

Domain Adaptation for Document Image Binarization via Domain Classification

Carlos GARRIDO-MUNOZ^{a,1}, Adrián SÁNCHEZ-HERNÁNDEZ^a,
Francisco J. CASTELLANOS^a and Jorge CALVO-ZARAGOZA^a

^a*Dept. of Software and Computing Systems, University of Alicante, Alicante, Spain*

Abstract. Binarization represents a key role in many document image analysis workflows. The current state of the art considers the use of supervised learning, and specifically deep neural networks. However, it is very difficult for the same model to work successfully in a number of document styles, since the set of potential domains is very heterogeneous. We study a multi-source domain adaptation strategy for binarization. Within this scenario, we look into a novel hypothesis where a specialized binarization model must be selected to be used over a target domain, instead of a single model that tries to generalize across multiple domains. The problem then boils down to, given several specialized models and a new target set, deciding which model to use. We propose here a simple way to address this question by using a domain classifier, that estimates which of the source models must be considered to binarize the new target domain. Our experiments on several datasets, including different text styles and music scores, show that our initial hypothesis is quite promising, yet the way to deal with the decision of which model to use still shows great room for improvement.

Keywords. Document Image Binarization, Deep Neural Networks, Unsupervised Domain Adaptation, Domain Classifier

1. Introduction

Binarization represents a key role in many Document Image Analysis (DIA) workflows [1,2], as it helps to reduce the complexity of the image by highlighting the relevant information (i.e., the ink). This process also enables the use of specific procedures involving morphological operators, connected-component search, or histogram analysis, among others. Given its importance in the field of DIA, there is a vast amount of existing literature concerning document image binarization, including surveys and reviews [3].

A straightforward procedure for binarization is that of thresholding, in which all pixels under a certain value are set to 0, being set to 1 otherwise. For example, Otsu's algorithm automatically estimates a global threshold for a given input image [4]. In contrast to global thresholding, there also exist algorithms that compute a different threshold for each pixel depending on its local neighborhood, such as Niblack's [5], Sauvola's [6] and that proposed by Wolf et al. [7]. More complex algorithms for document image binarization have also been proposed, such as those by Gatos et al. [8], Su et al. [9] or Howe

¹Corresponding Author: Carlos Garrido-Munoz, Dept. of Software and Computing Systems, University of Alicante, Alicante, Spain; E-mail: cgm156@alu.ua.es

[10], all of which are based on different image processing workflows comprising several steps.

Supervised learning has been also considered for document image binarization. A classification-based approach consists in querying every pixel of the image, performing a feature extraction, and then using a learned model to predict its category. Within this formulation, both classical Multi-Layer Perceptron [11] and Convolutional Neural Networks [12] have been studied.

More recently, image-to-image models based on Fully Convolutional Neural Networks (FCN), which were first proposed for semantic segmentation [13], have been applied to document image binarization [14]. The FCN takes an input grayscale or color image and directly provides the probability of each pixel to be foreground or background in just one step. According to one of the latest Competition on Document Image Binarization [15]—a common benchmark for this task—the use of FCN can be considered the current state of the art, as most of the best methods are based on this idea. Therefore, we will use this kind of neural network as the backbone of our methodology.

It is convenient to emphasize, however, that it is very difficult for the same model to work successfully in a number of document styles, since the set of potential domains is very heterogeneous. Previous work demonstrated that the use of a model over a different type of manuscript to that for which it was trained noticeably decreased the performance [16].

The common idea to deal with the situation mentioned above is to train a single model with datasets of different types, with the hope that it will generalize to unknown domains successfully. However, our hypothesis is that a specialized model could better deal with the binarization of a target domain. The open question, therefore, is how to decide which of these specialized models should be used or to what extent one should trust their predictions, conditioned to the characteristics of the target set to be binarized. We study here a straightforward way to address this question by using a domain classifier, which determines the suitability of each specialized binarization model for the new target domain.

Our experiments on several datasets, including different text styles and music scores, show that our initial hypothesis is quite promising, yet the way to deal with the decision of which model to use still shows great room for improvement.

The rest of the work is structured as follows: in Section 2, we thoroughly describe our methodology; Section 3 specifies our experimental setup are given; in Section 4, we report the results attained, along with analysis and discussion; in Section 5 a qualitative evaluation is discussed; and finally, in Section 6, we present the conclusions, including some potential ideas for future work.

