A transformer-based siamese network for word image retrieval in historical documents

Abir Fathallah, Mounim El Yacoubi, Najoua Essoukri Ben Amara

A Transformer-based Siamese Network For Word Image Retrieval In Historical Documents

1st Abir Fathallah
Université de Sousse, Institut Supérieur de l'Informatique et des Techniques de Communication,
LATIS- Laboratory of Advanced Technology and Intelligent Systems,
4011, Sousse, Tunisie;
Email: abir.fathallah@telecom-sudparis.eu

2nd Mounîm A. El-Yacoubi
Samovar, Telecom SudParis,
Institut Polytechnique de Paris
91120 Palaiseau, France
Email: mounim.el_yacoubi@telecom-sudparis.eu

3rd Najoua Essoukri Ben Amara
Université de Sousse, Ecole Nationale d'Ingénieurs de Sousse,
LATIS- Laboratory of Advanced Technology and Intelligent Systems,
4023, Sousse, Tunisie;
Email: najoua.benamara@eniso.rnu.tn

***Abstract*—The increasing availability of digitized historical documents has sparked a need for effective information processing tools to extract the valuable information contained within them. Word spotting, an area of focus in historical document analysis, involves identifying specific words within images of documents. In this paper, we propose a novel approach for word spotting in historical Arabic documents, utilizing improved feature representations for learning word images. More precisely, we put forward an end-to-end approach for generating word image descriptors, based on the Siamese vision transformer architectures. The model learning is guided by a contrastive loss objective. Additionally, we carry out transfer learning techniques by leveraging knowledge acquired from two distinct source domains to generalize model learning. The proposed approach utilizes the embedding space to evolve the word spotting system by projecting the query word image and all reference word images into the embedding space, where their similarity is determined based on their corresponding embedding vectors. Our method is evaluated on the historical Arabic VML-HD dataset and the results indicate that our approach significantly outperforms state-of-the-art methods.**

***Index Terms*—Historical Arabic documents, Word spotting, Vision transformer, Siamese network, Transfer learning, Learning representation.**

I. Introduction

The preservation and protection of historical Arabic documents (HADs) pose a significant challenge, as access to these precious cultural heritage resources is often limited. This is due, in part, to the vast number of pages that must be scanned and stored on various servers. Furthermore, the digital form in which these historical documents are available is not readily amenable to automatic interpretation by computer vision techniques and thus requires pre-processing to convert the documents into a more readable format.

The analysis and processing of historical documents present a significant challenge, largely due to the degraded quality of the manuscripts due to aging and various forms of degradation over time. Automated processing of these documents is crucial for making the vast amount of information contained within them accessible to both human readers and computer vision systems. Among the most challenging historical documents to process are HADs, owing to the complexity of the Arabic script and the variety of its forms. Many old Arabic texts contain additional characters and diacritical marks, further complicating the processing of these documents.

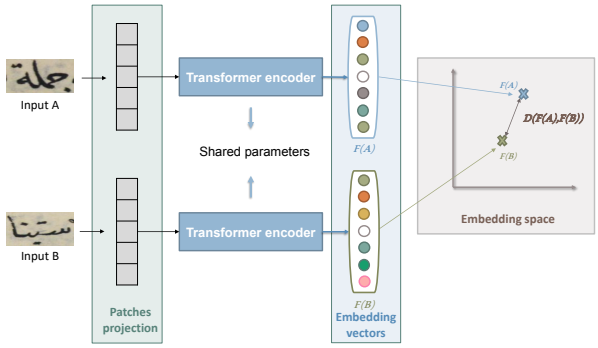

Fig. 1. The method involves mapping two input word images into a common feature space via the use of transformers and utilizing a contrastive loss function to train the model. The Siamese transformer architectures allow the model to learn the similarities and differences between the two input word images, enabling it to effectively identify and match word images in the common feature space.

In recent years, there has been a growing trend in the field of HAD analysis toward the application of word spotting techniques in image processing. The complexity of the Arabic script presents a significant challenge in accurately locating occurrences of a specific word within a large set of document images. Therefore, there is a need for effective approaches for retrieving query words in such documents. Two main approaches for word retrieval in HADs have been proposed, depending on the representation of the query word. The first approach, known as the Query-by-Example (QbE) method, involves providing the query word as an image. The second

approach, known as the Query-by-String (QbS) method, utilizes a text string to indicate the query word.

However, extracting features from historical images is a challenging task, as the images are often of poor quality and may be degraded. A current strategy to address this issue involves representing the data in a new embedding space. This approach has been found to improve the performance of QbE methods, as the new embedding space allows for the extraction of more robust features that are better suited for the task of word retrieval in HADs.

