UNSUPERVISED DOMAIN ADAPTATION FOR DOCUMENT ANALYSIS OF MUSIC SCORE IMAGES

Francisco J. Castellanos **Antonio-Javier Gallego** **Jorge Calvo-Zaragoza**
Department of Software and Computing Systems, University of Alicante, Spain
[fcastellanos, jgallego, jcalvo]@dlsi.ua.es

ABSTRACT

Document analysis is a key step within the typical Optical Music Recognition workflow. It processes an input image to obtain its layered version by extracting the different sources of information. Recently, this task has been formulated as a supervised learning problem, specifically by means of Convolutional Neural Networks due to their high performance and generalization capability. However, the requirement of training data for each new type of document still represents an important drawback. This issue can be palliated through Domain Adaptation (DA), which is the field that aims to adapt the knowledge learned with an annotated collection of data to other domains for which labels are not available. In this work, we combine a DA strategy based on adversarial training with Selectional Auto-Encoders to define an unsupervised framework for document analysis. Our experiments show a remarkable improvement for the layers that depict particular features at each domain, whereas layers that depict common features (such as staff lines) are barely affected by the adaptation process. In the best-case scenario, our method achieves an average relative improvement of around 44%, thereby representing a promising solution to unsupervised document analysis.

1. INTRODUCTION

Optical Music Recognition (OMR) is a computational process that aims to read the music notation from scanned documents and export their content to a structured digital format [1]. The countless number of music manuscripts scattered around the world, along with their high variability due to the different engravings, writing styles, ink colors, notations, or even the period in which they were written, represents a great obstacle to tackle this task in a simple way. In addition, physical formats are inevitably associated with page degradation over time, which is one of the motivations for digitizing them.

Given the complexity of OMR, the process is typically divided into a series of sequential tasks with partial goals.

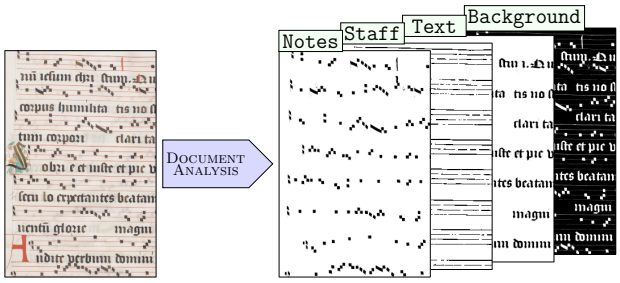


Figure 1: Overview of the document analysis process for music score images.

Document analysis is usually one of the most important tasks, where the relevant elements that make up the image content are recognized and extracted as different layers of information, e.g. by classifying each pixel into a set of categories such as staff lines, music notes, lyrics or background [2, 3], as shown in Figure 1.

Recent advances in machine learning, and particularly in Deep Neural Networks (DNNs), have opened up opportunities to carry out OMR processes effectively [4]. However, in spite of their high performance and demonstrated generalization capability in multiple tasks, this formulation brings an important drawback: the need for training data. Indeed, this is a common issue associated with machine learning, which requires labeling (often manually) a representative part of the data. However, the large number of manuscripts to be digitized contributes to making this an unaffordable task, so it is of great interest to reformulate this supervised problem to an unsupervised one.

Domain adaptation (DA) is a field that studies how to adapt the knowledge learned from a labeled collection of data—source domain—to another related, but different one—target domain—in an unsupervised manner. The idea behind this is the learning of domain-invariant features or a common representation between the source and the target domains. In this way, a model is able to process images from the target domain without using ground-truth information of that domain, thus eliminating the requirement for labeling images given a new domain. Note that, in the DA context, although the source domain labels are available, as the goal is to adapt to the target domain (for which there are no labels) this type of problem is considered unsupervised [5].

In this paper, an unsupervised approach based on Selectional Auto-Encoders (SAEs) and adversarial training

by means of a Gradient Reversal Layer (GRL) is proposed to carry out document layout analysis. The goal of this proposal is to recognize different layers of information—such as staff lines, notes, lyrics, and background—without having to manually label images of each new domain. The goodness of our approach is assessed through experiments with corpora of different music notations, reporting a substantial improvement in the performance depending on the layer at issue.

