includes edges in the FG, which is a reasonable choice to make. From Eqs(2,3) it can be inferred that there are five parameters that can be associated with the energy function in Eq(1); they are Guassian radius r, label switching cost ϕ, discontinuity cost c, and thresholds (t_{hi}, t_{lo}) associated with Canny edges E_{ij}. Of all the parameters mentioned c and t_{hi} are critical for performance. It is worth noting that both these parameters are associated with edge information and smoothness term of the energy function.

2.2 Topological clustering approach

This method is similar in spirit to Lelore and Bouchara [6] where image pixels are divide into three classes: ink, background and unknown. The FG and BG cluster help define the unknown pixels as belonging to one of the two classes. In this approach an ordered triplet (I, dI_h, dI_v) corresponding to pixel intensity, gradient in horizontal and vertical direction respectively are used to represent every pixel in a three-dimensional space defined as the *binarization space* denoted

by $\mathscr{S}$ in Eq(4). It can be noted that the region $\mathscr{S}$ is bounded as shown below

$$\mathscr{S} = \begin{cases} 0 \leq I \leq 255, \\ -255 \leq dI_h \leq 255, \\ -255 \leq dI_v \leq 255. \end{cases} \tag{4}$$

This theoretical framework helps to define a *topology* [1] within this space with a natural way of defining hierarchical clusters with neighborhood constraints. Fig.2 shows the space $\mathscr{S}$ for a typical image from the DIBCO dataset with points in red and blue representing the FG and BG pixels, respectively. The FG and BG clusters are encircled with ellipses for a *core-object* denoted in * for a given ε-reachable neighborhood N_ε. For all practical purposes with a certain abuse of notation we can think of *core-object* as choice of cluster center with ε denoting the maximum distance separating a pixel from core point in $\mathscr{S}$. The advantage in this approach is that it allows to iterative refine the cluster hierarchically with density of points packed in it. Upon using the core-objects as estimates for source-sink, the information about the cluster associated with core objects help in including only the relevant edges for the graph cut to make a better segmentation of FG from BG thus improving the result.

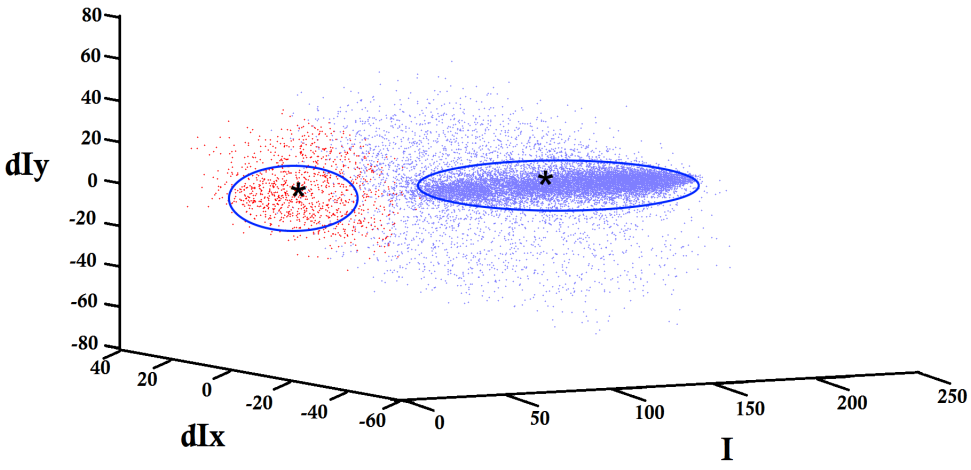

Fig. 2. Distribution of FG and BG pixels as points in binarization space in red and blue respectively, for the file *'2009_H02.bmp'* from the DIBCO dataset.

