



INSTITUT
POLYTECHNIQUE
DE PARIS



TELECOM
SudParis



NNT : 2023IPPAS004

Thèse de doctorat

Contributions for indexing Historical Arabic Documents using Deep Learning approaches

Thèse de doctorat de l'Institut Polytechnique de Paris en cotutelle avec
l'Université de Sousse
préparée à Télécom SudParis

École doctorale n°626 Institut Polytechnique de Paris (IPP)
Spécialité de doctorat: Signal, Images, Automatique et robotique

Thèse présentée et soutenue à Paris, le Date, par

ABIR FATHALLAH

Composition du Jury :

Mehdi Ammi Professeur, Université de Paris 8, (LIASD)	Président
Laurence Likforman-Sulem Professeur, Télécom Paris, (LTCI)	Rapporteur
Mohamed Adel Alimi Professeur, École nationale d'ingénieurs de Sfax, (REGIM)	Rapporteur
Afef Kacem Maître de conférences, Ecole Supérieure des Sciences et Techniques de Tunis, (LaTICE)	Examineur
Mehdi Ammi Professeur, Université de Paris 8, (LIASD)	Examineur
Najoua Essoukri Ben Amara Professeur, École nationale d'ingénieurs de Sousse, (LATIS)	Directeur de thèse
Mounim A. El yacoubi Professeur, Télécom SudParis, (SAMOVAR)	Directeur de thèse

Résumé

Avec les énormes progrès technologiques de ces dernières années, la quantité de documents historiques numérisés, tant manuscrits qu'imprimés, a considérablement augmenté. Il est évident que les documents historiques numériques ne sont pas faciles à traiter dans leur forme originale, mais ils doivent être transformés en une forme lisible afin d'être compris automatiquement par les outils de vision par ordinateur. Le repérage de mots est une tâche importante pour comprendre et exploiter le contenu des documents en créant des index. Il s'agit d'une technique de recherche d'informations qui vise à identifier toutes les occurrences d'un mot de requête dans un ensemble de documents (par exemple, un livre). Dans la tâche de repérage de mots, l'entrée est un ensemble de documents non indexés et la sortie est une liste de mots classés en fonction de leur similarité avec le mot de requête. Cela permet un accès en ligne rapide et facile aux documents du patrimoine culturel et offre d'autres possibilités d'étudier ces ressources.

La présente thèse de doctorat porte sur le problème du repérage des mots dans les documents historiques. La première contribution de ce travail est le développement d'un espace de représentation d'images de mots basé sur la combinaison de réseaux convolutifs et de pertes de triplets. Ensuite, les distances de similarité sont appliquées pour établir une correspondance entre les mots de la requête et tous les mots présents dans les documents historiques. La deuxième contribution de cette thèse présente une méthode améliorée de construction d'un espace de représentation pour un modèle de repérage de mots grâce à l'adoption de plusieurs stratégies d'amélioration. Ces stratégies comprennent des étapes de prétraitement, l'apprentissage par transfert, l'extraction de triplets en ligne et des techniques de sélection de triplets semi-durs. La troisième contribution vise à améliorer les performances de repérage des mots en développant un modèle conditionnel génératif basé sur un réseau adversatif pour générer des images de documents propres à partir d'images fortement dégradées. Ce modèle d'amélioration traite de diverses tâches de dégradation telles que les filigranes et la dégradation chimique, dans le but de produire des images de documents hyper-propres et des performances de récupération de détails fins. Dans la dernière contribution, nous proposons l'utilisation d'une architecture de Vision Transformer pour la génération de représentations mot-image. L'approche utilise la perte de triplets comme critère d'optimisation et incorpore l'apprentissage par transfert de deux domaines distincts pour améliorer la performance de la représentation mot-image.

Toutes ces contributions sont évaluées sur de nombreuses bases de données publiques qui fournissent différents défis de documents historiques. Les résultats expérimentaux obtenus dans la tâche de repérage de mots pour les documents historiques se comparent favorablement à de nombreuses méthodes récentes de l'état de l'art.

Mots clés: Documents historiques, documents dégradés, indexation, repérage de mots, perte de triplets, amélioration de documents, correspondance, réseaux adversaires génératifs, espace de représentation, Vision Transformer, distances de similarité.

Abstract

With the enormous technological advances of recent years, the amount of digitized historical documents, both handwritten and printed, has increased. It is well known that digital historical documents are not easily processed in their original form, but they need to be transformed into a readable form in order to be automatically understood by computer vision tools. Word spotting is an important task to understand and exploit document contents by creating indexes. It is an information retrieval technique that aims to identify all occurrences of a query word in a set of documents (for example, a book). In the word spotting task, the input is a set of unindexed documents and the output is a ranked list of words according to their similarity to the query word. This allows quick and easy online access to cultural heritage materials and provides further opportunities to investigate these resources.

The present PhD thesis investigates the problem of word spotting in historical documents. The first contribution of this work is the development of embedding space for word image representation based on the combination of convolutional networks and triplet loss. Subsequently, similarity distances are employed to match query words with all words present in the historical documents. The second contribution of this thesis presents an improved method for constructing an embedding space for a word spotting model through the adoption of multiple enhancement strategies. These strategies include preprocessing steps, transfer learning, online triplet mining, and semi-hard triplet selection techniques. The third contribution aims to enhance word spotting performance by developing a conditional generative adversarial network-based model for generating clean document images from highly degraded images. This enhancement model addresses various degradation tasks such as watermarks and chemical degradation, with the goal of producing hyper-clean document images and fine detail recovery performance. In the final contribution, we propose the utilization of a vision transformer architecture for the generation of word-image representations. The approach utilizes triplet loss as the optimization criterion and incorporates transfer learning from two distinct domains to improve the performance of the word-image representation.

All these contributions are evaluated on many public databases that provide different challenges of historical documents. The obtained experimental results in the word spotting task for historical documents compare favorably with many recent state-of-the-art methods.

Keywords: Historical documents, Degraded documents, Indexing, Word spotting, Triplet

loss, Document enhancement, matching, Convolutional network, Generative adversarial networks, Embedding space, Vision transformer, Similarity distances.