2. Methodology

Image binarization is a function $\mathcal{B} : [0, 255]^{(h \times w \times c)} \rightarrow \{0, 1\}^{(h \times w)}$ that converts an image with a defined height h , width w , and c channels into its binary version.

As mentioned above, the state of the art considers image-to-image FCN to approximate \mathcal{B} . Typically, these methods assume that the distribution of the data to be binarized is similar to that used for training, which is not very useful in many real cases. We here propose a strategy to deal with this situation.

Our methodology assumes an unsupervised multi-source domain adaptation scenario, where there is a collection of training sets $\{\mathcal{S}_1, \mathcal{S}_2, \dots\}$, each representing a source labeled domain with pairs of document images and their perfectly binarized versions, being $\mathcal{X}_S^i = [0, 255]^{h_s^i \times w_s^i \times c}$ the i -th image from the training sets and $\mathcal{Y}_S^i = \{0, 1\}^{h_s^i \times w_s^i}$ its corresponding ground-truth data. The goal, however, is to provide a binarization strategy that works successfully over a completely unknown target set, which is only presented at the test time.

In our strategy, a specialized binarization model is first trained for each of the source domains. Our hypothesis is that a new target set should be binarized with the best model among those of the previous step. We propose a domain classifier for this decision that knows how to distinguish between all these source sets. This classifier is later used to classify a test sample, and we then provide a binarization accordingly through a specific decision mechanism.

Below we describe thoroughly the three elements involved in our methodology: the binarization model, the domain classifier, and the decision mechanism.

2.1. Selectional Auto-Encoders

Although our strategy is independent of the underlying binarization model considered, as long as it is based on machine learning, in this work we shall implement the Selectional Auto-Encoder (SAE) model, proposed in the work of Calvo-Zaragoza and Gallego [14].

An SAE model is trained to perform a function such that $b : \mathbb{R}^{(h \times w \times c)} \rightarrow [0, 1]^{(h \times w)}$. That is, it learns a score map over a $h \times w \times c$ image that preserves the input shape. The score value (or neural activation) of each pixel depends on whether the pixel belongs to the foreground or the background.

This type of architecture is typically implemented as an FCN, and so the prediction can be done through successive convolutions and sampling operations, without any fully-connected layer. The last layer consists of a set of neurons that predict a value in the range of $[0, 1]$, depending on the selectional level of the corresponding input pixel. A graphical illustration of this configuration is shown in Figure 1.

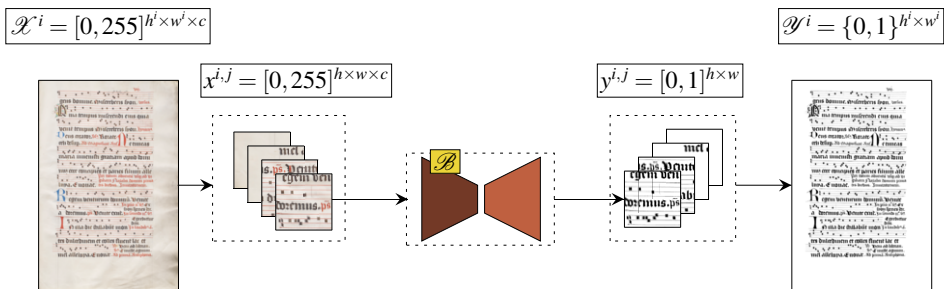


Figure 1. Schema of an SAE model in document image binarization. Note that \mathcal{X}^i is the i -th image from the source set while training, or the target set otherwise.

The weights of the SAE are learned through a training process with the aim at binarizing document images. The training stage consists of providing examples of images and their corresponding binarized ground-truth data. Since an SAE is a type of feed-forward network, the training process can be carried out by conventional means [17].

Once the SAE has been properly trained, an image can be parsed, after which a score level is assigned to each input pixel. In practice, the network barely outputs either 0 or 1 but an intermediate value. Therefore, a thresholding process is still necessary to convert the obtained neural activations into actual binary values. Since the network already takes into account the context of each pixel in its internal operation, a single global threshold is sufficient. In practice, unless otherwise stated, a threshold value of 0.5 can be assumed.

Furthermore, it might happen that the size of the input document is higher than the input layer of the SAE. In such a case, we simply split the input document into pieces of the size expected by the network and parse them independently. Then, to reconstruct the original image, we assemble the independent pieces, without further processing.

2.2. Domain Classifier

At this point, we have a way to train a specialized binarization model for each source domain (\mathcal{B}_i). However, when a new test sample must be binarized, we do not know which of them should be used. To address this issue, our strategy includes a domain classifier that is responsible for estimating which of these specialized models is more appropriate at each case.