The objective of deep embedding approaches in historical document analysis is to construct an embedding space that effectively transforms the input image into new representations by selecting and extracting pertinent features [1], [2]. The recent advancements in deep architectures, particularly vision transformers, have facilitated the creation of more accurate embedding representations. The transformer architecture, first introduced by [3], has been widely adopted in natural language processing tasks, and more recently, in several computer vision tasks such as image classification. This is particularly noteworthy as transformer-based models for computer vision are built on a different set of inductive biases than the commonly used convolutional architectures, which implies that they may provide new solutions and overcome the limitations of traditional convolutional architectures.

The scarcity of annotated public datasets for historical documents is a pressing concern in the research field of HADs. Despite the availability of numerous public datasets for document analysis, the limited number of annotated public historical documents presents a challenge in evaluating the efficacy of word spotting approaches. In light of this limitation, the application of transfer learning techniques has emerged as a viable solution [4], [5]. It enables the transfer of knowledge from relevant source data to improve specific target tasks [6]–[8].

In the same context, in this paper, we investigate the potential of transfer learning as a means to leverage learned features from previous document datasets to improve the representation of embedding features of Arabic word images. Specifically, we propose a Siamese transformer-based approach for word spotting in HADs. We provide an in-depth analysis of the adaptation of transfer learning techniques and their interaction with transformers. In particular, we investigate the transfer of knowledge from two datasets, which have different characteristics: historical documents written in English and handwritten documents written in Hebrew. As shown in Fig.1, our model involves a Siamese architecture with a backbone of transformers which presents a promising avenue for research, as the use of a transformer-based approach in combination with the Siamese architecture may enable the model to learn the similarities and differences between word images and effectively identify and match them in a common feature space, thereby improving the performance of word retrieval task in historical documents. The key contributions of this work are highlighted below.

- We introduce a simple contrastive loss function to train Vision Transformer (ViT) for word image spotting in Arabic historical documents,
- We explore the use of transfer learning by utilizing two different source domains, Hebrew manuscript documents and English historical documents, and transferring the learned features to the target domain of Arabic historical documents,
- Accordingly, we report the new state of the art on three widely used benchmarks for word spotting in historical documents.

The remainder of this paper is organized as follows: In Section II, we provide an overview of existing techniques for word retrieval in historical documents, as well as a brief summary of the related work in transfer learning and ViT approaches. In Section III, we introduce our proposed approach for word spotting in historical documents using Siamese transformers-based transfer learning. The experimental design, results, and error analysis are discussed in Section IV. Finally, in Section V, we present the conclusions of this study, as well as potential future research directions.

II. Related Work

In this section, we present an overview of the field of word spotting in historical documents, including a summary of the state-of-the-art techniques and recent developments.

A. Word spotting in historical documents

Word spotting in historical document images can be utilized to exploit document content in the digital form [9]. 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 [1], [2], [4], [5], [10], [11]. 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 [1], [2], [12], [13] and segmentation-free approaches which involve the selection of patches from the input document through a sliding window technique or template matching [14], [15]. Several researchers are interested in improving the word spotting process in historical documents. The authors in [16] proposed an enhanced internal structure hierarchical classifier. Other methods, were employed to enhance the performance of the word retrieval approaches introduced in [17], [18]. Using the method proposed in [19], a data augmentation process was applied to improve word spotting performance. Meanwhile, the authors of [18] suggested a pre-training CNN architecture by using the synthetic dataset [20] which proved an improvement in word spotting performance, even though the training samples were few.

On the other hand, the authors of [17] 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.

B. Transfer learning for historical documents

To enhance the content exploitation processes in historical documents, many approaches have aimed at developing better word spotting models by applying transfer learning. For detecting similarity between two different documents written by two different authors, the HWNet [20] took advantage of a convolutional network architecture. The HWNet v2 [21], provided an adaptation of the ResNet-34 architecture featuring region-of-interest pooling layers facilitating the reading of images of varying sizes. This database held approximately one million word images, which avoided the need for any data augmentation methods. The authors in [22] investigated the handwriting recognition issue in the absence of any data training for historical documents. Indeed, the authors felt that several difficulties may have occurred as a result of considerable differences in the time period and geographical area, which often affected handwriting style.

The transductive transfer learning approach [23] is particularly interesting. It aims at domain adaptation by using various sources and target data for the same task. In this paper, we intend to apply this method and push it further by increasing the number of annotated data used as sources and by introducing parameter transfer.

C. Vision Transformer

The transformer architecture was first published in [3] to handle sequential data in natural language processing. Later, various studies investigated the success of Transformer in computer vision by introducing the sequence of feature maps taken from CNNs [24]–[26]. In 2020, Google designed ViT [27], a simple transformer applied directly to a sequence of image patches for classification. Some variants of ViT have achieved considerable success. For example, authors in [28] proposed a hierarchical ViT based on staggered windows. On the other hand, for providing a dense prediction, authors in [29] suggested a ViT in the form of a pyramid. In the context of document processing, transformer architectures were specifically developed and chosen [30]–[33]. The authors in [34] suggested an end-to-end solution based on a transformer operating at the paragraph level. The proposed transformer model aimed to recognize named entities in handwritten documents. A word spotting system relying on the QbE and QbS approaches has been provided in [35]. In order to build a representation to encode both forms of words (texts and images), they adopted the strengths of convolutions and transformer layers.