List of Acronyms

- AP** Average Precision
- BCCs** Basic Connected Components
- BoC** Bag-of-Characters
- BoW** Bag of Visual Words
- CNNs** Convolutional Neural Networks
- CTC** Connectionist Temporal Classification
- DTW** Derivative Dynamic Time Warping
- DCToW** Discrete Cosine Transform of Words
- DNN** Deep Neural Network
- DRD** Distance Reciprocal Distortion
- DTW** Dynamic Time Warping
- EDM** Euclidean Distance Mapping
- FCN** Fully Convolutional Network
- FESs** Fuzzy Expert Systems
- FN** False Negatives
- GANs** Generative Adversarial Networks
- GHT** Generalized Hough Transform
- GPUs** Graphics Processing Units
- HADs** Historical Arabic Documents

HGT Hough Generalized Transform
HHD Hebrew Handwritten Dataset
HMM Hidden Markov Model
HOG Histograms Of Oriented Gradients
LDA Linear Discriminant Analysis
LGH Local Gradient Histogram
LSTM Long Short-Term Memory
mAP mean Average Precision
MLP Multi Layer Perceptron
NCC Normalized Cross Correlation
PBCS Pyramid of Bidirectional Character Sequences
PCA Principal Component Analysis
PHOC Pyramidal Histogram Of Character
PSNR Peak Signal-to-Noise Ratio
 F_{ps} pseudo-F-measure
QbE Query-by-Example
QbS Query-by-String
ResNets Residual Networks
ReLU Rectified Linear Units
RLSA Run Length Smoothing Algorithm
RNN Recurrent Neural Networks
SIFT Scale-invariant Feature Transform
SPOC Spatial Pyramid of Characters
SSD Sum-of-Squares-Distances
SVM Support Vector Machine
TP True Positives
TV Total Variation
ViT Vision Transformer

VML-HD Visual Media Lab Historical Documents dataset

Contents

Résumé	i
Abstract	iii
List of Acronyms	v
List of Figures	xi
List of Figures	xii
List of Tables	xiv
List of Tables	xv
1 General Introduction	1
1.1 Introduction	2
1.2 Historical documents	2
1.3 Word spotting in historical documents	3
1.4 Motivation and context	4
1.5 Thesis contributions	6
1.6 Thesis outline	7
1.7 Publications and current submissions	10
2 State of the art	11
2.1 Introduction	12
2.2 Historical documents	12
2.3 Arabic language characteristics	13
2.4 Word spotting	16
2.4.1 Holistic analysis techniques	18
2.4.2 Analytical analysis techniques	24
2.4.3 Word spotting system	28

2.4.4	Deep Learning in Word Spotting	31
2.5	Datasets	33
2.5.1	BH2M	33
2.5.2	IFN/ENIT	33
2.5.3	GRPOLY-DB	35
2.5.4	Hebrew Handwritten dataset	35
2.5.5	CFRAMUZ	35
2.5.6	HADARA80P	35
2.5.7	VML-HD	36
2.5.8	AMADI-LontarSet	36
2.5.9	George Washington	37
2.5.10	Hebrew Handwritten dataset	37
2.6	Evaluation protocols and performance measures	37
2.7	Conclusion	39
3	Word Spotting based Triplet-CNN in Historical Arabic Documents	40
3.1	Introduction	41
3.2	Deep learning for word spotting	41
3.3	Proposed approach for word spotting in historical Arabic document	44
3.3.1	Embedding space construction	46
3.3.2	Word spotting method	48
3.4	Experimental Results	49
3.4.1	Dataset	49
3.4.2	Evaluation Protocol	50
3.4.3	Experimental results	51
3.4.4	Further analysis	53
3.5	Conclusions	56
4	Enhancement Strategies for Word Spotting in Historical Documents	58
4.1	Introduction	59
4.2	Related Work	60
4.2.1	Word spotting	60
4.2.2	Transfer Learning	61
4.2.3	Triplet loss	62
4.3	Enhancement strategies for word spotting in historical documents	63
4.3.1	Word spotting process	64
4.4	Experiments	67
4.4.1	Experimental setup	67
4.4.2	Results	68
4.5	Qualitative Evaluation Analysis	70
4.6	Ablation Study	71
4.7	Conclusion	73

5	Enhancement of Historical Document Images via Generative Adversarial Networks	74
5.1	Introduction	75
5.2	Research related to enhancing degraded document images	75
5.2.1	Degraded document enhancement	76
5.2.2	Generative adversarial networks for image-to image transform	78
5.3	Proposed Method	80
5.3.1	Generator architecture	80
5.3.2	Discriminator architecture	81
5.3.3	Loss functions of proposed GAN	81
5.4	Experiments	83
5.4.1	Datasets	83
5.4.2	Experimental setup	83
5.4.3	Results	85
5.5	Conclusion	93
6	A Triplet Vision Transformers for Word Spotting in Historical Documents	94
6.1	Introduction	95
6.2	Related work	96
6.2.1	Vision transformer	96
6.3	Proposed approach	97
6.3.1	Pre-processing	98
6.3.2	Enhancement based Transfer Learning	98
6.3.3	Transformer architecture	99
6.3.4	Triplet loss	101
6.3.5	Embeddings matching	101
6.4	Experiments	102
6.4.1	Experimental setup	102
6.4.2	Results and discussions	103
6.4.3	Error Analysis	105
6.4.4	Ablation Study	107
6.5	Conclusion	108
7	Conclusion and perspectives	109
7.1	Conclusion	110
7.2	Perspectives	112
	Bibliography	114

List of Figures

1.1	Examples of HADs in their original form.	3
1.2	Architecture of typical camera-based document image retrieval system. . .	4
1.3	Report structure.	9
2.1	Example of degraded historical documents.	13
2.2	Example of some Arabic calligraphy writing styles.	14
2.3	Example of some Arabic calligraphy writing styles.	15
2.4	Different Arabic letters position in the word.	15
2.5	An overview of word spotting in historical documents.	17
2.6	An overview of the different holistic methods.	18
2.7	Synthetically generated query word [Konidaris et al., 2007].	19
2.8	(a) Corners detected with the Harris Corner detector on two gray level images. (b) Recovered correspondences in two word images [Rothfeder et al., 2003].	21
2.9	An overview of the different holistic methods.	22
2.10	An overview of the different analytical methods.	24
2.11	Use of a sliding window for extracting small fragments.	25
2.12	The process of extracting LGH features for a small overlapping windows of one word [Rodríguez-Serrano and Perronnin, 2009].	27
2.13	General word spotting system architecture.	28
2.14	An example of BH2M database at each level segmentation.	33
2.15	(a) Examples of GRPOLY-DB-MachinePrinted grayscale and color page images. (b) Examples of GRPOLY-DB-Handwritten at text line and word level [Gatos et al., 2015].	35
2.16	Depicting the distinct writing styles in the dataset: (a) and (b) introduce the word "petite" in a first style while (c) and (d) present the same word in a second style. (1) and (2) present two pages of different novels from the CFRAMUZ dataset.	36
2.17	(a) Sample images of palm leaf manuscript, (b) Samples of word annotated images [Kesiman et al., 2016].	37

