

Transfer Learning for Word Spotting in Historical Arabic Documents Based Triplet-CNN

Abir Fathallah^{1,2} ^a, Mounim A. El-Yacoubi² and Najoua Essoukri Ben Amara³

¹Université de Sousse, Institut Supérieur de l'Informatique et des Techniques de Communication, LATIS-Laboratory of Advanced Technology and Intelligent Systems, 4023, Sousse, Tunisia

²Samovar, CNRS, Télécom SudParis, Institut Polytechnique de Paris, 9 rue Charles Fourier, 91011 Evry Cedex, France

³Université de Sousse, Ecole Nationale d'Ingénieurs de Sousse,

LATIS-Laboratory of Advanced Technology and Intelligent Systems, 4023, Sousse, Tunisia

Keywords: Historical Arabic Documents, Word Spotting, Transfer Learning, Learning Representation.

Abstract: With the increasing number of digitized historical documents, information processing has become a fundamental task to exploit the information contained in these documents. Thus, it is very significant to develop efficient tools in order to analyze and recognize them. One of these means is word spotting which has lately emerged as an active research area of historical document analysis. Various techniques have been suggested successfully to enhance the performance of word spotting systems. In this paper, an enhanced word spotting approach for historical Arabic documents is proposed. It involves improving learning feature representations that characterize word images. The proposed approach is mainly based on transfer learning. More precisely, it consists in building an embedding space for word image representations from an online training triplet-CNN, while performing transfer learning by leveraging the varied knowledge acquired from two different domains. The first domain is Hebrew handwritten documents, the second is English historical documents. We will investigate the impact of each domain in improving the representation of Arabic word images. As a final step, in order to evolve the word spotting system, the query word image along with all the reference word images will be projected into the embedding space where they will be matched according to their embedding vectors. We evaluate our method on the historical Arabic VML-HD dataset and show that our method outperforms significantly the state-of-the-art methods.

1 INTRODUCTION


It is not always convenient to access the precious cultural heritage of historical Arabic documents (HADs). Protection and preservation of these valuable resources require scans of innumerable pages which are stored on different servers. Historical digital documents are not easily processed in their native form but must be transformed into a readable form in order to be effectively and automatically interpreted by computer vision.

One important emerging field in HADs is word retrieval, which has attracted the interest of many researchers over the past few decades. With the complexity of the Arabic script, it is a challenging task to locate the occurrences of a particular word in a large set of historical document images. Thus, there is a need for an effective approach to locate the query

word in such a document. There are two different main approaches for retrieving a query word in a set of documents depending on the query representation. One is the Query-by-Example (QbE) method in which the query word is given as an image. The second method is the Query-by-String (QbS) method, in which the query word is indicated by a text string. As we already know, it is not always easy to get access to transcriptions of all historical documents, so we are interested in the QbE. However, extracting features from historical images is a challenging task.

The objective of deep embedding approaches is to construct an embedding space that ensures the transformation of the input image into new representations by selecting the pertinent features.

This problem has been extensively explored on historical document datasets using deep learning networks (Barakat et al., 2018; Fathallah et al., 2019). Unlike the existing works for improving the perfor-

^a  <https://orcid.org/0000-0003-0433-1029>

mance of HAD applications, such as architecture enhancement or data pre-processing, Transfer Learning is an appropriate option for adopting knowledge from pertinent source data to enhance specific target tasks (Cui et al., 2018; Ye and Shen, 2020; Rajeswar et al., 2022). More specifically, insights (characteristics, weights, etc.) can be leveraged from already trained models for forming new models and even solving issues such as providing less data for the new task. As such, HADs are challenging, especially in the context of deep neural networks where most models that address complex problems require large amounts of data, given the time and effort required to label these data. This drives Transfer Learning, looking beyond specific tasks and domains to explore how to draw on the knowledge of pre-trained models by applying them to solve new issues.