In light of the considerable success of transformers in many domains, in this paper, we present a Siamese transformer model based on transfer learning for word spotting in historical documents. The Siamese transformer architecture is exploited to build an embedding space for image representations to find specific words in a historical document. More details on the proposed approach are given in section III.

III. Proposed Method

In this section, we investigate the problem of word spotting enhancement using the transfer learning technique where the goal is to exploit and leverage the knowledge of feature representations from the source domain to the target domain. The ViT architecture is used to extract pertinent features from word images that will be considered as their new representations in the constructed embedding space. 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 representations.

Fig.2 shows the flowchart of our proposed approach. More specifically, the Siamese transformer takes as input a pair of word images. The idea involves training the Siamese transformer-based contrastive loss to extract features along with transferring information from source domains D_{S1} and D_{S2} to the target domain D_T. Then, each word image is represented efficiently by a feature vector named (embedding vector).

A. Enhancement-based Transfer Learning

transfer learning [36] involves the ability to leverage the existing prerequisite knowledge provided previously by the source learner in the target task. A domain, $\mathcal{D}$, is denoted by a tuple of two elements that consists of the feature space, $\mathcal{X}$, along with the marginal probability, $\mathcal{P}(\mathcal{X})$, given that $\mathcal{X}$ corresponds to a particular point in the sample, where $\mathcal{X} = \{x_1, ..., x_n\} \in \mathcal{X}$. In this way, a mathematical description of the domain is given as $\mathcal{D} = \{\mathcal{X}, \mathcal{P}(\mathcal{X})\}$. transfer learning includes using a domain ($\mathcal{D}$) and a task ($\mathcal{T}$).

For our approach, $\mathcal{X}_1$ and $\mathcal{X}_2$ are the representations learned from two different source spaces: historical English documents and handwritten Hebrew documents, respectively. We note by x_i is the i^{th} term vector corresponding to some word and $\mathcal{X}$ is the sample of word image employed for training. For a particular domain, $\mathcal{D} = \{\mathcal{X}, \mathcal{P}(\mathcal{X})\}$, a task $\mathcal{T}$ consisting of a label space $\mathcal{Y}$ and a conditional probability distribution $\mathcal{P}(\mathcal{Y}|\mathcal{X})$ learned generally on the basis of learning data formed of pairs $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. Additionally, we note that $\mathcal{D}_{\mathcal{S}_1}$ refers to the first source domain, historical English documents, and $\mathcal{D}_{\mathcal{S}_2}$ corresponds to the second source domain, handwritten Hebrew documents, while $\mathcal{D}_{\mathcal{T}}$ represents the target domain historical Arabic documents. All domains are assigned the same task $\mathcal{T}$. Given a source domain $\mathcal{D}_{\mathcal{S}_1}$ and$\mathcal{D}_{\mathcal{S}_2}$, a corresponding source tasks $\mathcal{T}_{\mathcal{S}_1}$ and $\mathcal{T}_{\mathcal{S}_2}$, as well as a target domain $\mathcal{D}_{\mathcal{T}}$ and a target task $\mathcal{T}_{\mathcal{T}}$, the objective of transfer learning now is to enable us to learn the target conditional probability distribution $\mathcal{P}(\mathcal{Y}_{\mathcal{T}}|\mathcal{X}_{\mathcal{T}})$ in $\mathcal{D}_{\mathcal{T}}$ with the information gained from $\mathcal{D}_{\mathcal{S}}$ and $\mathcal{T}_{\mathcal{S}}$ where $\mathcal{D}_{\mathcal{S}_1} \neq \mathcal{D}_{\mathcal{S}_2} \neq \mathcal{D}_{\mathcal{T}}$ and $\mathcal{T}_{\mathcal{S}_1} = \mathcal{T}_{\mathcal{S}_2} = \mathcal{T}_{\mathcal{T}}$.

The main aim is to identify suitable feature representations that can be transmitted from the multi-source domains to the target domain.

