CT-Net
He, Sheng; Schomaker, Lambert

Published in:
Pattern recognition

DOI:
10.1016/j.patcog.2021.108010

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2021

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
CT-Net: Cascade T-shape deep fusion networks for document binarization

Sheng He\textsuperscript{a,*}, Lambert Schomaker\textsuperscript{b}

\textsuperscript{a} Department of Radiology, Boston Children’s Hospital, Harvard Medical School, Harvard University, Boston MA 02215, USA
\textsuperscript{b} Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, PO Box 407, Groningen 9700 AK, The Netherlands

A R T I C L E   I N F O

Article history:
Received 14 June 2020
Revised 19 October 2020
Accepted 27 April 2021
Available online 5 May 2021

Keywords:
Cascade T-Net
Binarization
Enhancement
DIBCO
Convolutional neural networks

A B S T R A C T

Document binarization is a key step in most document analysis tasks. However, historical-document images usually suffer from various degradations, making this a very challenging processing stage. The performance of document image binarization has improved dramatically in recent years by the use of Convolutional Neural Networks (CNNs). In this paper, a dual-task, T-shaped neural network is proposed that has the main task of binarization and an auxiliary task of image enhancement. The neural network for enhancement learns the degradations in document images and the specific CNN-kernel features can be adapted towards the binarization task in the training process. In addition, the enhancement image can be considered as an improved version of the input image, which can be fed into the network for fine-tuning, making it possible to design a chained-cascade network (CT-Net). Experimental results on document binarization competition datasets (DIBCO datasets) and MCS dataset show that our proposed method outperforms competing state-of-the-art methods in most cases.

© 2021 Elsevier Ltd. All rights reserved.

1. Introduction

Document binarization, which is a common pre-processing step, aims to classify each pixel in a degraded image into either text or background. Document binarization is a fundamental research topic for document processing approaches including layout analysis [12], historical document analysis [3–5] and word spotting [6,7]. The quality of document images and their binarization maps affect the performance of high-level processing tasks, especially for methods which extract features based on binarized images [3]. Therefore, a high-quality and accurate binarization of a degraded image can boost the performance of the final task in the workflow. The problem of binarization attracts interest in the field, leading to nine document image binarization contests (DIBCO) from the year 2009 and several document images with ground-truth are released every year (except 2015).

Document binarization is a challenging problem because most documents suffer from various degradations such as pitch-black or pure-white margins; feeble contrast between ink and parchment; smear; stain; uneven illumination and pen strokes; artifacts and bleed-through [8,9]. Examples are shown in Fig. 1. These degradations affect the results of edge detection, intensity distribution computation and stroke-width estimation, which are the basic steps of traditional binarization methods [10–13]. Therefore, the performance of these traditional binarization methods is limited and none of them can deal with all types of degradations.

Recently, the performance of document binarization has been greatly improved by using the convolutional neural network [8,14], which has a large information capacity and it can correct various degradations learned from the training set. Document binarization is similar to the semantic segmentation problem [15] and the neural networks used for semantic segmentation can also be used for binarization, such as the fully convolutional network [16], hierarchical deep-supervised network [14] inspired by holistic edge detection (HED) [17] and U-Net [18,19].

1.1. Motivation

There are two types of tasks related to document binarization: (1) Binarization task. Most methods consider the document image binarization as a single task and train the neural network to directly classify each pixel into either text or background [8,14,20]. (2) Enhancement task. The aim of an enhancement task is to train a model to improve the quality of the input image and output an enhanced version which is locally uniform [21,22]. A neural network for enhancement should learn the degradations which are present in the original image, yielding a clear and enhanced image [19]. In this paper, we train a neural network to jointly learn...
the binarization and enhancement task and transfer the learned degradations from the enhancement task to the binarization task to improve the performance. Training the neural network jointly for enhancement and binarization is expected to improve the performance of binarization because the learned features are shared between two tasks and the risk of over-fitting for binarization is reduced. In addition, features learned for enhancement contain the degradation information and adapting these features inspired by the deep-adaptive learning approach [23] to binarization can further improve the performance.