To achieve this goal, we use a very straightforward premise: we train a classifier that predicts which of the source domains a certain sample belongs to. This is simple to carry out, as we can create ground-truth data for domain classification by taking samples of the images from each source domain. Then, the model can be trained to categorize them as samples of the domain they were taken.

Among the options available to build this classifier, we consider Convolutional Neural Networks, because of their successful performance in image processing tasks. Note that a neural network does not fully categorize the input but produces a weight $\hat{w} \in [0, 1]$ (output activation) for each possible category—domains, in this case. The weights represent the probability of the input sample belonging to each possible domain. As we describe in the next section, we shall propose alternatives for taking advantage of \hat{w} to select how the final binarization is performed.

2.3. Combination Mechanism

The only thing left to complete our strategy is to decide how to combine the binarization results provided by each of the specialized SAE models and the probabilities \hat{w} provided by the domain classifier.

From the received information, we must produce a single binarization result, which represents the actual output of our domain adaptation approach. We consider two alternatives for the combination of information:

- Class-based selection: the domain with the highest probability according to the domain classifier is selected for binarization.
- Weighted combination: the final score of each pixel is computed as a linear combination of the prediction by each specialized SAE, weighted by their \hat{w} .

The first option follows a *winner-takes-it-all* approach, whereas the second one ensembles the predictions of each specialized SAE according to their estimated confidence for the test data.

For the sake of clarity, we provide a general overview of our strategy in Figure 2, including the different elements and the data flows.

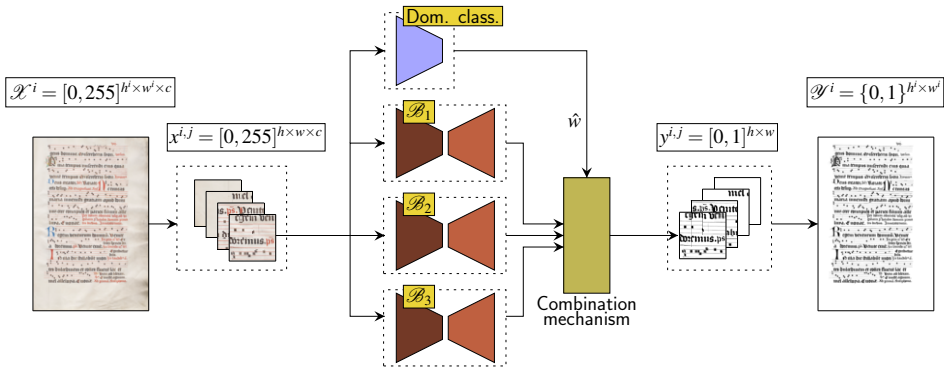


Figure 2. Framework schema proposed for multi-source domain adaptation for document image binarization.

3. Experimental setup

This section details the experimentation carried out to evaluate the proposed approach. First, the corpora and metrics considered are described, and second, the experimental setup is detailed, with the architecture of the neural networks considered and their hyperparameterization.

3.1. Corpora

In order to assess our proposal, documents of different types have been considered, including text documents and music score images. Figure 3 includes representative examples from each and Table 1 shows the details of each one.

- DIB: set of images of handwritten Latin text documents from the Document Image Binarization Contest [18], annually held from 2009. We combined the collections from 2009 to 2016 as a single corpus for our experiments.
- PHI: collection of scanned images of Persian manuscripts from the Persian Heritage Image Binarization Competition [19] as a different text domain.
- SAL and EIN: two collections of high-resolution images of scanned documents that contain lyrics and music scores in neumatic notation. Specifically, those images from Salzinnes Antiphonal (CDM-Hsmu 2149.14)² and Einsiedeln, Stiftsbibliothek, Codex 611(89)³, respectively.

²<https://cantus.simssa.ca/manuscript/133/>

³<http://www.e-codices.unifr.ch/en/sbe/0611/>



Figure 3. Examples of document patches from the corpora considered.

Table 1. Details of the corpora. The columns represent the corpus, the number of images, the average resolution of the images, and finally, the average amount of ink pixels with respect to the entire images, respectively.