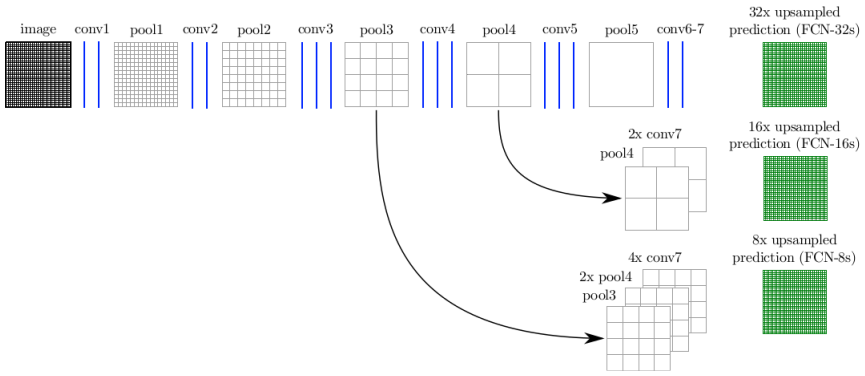

Fig. 3. Architecture learns to combine coarse, high layer information with fine, low layer information. Pooling and prediction layers are shown as grids that reveal relative spatial coarseness, while intermediate layers are shown as vertical lines. FCN-32s, FCN-16s, FCN-8s represent the prediction stride corresponding to pixels in a single step.

2.3 Proposed method

From Howe's method and the clustering approach, it can be inferred that results of graph-cut can benefit a lot from an approach that can produce better edge estimates. To meet this end we propose an approach to assist the graph-cut through source-sink and edge estimates from a *Fully Convolutional Neural Network* (FCNN). The idea is built from the fact that FCNNs have been very effective in producing state-of-the-art results in semantic segmentation [14]. The binarization results can be posed as a semantic segmentation problem with two class instances. Fig.3 shows the architecture of a FCNN as a *single-stream* and is taken from [14]. It learns to combine coarse, high layer information with fine, low layer information. Pooling and prediction layers are shown as gray grids, reveal relative spatial coarseness, while intermediate convolutional layers are shown in blue as vertical lines.

The output from strided convolution and a pooling layer is a down sampled version of the input affecting the resolution at the final layer. To match output the labels at the same resolution as the input, upsampling layers are incorporated. Depending on the number of upsampling layers used, the FCNN in [14] can have three variations of stream-nets. Referring to Fig.3 the first row of an FCN-32s: depicts a single-stream net, it upsamples stride 32 predictions back to pixels in a single step. Second row FCN-16s: combines predictions from both the final layer and the pool4 layer, at stride 16, it allows the network to predict finer details, while retaining high-level semantic information. Third row FCN-8s: combines additional predictions from pool3, at stride 8, to provide further precision. Combining the low level features with high level semantic information is achieved through skip connections that allow for a better gradient propagation while training. A trained network can be used to predict the probability of the

pixel belonging to FG/BG class can be obtained soft-max layer with the labeled pixels based on the maximum probability of a pixel belonging to a class.

All the observations made so far have been incorporated into our proposed method as follows:

- The labels obtained from FCNN output act as very good source and sink estimates.
- The network performs very well in estimating the background, but gives a very conservative estimate of the foreground labels. This is due to the class imbalance between FG and BG pixels, as the FCN is optimized for "overall accuracy" on class label prediction.
- The probability map of the BG class thresholded on mean BG class probability is good indication of FG/BG separation.
- A graph is constructed by including edges within this thresholded probability mask give a good estimate of the FG spread as shown in Fig.4(a),(b).
- These source, sink and edge map serve as very reasonable estimates for a graph-cut algorithm to be applicable.

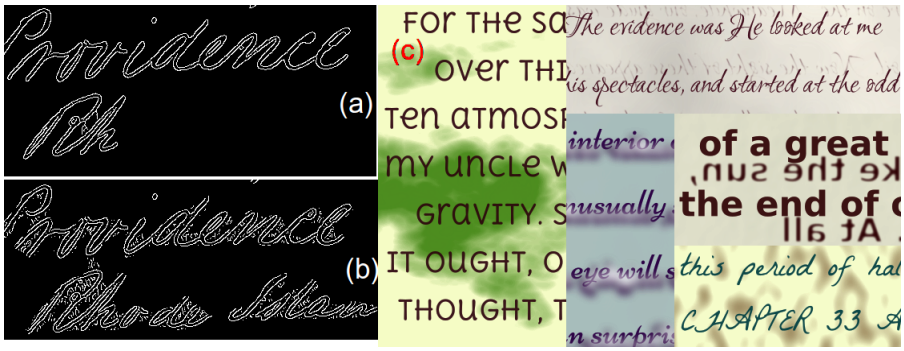

Fig. 4. Fig (a) shows edge map from Canny threshold picked through optimization as described in [5], fig (b) shows the pruned edge map from the class probabilities, fig (c) shows few samples from synthetic text data

3 Experiments

3.1 DIBCO Data Processing

The experiments were conducted on the *DIBCO* [12] datasets for binarization consisting of 76 images. The dataset was divided into training, validation with 70-30 split. The and ground truth were converted into 500×300 pixels of cropped images with overlap of 100 pixels horizontally and vertically to augmented data in order to create more data for training. The cropped size 500×300 permitted

the images to fit into memory in Convolutional Architecture for Fast Feature Embedding framework [15] (commonly known as CAFFE). The model was then initialized with weights from pre-trained model on PASCAL-VOC dataset available at [16].