The proposed framework shown in Fig. 2 implicitly decomposes the binarization into two steps, corresponding to two neural networks: $N_b$ is the residual network to learn the enhanced image $e$ and $N_e$ is the network to learn the binarized image $b$, given the input image $x$. The enhanced image $e$ can be denoted by: $e = N_e + x$, thus the network $N_e$ learns the degradations (the differences between the degraded and clear images) in the image $N_e = e - x$. We transfer the deep features learned from the network $N_e$ to the binarized network $N_b$ by the add operation, making the network $N_b$ indirectly learning from $x + N_e = x + e - x = e$ which is the enhanced image. The network $N_e$ learns the degradations which are transferred to the binarization network $N_b$ to correct the deep features of the $N_b$. Thus, unlike the traditional neural network learning the binarization directly from the input image [14,20], our network $N_b$ yields good performance by learning the binarization from the corrected or enhanced (degradation-free) deep features learned by the $N_e$.

1.2. Method

Although there are different ways to implement the framework shown in Fig. 2, in this paper, we propose a T-shape network (T-Net) which performs these two tasks within one neural network, similar to the multi-task learning problem [24] which aims to model related tasks jointly. The T-Net receives the degraded image as input and outputs two maps: one is the binarization map and another one is the enhancement map. The difference between document enhancement and binarization is that document enhancement aims to improve the perceptual quality by removing or correcting various degradations [19] and the output image of enhancement is an improved version of the input image, which it is not guaranteed to be binarized.

Since the enhanced image can be considered as an improved version of the input image, it can also be used as the input for fine-tuning, making it possible to design a cascade neural network to deal with various types of degradations iteratively. The cascade approach presents a powerful architecture for boosting performance iteratively and has been widely used in various tasks [25–27]. Cascade neural networks usually have multiple stages and the earlier stages can correct the easy degradations whereas the later stages can pay more attention to the hard degradations in the document images [28].

Inspired by the multiple-task and cascade-learning methods, we propose a Cascade T-shape neural network (CT-Net) for document binarization. The CT-Net consists of several T-Nets and each T-Net has three branches, one branch is the encoder which converts the input images into features maps in different levels and the other two branches are two decoders corresponding to the binarization task and the enhancement task. The encoder is shared between two tasks and two different decoders learn specific features for each task. In addition, as inspired by the deep adaptive learning method [23], we also transfer the features learned for enhancement to the features for binarization, named CTada-Net. As discussed above, the learned features in the decoder for enhancement contain the degradation information of the input image, which makes the decoder of binarization to learn indirectly from the enhanced features and thus improve the performance.

The remainder of this paper is organized as follows. Section 2 covers a background overview of document binarization. In Section 3, we describe our proposed methods in detail. Experimental results are given in Section 4 and conclusions are presented in Section 5.

2. Related work

There are two groups of binarization methods proposed in the literature: traditional methods which are based on image processing techniques, usually at the single pixel-intensity level, and deep learning methods which apply convolutional neural networks to learn an end-to-end model for binarization while exploiting detailed regional image information. Traditional methods usually compute a threshold on each pixel while the trained deep models predict the label of each pixel in a document image.

2.1. Traditional threshold-based methods

The classical threshold-based method for binarization is the Otsu method [29], which computes a global intensity threshold
to minimize the intra-class variance on the whole document images and it provides a good performance if the histogram of the intensity of document images has bimodal distribution. The global threshold method does not work well on document images with high degradations and uneven distributions. Therefore, local threshold methods have been proposed, such as Niblack [30] and Sauvola [10], which compute the threshold on each pixel, based on statistical information in a local patch.

In [31], a local adaptive threshold method is proposed which relies on a background surface estimation for computing the threshold on each pixel. AdOtsu [12] introduces an estimated background map to determine whether to use the Otsu's threshold or not on a local patch because using Otsu directly on small patches is impossible due to the fact that it always separates pixels into two classes, which makes large errors on background patches without any text. T-transition pixels, which is a generalization of edge pixels, are defined and used for binarization in [32] since texts in document images usually have strong edges. The binarization methods proposed in [11,33,34] use the detected edge, combined with an adaptive contrast map computed by using local maximum and minimum [35] and ensemble strategy [36] which is tolerant to background and text variation caused by variant degradations, to determine the threshold on each pixel. The method proposed by Jia et al. [13] uses the structural symmetric pixels to compute the local threshold determined by the voting result of multiple thresholds on each pixel.

Howe [37] defines the binarization problem as an optimization problem based on a global-energy function whose data-fidelity term relies on the Laplacian of the image intensity and the smoothness term is related to edge discontinuities. Later, Howe [38] proposes a automatic parameter tuning method for binarization. Nafchi et al. [39] use phrase congruency-based features maps and Otsu’s threshold for binarization. Recently, a game theory inspired binarization is proposed in [40], which uses a two-player, non-zero-sum, non-cooperative game to extract the local information, followed by a K-means method for text classification.