Corpus	Pages	Size	Ink
<i>Text documents</i>			
DIB	86	659 × 1 560 px.	7.2%
PHI	15	1 022 × 1 158 px.	9.2%
<i>Music documents</i>			
SAL	10	5 100 × 3 200 px.	19.2%
EIN	10	5 550 × 3 650 px.	20.0%

For the experimentation, the images from the corpora have been configured with the 5-fold cross validation technique with three partitions, for training, validating and testing with 60%, 20% and 20% of the whole collections, respectively.

3.2. Metrics

The binarization issue is a two-class problem in which the pixels are classified as foreground or background. However, as seen in Table 1, given the imbalanced scenario in this context, the assessment requires to use metrics that do not bias towards the majority class—background in this case. For this, we consider the calculation of the *F-measure* (F_1), which is mathematically defined as

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}, \quad (1)$$

where TP represents the *True Positives* or correctly classified pixels, FP is *False Positives* or type I errors and FN stands for the *False Negatives* or type II errors. Note, that in our context, we assume the foreground as the positive class.

In addition to computing this metric for the binarization predictions, we also consider the multi-class version of this metric to assess the isolated performance of the domain classifier proposed.

3.3. Hyper-parameterization

Concerning the SAE architecture, it is divided into two parts: an encoder, composed of consecutive blocks of convolutional layers and down-sampling operators, and a decoder,

which includes blocks of convolutional layers and up-sampling operators, as many up-sampling operations as there are down-sampling ones in the encoder stage. For our experiments, we define an encoder block as two consecutive convolutional layers with the same number of filters in both and followed by a max-pooling operator of 2×2 px. In addition, after all these encoder blocks, two additional convolution layers are included with 128 filters each. Concerning the decoder block, it is defined as an up-sampling operator of 2×2 px. and two convolutional layers with a symmetric number of filters with respect to the encoder blocks. We particularly considered two blocks for both the encoder and decoder and a residual connection between the second blocks of each one after the convolution layers.

It should be noted that the number of filters is increased with the number of blocks in the encoder stage, and decreasing in the decoder one, in a factor of 2, with an initial number of 32 filters. We considered using a kernel shape of 3×3 px. for these layers and a Rectified Linear Unit (ReLU) activation for each convolution. Finally, another convolutional operation with 32 filters and a kernel size of 15 is performed to eventually finish in a 1×1 convolutional layer of one filter with Sigmoid activation to obtain the score result for each pixel of the input image.

In addition to the SAE model, we also need to define the domain classifier used for our proposal, which is the one responsible for determining which of the known domains—sources—most closely resembles to the new unlabeled domain. The architecture considered for this purpose contains three sequential blocks of a convolution layer, a max-pooling operator of 2×2 px. and a dropout with 0.2 of probability. The first block uses 32 filters with a kernel of 5×5 px. whereas the remaining two convolutions contain 64 filters with the same kernel. After these blocks, a flatten operation is performed followed by a fully-connected layer with 64 neurons. Finally, another fully-connected layer is used, with as many neuron as known domains, with a softmax activation. The rest of the convolutions use the ReLU activation function.

As aforementioned in Section 2, both the SAE model and the domain classifier do not process the full image at once, but they deal with patches in which the images are split. After informal experiments, we consider a size of 256×256 px. for these patches. Moreover, the models are trained for 300 epochs maximum, with a batch size of 32 and checking the models in each epoch with the validation partitions to keep the best configuration. It is worth highlight that the learning process is carried out by means of the Adam optimizer [20] in the case of the binarization model, and stochastic gradient descent [21] for the domain classifier. Note that the test partition is not used at all for the training process.

It is worth mentioning that, in our experiments, we have considered RGB images (that is, $c = 3$). However, any other color/grayscale space could be considered as long as the model is trained and used consistently.

3.4. Competing Approaches

For the sake of clarity in the analysis of the results, in this section, we here describe and denote each scenario considered in our experiments.

First, we considered two possible baseline experiments: the case in which only a single source domain is used and the one in which a single model is trained with all the source training sets. Henceforth, we denote these baseline cases as *Single-domain* and *Multi-domain*, respectively.

Another scenario to be studied is that in which our proposals are considered. As aforementioned, given a new image, a domain classifier recommends which of all the available specialized models should be used to binarize the image. The recommendation of the classifier is used in two different modes (see Section 2.3): the first one would be the case in which it decides the binarization model to be used according to the classification, and the second one, that obtains the probability of that image to belong to each source domains for then weighting the probabilities obtained for each binarization model. These two proposals are referred to in the rest of the paper as *Class-based selection* and *Weighted combination*, respectively.