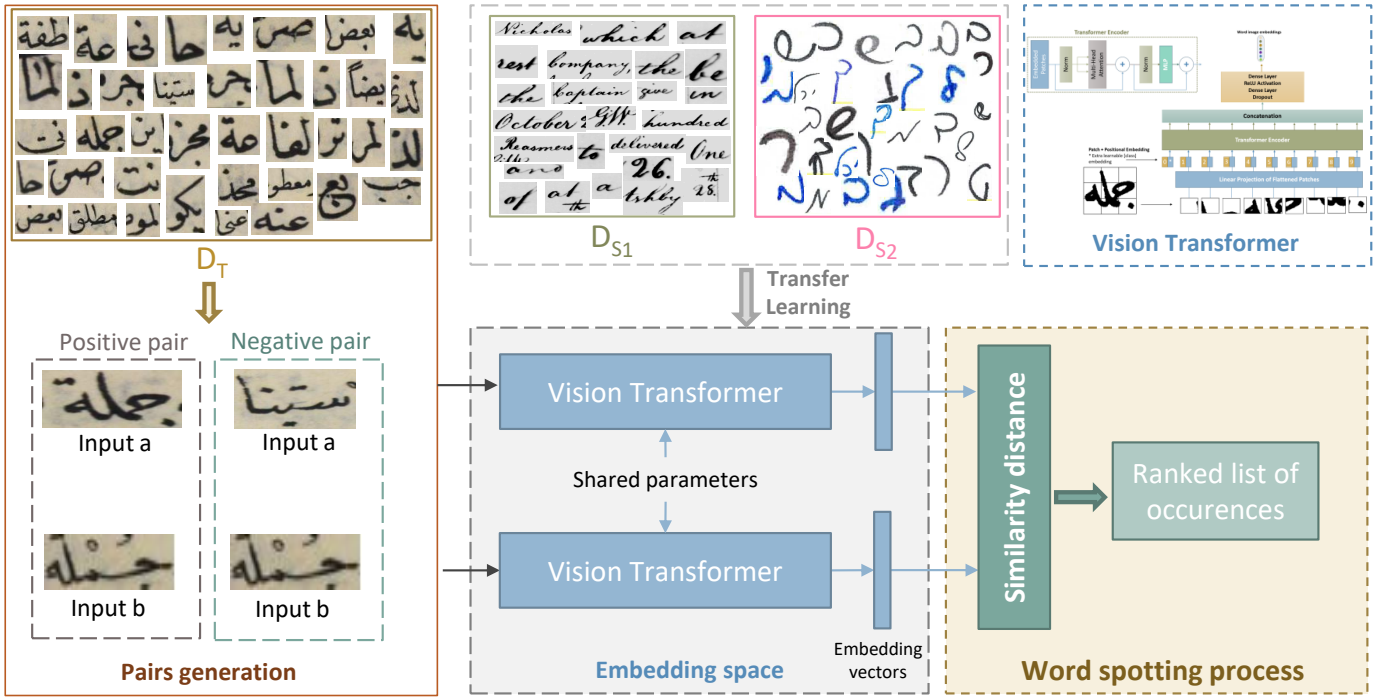

Fig. 2. The proposed approach based on Siamese transformer: transfer learning is applied from historical English and handwritten Hebrew documents to Arabic documents.

B. Siamese vision transformer for word image representation

The employed ViT architecture is depicted in Figure 2. It is inspired from [27]. It incorporates a patch embedding generation and a transformer encoder for representing the word image. First, the input image is subdivided into M patches of fixed size (e.g., 16×16). Then each patch is projected in a linear way into M vector chips and subsequently used as input to the transformer in a permutation-invariant manner. Location prior is introduced by adding a learnable 1-D location encoding vector to the input tokens. An additional learnable token is added to the input sequence such that its consistent output token serves as a global representation of the image. The transformer consists of L layers, with each layer consisting of two main blocks: a multi-headed self-attention (MSA) layer, which performs a self-attention operation on different projections of the input tokens, and a feed-forward network (FFN). Both the MSA and FFN layers are preceded by a layer normalization and followed by a skip connection. For more details, this paper refers the reader to [27].

In the second step, the distance between the outputs of the transformer branches is measured using the Euclidean distance with the aim to minimize the contrast loss [37]. Let x_1, x_2 denote a pair of word images. Let y represent the label of this pair, where y = 0 if their distance is small and y = 1 if their distance is large. Then the network learns the parameters θ of the distance function D will be defined by Eq.(1).

$$D_\theta(x_1, x_2) = \|\mathcal{H}_\theta(x_1) - \mathcal{H}_\theta(x_2)\|_2 \tag{1}$$

Where $\mathcal{H}_\theta$ represents the output function of the transformer. In this case, the contrastive loss function $\mathcal{L}$ is then introduced by Eq.(2).

$$\mathcal{L}(\theta, y, x_1, x_2) = \frac{1}{2}[(1-y)(D_\theta)^2 + (y)[\max(0, m - D_\theta)]^2] \tag{2}$$

where $m > 0$ refers to a margin.

The ViT architecture is used to extract pertinent features from word images that will be considered as their new representations in the constructed embedding space. In this work, we apply the same transformer architecture and its parameters to all used datasets. The main idea is intended to train the Siamese transformer for extracting features from Arabic word images by transferring knowledge acquired from previous datasets. Precisely, an embedding space is constructed to represent word images. This latter must be able to withstand inter-class similarity and intra-class variance. For all employed datasets, the same training strategy is used.
Thus, the word spotting process is applied to the Arabic database twice using English and Hebrew datasets.