The rest of the paper is organized as follows. A review of related work is discussed in Section 2. The formulation of the problem and the description of the methodology are included in Section 3. The experimental setup and the empirical results are reported and analyzed in Section 4. A complementary qualitative evaluation is performed in Section 5. Finally, Section 6 summarizes the main conclusions, pointing out some potential future work.

2. RELATED WORK

Document analysis is a well-known stage within OMR [6], already studied in the literature with different strategies. Traditionally, this problem was addressed by dividing the task into several smaller consecutive steps. An example is binarization—used to split foreground and background information—for which we can also find different solutions, including traditional algorithms [7–9] or even specific approaches for music documents [10, 11]. There are also works in which staves and lyrics are split so that they can be processed separately, such as [12]. Another common step is the staff-line removal, where staff lines are eliminated to isolate the music symbols and make easier their classification. Dalitz et al. [13] reviews traditional methods, however, this is an active research field in which new work continually appears [14, 15].

More recently, there is a tendency to formulate document analysis as a machine learning problem. Given its performance and efficiency, the SAE architecture has been explored for related purposes, such as staff-line removal [16]. In addition, a SAE-based framework [17] was proposed to detect different layers of information by training a set of models to recognize each layer separately. However, although these approaches are usually aligned to high performance and generalizability, they entail a drawback derived from supervised learning: the need to manually label a portion of each manuscript to generate training data.

DA aims to palliate this issue by adapting the knowledge learned from a labeled (or source) manuscript to process another but related unlabeled (or target) manuscript in an unsupervised fashion. The adversarial training highlights within this field, which is an adaptation strategy in which different neural networks—or parts of them—are configured as opposing sides, with the aim of learning a common representation equally applicable to both domains. A relevant example is Domain-Adversarial Neural Network (DANN) [18], which presents a categorical neural network combined with a special type of layer named Gradient Reversal Layer (GRL). This layer aims to learn domain-invariant features to perform the DA process.

In this work, we propose to extend the SAE-based supervised framework proposed in [17], in order to combine it with the GRL so that it performs the learning of domain-invariant features in an unsupervised fashion. While this idea has been proven successful for document binarization [19], we study here its performance for the document analysis of music score images.

3. METHOD

3.1 Problem formulation

Let $\mathcal{S}$ be an annotated or *source* domain composed by a set of images with their corresponding ground truth ($\mathcal{X}_\mathrm{S}$, $\mathcal{Y}_\mathrm{S}$), where $\mathcal{X}_\mathrm{S}$ contains the scanned images of documents represented as $\mathcal{X}_\mathrm{S}^i = [0, 255]^{h_\mathrm{s}^i \times w_\mathrm{s}^i \times c}$, being $\mathcal{X}_\mathrm{S}^i$ the i-th image within $\mathcal{X}_\mathrm{S}$ with height h_s^i px., width w_s^i px., and c channels, being $c = 1$ for grayscale and $c = 3$ for colored images; and $\mathcal{Y}_\mathrm{S}$ standing for a pixel-wise annotation of each $\mathcal{X}_\mathrm{S}^i$ image for a particular layer, with $\mathcal{Y}_\mathrm{S}^i = \{0, 1\}^{h_\mathrm{s}^i \times w_\mathrm{s}^i}$, where 1 represents the foreground—or ink—of the layer at issue (e.g. staff lines, notes, text, etc.) and 0 the background.

Let $\mathcal{T}$ be a non-annotated or *target* domain that consists of a series of images $\mathcal{X}_\mathrm{T}$, being $\mathcal{X}_\mathrm{T}^k = [0, 255]^{h_\mathrm{t}^k \times w_\mathrm{t}^k \times c}$ the k-th image with no labeling data available.

3.2 Document analysis framework with SAEs

Our approach builds upon the supervised state-of-the-art document analysis framework proposed by Castellanos et al. [17], which processes the input images to classify each pixel into a set of possible categories—`staff lines`, `notes`, `text` and `background`. This method is based on a series of SAE models—namely four models, one per layer—trained to individually recognize each layer of information in a supervised fashion. Note that the number of models represents the number of layers of information of interest, so the method could easily be extended by using as SAEs as layers to detect.