3.1	Concept of embedding-space in learning-feature representation.	42
3.2	Principal steps the proposed approach.	46
3.3	Triplet-CNN architecture proposed in our framework.	47
3.4	Word spotting flowchart.	48
3.5	(a) Example of pages from each book.	49
3.6	(b) Example of segmented data.	49
3.7	Examples of images extracted from the VML-HD database.	49
3.8	Comparison of the performances of the Siamese and Triplet networks based on $P@K$ metric.	52
3.9	Comparison of the performances of the Siamese and Triplet networks based on mAP metric.	53
3.10	Comparison of performances of embedding methods based on $P@K$ metric.	55
3.11	Comparison of performances of embedding methods based on mAP metric.	56
4.1	Illustration of general steps presented in our proposed approach: (A) An enhancement method based on conditional GANs, (B) Document images pre-segmented using their annotation XML files, (C) Triplet-CNN for feature extraction and embedding space construction, (D) Similarity distances to match embeddings of query word and each reference word.	63
4.2	Proposed word spotting steps for historical documents.	64
4.3	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. On the top, the images without enhancement while on the bottom, the images with enhancement.	71
5.1	Examples of documents: (a): Degraded documents, (b): A document including a big stamp.	76
5.2	Proposed GAN architecture for document enhancement process.	80
5.3	Example of stacked patches of size 256×256 pixels during training phase of our proposed model.	85
5.4	Example of degraded documents enhancement by our EHDI and DE-GAN on sample PR08 from DIBCO-2013 [Souibgui and Kessentini, 2020].	86
5.5	Example of enhancing degraded documents by our EHDI.	88
5.6	Example of enhancing degraded documents by our proposed model.	89
5.7	Qualitative binarization results on sample 16 in DIBCO 2017 dataset. Here, we compare the results of our proposed model with the winner’s approach [Pratikakis et al., 2017] and DE-GAN [Souibgui and Kessentini, 2020].	90
5.8	Example of qualitative results on HADARA dataset compared to DE-GAN enhancement approach [Souibgui and Kessentini, 2020].	91
5.9	Example of qualitative enhancement results produced by different models of the sample (9) from the H-DIBCO 2018 dataset.	92
6.1	Proposed approach based on triplet transformer: TL is applied from historical English (D_{S1}) and handwritten Hebrew documents (D_{S2}) to Arabic documents (D_T).	98
6.2	Vision Transformer architecture for word image representations.	99
6.3	Flowchart of the proposed TripTran architecture.	102

6.4	Results on VML-HD and GW datasets in terms of mAP metric using Euclidean distance: Comparison with state-of-the-art methods.	106
6.5	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.	107

List of Tables

1.1	Thesis under an agreement for a joint supervised doctorate agreement	6
2.1	Summary of related works to word spotting.	34
3.1	First book dataset partition for model training.	50
3.2	Results on VML-HD dataset according to $P@K$	52
3.3	Results on VML-HD dataset according to mAP	52
3.4	Results according to $P@K$	55
3.5	Results according to mAP	55
4.1	Partitioning of datasets according to the level of the word classes and number of images per class.	67
4.2	Results of our proposed model on the VML-HD dataset according to P@K and mAP metrics.	69
4.3	Results of our proposed model on different datasets according to P@K and mAP metrics.	69
4.4	Results on VML-HD according to P@K and mAP metrics using Euclidean distance: Comparison with state-of-the-art methods.	70
4.5	Results on VML-HD according mAP metric using Euclidean distance. . . .	72
5.1	Results of our proposed EHDI on DIBCO 2013 dataset.	85
5.2	A comparative review of competitor approaches of DIBCO 2018 on DIBCO 2017 and DIBCO 2018 Datasets.	87
5.3	Results of our proposed model on DIBCO 2018 dataset.	90
6.1	Partitioning of datasets according to the level of word classes and the number of images per class.	103
6.2	Results of our proposed TripTran model on VML-HD dataset according to P@K and mAP metrics.	103
6.3	Word spotting results in terms of P@k metric for some query word images used in our experiment.	104

6.4	Results of our proposed model on different datasets according to P@K and mAP metrics.	104
6.5	Results on VML-HD according to P@K and mAP metrics using Euclidean distance: Comparison with state-of-the-art methods.	105
6.6	Results on VML-HD according to mAP metric using Euclidean distance.	107

CHAPTER 1

General Introduction

1.1	Introduction	2
1.2	Historical documents	2
1.3	Word spotting in historical documents	3
1.4	Motivation and context	4
1.5	Thesis contributions	6
1.6	Thesis outline	7
1.7	Publications and current submissions	10

1.1 Introduction

The present chapter serves as an introduction to the motivation and objectives of this Ph.D. thesis. The research field of document analysis and indexing, with a specific focus on the analysis of historical documents, is briefly introduced. This is followed by an overview of the word retrieval architecture, and its relevance to historical documents is discussed.

The general context of the Ph.D. thesis is then presented, highlighting the main difficulties and challenges encountered in the field. Finally, the specific objectives and contributions of this work are outlined. This includes the identification of research gaps in the field and the proposed solutions to address them, with the goal of advancing the state-of-the-art in document analysis and indexing, particularly for historical documents.

1.2 Historical documents

The paper document has been the most effective means of communication throughout history, and as such, represents a valuable cultural heritage. It provides insight into both tangible and intangible cultural aspects. Historical archives, in particular, often contain handwritten documents that can include manuscripts written by notable scientists, writers, or artists, as well as letters, official forms, and administrative records that can aid in reconstructing historical sequences in a particular place or time.

The advancement of electronic storage and digitization technology has enabled the digitization of historical documents for cultural heritage preservation and analysis. This allows for important knowledge from historical document collections in museums, archives, and libraries to be easily accessed by the general public, while also preserving the documents from further degradation.

Given the advantages that computer systems offer in terms of storage capacity, retrieval, transmission, and automatic processing of documents, it is necessary to establish a document management system that is specifically tailored to Historical Arabic Documents (HADs). This is particularly challenging due to the complex nature of Arabic writing and the poor quality of many ancient historical documents. Developing an indexing and search mechanism that is based on the digital availability of these documents is essential to effectively process and utilize HADs.

This thesis presents a comprehensive examination of Arabic historical documents, an area of study that has garnered significant attention in recent years. According to statistics reported in [Saabni and El-Sana, 2013], more than 90 million documents written in Arabic script have been produced between the 7th century and the 14th century. Out of these, an estimated seven million documents, spanning a diverse range of disciplines, have been preserved over time. These historical documents are visualized in Figure 1.1, which provides some examples of Arabic historical document images.



Figure 1.1: Examples of HADs in their original form.

1.3 Word spotting in historical documents

The automatic processing of HADs is a crucial endeavor aimed at making the vast content of these documents easily accessible to the public community.

In recent years, word spotting has emerged as a popular trend in the image processing of HADs, attracting the attention of numerous researchers in the field. However, the task of locating the occurrences of a particular word in a large collection of historical document images is a challenging undertaking, due to the complexities inherent in the Arabic script. A word spotting system takes as input a collection of document images and a query and produces as output a list of document image regions, typically ranked based on query similarity.