In the same context, in this paper, we investigate Transfer Learning as a possible way to exploit learned features from previous document datasets to train new model for word spotting in HADs in order to improve embedding features representation of Arabic word images. Our contribution consists in exploiting the knowledge acquired from two datasets with different aspects: historical documents written in English and handwritten documents written in Hebrew.

Precisely, we first train a base network on a source dataset (source domain), and then we repurposed the learned features to a second target network to be trained on a target dataset (target domain). In our case, the source domain is two different datasets: historical document dataset written in English and Hebrew handwritten dataset. The target domain is HADs. Transfer Learning is introduced as an optimization tool that allows rapid progress and enhanced performance when training the target task. In our proposed approach, we used Triplet-CNN (Fathallah et al., 2019) to build the embedding space.

The remainder of the paper is structured as follows: in section 2, an overview of existing word retrieval techniques designed for historical documents and a brief summary of existing types of Transfer Learning approaches are provided. Then, in section 3, we introduce a Transfer Learning-based enhancement strategy for improving the performance of the Triplet-CNN model (Fathallah et al., 2019). The experiments and error analysis of the obtained results are discussed in section 4. Finally, the section 5 reports the conclusions and some potential perspectives.

2 RELATED WORK

In this section, we will give an overview of word spotting in historical documents as well as Transfer Learning approaches.

2.1 Word Spotting in Historical Documents

Word spotting in historical document images can be utilized to exploit documents content in digital form (Fathallah et al., 2020). It delineates different query word occurrences in such document sets. Given adequate and well-annotated data on historical documents, Convolutional Neural Networks (CNNs) have achieved revolutionary progress in word retrieval tasks (Barakat et al., 2018; Mhiri et al., 2019; Fathallah et al., 2019; Mohammed et al., 2022). As for historical documents, word spotting techniques already existing in the literature are divided into two main categories: segmentation-based approaches that aimed to split the input document into words and sub-words (Fathallah et al., 2019; Zagoris et al., 2014; Barakat et al., 2018) and segmentation-free approaches which involve the selection of patches from the input document through a sliding window technique or template matching (Konidaris et al., 2016).

Several researchers are interested in improving the word spotting process in historical documents. For word spotting handwritten documents, authors in (Khayyat and Suen, 2018) proposed an enhanced internal structure hierarchical classifier. They combined support vector machines and regularized discriminant analysis classifiers to increase the performance of closed lexicon word spotting systems. They achieved an enhancement in precision rate of 4%. Other methods, which were employed to enhance the performance of the word retrieval approaches introduced in (Westphal et al., 2020; Sudholt and Fink, 2018; Gurjar et al., 2018).

On the other hand, the authors of (Westphal et al., 2020) illustrated effective sample selection approaches. Within the word spotting training step, by using the character pyramid histogram representation, they decreased the amount of training data required. Alternatively, other approaches (Can and Kabadayi, 2020; Khayyat and Elrefaei, 2020) for processing historical documents are based on transfer learning which seeks to build on the information gained from the source task to the target task.

In the same context, we propose in this paper an enhanced Triplet-CNN word spotting approach presented in (Fathallah et al., 2019). In (Fathallah et al., 2019), the offline training method is applied. It in-

volves the process of selecting all the triplets appearing in the training set and then proceeding to train with the generated triplets. Differently from the previous works in literature, we employ the online learning method which consists in dynamically selecting the triplets to mitigate the poor computational efficiency and parameter non-convergence issues. Moreover, we Transfer Learning in our approach is employed in the word spotting process.

3 PROPOSED METHOD

In this section, we investigate the problem of word spotting enhancement using Transfer Learning technique where the goal is to exploit and leverage the knowledge of feature representations from the source domain to the target domain.

3.1 Enhancement Based Transfer Learning Approach

In order to exploit the knowledge of other languages and to benefit from the progress of research on Latin scripts rather than Arabic, two different languages, Hebrew and English, have been chosen to be evaluated. Then, we intend to study the impact of each language on the improvement of Arabic feature representation. Figure.1 shows the flowchart of our proposed approach. It consists of two propositions: First, Transfer Learning based on historical English documents George Washington (GW) and Transfer Learning focused on Hebrew Handwritten Documents (HDD).