2.2. Deep learning-based methods

With the development of deep learning, convolutional neural networks have also been used for binarization, exploiting regional texture. A fully convolutional network, which is originally proposed for semantic segmentation in natural images [15], is adapted for the binarization task in [16] with a pseudo F-measure loss. Inspired by HED [17], a hierarchical deep-supervised network (DSN) architecture [14] is proposed to learn the labels of each pixel at different feature levels. A primal-dual network (PDNet) [8] which combines an energy minimization function and convolutional network learned features for binarization. In [18], a selection auto-encoder network is used to predict a selection of values for each pixel’s confidence value to determine whether the pixel belongs to text or background. Zhao et al. [20] formulate the binarization as an image-to-image translation task and use the conditional generative adversarial networks (cGANs) to combine the multi-scale information for document image binarization.

Unlike the previous works which use a deep neural network to predict the label of each pixel directly, the DeepOtsu [19] method trains a network to output an enhanced image, which corrects the degradations in the input image in a gradual manner. The final binarization map can then be obtained from the enhancement map by applying the global Otsu method.

The proposed method in the current paper is also a deep learning-based method, based on a vanilla version of U-Net. However, the algorithm is modified to address the dual tasks of binarization and enhancement simultaneously, instead of requiring a separate final Otsu stage as in [19]. In addition, the proposed method here decomposes the binarization into two steps: enhancement and binarization. The enhancement step learns the degradations which is transferred to the binarization step, making the binarization based on the degradation-free deep features.

3. The proposed method

This section describes the proposed CT-Net and Ctada-Net neural networks for document binarization and enhancement in a multi-task learning scenario.

3.1. Problem formulation

We used $x$ to denote the input degraded image, $b$ to denote the output binarized map and $e$ to denote the enhancement version of $x$. Both the binarization and enhancement tasks can be regarded as an image-to-image translation problem, which can be solved by the typical U-Net [20,41].

In this paper, a T-shape neural network is proposed for both enhancement and binarization tasks (shown in Fig. 3), named as T-Net, which is defined as:

$$
c. \tilde{m} = E(x)$$
$$e. \tilde{n} = D_b(c, \tilde{m}) + x$$
$$b = D_b(c, m)$$

where $E$ is the encoder which contains five convolutional layers and coverts the input image into feature maps $\tilde{m}$ and $c$, $e$ is the context and contains the semantic information of the input image and $\tilde{n} = m_1, \ldots, m_n$ are the intermediate spatial feature maps from the encoder with different resolutions, $D_b$ and $D_e$ are the decoders for binarization and enhancement, respectively. More details can be found in Fig. 3. Our proposed T-Net can be considered as the generalized U-Net in the multi-task scenario.

Compared to the vanilla U-Net [41], our proposed T-Net has another branch $D_e$ for enhancement, which also uses the context information $c$ and feature maps $\tilde{m}$ from the encoder branch. In other words, the two tasks share the same encoder $E$ which learns the shared representation but use different decoders $D_e$ and $D_b$ which learn the task specific representation. Note that we added the original input $x$ on the enhancement decoder $D_e$ so it can be rewritten as $D_e(c, \tilde{m}) = -(x - e)$, which means that the decoder $D_e$ learns the negative degradations $d = -(x - e)$ which are the differences between the clear and enhanced image $(e)$ and degraded image $(x)$, similar to DeepOtsu [19]. As discussed in [23,42], sharing parameters among multi-tasks can greatly reduce the risk of overfitting on a specific task.

In T-Net, the decoders $D_b$ and $D_e$ learn specific features for each task without any interaction. However, the information learned by $D_e$ for enhancement should be also helpful for binarization since the task of enhancement is to remove the degradations and obtain a clear and uniform image. In order to allow the network to learn the relationship between these two tasks, we also applied the shortcuts from the decoder $D_e$ to the decoder $D_b$ to improve the information flow, inspired by the multi-task learning [23,24]. The T-Net with the task adaption, named Tada-Net, can be denoted as:

$$c. \bar{m} = E(x)$$
$$e. \bar{n} = D_b(c, \bar{m}) + x$$
$$b = D_b(c, \bar{m} + \bar{n})$$

where $\bar{n}$ represents the spatial feature maps generated by the decoder $D_b$ during recovering the enhancement image $e$. We added the intermediate spatial feature maps $\bar{m}$ and $\bar{n}$ in the decoder $D_b$ since $\bar{n}$ is from the decoder $D_b$ which contains the negative degradation information $(D_e(c, \bar{m}) = -(x - e))$. Thus, $\bar{m} + \bar{n}$ can remove
Fig. 3. The overall architecture of the proposed T-Net. It has three branches, encoder $E$ and decoder $D_t$ for enhancement task and $D_b$ for binarization task. Short cuts are applied between encoder and decoders. In the encoder, we use max-pooling for downsampling and in decoder we use deconvolutional operations for upsampling.