As depicted in Figure 1.2, the overall process of word spotting utilizes two distinct query modalities: query-by-example and query-by-string. In the query-by-example scenario, the user is required to provide an exemplar image of the desired query word, necessitating the identification and selection of a specific instance of the query. In contrast, the query-by-string scenario allows the user to input a textual query via keyboard, with the system subsequently highlighting and prioritizing relevant regions within the document image in the ranked retrieval list.

The fundamental structure of a word spotting system closely resembles that of an information retrieval system. The system is initially provided with a database of document images, which are made searchable through the use of image region representations in the form of features. These features, which are numerical representations that encapsulate information pertinent to word image retrieval, serve as the basis for retrieval regardless of the query modality utilized. In the query-by-example scenario, it is sufficient to model the visual appearance of the query word at the word level. This approach uses the exemplar image provided by the user as a reference for matching against regions within the document image database. On the other hand, for the query-by-string scenario, a more detailed approach is necessary. In this case, modeling the visual appearances of individual characters within the query string allows for a more comprehensive comparison with regions within the document image. To generate the ranked retrieval list, scores representing the similarity between the query and the document image regions are calculated and utilized. These scores are typically based on a comparison of the visual features extracted from the query and the document im-

age regions.

In order to effectively identify relevant regions within a document image, it is often necessary to first identify and segment the text within the image into individual words. This is typically a straightforward task for printed documents that utilize standardized fonts, as the gaps between words are often larger than those between characters within words. However, this assumption may not hold true for handwritten or historical documents, as the visual variability of the text can be substantial even when produced by a single writer. In these cases, the use of a segmentation-based retrieval pipeline can result in errors, leading to inaccurate retrieval results.

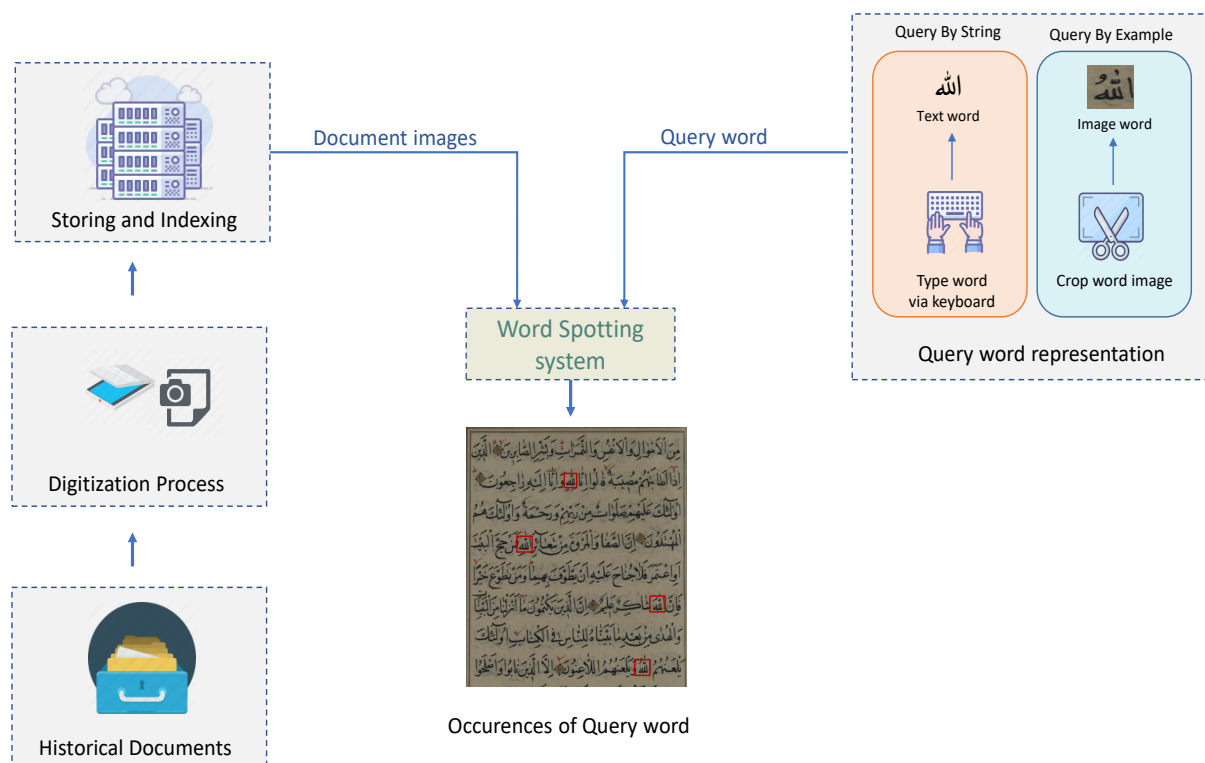


Figure 1.2: Architecture of typical camera-based document image retrieval system.

1.4 Motivation and context

The study of history is a continuous process that requires constant awareness of past events in order to gain insight into potential future outcomes. Digital archives, through their ability to preserve historical documents and make them accessible to a broad audience, have become crucial tools in this endeavor. However, as the number of digital document images continues to grow at a rapid pace, the need for automated processing methods to extract and understand their contents has become increasingly pressing. One of the most challenging and significant areas of research in document analysis is word spotting, which aims to support the exploration of historical document images through the use of automatic information extraction techniques. There are various methods that have been developed for this

purpose, many of which focus specifically on text-based documents. Obtaining a full-text transcription of a document image, complete with annotations indicating the precise locations of letters and words, is a highly desirable outcome in document analysis. However, this is often a challenging task due to the difficulty of obtaining annotated sample data. These data are typically created manually, making the process time-consuming and labor-intensive. Furthermore, historical document collections are unique in their visual characteristics, meaning that annotated samples from one collection cannot be used to develop recognition tools for other historical collections. This necessitates repeated manual annotation efforts. While manual transcription is often required as a prerequisite to the application of automatic processing methods, it may be more efficient in certain cases to transcribe the entire document collection manually, in order to minimize the time and effort required for annotation.

Word spotting offers a compromise in terms of the need for annotated sample data, as it allows for a wide range of annotation requirements. This includes systems that work with synthesized annotated samples and do not require manual labeling, systems that only require a single annotated sample of the target word, and models that are estimated using large volumes of annotated word images. By basing word retrieval on a single annotated sample of the target word, it is possible to retrieve document images with a visually similar appearance. However, if a large set of annotated samples is available, it can lead to improved retrieval performance. It is crucial, however, that these annotated samples remain textually representative of the document images. Synthesizing sample data that fulfill these requirements is a significant challenge.