More specifically, as in the first scenario, the source domain contains word images written in English and the target domain contains word images written in Arabic. Thus, from historical documents presented in the source domain, the idea involves training the Triplet-CNN to extract features from Latin script. Each word image is represented efficiently by a feature vector. Then, at the training phase on the Arabic dataset, the extracted features from the English document are transferred in order to enhance the embedding features of Arabic word images used in the target documents.

In this context, we introduce, in the second scenario, a Transfer Learning from Hebrew handwritten documents to target documents. The leveraged features contributed in extracting a better embedding features representation for word images.

3.2 Training Deep Network on GW and HDD Datasets

The CNN architecture is used to extract pertinent features from word images that will be considered as their new representations in the constructed embedding space. For all employed datasets, the same training strategy is used. As previously mentioned, a CNN architecture is used to extract features from word images. We used the same Triplet-CNN introduced in (Fathallah et al., 2019). The detailed architecture of Triplet-CNN is shown in Figure.2. It is composed of three CNN instances with shared parameters that take as input three-word images (an anchor, a positive sample and a negative sample) having a size of 60×110 pixels. The CNN architecture is composed of five convolutional layers followed by a rectified linear units activation layer. It also consists of four pooling layers. In addition, the CNN ends with a fully connected layer of 1024 neurons that represents the feature embedding vector.

In this work, we apply the same Triplet-CNN architecture and its parameters to all used datasets as shown in Figure.2. To train each Triplet-CNN architecture, a triplet dataset must be provided. A triplet is given by three instances: an anchor (a), a positive sample (p) from the same anchor class and a negative sample (n) from different classes. We note the function of the feature embedding network by G , which projects an input image x from raw pixel space into embedding space \mathbb{R}^L . From the training set a combination of image triplet $\langle a, p, n \rangle$ is produced. We denote the set of all possible combinations of obtained triplets by $T = (a^{(i)}, p^{(i)}, n^{(i)})$. In order to minimize the distance between the anchor and a positive sample and maximize the distance between the anchor and a negative sample, the triplet loss function $L(T)$ proposed in (Schroff et al., 2015) and defined in equation Eq.(1) must be minimized:

$$L(T) = \sum_{(a,p,n) \in T} \max(0, \|G(a) - G(p)\|_2^2 - \|G(a) - G(n)\|_2^2 + \alpha) \quad (1)$$

where α represents the margin enforced between positive and negative pairs.

3.3 Applying Transfer Learning on Arabic Documents

The main idea is intended to train the Triplet network for extracting features from Arabic word images by transferring knowledge acquired from previous datasets. We train the same Triplet-CNN (Figure.2) on the VML-HD dataset.