Fig. 4. The overall architecture of the proposed Tada-Net. The structure of the encoder $E$ is same to the one in Fig. 3. The dashed lines denote the short cuts from the decoder $D_b$ to $D_t$.

the learned degradations from the input feature map $\tilde{m}$, improving the performance of binarization. Fig. 4 shows the structure of Tada-Net.

As mentioned in [19], the enhancement image $e$ is the latent uniform and clear version of the input image and recovering it allows the network to learn the degradations and binarization iteratively. More precisely, the enhancement map $e$ may still contain some slight degradations and it can be fed into the network for further processing, which makes it possible to design a successful architecture for binarization in a cascade way. In this paper, we extended the T-Net into cascade T-Net, named CT-Net, which is defined as:

$$e^{i+1}, b^{i+1} = T_i(e^i), e^0 = x$$

(3)

where $T_i(x)$ can be either T-Net (CT-Net) or Tada-Net (CTada-Net). $e^i$ is the $i$-th enhancement map of T-Net and $e^0 = x$ is the original input image. Fig. 5 shows the framework of the proposed CT-Net.

3.2. Network configuration

Inspired by the U-Net structure [41], the encoder of the CT-Net contains five convolutional layers with channel size [32,64,128,256,512], followed by batch normalization, ReLU, and max-pooling (2×2, stride:2) layers. For decoders, the deconvolutional layers are used for upsampling feature maps which is concatenated with the corresponding spatial feature maps in $\tilde{m}$ from the encoder branch, followed by a convolutional layer for feature combination. The size of all kernels used in convolutional and de-convolutional layers is 3×3, except the last layer which uses 1×1 kernel to compute the output map. The input color image size is set to $h \times w \times 3$, where $h$ and $w$ are height and width. The output layer of the enhancement has the same size as input. For binarization output, the size is $h \times w \times 1$, in which the value on each pixel denotes the probability computed by a sigmoid function.

The loss of the enhancement is defined as:

$$L_e = \frac{1}{n} \sum |e_t - e|$$

(4)

where $e_t$ is the ground-truth of the enhancement image, $e$ is the output of the network and $n$ is the total number of pixels in the image. The loss of the binarization is the typical cross-entropy loss, given by:

$$L_b = -\frac{1}{n} \sum b_t \cdot \log(p) + (1 - b_t) \cdot \log(1 - p)$$

(5)

where $b_t$ is the ground-truth of the binarization map and $p$ is the probability from the neural network. In fact, each T-Net network in CT-Net can be trained separately or jointly. In this paper, we trained the network jointly and the final training loss is given by:

$$L_{\text{joint}} = \sum_{i=1}^{N-3} (L_{e_i} + L_{b_i})$$

(6)

where $L_{e_i}$ and $L_{b_i}$ are the enhancement and binarization losses in the $i$-th T-Net and $N = 3$ is the total number of the T-Net.

4. Experimental results

4.1. Datasets and evaluation metrics

We evaluated the proposed method on nine DIBCO benchmark datasets from the document binarization competitions, including DIBCO’09 [43], H-DIBCO’10 [44], DIBCO’11 [45], H-DIBCO’12 [46], DIBCO’13 [47], H-DIBCO’14 [48], H-DIBCO’16 [49], DIBCO’17 [50], H-DIBCO’18 [51]. The number of documents on each dataset is shown in Table 1. We used the leave-one-dataset-out strategy inspired by [8,20] for evaluation. More precisely, when evaluating on a particular DIBCO dataset, all rest of datasets are used for training. The training set also includes images from the Bickly-diary dataset [52], Persian Heritage Image Binarization dataset (PHIDB) [53] and Symchromedia Multispectral dataset [54], which were also used for training the networks [14,19].