Word spotting is a practical solution in the development of a comprehensive transcription recognition system. Its flexibility in terms of annotation requirements allows for its application to document images even in cases where no annotation is initially provided. The collected search results can then be used to generate a set of annotated samples, providing a solid foundation for the estimation of a complete search system. The ability to generate annotated samples through the use of a word spotting system allows for the development of a transcription recognition system with minimal manual annotation effort. In this way, word spotting can be considered a valuable approach for the analysis of historical document images and for the development of automated information extraction methods.

In this thesis, we investigate the utilization of word spotting as a means of automating the analysis of HADs. Specifically, we examine the potential of word spotting techniques to support the information extraction process and explore various methods for improving the performance of word spotting in the context of HADs. Through a detailed examination of the challenges and limitations of current approaches, we aim to identify opportunities for further research and development in this area. Ultimately, our goal is to demonstrate the potential of word spotting as a powerful tool for unlocking the knowledge contained within HADs and making it more widely accessible.

This thesis is undertaken as a joint supervised doctorate agreement between the University of Sousse in Tunisia and the Institut Polytechnique of Paris. The research is conducted within the framework of the SID team (Signal, Image and Documents) of the LATIS (Laboratory of Advanced Technology and Intelligent Systems) laboratory, in collaboration with the National Archives of Tunisia and the SAMOVAR laboratory (Distributed Services, Architectures, Modeling, Validation, and Network Administration). The project is aligned with the general research activities of these institutions and is aimed at advancing the field of automatic processing of historical Arabic documents through the application of word spotting

techniques. The details of the project are provided in Table 1.1.

Table 1.1: Thesis under an agreement for a joint supervised doctorate agreement

	France	Tunisia
Date of first registration	March 2019	November 2018
Diploma specialty	PhD in Signal, Image, Automation and Robotics	PhD in computer sciences
Thesis supervisor	Pr. Mounim A. El Yacoubi	Pr. Najoua Essoukri Ben Amara
Research laboratory	SAMOVAR	LATIS
Establishment structure	Télécom SudParis	ISITCom Hamem Sousse
Doctoral school	Institut Polytechnique of Paris	University of Sousse

1.5 Thesis contributions

The main contributions of this dissertation are summarized in the following.

- In this research, we have proposed a word spotting system specifically tailored for HADs. A key focus of the study is the examination of the impact of feature representation learning on the performance of word spotting in historical documents. To that end, we devised a novel word spotting method, which utilizes a *triplet-loss* based representation learning approach, effectively addressing the threshold tuning challenge inherent in Siamese networks. The proposed method demonstrates a significant improvement in performance for word spotting in historical Arabic documents.
- Historical documents often exhibit poor visual quality due to various degradation issues, which negatively impact feature representation performance. To address this challenge, we have proposed a novel approach utilizing a conditional generative adversarial architecture for document enhancement. This approach aims to generate a high-quality, visually clear image document from a degraded input and will be employed as a pre-processing step in a subsequent framework we propose.
- The triplet-CNN approach has been found to be less efficient in offline learning scenarios, as a complete iteration over the training set is required to generate triplets, and the triplets must be regularly updated. To address these limitations, we have investigated an online learning approach for feature extraction from word images using a triplet-CNN. Our proposed method employs an online learning procedure and a semi-hard triplet selection strategy to enhance the performance and efficiency of the model.
- While a plethora of public datasets exists for document analysis, there is a scarcity of annotated public historical documents for evaluating word spotting approaches. To mitigate the issue of limited annotated data and improve the representation of word images, we have employed a transfer learning technique for word spotting model generalization. This technique involves utilizing knowledge acquired from similar and dissimilar source domains to improve the performance of the model.

- The recent proliferation of vision transformer approaches in various applications has led to increased utilization of their encoding-decoding architectures for data representation. In light of this trend, we have presented a new word spotting technique for historical document images based on a triplet transformer. Specifically, our goal is to construct an embedding space for word image representation through the implementation of triplet transformer architectures.

In this thesis, thorough experimental evaluations were conducted on various datasets to assess the efficacy of the proposed approaches. These evaluations revealed that the suggested approaches exhibit both high levels of accuracy and robustness, providing strong evidence for their effectiveness.

1.6 Thesis outline

The chapters in this dissertation are meticulously organized in a manner that seamlessly flows and enhances the overall understanding of the research.

- In **Chapter 2**, we first provide a comprehensive review of the current state of research on various challenges associated with the Arabic script. This is followed by an overview of word spotting approaches, which are classified into two main categories: segmentation-free and segmentation-based approaches. We then present an overview of public datasets designed for word spotting, as well as a discussion of the commonly used metrics for evaluating word spotting systems. This information will serve as the foundation for our subsequent research.
- In **Chapter 3**, we present our first contribution which is an approach that utilizes a triplet-CNN to address the challenges of word spotting in historical document images. This approach is composed of two distinct stages: the construction of an embedding space using the triplet-loss and word spotting based on the embedding features. The primary objective of the first stage is to develop a feature representation space through the use of the triplet-loss. The second stage involves applying similarity distances to match embedding features for the purpose of word spotting.
- In **Chapter 4**, we introduce an updated approach for word spotting in historical document images that is based on a triplet-CNN. We investigate the impact of document enhancement, triplet mining, and transfer learning on the performance of the word spotting process in historical documents. Specifically, we utilize a specific triplet mining strategy that is based on the triplet loss, and we explore the effect of transfer learning from similar domains on the performance of the learning model. Our approach aims to improve the representation of word images and enhance the overall performance of the word spotting process in historical documents.
- In **Chapter 5**, we present a novel approach for enhancing the visual quality of degraded historical documents. Our method utilizes a generative adversarial architecture to generate a clean version of the degraded image. This task is treated as an image-to-image conversion process, where the goal is to learn a mapping from a degraded document image to a clean version of the image. By training this model, we aim to

improve the quality of historical documents to serve as a pre-processing step in our proposed word spotting framework.

- In **Chapter 6**, we present a novel end-to-end triplet transformer approach for the task of word retrieval in historical documents. Our approach consists of three main stages: pre-processing, feature representation, and similarity measurement. In the pre-processing phase, we utilize generative models to enhance the visual quality of the historical documents, thus improving the overall performance of the word retrieval task. The feature representation phase employs a transformer-based triplet loss to construct an embedding space, in which each feature is represented as an embedding vector. Finally, to achieve a high degree of matching accuracy, similarity distances are used to measure the similarity between embedding vectors, thus allowing for efficient retrieval of target words.
- In **Chapter 7**, we provide a comprehensive summary of the main contributions and conclusions of our study. In the first section, we succinctly summarize the key findings and innovations presented in the preceding chapters. We then proceed to discuss potential avenues for future research, highlighting opportunities for further exploration and elaboration of the ideas presented in this work.