3.2 Training

The FCN-8 architecture was used as it has a better receptive field which translates to accurate pixel labeling for binarization. In order to train the network on binarization data, the weights for layers till FC7 are loaded from the pre-trained model and layers beyond FC7 are trained on the DIBCO dataset using the ground truth labels. The training was continued for 150,000 iteration till an accuracy of 75% was obtained for mean Intersection over Union (mIoU) for predicted vs. ground truth segmented regions.

3.3 Synthetic Data Processing

In order to eliminate any bias that could arise due to over-fitting the network to data, the experiments were repeated by training the network on synthetic data. Documents resembling historical handwritten and printed material were generated synthetically. Various filters were applied to resemble background textures and degradations in the parchment. The text was generated using handwriting and machine printed fonts from the Google™fonts [17]. Fig.4(c) shows few cropped images from the synthetic dataset. The results from binarization on DIBCO dataset using the network trained on the synthetic dataset are presented in Table-1 in FC(S) column to compare with a network trained in DIBCO dataset in FC column.

4 Results

The trained network as described previously was then used to predict the labels for the test data as shown in Fig.5. The results as shown in R_3 do a good job of estimating the back-ground very accurately. However the fore-ground estimate is quite conservative as to where the ink/text is present. These results have been used in conjunction with a min-cut max-flow based segmentation [3] to improve the results over state-of-the-art approach in DIBCO 2013-2016 binarizations; were based on Howe's method [5], [4] or its variants in some form.

The Table 1. provides quantitative evaluation of Howe's (Howe), Topological Clustering (TCl), FCNN+GraphCut trained on DIBCO dataset (FC) and FCNN+GraphCut trained on synthetic data set (FC(S)) comparing few of the DIBCO metrics [12] for brevity. Results shown are for the files which have shown more then 10% gain on any one of the metrics in absolute scale. A high score on F-Measure, Peak Signal to noise ratio (PSNR) and a low Distance Reciprocal Distortion Metric (DRD)[7] is desirable for an algorithm which are defined in