Four metrics which were widely used in the contests [49–51] were used to assess the performance and quantitatively compared with state-of-the-art methods, including the F-measure (FM), pseudo F-measure ($F_p$), peak signal-to-noise ratio (PSNR) and distance reciprocal distortion metric (DRD). For fair comparison, we used the evaluation tool provided by the DIBCO competitions to compute these metric scores in this paper.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Handwritten</th>
<th>Printed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIBCO’09 [43]</td>
<td>5</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>H-DIBCO’10 [44]</td>
<td>10</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>DIBCO’11 [45]</td>
<td>8</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>H-DIBCO’12 [46]</td>
<td>14</td>
<td>-</td>
<td>14</td>
</tr>
<tr>
<td>DIBCO’13 [47]</td>
<td>8</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>H-DIBCO’14 [48]</td>
<td>10</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>H-DIBCO’16 [49]</td>
<td>10</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>DIBCO’17 [50]</td>
<td>20</td>
<td>-</td>
<td>20</td>
</tr>
<tr>
<td>H-DIBCO’18 [51]</td>
<td>10</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>95</td>
<td>21</td>
<td>116</td>
</tr>
</tbody>
</table>
The overall architecture of the proposed CT-Net with three T-Nets. The output of the enhancement map on T-Net-1 can be considered as the improved version of the original image which can be fed into T-Net-2 for fine-tune. This is the same for T-Net-3. T-Net-n are identical T-Nets with different parameters.

Training samples (red box left column) with their corresponding enhancement ground-truth (middle column) and binarization ground-truth (right column). All patches have the same spatial size of 256 × 256. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Implementation details

Ideally, the entire image instead of patches can be processed by the proposed CT-Net. However, due to the limitation of GPU memory it is very hard to work with arbitrary-sized inputs. The most popular way is to crop small patches and resize them to a fixed width and height [18,20]. The most common size is 256 × 256, which is also adopted in the proposed CT-Net. For each image on DIBCO datasets, we cropped image patches with 256 × 256 with 10% overlap horizontally and vertically. For scale augmentation, we also sampled patches with the scale factor (0.75, 1.25, 1.5) based on the input size of the network. In addition, the global patches whose sizes are equal to the minimum of the entire image width and height are cropped and resized to 256 × 256 for training and testing, inspired by Vo et al. [14]. The network is initialized by the Xavier method [55] and is trained by the Adam optimizer [56] with an initial learning rate 0.0001, which is decreased to half at every 50,000 iterations. The number of training iterations is 200,000 with the batch size 10 due to the limitation of the memory.

The ground-truth of the binarization is provided in DIBCO datasets while the ground-truth of the enhancement is built by using the same method in [19]. The value of each pixel on the enhancement ground-truth image is computed as the average pixel value within the same label within the patch. The label of each pixel can be obtained from the ground-truth of the binarization. Fig. 6 shows several training samples used in this paper for training the neural networks.

During testing, the patches are cropped from the test images with the same method used to crop patches for training. The predicted binarized maps from the trained networks are stitched to the original document size by averaging values over the overlapping patches. Therefore, the value of each pixel on the predicted maps is the probability in [0,1]. The final binarization map is computed with a global threshold T based on the predicted probability map. Unlike the work in [14] which uses a fix T learned from a separated training set for all documents, we used Otsu [29] to compute the threshold T for each document. It is very efficient to compute the Otsu threshold and each image has its own threshold depending on the quality of predicted probability maps from the network.

4.3. Performance evaluation

In this paper, we stacked three T-Nets due to the trade-off between performance and model complexity. It has been shown in DeepOtsu [19] that the performance can be improved by cascading at least three basic components (T-Net) in the network. We evaluated the performance of binarization results from each T-Net and Tada-Net in the cascaded CT-Net and CTCT-Net. Fig. 7 shows the average scores of F-measure and DRD overall 116 binarization maps with different n of CT-Net-n and CTCT-Net-n. From the figure we can see that the performance in terms of F-measure and DRD increases as stacking more T-Net in the CT-Net. Practically, we also found the similar trends in terms of Fps and PNSR. Thus, we can conclude that cascading different T-Nets can iteratively improve the performance. In addition, the CTCT-Net provides better performance than CT-Net, which demonstrates the powerful of the adaptive shortcuts from the enhancement task to the binarization task.

We roughly divided all documents in DIBCO datasets into eight groups, according to main degradations contained in document images as shown in Fig. 1. The average F-Measure scores of the proposed CT-Net-3 and CTCT-Net-3 is shown in Fig. 8, from which we
can see that the performance of CT_{ada}-Net-3 outperforms CT-Net-3 in documents with most degradations except textual background and smear. Note that the CT_{ada}-Net-3 provides much better results than CT-Net-3 on documents with the degradation of black margins.