This architecture consists of two parts: an encoder, in which data are processed by a series of consecutive convolution and down-sampling layers, and a decoder, composed of convolution and up-sampling layers, as many as down-sampling layers are in the encoder. The output of the SAE model is a probabilistic map of the same size as the input, but with only one channel, in which the probability of each pixel belonging to a specific layer is computed. This scheme can be successfully trained in a supervised manner when there are ground-truth data available for a portion of the collection to be processed.

As mentioned above, the framework processes each category separately with the aim of modeling specialized SAEs that detect individual layers of information to eventually be processed or combined. In order to combine the individual decisions to provide an actual document analysis result, we eventually label each pixel as that category for which its SAE retrieves the highest probability. This combination is mathematically defined as

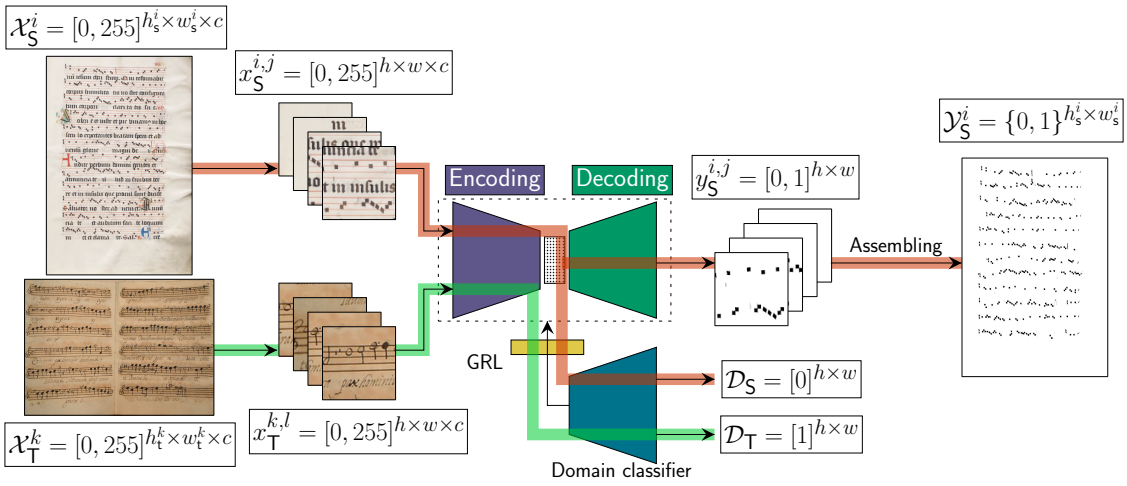


Figure 2: Scheme of our approach for the case of recognition of music notes. It would be repeated for each new layer to be considered by using its own SAE model trained with the source ground-truth data $\mathcal{Y}_{\mathrm{S}}$ for each layer.

$$\mathcal{Y}_{\mathrm{T}}^{k} = \arg\max_{\mathrm{c} \in \Sigma} P(\mathrm{c} \mid \mathcal{X}_{\mathrm{T}}^{k})$$

where Σ represents the entire set of classes or layers of interest.

It should be noted that the SAE model must be trained by patches, therefore, each input image is split into a set of chunks of $h \times w \times c$ px. that are individually processed. Therefore, once the prediction is made, it is necessary to assemble all these patches to finally build the full layered image.

3.3 Unsupervised domain adaptation approach

The method presented in the above section can successfully deal with the document analysis task when there are representative training data of the collection to be processed. However, this requires manually labeling some images of each target manuscript to provide enough training data for the learning process. Within this context, DA plays an important role to enable the application of recognition models when there are no annotations but for one source domain $\mathcal{S}$.

Our approach addresses the problem of how to adapt a document analysis model for processing an unlabeled target collection $\mathcal{T}$. Given such conditions, the model must be adapted in an unsupervised way. We propose the use of GRL [18], originally designed for classification tasks, to face this challenge.