Finally, since our approach is based on the decision of using a specific model or the weighted ensemble, we also include a study to analyze the upper bound that could be achieved with our hypothesis. This is artificially assessed by selecting always the best binarization model for each test patch. This experiment is referred to as *Ideal-patch*. Another ideal case to be analyzed is the case in which the domain classifier selects the best specialized model given a full domain, so that the same binarization model is applied for all the patches. We denote this experiment as *Ideal-global*. Note that these experiments are only used as a reference to measure the potential of our premise, but they do not represent real results.

4. Results

In this section, we present the final results achieved by our approach. Moreover, we compare them with the ones obtained by baseline methods, described in Section 3.4. Our results are reported in Table 2.

As can be observed, comprehensive experiments have been carried out, reporting all possible combinations of source and target sets for each method. Note that the *Ideal-global* is not directly included in the table because its results can be inferred by selecting the best model for each target in the *Single-domain* experiments.

Concerning the baseline cases, i.e. *Single-domain* and *Multi-domain*, we observe that, as expected, the models trained with multiple sources are significantly more robust. Generally, we realize that the *Single-domain* method is not adequate for unsupervised experiments except in certain cases. For example, in the case in which SAL is the source, the supervised results achieve 96% of F_1 , however, the results for the unsupervised ones are severely worsened, obtaining between 19.7% and 72.2% of performance. Note that this last figure matches with that obtained with EIN, which in turn, matches in document type with the source domain, since both contain music notation. Given that the remaining domains are collections of text manuscripts, important differences can be found in their content compared to the music documents. This might explain this high degradation in the binarization result. However, when a text manuscript—DIB or PHI—is used as the source, the degradation in the performance when dealing with music documents is not as accentuated. This may be attributed to the variability in the content of the images within the text collections, being the images of music documents more uniform. Therefore, this characteristic could be to blame for these disparate behaviors.

However, when multiple labeled domains are aggregated in the training process of a single domain (*Multi-domain*), the results are generally more robust and stable, with the exception of the two text manuscripts—DIB and PHI. When they are considered as unlabeled

Table 2. Average results in terms of F_1 (%) for different source combinations. The “Avg. on target” column represents the average results among the target testing domains, i.e. the unsupervised experiments.

Method	Testing data				Avg. on target	
	Training data	SAL	EIN	DIB		PHI
<i>Single-domain</i>						
SAL		96.0	72.2	28.9	19.7	40.3
EIN		94.9	91.1	40.9	47.2	61.0
DIB		89.6	85.3	87.3	86.2	87.0
PHI		86.7	84.9	79.0	88.9	83.5
<i>Multi-domain</i>						
EIN-DIB-PHI		94.7	91.2	87.6	84.9	94.7
SAL-DIB-PHI		95.9	90.0	88.8	85.8	90.0
SAL-EIN-PHI		96.0	91.3	58.5	83.7	58.5
SAL-EIN-DIB		95.7	91.3	88.2	79.2	79.2
<i>Class-based selection</i>						
EIN-DIB-PHI		92.9	90.3	79.8	81.7	92.9
SAL-DIB-PHI		95.3	84.7	77.1	88.4	84.7
SAL-EIN-PHI		96.2	91.0	58.3	83.7	58.3
SAL-EIN-DIB		95.7	90.8	84.7	84.4	84.4
<i>Weighted combination</i>						
EIN-DIB-PHI		93.3	90.3	81.6	89.2	93.3
SAL-DIB-PHI		95.3	86.2	78.5	89.3	86.2
SAL-EIN-PHI		96.2	91.0	62.2	84.7	62.2
SAL-EIN-DIB		95.8	90.9	86.5	85.4	85.4
<i>Ideal-patch</i>						
EIN-DIB-PHI		95.2	91.5	89.0	91.0	95.2
SAL-DIB-PHI		96.4	88.7	88.7	90.9	88.7
SAL-EIN-PHI		96.6	91.5	80.8	89.1	80.8
SAL-EIN-DIB		96.6	91.5	88.4	86.6	86.6

beled domains in unsupervised experiments, they obtain the worst results found in *Multi-domain* ($F_1 = 58.5\%$ and $F_1 = 79.2\%$, respectively). This situation can be attributed to the same reason explained in the *Single-domain*, i.e. because of the high variability of these text manuscripts.

With all this, we can conclude that the *Single-domain* is able to obtain good results for specific domains, but it is way less generalizable than *Multi-domain*.