The illustration presented in Figure 1.3 serves to demonstrate the positioning of our research endeavors within the broader context of the word spotting process in historical documents. It provides a visual representation of how our contributions have been integrated into the overall methodology and highlights their significance in advancing the field. This representation serves to provide a clear and concise overview of the research and its results and helps to contextualize and enhance the understanding of the findings presented in the report.

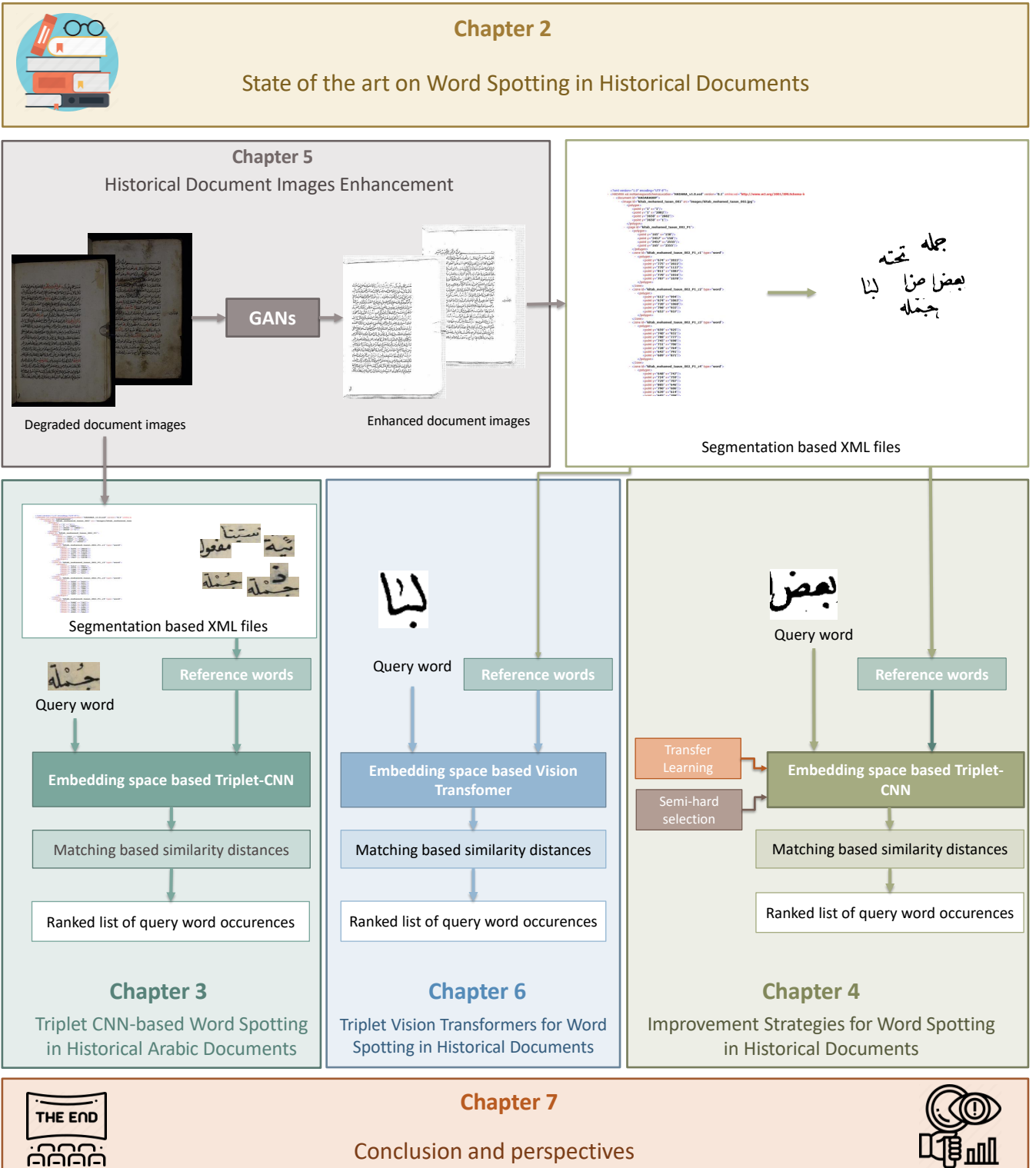


Figure 1.3: Report structure.

1.7 Publications and current submissions

- **Conference papers:**

Abir Fathallah, Mohamed Ibn Khedher, Mounim A. El Yacoubi, and Najoua Essoukri Ben Amara, “Triplet CNN-based Word Spotting of Historical Arabic Documents”, In the 26th International Conference on Neural Information Processing of the Asia-Pacific Neural Network Society (ICONIP). Sydney, Australia, 2019. [**Rank A**]

Abir Fathallah, Mohamed Ibn Khedher, Mounim A. El Yacoubi, and Najoua Essoukri Ben Amara, “Evaluation of Feature-Embedding Methods for Word Spotting in Historical Arabic Documents “, In the 17th IEEE International Multi-Conference on Systems, Signals & Devices (SSD). Monastir, Tunisia, 2020. [**IEEE Scopus**]

Abir Fathallah, Mounim A. El Yacoubi, and Najoua Essoukri Ben Amara, “EHDI: Enhancement of Historical Document Images via Generative Adversarial Network “, In the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP). Lisbon, Portugal, 2023. [**Rank B**]

Abir Fathallah, Mounim A. El Yacoubi, and Najoua Essoukri Ben Amara, “Transfer Learning for word spotting in Historical Arabic Documents based Triplet-CNN “, In the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP). Lisbon, Portugal, 2023. [**Rank B**]

Abir Fathallah, Mounim A. El Yacoubi, and Najoua Essoukri Ben Amara, “A Transformer-based Siamese Network For Word Image Retrieval In Historical Documents “, In the 17th International Conference on Document Analysis and Recognition (ICDAR). San José, California, USA, 2023. [**submitted**]

- **Journal papers:**

Abir Fathallah, Mounim A. El Yacoubi, and Najoua Essoukri Ben Amara, “Enhancement Strategies for Word Spotting in Historical Arabic Documents “, In International Journal on Document Analysis and Recognition (IJ DAR). [**Quartile: Q1, IF= 3.870**]. (under review)

Abir Fathallah, Mounim A. El Yacoubi, and Najoua Essoukri Ben Amara, “TripTran: A Triplet Vision Transformer for Word Spotting in Historical Arabic Documents “, In Pattern Analysis and Applications journal (PAAA). [**Quartile: Q2, IF= 2.307**]. (under review)

CHAPTER 2

State of the art

2.1	Introduction	12
2.2	Historical documents	12
2.3	Arabic language characteristics	13
2.4	Word spotting	16
2.4.1	Holistic analysis techniques	18
2.4.2	Analytical analysis techniques	24
2.4.3	Word spotting system	28
2.4.4	Deep Learning in Word Spotting	31
2.5	Datasets	33
2.5.1	BH2M	33
2.5.2	IFN/ENIT	33
2.5.3	GRPOLY-DB	35
2.5.4	Hebrew Handwritten dataset	35
2.5.5	CFRAMUZ	35
2.5.6	HADARA80P	35
2.5.7	VML-HD	36
2.5.8	AMADI-LontarSet	36
2.5.9	George Washington	37
2.5.10	Hebrew Handwritten dataset	37
2.6	Evaluation protocols and performance measures	37
2.7	Conclusion	39

2.1 Introduction

This chapter presents the state of the art of indexing Historical Arabic Documents (HADs). We first present the importance of historical documents and their several challenges. Then, the same presentation is made for the HADs. Following this, different automatic processing systems for historical documents are discussed, in particular word retrieval systems. The next section is devoted to a set of public datasets designed for the word retrieval process. Finally, some metrics for evaluating the word retrieval process are highlighted.