The GRL-based approach makes use of adversarial training to penalize domain-specific features in order to train a neural network model capable of dealing with images from $\mathcal{S}$ or $\mathcal{T}$, indistinctly. This special layer is connected to a domain classifier that takes advantage of the only information available for $\mathcal{T}$: the certainty that the $\mathcal{X}_{\mathrm{T}}$ images belong to a different domain than the source. Hence, the domain classifier shall try to identify whether, given an image $\mathcal{X}_{\mathrm{S}}^{i}$ or $\mathcal{X}_{\mathrm{T}}^{k}$, it belongs to $\mathcal{S}$ or $\mathcal{T}$. This classifier, therefore, will look for domain-specific features that allow the images of both domains to be differentiated. However, as it is connected to the SAE architecture through the GRL, the gradients calculated as consequence of this classification are reversed in the training process. That is, GRL penalizes the domain-specific features found by the domain classifier, thus achieving a SAE model which focuses on domain-invariant features. Note that GRL includes a hyper-parameter λ to adjust the contribution of the domain classifier in the training process, to be empirically studied according to the task.

A graphical outline can be found in Figure 2 with an example of this method for a single layer, that of note symbols. The idea of our approach is to use independent SAEs, each one trained with the ground truth of $\mathcal{S}$ for a specific layer. Thus, our approach applies document analysis through four SAE models, one for each layer of information, and finally combines these results based on the probability of the output layer.

4. EXPERIMENTAL SETUP

4.1 Corpora

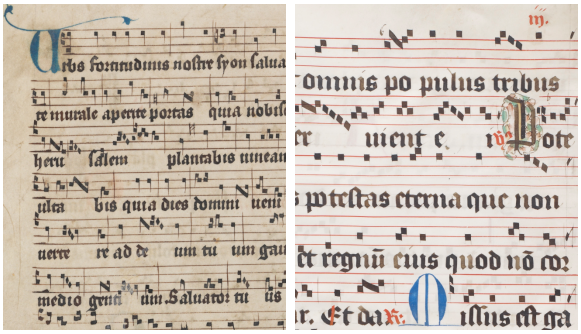
For our experiments, we selected three corpora manually labeled for the considered layers. The details of these datasets are listed below (some examples of images can also be found in Figure 3):

- EINSIEDELN: collection of 10 music documents in Neumatic notation, specifically those of Einsiedeln, Stiftsbibliothek, Codex 611(89) [1] with an average size of 6 496 × 4 872 px.
- SALZINNES: set of 10 music score images in Neumatic notation with an average resolution of 5 847 × 3 818 px., of Salzinnes Antiphonal (CDM-Hsmu2149.14) [2] .
- CAPITAN: 10 images from a complete *Missa* of the second half of the 17th century [20] in Mensural notation with an average size of 2 126 × 3 065 px.

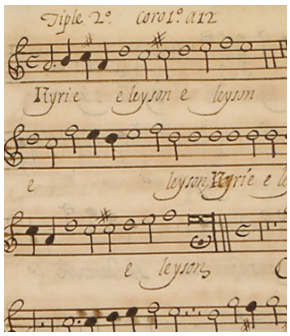
[1] http://www.e-codices.unifr.ch/en/sbe/0611/
[2] https://cantus.simssa.ca/manuscript/133/

Table 1: Description of the SAE architecture considered, implemented as a Fully-Convolutional Network (FCN). Notation: Conv(f,h_c,w_c,a) represents a convolution operator of f filters, with kernels of $h_c \times w_c$ pixels and an a activation function; MaxPool(h_p,w_p) indicates a max-pooling operator with a $h_p \times w_p$ kernel; UpSamp(h_u,w_u) stands for an up-sampling operator of $h_u \times w_u$ px.; *ReLU* and *Sigmoid* denote Rectifier Linear Unit and Sigmoid activations, respectively.

Input	Encoding	Decoding	Output
$[0, 255]^{256\times256}$	Conv(64,3,3,ReLU) MaxPool(2,2)	Conv(64,3,3,ReLU) UpSamp(2,2)	$[0, 1]^{256\times256}$
	Conv(64,3,3,ReLU) MaxPool(2,2)	Conv(64,3,3,ReLU) UpSamp(2,2)	
	Conv(64,3,3,ReLU) MaxPool(2,2)	Conv(64,3,3,ReLU) UpSamp(2,2)	
		Conv(1,3,3,Sigmoid)	



(a) EINSIEDELN (b) SALZINNES



(c) CAPITAN

Figure 3: Examples of some representative regions of the images from the corpora.