C. Embedding matching process

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

IV. Experimental Results

A. Databases

The proposed Siamese transformer model is evaluated on the Visual Media Lab Historical Documents dataset (VML-HD) [38] which is one of the largest available datasets 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 [39]. It contains 20 pages segmented into 4860 words.
- **Hebrew Handwritten dataset (HHD)**
  The HHD dataset [40] 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.

B. Experimental Setup

1) Databases partition: In our proposed approach, we use a specific partition for each data set as described in Table I. To have a suitable comparison with [1], [2], 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.

TABLE I
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

The learning process is optimized by the stochastic gradient descent algorithm with a learning rate of 10^{-3} and a batch size of 512. All parameter values are empirically chosen. For all experiments, the PyTorch framework is used and the model is trained on an NVIDIA Quadro RTX 6000 GPU with 24 GB of RAM.

2) Evaluation protocol: To assess performance, we have chosen two metrics: $P@K$ and mAP. $P@K$ (Precision at the top-K-retrievals) measures the precision of the top K retrieved items and is defined as the proportion of relevant items among the highest K scores produced by the model (as described in [41]). mAP (mean Average Precision) represents the average precision across all relevant items, as defined in [42]. In the calculation of $P@K$, only the top K scores are considered as depicted in (Eq.(3)).

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

The calculation of $P@K$ involves considering the first K words retrieved by the system, represented as $Res[1..K]$, and the set of relevant words, represented as Rel. The evaluation of $P@K$ presents results for the first to fifth ranks. The calculation of mAP is based on the precision scores obtained over all queries and all ranks K.

C. Performance Comparison

The objective of this section is to evaluate the performance of our proposed word spotting model. To further highlight our proposed approach for embedding feature representations, we evaluate the results on five books of the VML-HD dataset. We first report the retrieval accuracy obtained by applying transfer learning using the GW dataset and then, using the HHD dataset.

TABLE II
The P@K performance of the state-of-the-art methods and different transfer learning provided in our proposed method on VML-HD dataset.

Method	Book	P@1	P@2	P@3	P@4	P@5
[1]	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
	mP@k	**0.93**	**0.92**	**0.90**	**0.90**	**0.90**
[2]	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
	mP@k	**0.92**	**0.92**	**0.91**	**0.92**	**0.89**
TL-GW	Book 1	**1.00**	**1.00**	**1.00**	0.98	0.96
	Book 2	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	Book 3	1.00	**1.00**	0.99	0.97	0.97
	Book 4	**1.00**	**1.00**	**0.98**	**0.98**	**0.97**
	Book 5	**0.99**	**0.97**	**0.96**	**0.96**	**0.95**
	mP@k	**0.99**	**0.99**	**0.98**	**0.97**	**0.97**
TL-HHD	Book 1	**1.00**	**1.00**	**0.98**	**0.98**	**0.97**
	Book 2	**1.00**	**1.00**	**1.00**	0.98	0.98
	Book 3	**1.00**	**0.95**	**0.97**	0.96	0.93
	Book 4	0.96	1.00	0.98	0.94	0.97
	Book 5	**0.93**	0.94	0.94	0.93	0.94
	mP@k	**0.97**	**0.97**	**0.97**	**0.95**	**0.95**

Table II presents the results of the P@K metric for the top five ranks. The results demonstrate that the proposed TL-GW (Transfer Learning on the GW dataset) and TL-HHD (Transfer Learning on the HHD dataset) approaches exhibit a slight improvement over the state-of-the-art Siamese [1] and Triplet [2] methods, as evaluated on the VML-HD dataset.

Table III presents a comprehensive comparison of performance results utilizing the mAP metric on the VML-HD dataset. Many different observations are proposed to analyze the impact of the transfer learning technique on the word

TABLE III
THE MAP PERFORMANCE OF THE STATE-OF-THE-ART METHODS AND DIFFERENT TRANSFER LEARNING IN OUR METHOD ON EACH BOOK FROM THE VML-HD DATASET.

	[1]	[2]	TL-GW	TL-HHD
Book1	0.66	0.74	0.89	0.86
Book2	0.60	0.67	0.87	0.83
Book3	0.72	0.81	0.88	0.82
Book4	0.67	0.75	0.89	0.80
Book5	0.68	0.69	0.84	0.81

spotting system. Firstly, according to $P@1$ in Table II, our both proposed TL-GW and TL-HHD approaches provide an average enhancement of 6% and 7% on all VML-HD books compared to Siamese in [1] and Triplet in [2], respectively. In addition, at ranks 2 and 3, an improvement of 6% and 7% is respectively achieved. We can note that the transferred knowledge from different datasets and ViT architecture contributed significantly to training the model on VML-HD for better feature embedding representations. Secondly, according to mAP values in Table III, the proposed approach increased the model performances with 23% and 15% on first book books compared to [1] and [2]. We can state that our model performed also very well on the other books. Furthermore, from a different point of view, the combination of the ViT and transfer learning improves the word retrieval system by constructing an appropriate embedding space for word image representations. The strong transformer architecture along with transfer learning technique-based multi-source domains represent a suitable solution to improve word spotting tasks in historical Arabic documents. 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.