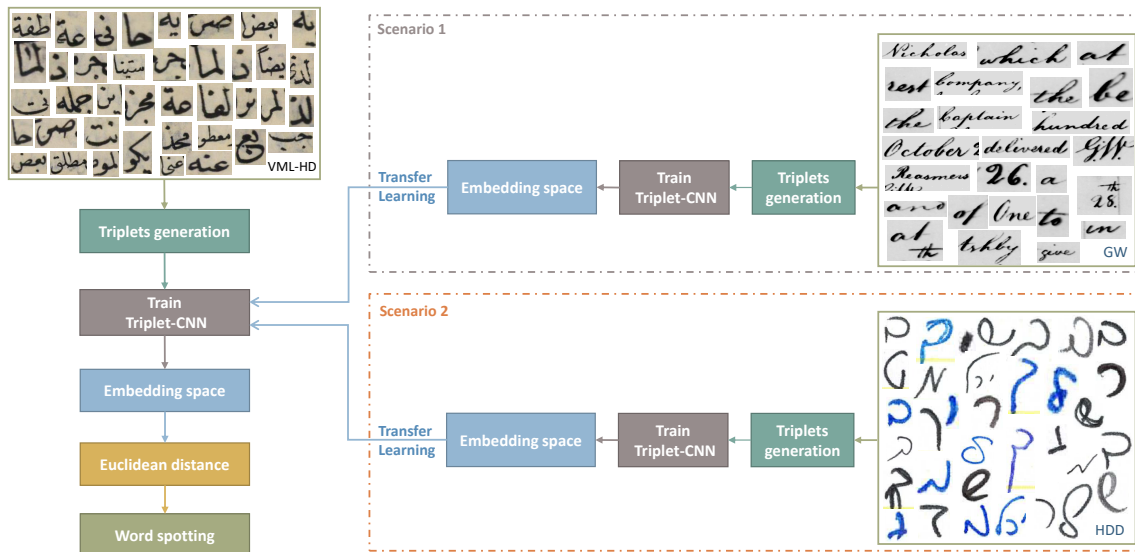


Figure 1: The proposed approach based Triplet-CNN architecture: Transfer Learning from historical English and handwritten Hebrew documents to Arabic document.

Precisely, an embedding space is constructed to represent word images. This latter must be able to withstand inter-class similarity and intra-class variance.

Transfer Learning is performed from two different datasets. Thus, the word spotting process is applied on the Arabic database twice using English and Hebrew datasets. Once the trained model is formed, the word spotting phase is performed by projecting the query word and pre-segmented dataset on the embedding space to obtain their new feature embedding, then, a Euclidean distance is computed to measure the similarity of the extracted embedding features from the query word and all dataset words as shown in Eq.(2).

$$\text{distance}(G(q), G(r)) = \|G(q) - G(r)\|^2 \quad (2)$$

where: q is the query word, r is a reference word.

Finally, word spotting produces a list that represents the retrieved words ranked according to their similarity to the query word.

4 EXPERIMENTAL RESULTS

4.1 Databases

The proposed triplet model is evaluated on the Visual Media Lab Historical Documents dataset (VML-HD) (Kassir et al., 2017). It is one of the largest available dataset in historical documents. It is consisting of five books in Arabic with a total of 680 pages written from the XI^{th} to the XV^{th} centuries onward by five

writers. For transfer Learning task, two databases are employed:

- **George Washington (GW):** The George Washington dataset is a collection of historical documents written in English by George Washington and his assistants (Rath and Manmatha, 2007). It contains 20 pages segmented into 4860 words with 1124 different transcriptions, providing ground truth at page, text line and word levels.
- **Hebrew Handwritten Dataset (HHD):** The HHD dataset (Rabaev et al., 2020) contains around 1000 handwritten forms written by different writers and accompanied by their ground truth at character, word and text line levels. The dataset contains 26 classes balanced in terms of the number of samples. The train set contains 3965 samples, test set contains 1134 samples.

4.2 Experimental Setup

4.2.1 Databases Partition

In our proposed Transfer Learning approach, we use a specific partition for each data set as described in Table 1.

To have a suitable comparison with (Barakat et al., 2018; Fathallah et al., 2019), the same evaluation procedure on the VML-HD dataset is used. Only one book is used for the training phase, while the model was evaluated on all five books. There is no shared data between the train, validation and test sets. The model was evaluated on all five books.