4.4. Comparison with winner methods submitted to DIBCO datasets

In this section, we compared the best score per metric for any submission on each year and the proposed CT-Net-3 or CT_{ada}-Net-3 methods. Table 2 shows the comparison between the proposed method and the winner among competition submissions on each DIBCO dataset. From the table we can see that our proposed methods outperform the winner methods for all of nine data sets in terms of all the evaluation metrics. The average scores of F-measure over all nine datasets are 91.33 and 94.28 for the winner and proposed CT-Net respectively. The results show that our proposed CT-Net achieved, on average, a 3.2% improvement over the winner results of DIBCO datasets.

4.5. Comparison with state-of-the-art methods

In this section, the proposed methods are qualitatively evaluated on the nine DIBCO datasets and compared with traditional binarization methods including Otsu [29], Sauvola [10], Su et al. [11], Howe [38], Lelore et al. [57], Jia et al. [13], Mitianoudis et al. [58], GiB [40] and deep learning-based methods including Hierarchical DSN [14], PDNet [8], cGANs [20]. Tables 3, 4, 5 show the performance of the proposed method and the compared algorithms. Figs. 9-11 display several visual quality of binary results, highlighting the advantages and disadvantages of the proposed methods.

4.5.1. DIBCO’09 dataset

The dataset contains ten documents with various degradations such as bleed-through, smear or nonuniform background. From Table 3 we can see that our proposed CT_{ada}-Net-3 provides the best scores of F-measure, F_p, and PSNR. The cGANs [20] method which uses the cascaded conditional generative adversarial networks to generate the binarization image gives lowest visual distortion (best DRD score). Fig. 9 shows several examples of the proposed binary maps, revealing that our proposed methods can han-
dle the various degradations such as bleed-through and feeble contrast between ink and parchment. However, Fig. 9 also shows a problematic sample involving large character sizes with fat strokes in a deviant ink color. The main reason is that there was no similar training material in the training dataset.

### 4.5.2. **H-DIBCO’10 dataset**

Document images in this dataset are from historical collections and ambiguity exists in the text boundary locations. It is hard, even for humans, to label pixels near the boundary [9]. Therefore, most errors indeed are from pixels near text boundaries and weak ink traces. The best scores on this data set, for different metrics, are...
from different methods. The highest F-measure score is given by sGANs [20] while the proposed CT-Net-3 results in the best Fps, which has a better correlation with OCR [9]. The best performance in terms of PSNR and DRD is given by CTada-Net-3 which indicates that our proposed CTada-Net-3 method gives the lowest visual distortion.

4.5.3. DIBCO’11 and H-DIBCO’12 datasets
Both of our proposed CT-Net-3 and CTada-Net-3 provide the better results than other state-of-the-art methods on all four metric scores on these two data sets. Fig. 10 shows two images which contain variation of contrast. The binarization maps of our proposed methods have a clear background but lose some strokes on the very noisy regions.

Fig. 9. Binarization results on DIBCO’09. Left column shows the original images, the middle column shows the results of CT-Net-3 and the right column shows the results of CTada-Net-3. Black pixels are correctly classified text pixels and white pixels are correctly recognized background. Text pixels classified as background are marked in red while background pixels classified as text are show in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 10. Binarization results on DIBCO’11 and H-DIBCO’12. Left column shows the original images, the middle column shows the results of CT-Net-3 and the right column shows the results of CTada-Net-3. Black pixels are correctly classified text pixels and white pixels are correctly recognized background. Text pixels classified as background are marked in red while background pixels classified as text are show in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Most document images in DIBCO’11 dataset have a strong tural background. Thus, as discussed above, the CT-Net-3 provides better results than CTada-Net-3. However, document images in the H-DIBCO’12 dataset suffer from uneven distribution and weak ink traces, which can be solved efficiently by CTada-Net-3.

4.5.4. DIBCO’13 and H-DIBCO’14 datasets
Document images in DIBCO’13 contains large smear or show-through degradations. Although our proposed methods provide the best results over other methods, some strokes inside the smear are missed and noisy patterns from smear are retained in the final bi-
nization map. In these two datasets, the CT-Net-3 provides better performance than CT$_{ade}$-Net-3 since the document images suffer from the degradations of smear or weak ink traces which are hard to be corrected in the enhancement images, thus noises are introduced through the adaptive shortcuts in CT$_{ade}$-Net.

4.5.5. H-DIBCO’16 and DIBCO’17 datasets

Our proposed method provides the second best performance in terms of $F_m$ on the H-DIBCO’16 data set, which is lower than sGANs [20]. H-DIBCO’16 is a challenging dataset for binarization, compared to other DIBCO datasets which contains handwritten documents with various degradations such as bleed-through, uneven pen strokes and background.