With respect to our *Class-based selection* proposal, we observe also robust results, obtaining competitive results. Indeed, when one of the text manuscripts is unknown for the model—PHI—it achieves a F_1 of 84.4% for the unique unlabeled domain—PHI—whereas the *Multi-domain* baseline provides a result of 79.2%, obtaining, therefore, a relative improvement of 6.6%. In the rest of the unsupervised experiments, the proposal obtains slight reductions in the performance between 0.3% and 5.8% of relative degradation, so that there is room for improvement, mainly associated with the domain classifier performance, which shall be later analyzed.

Concerning our second proposal—*Weighted combination*—the results indicate that this method is more promising. All the unsupervised cases obtain better performance with respect to *Class-based selection*, with relative improvements from 0.4% and 6.7% of F_1 . In addition, compared with *Multi-domain*, both unsupervised experiments with a text manuscript as the target are significantly improved, with relative boosting of 6.3% and 7.8% for DIB and PHI, respectively. With respect to the music documents, the method also experiments certain degradation with relative figures between 1.5% and 4.2%. Therefore, with all this, the text manuscripts as the target domains may be consid-

ered as the main beneficiaries of our approach, whereas the music images do not find this boosting. Note, however, that the text manuscripts are those domains in which less performance is obtained by *Multi-domain*, so that, because of the high performance obtained in the music documents, this task becomes a major challenge to improve the baseline case.

Note that the cases in which our best proposal—*Weighted combination*—does not outperform the baselines are those where SAL and EIN (music corpora) are considered as testing data. Specifically, our proposal decreases the results for SAL from 94.7% to 93.3% and from 90% to 86.2% for EIN. These manuscripts are particularly difficult to process because of the high complexity of their content. They do not contain only one type of information, since there are staff lines, music notation, decorations and text. Precisely this variety is not provided by the text manuscripts considered in the experiments—DIB and PHI—so that the results of the individual models trained with these are not able to suitably detect such information and, therefore, the combination introduces errors. This leads us to conclude that the method will probably fail when few or none of the source domains are similar enough to the test one.

Furthermore, for the sake of the evaluation and analysis, the results that could be obtained if the domain classifier were always correct for each patch are also reported—*Ideal-patch*. We observe that, in this case, the unsupervised experiments are further improved, with relative figures for the text manuscript from 1.4% and 29.9% with respect to *Weighted combination*, representing a relevant improvement. If we compare the results with *Multi-domain*, we realize that the text domains experiment relative boosting between 9.3% and 38.1% in unsupervised learning, whereas concerning the music documents, the results reveal a relative improvement of 0.5% for SAL. Nevertheless, EIN do not exhibit the same situation, but a slight relative degradation of 1.4%.

Table 3. Average results in terms of F_1 (%) on all targets in each approach considered. The bold figure highlights the best result in a real scenario, while the underlined figure indicates the best ideal case—which is only provided as a reference.

Method	F_1
<i>Baseline cases</i>	
<i>Single-domain</i>	68.0
<i>Multi-domain</i>	80.6
<i>Our approaches</i>	
<i>Class-based selection</i>	80.1
<i>Weighted combination</i>	81.8
<i>Ideal cases</i>	
<i>Ideal-global</i>	86.4
<i>Ideal-patch</i>	87.8

Although not all cases are improved by our approaches, or even by the *Ideal-patch* experiment, we observe that, according to the average result obtained on unknown target domains, our approaches, and especially *Weighted combination*, may be beneficial to the binarization task. The results shown in Table 3 support this idea, since the baselines *Single-domain* and *Multi-domain* achieve 68% and 80.6% of average performance in unsupervised experiments, and our best model *Weighted combination* obtains 81.8%, thus overcoming the baseline cases. In addition, the potential of our approach is proved with the ideal experiments, reporting that, if the domain classifier were perfect, the average

results could be improved to 86.4% for the case in which the domain classifier selects the best specialized model for the entire domain—*Ideal-global*—and 87.8% for *Ideal-patch*, i.e. the case in which the domain classifier selects the best model for each patch. The latter is especially relevant, obtaining a relative improvement of almost 9% with respect to the best baseline, and demonstrating that our approach could perform this task better as long as the way of selecting the specialized model is more accurate.

Table 4. Average results in terms of F_1 (%) obtained by the domain classifier.