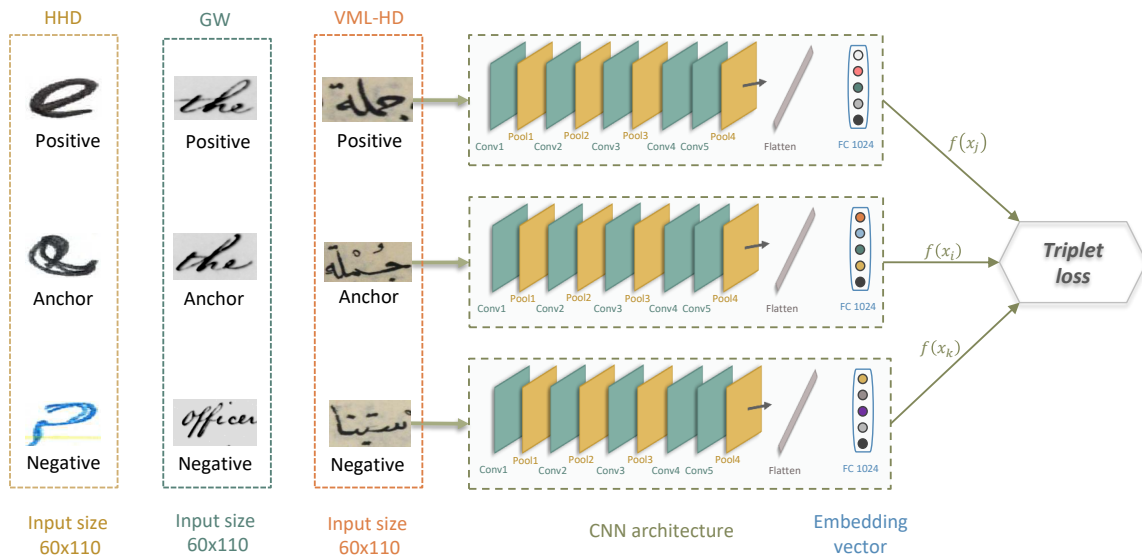


Figure 2: Triplet-CNN architecture employed for three datasets in order to extract embedding features from word images.

Table 1: Partitioning of datasets according to the level of word classes and the number of images per class.

Dataset	Sub-set	#words	#Samples/word
GW	Train	93	10
	Val	30	5
HHD	Train	27	20
	Val	27	10
VML-HD	Train	141	10
	Val	20	10
	Test	105	100

For all experiments, Keras backend Tensorflow framework is employed and the model is trained on NVIDIA Titan RTX GPU.

4.2.2 Evaluation Protocol

In order to evaluate performance, we select two performance assessments. The first one is $P@K$: Precision at the top- K -retrievals (Deng et al., 2011) and the second one is mAP : mean Average Precision (Everingham et al., 2010). To determine the K precision, the proportion of relevant items at a given rank K is calculated taking into account only the highest K scores given by the model (Eq.3).

$$P@K = \frac{|Res[1..K] \cap Rel|}{K} \quad (3)$$

where $Res[1..K]$ is the first K word retrieved by the system and Rel is the set of relevant words. In the evaluation, we present the results from first rank to the fifth ranks.

The mAP is calculated based on each precision score obtained over all queries and all ranks K .

4.3 Performance Comparison

In this section, we intend to assess the performance of word spotting using Transfer Learning technique. To better highlight the Transfer Learning technique in embedding features representation, we evaluate the results on five books of VML-HD dataset. We first report the spotting accuracy obtained by applying Transfer Learning using GW dataset and then, using HHD dataset.

Table 2 presents results according to $P@K$ metric following the top five ranks. From the obtained results, it is shown that the proposed TL-GW (Transfer Learning based GW dataset) and TL-HHD (Transfer Learning based HHD dataset) approaches slightly outperforms the state-of-the-art Siamese (Barakat et al., 2018) and Triplet (Fathallah et al., 2019) evaluated on VML-HD dataset.

Table 3 displays comparison performances results in terms of mAP metric on VML-HD dataset.

Many different observations are proposed to analyze the impact of the Transfer Learning technique on the word spotting system. Firstly, $P@1$, both proposed TL-GW and TL-HHD approaches provide an average enhancement of 2% over all books compared to Siamese and Triplet. In addition, at ranks 2 and 3, an improvement of 1% and 3% is respectively achieved. We can note that the transferred knowledge from different datasets contributed significantly to train the model on VML-HD for better feature embedding representations. Secondly, according to mAP values, the proposed Transfer Learning increased the model performances on Book 1, Book 2 and Book 5. But there is no improvement on Books 3 and 4.

Table 2: The P@K performance of the state-of-the-art methods and different Transfer Learning in our method on VML-HD dataset.