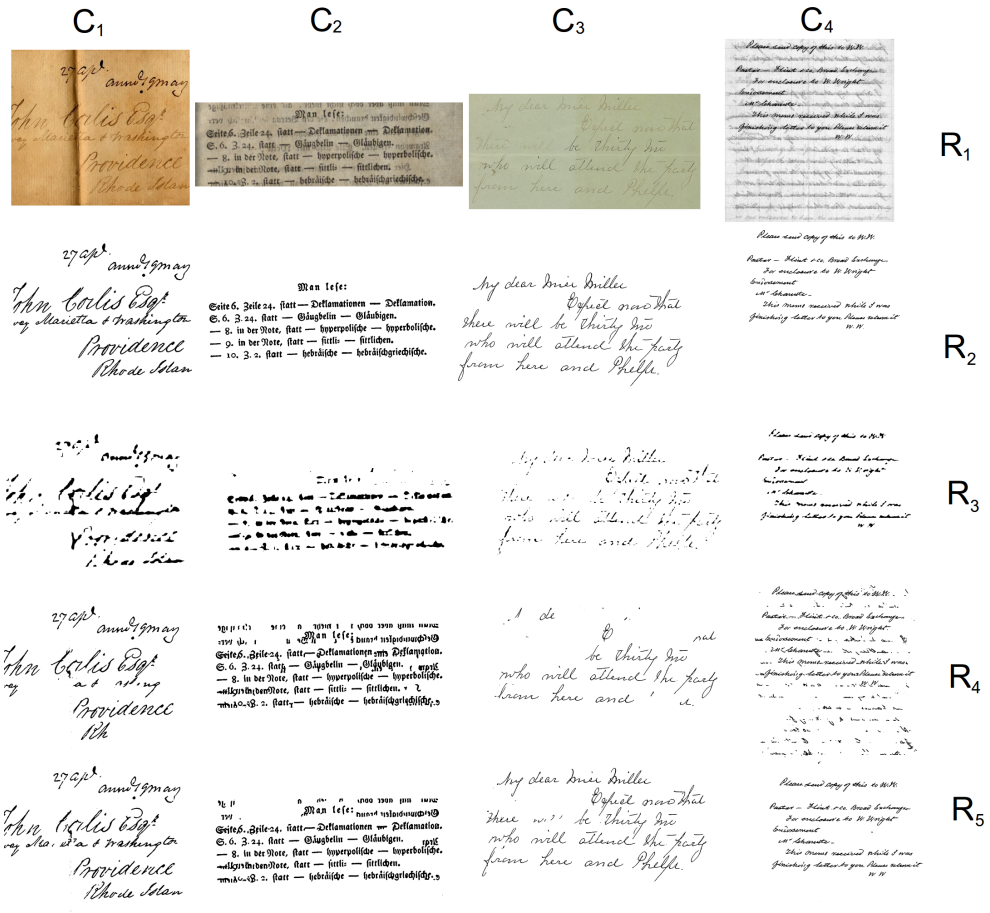

Fig. 5. Image Rows R_1, R_2, R_3, R_4 and R_5 shows the DIBCO images, ground truth, FCNN, Howe and FCNN+GraphCut outputs respectively for images shown in columns $C_1 - C_4$.

the appendix.

The F-Measure is a well understood metric for measuring classification accuracy where there is 2.1% gain in mean absolute percentage over the entire DIBCO dataset with 76 files, which is significant improvement. PSNR metric is useful when comparing with the topological clustering with FCNs as it can serve as a measure to associate unlabeled pixels in binarization space $\mathscr{S}$ to the exact classes where there is a 3.3% gain in mean relative scale. Both the metrics discusses are generic and DRD is more tailored for binarization of document images and there is a significant boost of 27.1% in results over mean relative

Table 1. Comparison of the results for F-Measure, PSNR, DRD

File Name	FMeasure				PSNR				DRD			
	Howe	TCl	FC	FC(S)	Howe	TCl	FC	FC(S)	Howe	TCl	FC	FC(S)
2011_HW6.png	78.9	76.1	**86.4**	43.9	15.6	14.4	**17.2**	12.7	5.50	8.6	**3.4**	11.7
2011_PR2.png	74.9	79.9	**80.1**	78.4	11.4	12.7	**12.7**	12.4	13.4	9.6	**9.3**	10.1
2012_H13.png	63.0	78.1	**87.4**	44.9	15.5	17.2	**19.2**	14.5	8.02	5.2	**2.9**	10.4
2013_HW5.png	69.3	85.7	**91.7**	86.9	15.9	20.1	**22.8**	21.1	19.1	6.1	**2.4**	4.0
2013_HW7.png	51.4	51.6	**74.3**	34.0	18.1	18.1	**20.0**	17.3	7.28	7.3	**4.5**	8.8
2013_PR6.png	70.0	72.8	**84.8**	79.5	10.3	10.9	**14.1**	13.1	21.9	18.9	**7.9**	10.1
Mean over DIBCO	87.7	88.3	**89.8**	75.1	17.8	18.0	**18.4**	16.0	4.2	4.0	**3.1**	7.2

scale. Fig.5 compares the results qualitatively for the methods to provide visual cue to the gain in results.

4.1 Conclusions and Future Work

The current experimental framework shows relevance of FCNNs in segmentation based tasks such as binarization in documents to obtain state-of-the-art results. The possibilities as stated below are speculations into intended directions to pursue our future research into this field. The current approach can be extended to other tasks such as layout analysis and de-noising of historical documents. The method can benefit greatly from training on synthetic data generated by mimicking various degradations through filter operation thus enabling zero shot learning. Though unable to achieve state-of-the-art results on synthetic data the preliminary results show potential for improvement and benefit along this direction. The graph-cut can be integrated as a loss layer into the network to provide end-to-end binarization. Also training the network on a modified loss layer based on DIBCO metrics may lead to improved results.

Acknowledgment

This project is a part of q2b, From quill to bytes, an initiative sponsored by the Swedish Research Council "Vetenskapsrådet D.Nr 2012-5743) and Riksbankens Jubileumsfond (R.Nr NHS14-2068:1) and Uppsala university. The authors would like to thank Fredrik Wahlberg and Tomas Wilkinson of Dept. of Information Tech., Uppsala University for their constructive criticism in improving the manuscript.