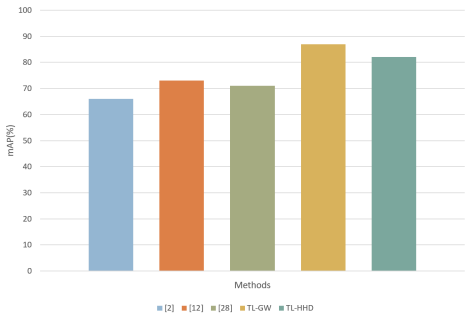

Fig. 3. Results on VML-HD dataset in terms of mAP metric using Euclidean distance: Comparison with state-of-the-art methods.

Additionally, in terms of mAP, our proposed method TL-GW is compared to different methods from the state of the art according to the VML-HD dataset. The same evaluation protocol is applied. As depicted in Fig.3, the result is again satisfactory: our model presents the best results obtained compared to [1], [4] and [2] according to the VML-HD dataset. Clearly, this figure confirms the advantage of the ViT representation over CNN and graph techniques.

D. Error Analysis

Our proposed approach 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 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.

Fig.4 introduces some examples of word classes that miss-retrieved. We display the query word in the blue box and the first 5 samples are retrieved then the correct occurrences rank of the query word is presented in the orange box. 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 transformer architecture performs well in representing data 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 incorrectly segmented 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 are missing precise letters or punctuation.
In addition, mislabeled data may be a reason for the error in the Siamese transformer model: in general, data labeling is a subjective task because it is provided by human judgments. For the VML-HD database, some word images are mislabeled. To conclude, this analysis showed that the error rate can be reduced by optimizing the transformer architecture used for feature extraction to be more efficient in distinguishing between similar images or through appropriate pre-processing with emphasis on incorrectly segmented word images.

E. Ablation Study

Our experimental results demonstrate that the proposed method achieves comparable performance with state-of-the-art techniques. However, to empirically demonstrate the advantages of TL-GW and TL-HDD, we conduct an ablation study on three widely used datasets: VML-HD, HADARA80P

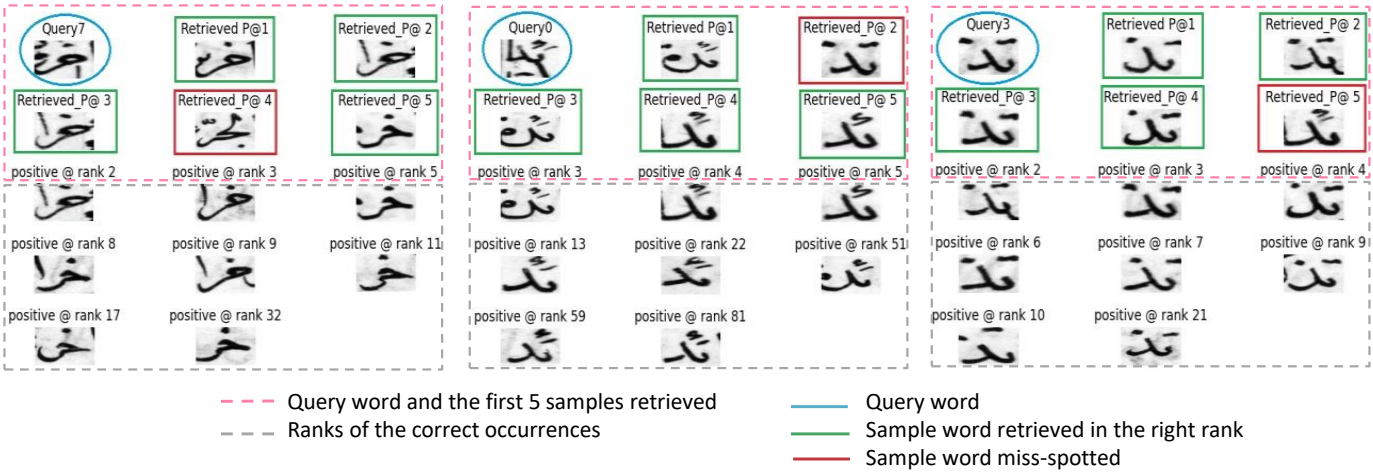

Fig. 4. Some examples of miss-retrieved words: Error analysis with displaying the first five ranks spotted (from P@1 to P@5) and the right occurrences ranks.