2.2 Historical documents

Most original historical documents are very valuable and irreplaceable. The exchange of these original treasures becomes impossible when they are preserved and protected by libraries, archivists and custodians. Any valuable historical document, especially any text, can deteriorate and be altered.

For priceless historical documents, whether they are corporate charters, national interests, or simply local traditions, archivists are carefully selected, prepared, and instructed in how to handle this priceless material. Digitizing historical documents offers a novel alternative to preserve valuable items from the past. Scanning such documents, allows them to be accessed from anywhere. Digitizing historical documents is also a wonderful way to track more recent texts, and preview trends in publishing or business. Through digitized documents, information previously inaccessible to common people, such as historical original manuscripts, hand-written notebooks and original rosters, becomes easily accessible. It is all being preserved for any personal access, and it is also easily accessible from your home.

A large number of historical documents is available in a handwritten form. The analysis of handwriting, as a particular computer vision task, poses many special challenges along with the standard issues. Among the greatest challenges in document image analysis is handling the high variability of manuscript writing with a robust representation model. The handwriting of a person has been observed to be influenced by many factors, including cultural context, education, lifestyle, physical conditions, etc. In some ways, this reflects the personality of each person and carries the individual characteristics.

Even if the same person writes the text consistently, it may be different under certain circumstances.

A considerable amount of work already has been and is still being done in providing automatic information retrieval services for historical documents. The processing of historical documents has been a very difficult area for many years and has become a very active research topic. It is still at the level of investigation and experimentation.

Important documentary collections currently exist in libraries, museums and other educational institutions. Historical records of ancient civilizations and national archives are typical rich examples that represent the heritage, history and dignity of nations. Indeed, old documents are weakened by time and, when they do not deteriorate naturally, suffer from being too often manipulated for consultation.

Librarians and owners of historical manuscript collections are very interested in everything related to their treasured documents. Despite their efforts and commitment, storage conditions are not always adapted to the requirements of these documents: Some old ink recipes

require quite specific storage conditions; temperature and humidity conditions should be well calculated in order to slow down the corrosion process; a document may be destroyed by its own ink if these conditions are not insured; specific storage conditions are required for certain documents that have been already very vulnerable. These conditions cannot always be ensured for diverse reasons related to the historical, social and political context of any region. Figure 2.1 presents some examples from degraded historical documents.

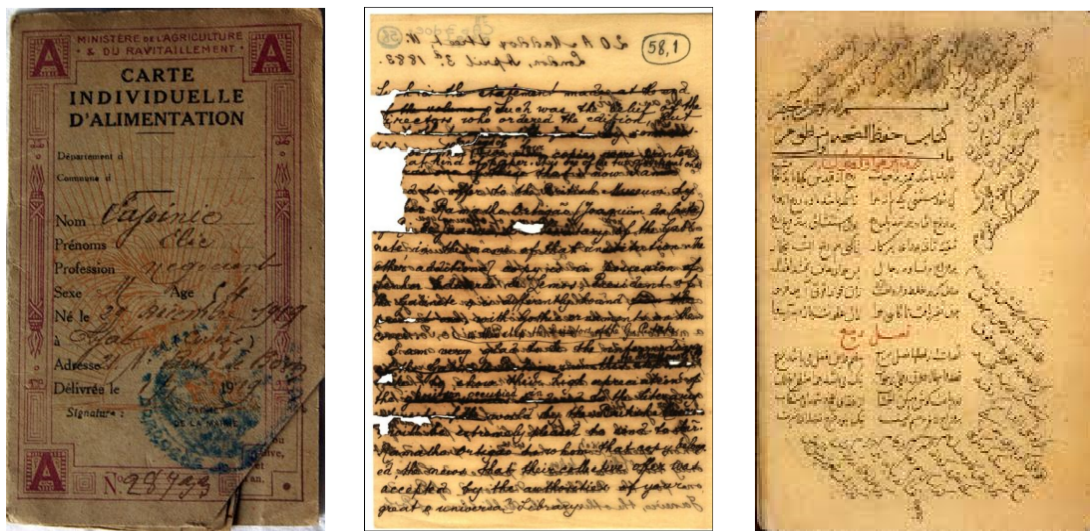


Figure 2.1: Example of degraded historical documents.

In order to overcome this problem, digitization has been taken as an effective solution. Digitization makes it possible to preserve the original documents, to share them with a large number of people, but it does not facilitate their access to consultation and research.

In fact, access to these collections requires effective indexing and search strategies. In most cases, indexes are created manually. If this approach is possible for a small number of documents, its cost and effort become very high for large collections. Automatic approaches are therefore desirable.

2.3 Arabic language characteristics

Arabic writing originally dates back to the first alphabets produced by the Phoenicians. The Phoenicians were located along the coastal areas of Syria, Lebanon and Palestine. As traders, the Phoenicians traveled the Mediterranean Sea and therefore their letters and scripts influenced all the nations of the Mediterranean as depicted in Figure 2.2.

The position of the Middle East as the center of the entire ancient world (east and west) was also a factor in the spread of the alphabet. For this reason, the Phoenician alphabet is considered the birthplace of both Arabic and Latin scripts. In 1300 BC, the primitive Phoenician alphabet (composed of 22 uncapitalized consonants that are written from right to left) originated in the coastal city of Byblos in Lebanon.