4.2 Metrics

Given the imbalance nature in the distribution of the classes, F-score (F_1) was considered for the evaluation of the method. In a two-class problem, it is defined as:

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

where TP, FP, and FN stand for *True Positives* or correctly classified elements, *False Positives* or type I errors, and *False Negatives* or type II errors, respectively. However, since the experiments will be conducted as a multiple-class problem, we considered reporting the results in terms of F_1 for each class and macro-F_1 [21] for a global evaluation, which is calculated as the average of the F_1 obtained for each class.

Considering that, in our context, the ground-truth data is often subjective—especially for edge pixels—and that there are also multiple thin elements—such as staff lines—we decided to use more suitable metrics for this task, such as those explained in [2]. Thus, we report the results in terms of this pseudo-F_1, henceforth ps-F_1. This metric considers as TP those pixels whose real class matches with the prediction in any vertically and horizontally adjacent pixel.

4.3 Hyper-parameterization

Since this paper addresses an unsupervised formulation for document analysis through a DA scheme, we considered the use of SAE as the basis of the model. Table 1 indicates a detailed description of the neural network, which shall be repeated for each layer of interest like in the state-of-the-art document analysis framework [17]. Note that, for simplification reasons, the input image is given in grayscale, although another color space might be used.

As described, an SAE is trained with patches extracted from the input images. These patches are randomly selected after each training epoch, for the sake of data variability. We considered patches of 256×256 px. Concerning the GRL, we connect it before the last convolutional block of the decoder, with $\lambda = 0.01$ and increments of 0.001 per epoch. These decisions were taken by informal testing. The convolutional weights are optimized by using the well-known stochastic gradient descent [22] with a batch size of 12. We carried out a pre-training step with only $\mathcal{S}$ for 50 epochs, before the GRL and target images become involved, up to a total of 300 epochs, taking 10 000 samples per epoch from each domain. Note that our approach extracts the same number of samples for each domain to properly balance them.

It is worth mentioning that data are divided into partitions for training, validating, and testing, with 60%, 20%, and 20% of the entire collections, respectively. The validation partition is used to choose the best model in $\mathcal{S}$, assuming the premise that learning domain-invariant features would allow to similarly process source and target images. Although this partitioning is only necessary for the source domain, we applied it in all cases for consistent evaluation.

Table 2: Average results, in terms of ps-F_1 (%), for the SAE-based framework—state of the art—and our document analysis approach based on GRL. The results are organized according to the music notation of $\mathcal{S}$ and $\mathcal{T}$. The best figures between both models are highlighted in bold.

$\mathcal{S} \rightarrow \mathcal{T}$	Framework	
	SAE-based	SAE-DANN-based
Neumatic → Mensural		
EINSIEDELN → CAPITAN	48.7	**60.0**
SALZINNES → CAPITAN	31.5	**55.6**
Avg.	40.1	**57.8**
Mensural → Neumatic		
CAPITAN → EINSIEDELN	44.7	**45.1**
CAPITAN → SALZINNES	**55.3**	53.5
Avg.	**50**	49.3
Avg.	45.1	**53.6**

4.4 Results

In this section, we assess our GRL-based approach and compare it with the state of the art [17] in unsupervised domain-adaptation scenarios. Note that our corpora contain different music notational systems—Neumatic and Mensural. Thus, we consider of great interest the applicability of our method to adapt images across different notational systems, depicting obvious differences at the graphic level and making document analysis very challenging.

Table 2 shows the average results for each pair of $\mathcal{S}$ and $\mathcal{T}$ considered for experimentation, in such a way that source and target manuscripts do not match in the type of music notation. Focusing on the first section of the experiments, when $\mathcal{S}$≡*Neumatic* and $\mathcal{T}$≡*Mensural*, we observe a clear improvement of the DA approach with respect to the state of the art. Our approach increases the ps-F_1 from 40.1% to 57.8% on average, which represents a substantial relative improvement of 44.1%.

Concerning the second section of results, those in which $\mathcal{S}$ contains pages in *Mensural* notation and $\mathcal{T}$ consists of *Neumatic* documents, we realize that two different situations are presented: the CAPITAN → EINSIEDELN case, with a slight improvement from 44.7% to 45.1%, and the CAPITAN → SALZINNES, in which the DA technique is not able to learn adequate features for $\mathcal{T}$, thus reducing until 2% approximately. Despite this drawback, we may consider it as marginal, representing changes barely perceptible in the layered resulting image with respect to the state-of-the-art method.