Sources	F_1
EIN-DIB-PHI	68.3
SAL-DIB-PHI	73.5
SAL-EIN-PHI	92.4
SAL-EIN-DIB	87.7

Since we saw that the domain classifier could be improved to closely match the results obtained in the *Ideal-patch* experiment, Table 4 shows a reference of the average performance provided by the domain classifier in controlled scenarios. We observe different behaviours depending on the domains involved. For example, the best case is presented when SAL, EIN and DIB are considered with a result of 92.4%. However, the worst result can be found when EIN, DIB and PHI are used, with only a performance of 68.3%. Indeed, the domain classifier reports the worst error rates when the two text manuscripts are considered, whereas the two cases in which the music domains SAL and EIN are used provide the best results. This reinforces the idea that the great variability of the text collections makes it difficult to carry out this classification task, likewise in the binarization performances, by adding an important degree of complexity in data. These values confirm that there exist great room for improvement that could be achieved with other neural architectures, more comprehensive hyper-parameter optimization or another combination mechanism.

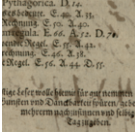
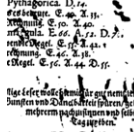
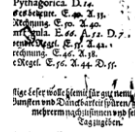
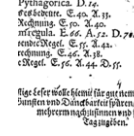
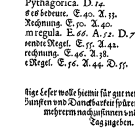
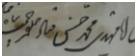
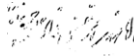
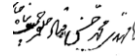

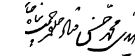
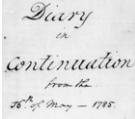
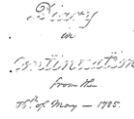
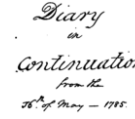
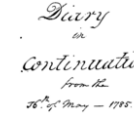
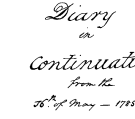
5. Qualitative Evaluation

To complement the previous experiments, we now report some representative examples of the benefits that our method can bring. Table 5 shows these selected examples, where we observe the potential of our approach compared with the *Multi-domain*, which is the most competitive baseline approach. Note that this table includes the binarization result by our best approach proposed in this work: *Weighted combination*.

Concerning the first example, the input image is a chunk from the DIB domain. After being processed by the *Multi-domain*, we observe that the method retrieves much of the foreground information, but also confuses several ink stains by labeling them as part of the foreground. These false-positive errors are slightly corrected by our method *Weighted combination*, whose result is a cleaner image. If we focus on the ideal case, we realize that the result that could be obtained by our approach is much closer to the ground-truth data, avoiding the aforementioned errors and obtaining a high-quality binarization result.

Concerning the two remaining examples, they correspond to images from PHI and DIB, respectively. Note the important difference of the input image in the third example compared with the first example, although both belong to the same domain. This explain

Table 5. Selected examples in which our approach is clearly beneficial for the binarization task. The columns, in order, represent the input images, the binarization performed by the best baseline method and our best approach, the ideal result that could be achieved and the ground truth, respectively.

	Input	Multi-domain	Weighted combination	Ideal-patch	GT
1					
2					
3					

the great variability within DIB described in Section 4. We observe that the *Multi-domain* barely detect the ink strokes, missing a lot of ink pixels by categorizing them as background. These false negatives are severely corrected but our approach, obtaining a highly reliable result. In both cases, we realize that the differences between the ideal experiment and our approach are slightly perceptible, but at least visually, all the relevant information is detected. Indeed, only some details are missing if these results are compared with the ground truth.

Although the comparison of the performance in the last section does not seem very significant, these graphic examples prove that our approach can be beneficial for binarization. This reinforces our premise that specialized models may carry out the binarization more reliably than a generic model trained with multiple domains. Thus, our approach is able to combine the advantages of the specialized models with the generalization capability of the more generic ones, since our domain classifier contributes with this characteristic to our approach.

6. Conclusions

In this work, we present an alternative to improve the generalization of binarization strategies based on machine learning. Our methodology assumes a domain adaptation scenario, for which we have several training corpora (referred to as source domains) and our goal is to improve the performance of the binarization when applied to a domain for which no data is available (target domain). Our approach consists of training different neural networks for binarization, each of which is specifically trained with the data of a single source domain. We then provide a final binarization building upon the specialized models and the output of a domain classifier, which determines the suitability of each base model for the target sample.

In our experiments, we consider four datasets of different typology and it is shown that our starting hypothesis is quite promising. The upper bound that can be achieved by selecting a specialized model clearly improves the best baseline result, which consists of training a single model with all the data from the source domains. However, our way of

selecting the best model for each sample is not very competitive, since the improvement with respect to the aforementioned baseline is not negligible.