Method	Book	P@1	P@2	P@3	P@4	P@5
Siamese	Book 1	1.00	0.95	0.90	0.92	0.91
	Book 2	1.00	0.98	0.95	0.95	0.96
	Book 3	0.95	0.93	0.95	0.92	0.91
	Book 4	0.90	0.90	0.89	0.89	0.89
	Book 5	0.81	0.88	0.84	0.85	0.83
	All	0.93	0.92	0.90	0.90	0.90
Triplet	Book 1	1.00	0.95	0.95	0.94	0.94
	Book 2	0.95	0.95	0.92	0.93	0.91
	Book 3	0.90	0.93	0.94	0.92	0.92
	Book 4	0.95	0.90	0.90	0.87	0.85
	Book 5	0.81	0.86	0.86	0.86	0.84
	All	0.92	0.92	0.91	0.92	0.89
TL-GW	Book 1	0.95	0.93	0.95	0.96	0.94
	Book 2	1.00	1.00	1.00	1.00	0.99
	Book 3	1.00	0.98	0.97	0.94	0.93
	Book 4	0.95	0.93	0.89	0.89	0.89
	Book 5	0.81	0.83	0.84	0.85	0.85
	All	0.94	0.93	0.93	0.93	0.92
TL-HHD	Book 1	1.00	1.00	0.98	0.98	0.96
	Book 2	1.00	1.00	0.98	0.98	0.96
	Book 3	0.95	0.95	0.95	0.94	0.83
	Book 4	0.95	0.93	0.92	0.90	0.87
	Book 5	0.81	0.81	0.83	0.86	0.85
	All	0.94	0.93	0.93	0.93	0.91

Table 3: The mAP performance of the state-of-the-art methods and different Transfer Learning in our method on VML-HD dataset.

Method	Siamese	Triplet	TL-GW	TL-HHD
Book1	0.66	0.74	0.82	0.80
Book2	0.60	0.67	0.79	0.78
Book3	0.72	0.81	0.78	0.72
Book4	0.67	0.75	0.69	0.68
Book5	0.68	0.69	0.77	0.70
All	0.66	0.73	0.77	0.74

We can state that transferred features from GW and HHD datasets performed better with word classes in Books 1, 2 and 5 due to the fact that data in these books are more eligible. Finally, from a different perspective, the Triplet-CNN word spotting system is enhanced with TL-GW approach with a *mAP* rate of 4%. However, we reached a slightly lower rate of (-3%) using TL-HHD. This can be explained by the fact that the GW dataset shares common characteristics with VML-HD, where both databases are historical in nature and the word images have a similar background and behavior. Whereas the HHD dataset is handwritten with a clean background.

4.4 Error Analysis

Our proposed approach-based Transfer Learning has demonstrated better performance on word retrieval in HADs. Despite its high performance, the word spotting process has shown miss-retrieved occurrences of some word images. The miss-retrieved in this case is defined as the difference in label between a given model prediction and its actual label. Error analysis involves examining examples of sets that the TL-GW model has miss-retrieved, in order to understand the underlying sources of errors. This can help to identify issues that require special treatment and to determine their priority. Thus, guidance for error handling can be provided. An error analysis is performed to outline the reasons why some images have been incorrectly spotted by the TL-GW model. The error analysis process is conducted with word images from the validation set of the VML-HD dataset where the P@K metric is considered to evaluate the model. The validation set consists of 20-word classes and each word class consists of 10 samples.

Figure.3 introduces some examples of word classes that are miss-retrieved. We display the query word in the blue box and the first 5 samples retrieved then the correct occurrences ranks of the query word presented in the orange box.