On the DIBCO’17 dataset, our proposed methods provide the best results than other methods. Document images in this dataset have a large image size and most of them suffer from the bleed-through degradations.

4.5.6. H-DIBCO’18 dataset

The special degradation on H-DIBCO’18 is dark margins near the page boundaries. The proposed CT$_{ade}$-Net-3 provides best results than traditional methods, such as Jia’s [13], Howe’s [38] and the deep learning-based cGANs [20]. Fig. 11 shows several examples, indicating that our proposed CT$_{ade}$-Net-3 can deal with the dark margin and uneven distribution of text pixels in document images.

4.5.7. Summary over nine DIBCO datasets

Table 6 presents the mean values of the four evaluation metrics over nine DIBCO datasets. We compared our methods with five traditional methods and two deep learning-based methods. From the table we can see that the proposed CT$_{ade}$-Net-3 provides the best scores of four metrics. The CT-Net-3 without adaption achieves second best results in terms of $F$-measure and PSNR. Fig. 12 shows the visual results of one document on the H-DIBCO’16 dataset from different methods. From the figure we can see that our proposed methods provide good results on weak ink traces.

4.6. Generalization to The MCS dataset

In this section, we evaluated the generalization of the proposed methods on the Monk Cuper Set (MCS) dataset, which was firstly introduced in our previous work [19] and images in this dataset were taken using iPhone. It contains 25 pages sampled from a historical collection with heavy bleed-through degradations and textural background. This data set is public available on the author’s website.

Table 7 shows the average results overall 25 document images from nine models trained using the DIBCO data sets. Our proposed CT-Net achieves better results than other methods, which demonstrates that the proposed methods generalize very well on other degraded historical documents.

4.7. Computing time analysis

The number of parameters of the proposed T-Net is the 1.5 times that of the standard U-Net [41] since our T-Net shares the same encoder and has different decoders for different tasks. The networks were trained using a single GPU card (16 GB memory), taking about 24 h. For testing, the time of inference of each patch is less one second and patches from different images can be processing in parallel with multiple GPUs.

---

Table 6

<table>
<thead>
<tr>
<th>Method</th>
<th>FM</th>
<th>$F_m$</th>
<th>PSNR</th>
<th>DRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st rank of contest</td>
<td>89.83</td>
<td>19.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Otsu [20]</td>
<td>82.50</td>
<td>86.17</td>
<td>16.62</td>
<td>11.35</td>
</tr>
<tr>
<td>Sauvola [10]</td>
<td>82.33</td>
<td>87.46</td>
<td>16.68</td>
<td>6.85</td>
</tr>
<tr>
<td>Jia et al. [13]</td>
<td>92.46</td>
<td>95.15</td>
<td>19.87</td>
<td>2.44</td>
</tr>
<tr>
<td>Howe [38]</td>
<td>92.53</td>
<td>93.98</td>
<td>20.06</td>
<td>2.86</td>
</tr>
<tr>
<td>Mittanouidis [58]</td>
<td>89.14</td>
<td>-</td>
<td>16.90</td>
<td>-</td>
</tr>
<tr>
<td>cGANs [20]</td>
<td>92.20</td>
<td>95.71</td>
<td>20.16</td>
<td>2.74</td>
</tr>
<tr>
<td>Tensmeyer [16]</td>
<td>93.19</td>
<td>94.92</td>
<td>20.42</td>
<td>2.24</td>
</tr>
<tr>
<td>Proposed CT-Net-3</td>
<td>93.47</td>
<td>94.47</td>
<td>20.90</td>
<td>2.57</td>
</tr>
<tr>
<td>Proposed CT$_{ade}$-Net-3</td>
<td><strong>94.03</strong></td>
<td><strong>96.17</strong></td>
<td><strong>20.99</strong></td>
<td><strong>2.00</strong></td>
</tr>
</tbody>
</table>

Note: * results are from Zhao et al. [20]. ▲ deep learning method.

---

1 https://www.ai.rug.nl/~sheng/
5. Conclusion

In this paper we have proposed a cascade T-shape network architecture for document binarization, which has two versions: C-TNet and C\textsubscript{tide}-Net. The T-Net has two outputs corresponding to a binarization map and an enhancement image, constituting a multitask learning framework. The CT-Net model is a network with three cascade T-Nets, and C\textsubscript{tide}-Net adapts the features learned for enhancement to the binarization task, which can improve the performance of binarization.