This phenomenon may be attributed to the fact that the filters of the neural network are not able to extract domain-invariant features by using only the ground truth of the unique labeled domain in these experiments—CAPITAN. For example, it should be noted that the complexity and variability of the types of music symbols in *Mensural* notation are considerably greater than those in *Neumatic* one, which presents very uniform symbols and very different to those in CAPITAN. Besides, the text within CAPITAN shows a certain degree of degradation and different contrast levels with respect to the rest of the ink, even within the same page. These aspects, as we shall show in Section 5, may hinder the learning of common features for both domains, since there are layers of information with strong differences between themselves.

Table 3: Average results for each layer of information. The figures are reported in terms of ps-F_1 (%). Note that the "Bg." column represents the background layer. The best results per layer are remarked in bold.

Framework	Staff	Note	Text	Bg.	Avg.
State of the art					
SAE-based [17]	80.6	39.0	8.9	51.8	45.1
Our approach					
SAE-DANN-based	**82.3**	**42.4**	**23.5**	**66.0**	**53.6**

In order to complement the analysis, Table 3 shows the average results for each layer of information obtained with the SAE-based framework and with our method. We can observe that the DA approach obtains average improvements for all the layers considered. It is worth mentioning that the staff lines in both music notations are very similar visually. Mainly, this is why the SAE specialized in recognition of staff lines from images of $\mathcal{S}$ may be able to deal with those staff lines of $\mathcal{T}$. Indeed, focusing on this layer, we realize that the SAE model obtains a high performance of 80.6%. Note that these results are obtained by unsupervised experiments, and also note that the ps-F_1 obtained for the staff layer precisely outperforms all the rest of the layers considered. This means that the SAE model without DA mechanisms is enough for extracting features from $\mathcal{S}$ capable of detecting staves from $\mathcal{T}$. In spite of this, we can observe a slight improvement, achieving 82.3% of ps-F_1.

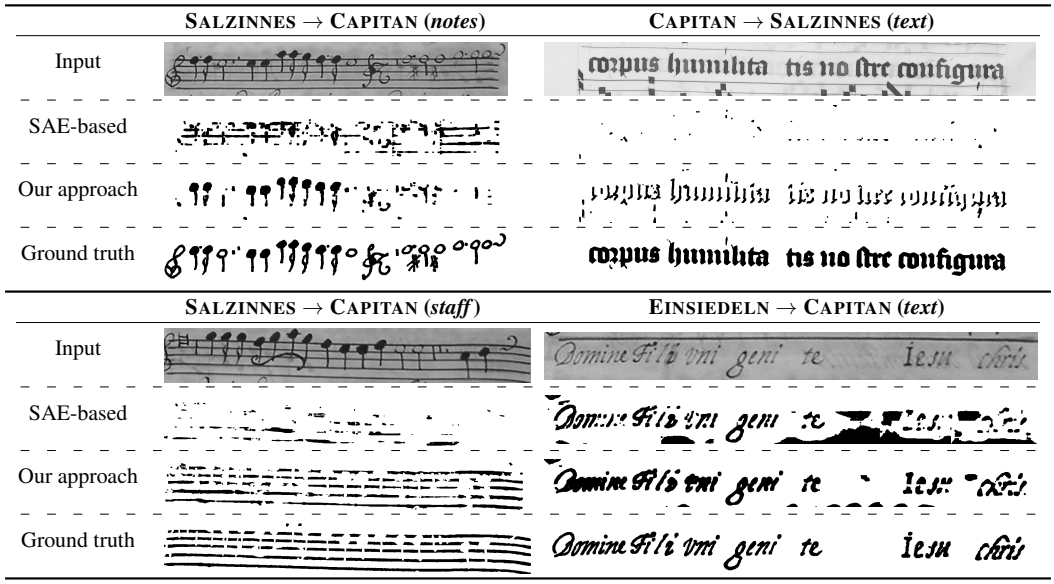
Concerning the rest of the layers, the increase of performance of the text and background layers is especially relevant. Although in the case of text, the performance may seem low with only 23.5% of ps-F_1, note the radical enhancement with a relative figure over 164%. Besides, the background layer also has a relative boost of 27%, supposing significant contributions for the unsupervised document analysis task. The global average also obtains an important increase in the results, with a relative enhancement of 18.8% with respect to the state of the art, thus supporting the idea of our proposal.