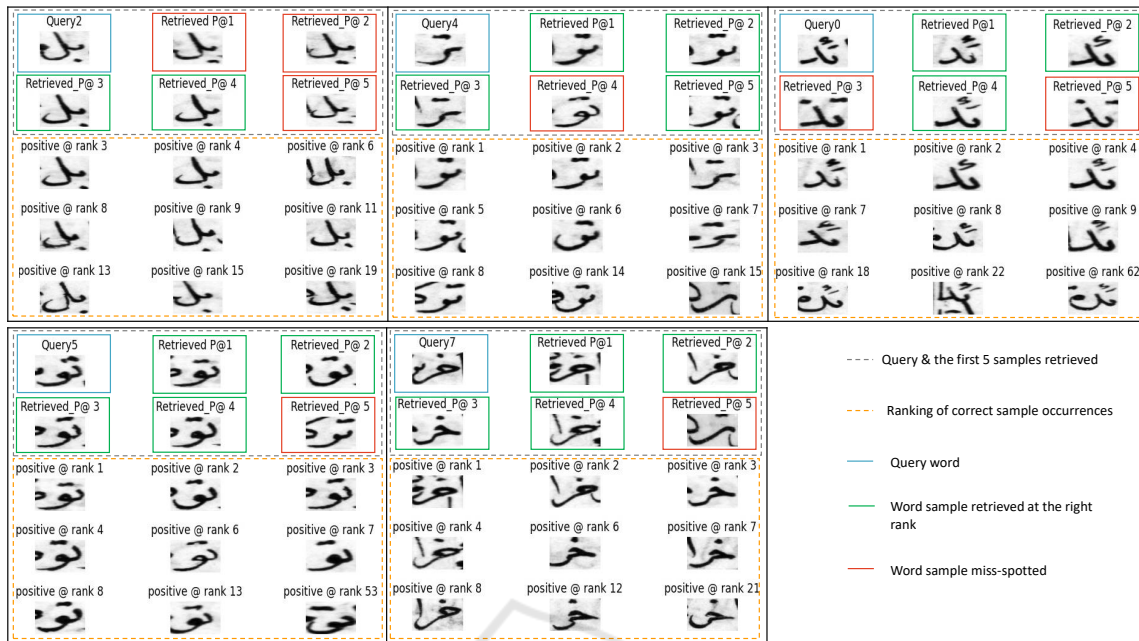


Figure 3: Some examples of miss-retrieved words: Error analysis with displaying the first five ranks spotted (from P@1 to P@5) and the right occurrences ranks.

Depending on the displayed words incorrectly retrieved by the enhanced model, several sources of errors can be identified. First of all, an ambiguity between similar word classes. It is commonly known that CNN works well when there is a clear separation between word classes and it is not the case in our dataset, where there are a lot of similar word classes. The similarity between the classes of the word consists in the style of writing of the Arabic letters that constitute a word and in the background color in each image.

Second, the source of error may come from the segmentation phase. In the VML-HD database, there are many missegmented word images, e.g. words that have additional letters or diacritics or also part letters of another preceding or the following word. On the other hand, many of the words that are missing precise letters or punctuation.

In addition, mislabeled data may be a reason for error in the Triplet-CNN model: in general, data labeling is a subjective task because it is provided by human judgments. In the VML-HD database, some word images are mislabeled. To conclude, this analysis showed that the error rate can be reduced by optimizing the CNN architecture used for feature extraction to be more efficient in distinguishing between similar images or through appropriate pre-processing with emphasis on missegmented word images.

5 CONCLUSION

In this paper, we proposed an enhancement approach for word spotting in HADs. Our approach is based on Transfer Learning method. It aimed to leverage knowledge acquired on a source dataset to better address a new target dataset. Our approach consists in two steps: training Triplet-CNN using English historical documents and Hebrew handwritten documents, and then, retraining the same deep network on HADs by transferring the learned features provided in the first step. The employed Triplet-CNN architecture aimed to build an embedding space for data representation using triplet loss for maximizing the distance between word images belonging to different classes and minimize the distance between word images belonging to the same class. The experimental results on the VML-HD dataset highlight a high performance of our proposed approach in order to enhance word spotting in HADs process. A detailed discussion on error analysis for TL-GW model was presented. It revealed that the most prominent source of error appears to be the similarity between word style writing and background.

There are various potential extensions of this research work. We intend to use Transfer Learning to exploit knowledge from the merging of two databases GW and HHD. Moreover, we plan to investigate do-

main adaptation techniques, such as based on adversarial networks, to enhance the data representation. Finally, we can evaluate our proposed approach by employing different matching algorithms to measure similarity between embedding features.