We have evaluated the proposed methods over nine widely-used competition DIBCO datasets and our proposed methods outperform all winner methods submitted to the competitions. In addition, our methods are superior to traditional and deep learning methods for binarization reported in the literature. An external historical document dataset is used to evaluate the generalization and the results show that our proposed methods generalize very well on these document images.

Future works include exploiting more complex structure of network and training different networks on documents with different degradations. Image enhancement, including traditional image operations and the trainable generative adversarial network, appears to be very important for boosting the performance of document binarization. It would be interesting to evaluate the effect of image enhancement on text recognition, in comparison to plain document binarization.

Table 7

Comparison of the mean values for different methods on the MDS test set. The best performance is highlighted in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>FM</th>
<th>FN</th>
<th>PSNR</th>
<th>DRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Otsu [29]</td>
<td>69.3</td>
<td>70.5</td>
<td>11.8</td>
<td>34.0</td>
</tr>
<tr>
<td>Sauvola [10]</td>
<td>75.8</td>
<td>76.9</td>
<td>13.1</td>
<td>21.5</td>
</tr>
<tr>
<td>Su et al. [11]</td>
<td>82.8</td>
<td>87.4</td>
<td>15.2</td>
<td>16.8</td>
</tr>
<tr>
<td>Jia et al. [13]</td>
<td>85.4</td>
<td>88.7</td>
<td>15.8</td>
<td>7.1</td>
</tr>
<tr>
<td>Howe [38]</td>
<td>85.6</td>
<td>89.1</td>
<td>15.8</td>
<td>6.4</td>
</tr>
<tr>
<td>Deep-Sauvola [19]</td>
<td>87.0</td>
<td>89.9</td>
<td>16.2</td>
<td>6.1</td>
</tr>
<tr>
<td>C-TNet-1</td>
<td>88.3</td>
<td>92.2</td>
<td>16.7</td>
<td>4.8</td>
</tr>
<tr>
<td>C-TNet-2</td>
<td>88.7</td>
<td>92.8</td>
<td>16.9</td>
<td>4.4</td>
</tr>
<tr>
<td>C-TNet-3</td>
<td>88.7</td>
<td>92.7</td>
<td>16.9</td>
<td>4.4</td>
</tr>
<tr>
<td>C\textsubscript{tide}-Net-1</td>
<td>88.8</td>
<td>92.9</td>
<td>16.9</td>
<td>4.4</td>
</tr>
<tr>
<td>C\textsubscript{tide}-Net-2</td>
<td>88.9</td>
<td>93.0</td>
<td>16.9</td>
<td>4.3</td>
</tr>
<tr>
<td>C\textsubscript{tide}-Net-3</td>
<td>89.1</td>
<td>93.2</td>
<td>17.0</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Note: * results are from He and Schomaker [19].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Thanks organizers of the DIBCO competitions. In addition, the authors would like to thank Zhenwei Shi to label the MDS dataset.

References


Sheng Schonaker is Professor in Artificial Intelligence at the University of Groningen and was the director of its AI institute ALICE from 2001 to 2018. His main interests are pattern recognition and machine learning problems, with applications in handwriting recognition problems. He has contributed to over 200 peer-reviewed publications in journals and books (h=21)[35], h=47(Citations). His work is cited in 23 patents. In recent years, his focus has been on continuous learning systems and bootstrapping problems, where learning starts using very few examples. Prof. Schonaker is a senior member of IEEE, a member of the IAPR, and a member of Dutch research program committees in e-Science (at the NWO), Computational science and energy (Shell/NWO/FOM). He received two IBM Faculty Awards (in 2011 and 2012) for the Monk word-retrieval system in historical manuscript collections using high-performance computing.

Lambert Schonaker is Professor in Artificial Intelligence at the University of Groningen and was the director of its AI institute ALICE from 2001 to 2018. His main interests are pattern recognition and machine learning problems, with applications in handwriting recognition problems. He has contributed to over 200 peer-reviewed publications in journals and books (h=21)[35], h=47(Citations). His work is cited in 23 patents. In recent years, his focus has been on continuous learning systems and bootstrapping problems, where learning starts using very few examples. Prof. Schonaker is a senior member of IEEE, a member of the IAPR, and a member of Dutch research program committees in e-Science (at the NWO), Computational science and energy (Shell/NWO/FOM). He received two IBM Faculty Awards (in 2011 and 2012) for the Monk word-retrieval system in historical manuscript collections using high-performance computing.