References

1. Ayyalasomayajula K.R., Brun A.: Document binarization using topological clustering guided Laplacian Energy Segmentation : Proceedings of ICFHR, 2014, pp. 523-528.

2. Bar-Yosef, I., Beckman, I., Kedem, K., Dinstein, I.: Binarization,character extraction and writer identification of historical Hebrew calligraphy documents. Int. J. Doc. Anal. Recogn. 9(2), 8999 (2007)
3. Boykov Y. and Kolmogorov V.: An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision, PAMI 26(9):11241137, September 2004.
4. Howe, N.: A Laplacian energy for document binarization:International Conference on Document Analysis and Recognition, pp. 610 (2011)
5. Howe, N.R.: Document Binarization with Automatic Parameter Tuning, International Journal on Documant Analysis and Recognition, doi:10.1007/s10032-012-0192-x, 2012
6. Lelore, T., Bouchara, F.: Super-resolved binarization of text based on FAIR algorithm. In: International Conference on Document Analysis and Recognition, pp. 839843 (2011)
7. Lu H., Kot A. C., Shi Y.Q.: Distance-Reciprocal Distortion Measure for Binary Document Images, IEEE Signal Processing Letters, vol. 11, No. 2, pp. 228-231, 2004.
8. Lu, S., Su, B., Tan, C.L.: Document image binarization using background estimation and stroke edges. Int. J. Document Anal. Recogn. 13(4), 303314 (2010)
9. Mishra, A., Alahari, K., Jawahar, C.V.: An MRF model for binarization of natural scene text. In: International Conference on Document Analysis and Recognition (2011)
10. Niblack W.:An Introduction to Digital Image Processing. Prentice-Hall, Englewood Cliffs (1986)
11. Otsu N.: A threshold selection method from gray level histograms, IEEE Trans. Systems, Man and Cybernetics, vol.9, pp 62–66, 1979.
12. Pratikakis, I., Gatos, B. and Ntirogiannis, K. : ICDAR 2011 document image binarization contest (DIBCO 2011), International Conference on Document Analysis and Recognition,15061510, 2011
13. Sauvola, N., Pietikainen, M.: Adaptive document image binarization. Pattern Recogn. 33(2), 225236 (2000)
14. Shelhamer E., Long J., and Darrell T.: Fully Convolutional Networks for Semantic Segmentation: arXiv:1605.06211, 2016
15. Yangqing J., Evan S., Jeff D., Sergey K., Jonathan L., Ross G., Sergio G., Trevor D.: arXiv preprint arXiv:1408.5093, Caffe: Convolutional Architecture for Fast Feature Embedding, 2014
16. http://dl.caffe.berkeleyvision.org/fcn8s-atonce-pascal.caffemodel
17. https://github.com/google/fonts

Appendix

F-Measure

$$F - Measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{5}$$

where, $Recall = \frac{TP}{TP+FN}$, $Precision = \frac{TP}{TP+FP}$, TP, FP, FN denote the True Positive, False Positive and False Negative values, respectively.

Peak Signal to Noise Ratio

$$PSNR = 10 \log \left(\frac{C^2}{MSE} \right) \tag{6}$$

where, $MSE = \frac{\sum_M^{i=1} \sum_N^{j=1} \left[I(i,j) - I'(i,j) \right]^2}{MN}$ PSNR is a measure of how close is one image to another. Higher the value of PSNR, more is the similarity between binarized image and the ground truth. Note that C equals to the difference between foreground and background, M, N are the width and height of the image respectively.

Distance Reciprocal Distortion Metric

The DRD Metric serves as a measure of visual distortion in a binary document images [7]. It correlates with the human visual perception and measures the distortion for all the $\mathbb{S}$ flipped pixels as follows:

$$DRD = \frac{\sum_{k=1}^{S} DRD_k}{NUBN} \tag{7}$$

where $NUBN$ is the number of non-uniform (gray pixels) 8x8 blocks in the ground truth (GT) image, and DRD_k is the distortion of the k-th flipped pixel that is calculated using a 5×5 normalized weight matrix W_{Nm} as defined in [7]. DRD_k equals to the weighted sum of the pixels in the 5×5 block of the GT that differ from the centered k-th flipped pixel at (x, y) in the binarization result image $\mathbb{B}$ as defined below:

$$DRD_k = \sum_{i=-2}^{2} \sum_{j=-2}^{2} |GT_k(i,j) - B_k(i,j)| \times W_{Nm}(i,j) \tag{8}$$