REFERENCES

- Barakat, B. K., Alasam, R., and El-Sana, J. (2018). Word spotting using convolutional siamese network. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 229–234. IEEE.
- Can, Y. S. and Kabadayı, M. E. (2020). Automatic cnn-based arabic numeral spotting and handwritten digit recognition by using deep transfer learning in ottoman population registers. *Applied Sciences*, 10(16):5430.
- Cui, Y., Song, Y., Sun, C., Howard, A., and Belongie, S. (2018). Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4109–4118.
- Deng, J., Berg, A. C., and Fei-Fei, L. (2011). Hierarchical semantic indexing for large scale image retrieval. In *CVPR 2011*, pages 785–792. IEEE.
- Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.
- Fathallah, A., Khedher, M. I., El-Yacoubi, M. A., and Amara, N. E. B. (2020). Evaluation of feature-embedding methods for word spotting in historical arabic documents. In *2020 17th International Multi-Conference on Systems, Signals & Devices (SSD)*, pages 34–39. IEEE.
- Fathallah, A., Khedher, M. I., El-Yacoubi, M. A., and Essoukri Ben Amara, N. (2019). Triplet cnn-based word spotting of historical arabic documents. *27th International Conference on Neural Information Processing (ICONIP)*, 15(2):44–51.
- Gurjar, N., Sudholt, S., and Fink, G. A. (2018). Learning deep representations for word spotting under weak supervision. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 7–12. IEEE.
- Kassis, M., Abdalhaleem, A., Droby, A., Alaasam, R., and El-Sana, J. (2017). Vml-hd: The historical arabic documents dataset for recognition systems. In *Arabic Script Analysis and Recognition (ASAR), 2017 1st International Workshop on*, pages 11–14. IEEE.
- Khayyat, M. and Suen, C. Y. (2018). Improving word spotting system performance using ensemble classifier combination methods. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 229–234. IEEE.
- Khayyat, M. M. and Elrefaei, L. A. (2020). Towards author recognition of ancient arabic manuscripts using deep learning: A transfer learning approach. *International Journal of Computing and Digital Systems*, 9(5):1–18.
- Konidakis, T., Kesidis, A. L., and Gatos, B. (2016). A segmentation-free word spotting method for historical printed documents. *Pattern analysis and applications*, 19(4):963–976.
- Mhiri, M., Desrosiers, C., and Cheriet, M. (2019). Word spotting and recognition via a joint deep embedding of image and text. *Pattern Recognition*, 88:312–320.
- Mohammed, H. H., Subramanian, N., Al-Maadeed, S., and Bouridane, A. (2022). Wsnet-convolutional neural network-based word spotting for arabic and english handwritten documents. *TEM*.
- Rabaev, I., Barakat, B. K., Churkin, A., and El-Sana, J. (2020). The hhd dataset. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 228–233. IEEE.
- Rajeswar, S., Rodriguez, P., Singhal, S., Vazquez, D., and Courville, A. (2022). Multi-label iterated learning for image classification with label ambiguity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4783–4793.
- Rath, T. M. and Manmatha, R. (2007). Word spotting for historical documents. *International Journal of Document Analysis and Recognition (IJ DAR)*, 9(2-4):139–152.
- Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.
- Sudholt, S. and Fink, G. A. (2018). Attribute cnns for word spotting in handwritten documents. *International journal on document analysis and recognition (ijdar)*, 21(3):199–218.
- Westphal, F., Grahn, H., and Lavesson, N. (2020). Representative image selection for data efficient word spotting. In *International Workshop on Document Analysis Systems*, pages 383–397. Springer.
- Ye, M. and Shen, J. (2020). Probabilistic structural latent representation for unsupervised embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5457–5466.
- Zagoris, K., Pratikakis, I., and Gatos, B. (2014). Segmentation-based historical handwritten word spotting using document-specific local features. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 9–14. IEEE.