[43], and GW. The objective of this study is to investigate the advantages of utilizing transfer learning from various source domains for the task of word spotting. Specifically, we aim to verify the effectiveness of transferring knowledge from different domains and its impact on the performance of the word spotting task.

TABLE IV
RESULTS ON VML-HD ACCORDING TO MAP METRIC USING EUCLIDEAN DISTANCE.

Model	TL-GW	TL-HDD	VML-HD	HADARA80P	GW
	✗	✗	0.79	0.76	0.83
Ours	✓	✗	0.87	0.83	0.91
	✗	✓	0.84	0.82	0.88
	✓	✓	**0.88**	**0.82**	**0.89**

All possible combinations of the two source domains are considered, in order to identify the most effective approach for improving the model's performance.

Following the obtained results in terms of all the employed databases, there is a drop in the overall MAP for each experiment. On the other hand, the best results are achieved if we apply the combination of the three proposed improvement strategies, as shown in the last row of Table IV.

The implementation of transfer learning from different domains, specifically TL-HDD and TL-GW, has resulted in a substantial enhancement in the mAP metric for VML-HD, HADARA80P and GW. Our findings show that the utilization of transfer learning has led to a 9%, 6% and 6% increase in mAP respectively. This can be attributed to the fact that transfer learning enables the transfer of knowledge from a source task to a target task, where the two tasks possess some level of similarity. By capitalizing on the knowledge acquired from the source task, the model is able to improve its performance on the target task, resulting in a higher level of precision. The improvement in mAP further confirms the efficacy of transfer learning in enhancing performance in related but dissimilar domains.

Transfer learning is a valuable technique for analyzing historical Arabic documents, as it allows for the utilization of pre-trained models to improve the performance of a word spotting model despite limited data availability. By leveraging the knowledge and expertise embedded in a pre-trained model that has been trained on a dataset similar to the historical documents, the word spotting model can benefit from a strong foundation and a deeper understanding of the structure, layout and patterns of Arabic text. This can lead to a more accurate and reliable word spotting model, capable of recognizing words in historical documents with greater precision.

V. CONCLUSION

In this paper, we proposed a Siamese transformer-based transfer learning approach for word spotting in HADs. It aimed to encode word images in new representations by leveraging knowledge acquired on a source dataset to better address a new target dataset. Our approach consisted in building an embedding space for word image representation using a Siamese transformer and TL based on two different source domains: English historical documents and Hebrew handwritten documents. The suggested method inputs a pair of word images extracts their representations and outputs the distance between them for distinguishing if the two input word images are belonging to the same class or to different classes. The experimental results on the VML-HD dataset highlight a high performance of our proposed approach in order to enhance word spotting in HADs.

There are various potential extensions of this research work. We intend to add a new recognition module for ensuring recognition and spotting simultaneously in HADs. Moreover, we plan to investigate domain adaptation techniques, such as those based on adversarial networks [44], to enhance the data representation. We can also evaluate our proposed approach by employing different matching algorithms to measure the similarity between embedding features.

REFERENCES

[1] B. K. Barakat, R. Alasam, and J. El-Sana, "Word spotting using convolutional siamese network," in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. IEEE, 2018, pp. 229–234.

[2] A. Fathallah, M. I. Khedher, M. A. El-Yacoubi, and N. Essoukri Ben Amara, "Triplet cnn-based word spotting of historical arabic documents." *27th International Conference on Neural Information Processing (ICONIP)*, vol. 15, no. 2, pp. 44–51, 2019.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] H. H. Mohammed, N. Subramanian, S. Al-Maadeed, and A. Bouridane, "Wsnet–convolutional neural network-based word spotting for arabic and english handwritten documents," *TEM*, 2022.

[5] R. Pramanik and S. Bag, "Handwritten bangla city name word recognition using cnn-based transfer learning and fcn," *Neural Computing and Applications*, vol. 33, no. 15, pp. 9329–9341, 2021.

[6] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *IEEE CVPR*, 2018, pp. 4109–4118.

[7] M. Ye and J. Shen, "Probabilistic structural latent representation for unsupervised embedding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5457–5466.

[8] S. Rajeswar, P. Rodriguez, S. Singhal, D. Vazquez, and A. Courville, "Multi-label iterated learning for image classification with label ambiguity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4783–4793.

[9] A. Fathallah, M. I. Khedher, M. A. El-Yacoubi, and N. E. B. Amara, "Evaluation of feature-embedding methods for word spotting in historical arabic documents," in *2020 17th International Multi-Conference on Systems, Signals & Devices (SSD)*. IEEE, 2020, pp. 34–39.

[10] Y. Serdouk, V. Eglin, S. Bres, and M. Pardoen, "Keyword spotting using siamese triplet deep neural networks," in *2019 International Conference on Document Analysis and Recognition*. IEEE, 2019, pp. 1157–1162.

[11] M. Mhiri, C. Desrosiers, and M. Cheriet, "Word spotting and recognition via a joint deep embedding of image and text," *Pattern Recognition*, vol. 88, pp. 312–320, 2019.