5. QUALITATIVE EVALUATION

To finalize the analysis of the results, we now shall discuss a qualitative evaluation for different scenarios. Table 4 gathers some selected examples.

As regards the first case, in which the document analysis carries out the extraction of notes in the SALZINNES → CAPITAN scenario, i.e. SALZINNES as source and CAPITAN as target, we observe that the SAE-based framework does not detect most of the elements. Indeed, it confuses parts of staff lines with notes, making the result not even close to what was expected. Oppositely, our method does differentiate the staff from the notes, obtaining a much more reliable result. Note that the detection still fails in many cases, particularly in those in which the sym-

Table 4: Selected examples for notes, text, and staff recognition cases. The table compares the extraction of the layer for the SAE-based framework and our approach. Input images and their corresponding ground-truth data are also provided.



bols are hollow. We attribute this issue to the fact that the source images do not contain hollow symbols since they contain *Neumatic* notation, so that the ground truth provided for the training does not include elements with common features to those involved in the hollow symbols presented in the *Mensural* notation. However, in spite of the clear differences between the symbols of both domains, the filled symbols are quite well recognized since SALZINNES also contains ink-filled symbols, but squared, obtaining, in general, a better recognition to that provided by the state of the art. These graphic differences can be seen in Figure 3.

Concerning the second case, we analyze an example of the detection of text when *Mensural* notation is used as source—CAPITAN—and the target manuscript is written in *Neumatic* notation—SALZINNES. As seen, the state of the art does not detect almost any pixel associated with this layer. Although the detection performed by our approach is not perfect, it does obtain a result that is much closer to the expected one. This could be used to detect the regions in which the texts are located to be individually processed by other mechanisms, for instance, Optical Character Recognition techniques.

Another example is the case in which staff lines must be identified. In this example—SALZINNES → CAPITAN—we observe that the staff-line retrieval is not properly performed by the SAE-based framework. Several parts of the staff are detected, but it would be quite difficult to reconstruct the lines due to the poor quality of the prediction. Conversely, our method can solve this issue by clearly improving the staff-line detection and obtaining a result closer to the ground truth. As shown in Table 3, the state-of-the-art framework achieves competitive average figures in terms of ps-F_1, however, note that this is a selected example in which the state of the art does not provide good results to visually analyze the capabilities of our method.

The last example in Table 4 provides another case of text recognition. In this scenario, the notation types of the $\mathcal{S}$ and $\mathcal{T}$ manuscripts are reversed with respect to the second example. Specifically, we use a *Neumatic* manuscript—EINSIEDELN—as source and the *Mensural* one—CAPITAN—as target. We observe that the state of the art recovers the text with various errors, losing part of the text and also confusing the background with text. The DA method improves this result by providing a much more accurate layered version. Note that, despite the great differences between the text of both domains, the method is able to adapt by searching for domain-invariant features in order to recognize these elements.

6. CONCLUSIONS

This work presents an unsupervised DA framework for document analysis of music score images, which builds upon existing approaches and the so-called Gradient Reversal Layer. The idea is based on learning domain-invariant features that allow transferring knowledge from a labeled source domain to an unlabeled target one.

Our experiments reveal that the approach is generally beneficial for the task at issue. The actual improvement depends on the layer of information considered. Substantial benefits are reported for the layers that depict particular features at each domain (such as text), whereas layers that depict common features (such as staff lines) are barely affected by the adaptation process. In addition, we observed that the source domain is indeed relevant to make the DA method successful. In the best case, our approach substantially increases the average performance up to 44% of relative improvement.

In light of these results, future work will focus on evaluating the method with more types of historical manuscripts, as well as studying the applicability of other DA techniques for document analysis, such as those based on Generative Adversarial Networks.

7. ACKNOWLEDGMENT

This work was supported by the University of Alicante project GRE19-04. The first author also acknowledges support from the "Programa I+D+i de la Generalitat Valenciana" through grant ACIF/2019/042.