As future work, we plan to close the gap between our results and those achieved when assuming that the specialized model is correctly selected (*Ideal*). There are different ways of pursuing this goal. For example, instead of classifying a sample as to which domain it belongs to, we could train a model that predicts which of the specialized models binarize a sample better. This may lead to better results, given that the classifier learns how to choose a binarization model, instead of choosing a domain. In addition, other ways of selecting the specialized model could be considered by means of domain similarity and the domain2vec approach [22].

Acknowledgements

This paper has been supported by Generalitat Valenciana through grant ACIF/2019/042 and project GV/2020/030, and Universidad de Alicante through project GRE19-04. The first two authors carried out this work as recipients of a grant from the Office for Educational Quality and Innovation of the University of Alicante, within the collaboration agreement with Banco de Santander S.A.

References

- [1] He S, Wiering M, Schomaker L. Junction detection in handwritten documents and its application to writer identification. *Pattern Recognition*. 2015;48(12):4036-48.
- [2] Giotis AP, Sfikas G, Gatos B, Nikou C. A survey of document image word spotting techniques. *Pattern Recognition*. 2017;68:310-332.
- [3] Sulaiman A, Omar K, Nasrudin MF. Degraded Historical Document Binarization: A Review on Issues, Challenges, Techniques, and Future Directions. *Journal of Imaging*. 2019;5(4):48.
- [4] Otsu N. A threshold selection method from gray-level histograms. *Automatica*. 1975;11(285-296):23-7.
- [5] Niblack W. An introduction to digital image processing. Strandberg Publishing Company; 1985.
- [6] Sauvola J, Pietikäinen M. Adaptive document image binarization. *Pattern Recognition*. 2000;33(2):225-36.
- [7] Wolf C, Jolion JM, Chassaing F. Text localization, enhancement and binarization in multimedia documents. In: *Proceedings of the International Conference on Pattern Recognition*. vol. 2; 2002. p. 1037-40.
- [8] Gatos B, Pratikakis I, Perantonis SJ. Adaptive degraded document image binarization. *Pattern Recognition*. 2006;39(3):317-27.
- [9] Su B, Lu S, Tan CL. Robust document image binarization technique for degraded document images. *IEEE transactions on image processing*. 2013;22(4):1408-17.
- [10] Howe NR. A laplacian energy for document binarization. In: *2011 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE; 2011. p. 6-10.
- [11] Kefali A, Sari T, Bahi H. Foreground-background separation by feed forward neural networks in old manuscripts. *Informatica*. 2014;38(4).
- [12] Calvo-Zaragoza J, Vigiensoni G, Fujinaga I. Pixel-wise binarization of musical documents with convolutional neural networks. In: *Fifteenth IAPR International Conference on Machine Vision Applications, MVA 2017, Nagoya, Japan, May 8-12, 2017*; 2017. p. 362-5.
- [13] Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017;39(4):640-51.
- [14] Calvo-Zaragoza J, Gallego A. A selectional auto-encoder approach for document image binarization. *Pattern Recognition*. 2019;86:37-47.
- [15] Pratikakis I, Zagoris K, Karagiannis X, Tsochatzidis LT, Mondal T, Marthot-Santaniello I. ICDAR 2019 Competition on Document Image Binarization (DIBCO 2019). In: *2019 International Conference on*

- Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019; 2019. p. 1547-56.
- [16] Castellanos FJ, Gallego AJ, Calvo-Zaragoza J. Unsupervised neural domain adaptation for document image binarization. *Pattern Recognition*. 2021;119:108099.
- [17] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*; 2010. p. 249-56.
- [18] Gatos B, Ntirogiannis K, Pratikakis I. ICDAR 2009 document image binarization contest (DIBCO 2009). In: *2009 10th International Conference on Document Analysis and Recognition*. IEEE; 2009. p. 1375-82.
- [19] Ayatollahi SM, Ziaei Nafchi H. Persian heritage image binarization competition (PHIBC 2012). In: *2013 First Iranian Conference on Pattern Recognition and Image Analysis (PRIA)*; 2013. p. 1-4.
- [20] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: *3rd International Conference on Learning Representations*. San Diego, USA; 2015. .
- [21] Bottou L. Large-scale machine learning with stochastic gradient descent. In: *Proceedings of the International Conference on Computational Statistics*. Springer; 2010. p. 177-86.
- [22] Peng X, Li Y, Saenko K. Domain2Vec: Domain Embedding for Unsupervised Domain Adaptation. In: *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*; 2020. p. 756-74.