[12] M. Kassis and J. El-Sana, "Word spotting using radial descriptor graph," in *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*. IEEE, 2016, pp. 31–35.

[13] K. Zagoris, I. Pratikakis, and B. Gatos, "Segmentation-based historical handwritten word spotting using document-specific local features," in *2014 14th International Conference on Frontiers in Handwriting Recognition*. IEEE, 2014, pp. 9–14.

[14] T. Konidaris, A. L. Kesidis, and B. Gatos, "A segmentation-free word spotting method for historical printed documents," *Pattern analysis and applications*, vol. 19, no. 4, pp. 963–976, 2016.

[15] X. Zhang, U. Pal, and C. L. Tan, "Segmentation-free keyword spotting for bangla handwritten documents," in *14th International Conference on Frontiers in Handwriting Recognition*. IEEE, 2014, pp. 381–386.

[16] M. Khayyat and C. Y. Suen, "Improving word spotting system performance using ensemble classifier combination methods," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018, pp. 229–234.

[17] F. Westphal, H. Grahn, and N. Lavesson, "Representative image selection for data efficient word spotting," in *International Workshop on Document Analysis Systems*. Springer, 2020, pp. 383–397.

[18] N. Gurjar, S. Sudholt, and G. A. Fink, "Learning deep representations for word spotting under weak supervision," in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. IEEE, 2018, pp. 7–12.

[19] S. Sudholt and G. A. Fink, "Attribute cnns for word spotting in handwritten documents," *International journal on document analysis and recognition (ijdar)*, vol. 21, no. 3, pp. 199–218, 2018.

[20] P. Krishnan and C. Jawahar, "Matching handwritten document images," in *European Conference on Computer Vision*, 2016, pp. 766–782.

[21] ——, "Hwnet v2: An efficient word image representation for handwritten documents," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 22, no. 4, pp. 387–405, 2019.

[22] J. Lladós, M. Rusinol, A. Fornés, D. Fernández, and A. Dutta, "On the influence of word representations for handwritten word spotting in historical documents," *International journal of pattern recognition and artificial intelligence*, vol. 26, no. 05, 2012.

[23] W. Pan, E. Xiang, N. Liu, and Q. Yang, "Transfer learning in collaborative filtering for sparsity reduction," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 24, 2010, pp. 230–235.

[24] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 244–253.

[25] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, 2020, pp. 213–229.

[26] E. Xie, W. Wang, W. Wang, P. Sun, H. Xu, D. Liang, and P. Luo, "Segmenting transparent object in the wild with transformer," *arXiv preprint arXiv:2101.08461*, 2021.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[29] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[30] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che *et al.*, "Layoutlmv2: Multi-modal pre-training for visually-rich document understanding," *arXiv:2012.14740*, 2020.

[31] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[32] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, "Docformer: End-to-end transformer for document understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 993–1003.

[33] J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, and F. Wei, "Dit: Self-supervised pre-training for document image transformer," *arXiv preprint arXiv:2203.02378*, 2022.

[34] A. C. Rouhou, M. Dhiaf, Y. Kessentini, and S. B. Salem, "Transformer-based approach for joint handwriting and named entity recognition in historical document," *Pattern Recognition Letters*, vol. 155, pp. 128–134, 2022.

[35] M. MHIRI, M. Hamdan, and M. Cheriet, "Handwriting word spotting in the space of difference between representations using vision transformers," *Available at SSRN 4113859*, 2022.

[36] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[37] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[38] M. Kassis, A. Abdalhaleem, A. Droby, R. Alaasam, and J. El-Sana, "Vml-hd: The historical arabic documents dataset for recognition systems," in *Arabic Script Analysis and Recognition (ASAR), 2017 1st International Workshop on*. IEEE, 2017, pp. 11–14.

[39] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 9, no. 2-4, pp. 139–152, 2007.

[40] I. Rabaev, B. K. Barakat, A. Churkin, and J. El-Sana, "The hhd dataset," in *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2020, pp. 228–233.

[41] J. Deng, A. C. Berg, and L. Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," in *CVPR 2011*. IEEE, 2011, pp. 785–792.

[42] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[43] W. Pantke, M. Dennhardt, D. Fecker, V. Märgner, and T. Fingscheidt, "An historical handwritten arabic dataset for segmentation-free word spotting-hadara80p," in *2014 14th International Conference on Frontiers in Handwriting Recognition*. IEEE, 2014, pp. 15–20.

[44] S. Yang, H. Qin, M. A. El-Yacoubi, and C. Liu, "Cross-modality domain adaptation for hand-vein recognition," in *2021 International Conference on Cyber-Physical Social Intelligence (ICCSI)*. IEEE, Dec. 2021. [Online]. Available: https://doi.org/10.1109/iccsi53130.2021.9736171