8. REFERENCES

[1] D. Bainbridge and T. Bell, "The challenge of optical music recognition," *Computers and the Humanities*, vol. 35, no. 2, pp. 95–121, 2001.

[2] J. Calvo-Zaragoza, F. J. Castellanos, G. Vigliensoni, and I. Fujinaga, "Deep neural networks for document processing of music score images," *Applied Sciences*, vol. 8, no. 5, p. 654, 2018.

[3] I. Fujinaga and G. Vigliensoni, "The art of teaching computers: The SIMSSA optical music recognition workflow system," in *27th European Signal Processing Conference, EUSIPCO, A Coruña, Spain, September 2-6*. IEEE, 2019, pp. 1–5.

[4] J. Calvo-Zaragoza, J. H. Jr., and A. Pacha, "Understanding optical music recognition," *ACM Comput. Surv.*, vol. 53, no. 4, Jul. 2020.

[5] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[6] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. S. Marçal, C. Guedes, and J. S. Cardoso, "Optical music recognition: State-of-the-art and open issues," *International Journal of Multimedia Information Retrieval*, vol. 1, no. 3, pp. 173–190, 2012.

[7] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, 2000.

[8] B. Gatos, I. Pratikakis, and S. J. Perantonis, "Adaptive degraded document image binarization," *Pattern Recognition*, vol. 39, no. 3, pp. 317–327, 2006.

[9] N. R. Howe, "Document binarization with automatic parameter tuning," *International Journal on Document Analysis and Recognition*, vol. 16, no. 3, pp. 247–258, 2013.

[10] T. Pinto, A. Rebelo, G. A. Giraldi, and J. S. Cardoso, "Music score binarization based on domain knowledge," in *5th Iberian Conference on Pattern Recognition and Image Analysis, Las Palmas de Gran Canaria, Spain*, 2011, pp. 700–708.

[11] Q. N. Vo, S. H. Kim, H. J. Yang, and G. Lee, "An MRF model for binarization of music scores with complex background," *Pattern Recognition Letters*, vol. 69, no. Supplement C, pp. 88–95, 2016.

[12] J. A. Burgoyne and I. Fujinaga, "Lyric extraction and recognition on digital images of early music sources," in *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR, Kobe, Japan, October 26-30*, 2009, pp. 723–728.

[13] C. Dalitz, M. Droettboom, B. Pranzas, and I. Fujinaga, "A comparative study of staff removal algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 753–766, 2008.

[14] J. Dos Santos Cardoso, A. Capela, A. Rebelo, C. Guedes, and J. Pinto da Costa, "Staff detection with stable paths," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 1134–1139, 2009.

[15] T. Géraud, "A morphological method for music score staff removal," in *International Conference on Image Processing*, 2014, pp. 2599–2603.

[16] A. Konwer, A. K. Bhunia, A. Bhowmick, A. K. Bhunia, P. Banerjee, P. P. Roy, and U. Pal, "Staff line removal using generative adversarial networks," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 1103–1108.

[17] F. J. Castellanos, J. Calvo-Zaragoza, G. Vigliensoni, and I. Fujinaga, "Document analysis of music score images with selectional auto-encoders," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR, Paris, France, September 23-27*, 2018, pp. 256–263.

[18] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[19] F. J. Castellanos, A.-J. Gallego, and J. Calvo-Zaragoza, "Unsupervised neural domain adaptation for document image binarization," *Pattern Recognition*, p. 108099, 2021.

[20] J. Calvo-Zaragoza, A. H. Toselli, and E. Vidal, "Handwritten music recognition for mensural notation: Formulation, data and baseline results," in *14th IAPR International Conference on Document Analysis and Recognition, ICDAR, Kyoto, Japan, November 9-15*, 2017, pp. 1081–1086.

[21] A. Özgür, L. Özgür, and T. Güngör, "Text categorization with class-based and corpus-based keyword selection," in *International Symposium on Computer and Information Sciences*, 2005, pp. 606–615.

[22] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of the 19th International Conference on Computational Statistics, COMPSTAT, Paris, France, August 22-27*. Springer, 2010, pp. 177–186.