Around 1000 BC, a new Aramaic alphabet emerged from the Phoenician of Aram (Syria and Mesopotamia), which constituted the language of the Aramaeans. The Nabatean script originated in the city of Petra, on the northern shore of the Red Sea (present-day Jordan) in

Modern Latin	A	B	C	D	E	F	Z	H		I	K	L	M	N		O	P		Q	R	S	T
Early Latin	A	B	<	>	E	F	Z	H		z	K	L	M	N		O	P		Q	R	S	T
Early Greek	Δ	Δ	Ϛ	Δ	Ξ	Α	Z	Β		z	κ	λ	μ	ν		ο	π		ϙ	ρ	σ	τ
Phoenician	𐤀	𐤁	𐤂	𐤃	𐤄	𐤅	𐤆	𐤇	𐤈	𐤉	𐤊	𐤋	𐤌	𐤍	𐤎	𐤏	𐤐	𐤑	𐤒	𐤓	𐤔	𐤕
Early Aramaic	𐤀	𐤁	𐤂	𐤃	𐤄	𐤅	𐤆	𐤇	𐤈	𐤉	𐤊	𐤋	𐤌	𐤍	𐤎	𐤏	𐤐	𐤑	𐤒	𐤓	𐤔	𐤕
Nabataean	𐤀	𐤁	𐤂	𐤃	𐤄	𐤅	𐤆	𐤇	𐤈	𐤉	𐤊	𐤋	𐤌	𐤍	𐤎	𐤏	𐤐	𐤑	𐤒	𐤓	𐤔	𐤕
Early Arabic	ل	ح	ع	ب	ا	ج	د	س	ك	م	ن	و	هـ	ز	ر	ط	ي	ف	ق	ص	ث	ج

Figure 2.2: Example of some Arabic calligraphy writing styles.

100 BC and expanded throughout the Middle East. In the year 100 AD, a newly developed Syriac alphabet (22 letters) appeared in Mesopotamia, which also evolved from Aramaic. The early Arabic alphabet was developed in Kufa (Iraq) during the middle of the first century. The ancient Kufi (archaic Kufi) included about 17 letter forms in which no diacritical points or accents were used. Later, diacritical points and accents were added to assist in pronunciation. The ancient Kufi (Archaic Kufi) consisted of about 17 letter forms without diacritical points or accents. Subsequently, the addition of diacritical points and accents assisted in pronunciation. and the number of Arabic letter forms increased to 29 (including Hamza). Following the birth of Islam, the reform of all Arabic scripts present in the Arab world was dictated by the Quran. In the 7th century, a single well-structured and unified Arabic script with 29 letters was designed for writing the holy texts of the Quran. The Quran was mainly written in the Kufi style, and later in the Naskh style. From its appearance in the Arab world, the Islamic conquests spread the Arabic alphabet throughout the Middle East, North Africa, and even Spain.

Since Arabic was the language of the Qur'an and therefore of God, all the countries involved were forced to use the Arabic language. Several Arabic cities and regions have developed Arabic calligraphic styles using various writing techniques and tools. Several Arab cities and regions have developed Arabic calligraphic styles using various writing techniques and tools. The most popular Arabic calligraphic styles include: Kufi (Ancient Kufi and Geometric Ornamented Kufi), Thuluth, Diwani and Diwani Djeli, Naskh, Persian, Ruqaa and Maghrebi. All the listed calligraphic styles are illustrated in Figure 2.3. The name Kufi is derived from the city of "Kufa" in Iraq. "naskh" action of scribes when they copied an Arabic text, and the name "Diwan" refers to the political documents called "Diwan" in Arabic. The name "Thuluth" originates from the names of many bamboo sticks used as writing tools. The name "Ruqaa" is inspired by the leather paper "Ruqaa" on which it was written. The Persian style takes its name from the Persian language. Nowadays, the main available typefaces include the "Naskh" or "Thuluth" style.

Like handwriting of different languages, the Arabic language has several characteristics that are:

The Arabic language is composed of 29 consonants and 11 vocalization marks that take the form of accents. However, the basic structure of the alphabet can be figured out in only

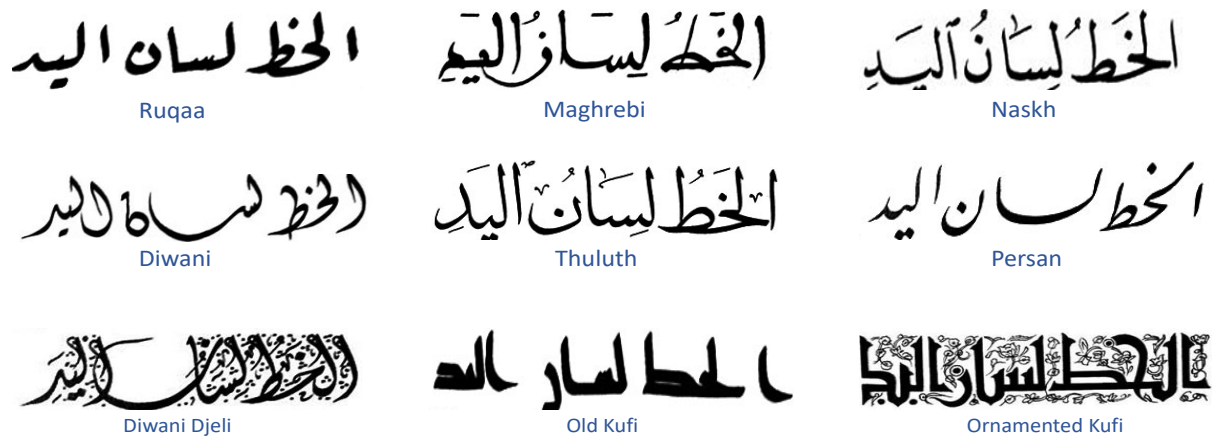


Figure 2.3: Example of some Arabic calligraphy writing styles.

19 main forms. As the letters have different shapes depending on their position in the word (beginning, middle, ending or isolated), the total number of glyphs reaches 106 (given that 23 letters are shaped in 4 different ways and 7 letters are shaped in 2 different ways). Adding the 2 essential ligatures of Lam Alef makes the total number 108.

- The Arabic language is composed of 29 consonants and 11 vocalization marks that take the form of accents. However, the basic structure of the alphabet can be figured out in only 19 main forms. As the letters have different shapes depending on their position in the word as shown in Figure 2.4 (beginning, middle, ending, or isolated), the total number of glyphs reaches 106 (given that 23 letters are shaped in 4 different ways and 7 letters are shaped in 2 different ways). Adding the 2 essential ligatures of Lam Alef makes the total number 108. Furthermore, as the Arabic alphabet is also employed in some other non-Arabic languages, further modifications of the letter have been undertaken to represent all additional non-Arabic phonetics. this brings the number of glyphs to 130.



Figure 2.4: Different Arabic letters position in the word.

- Contrary to Western scriptures which are written from left to right, Arabic is written right-to-left.
- There is no difference between handwritten letters and printed letters. The notions of capital letters and lowercase letters do not exist, so the writing is unicameral.
- Most of the letters are attached to each other, even in print, which gives the Arabic script the characteristic of cursivity.
- An Arabic character can contain a vertical line (TAA), an oblique line (KAF), or a zigzag (ALIF).
- Arabic characters have different sizes (height and width). As mentioned before, six letters do not attach themselves, therefore, a single word may be interspersed with one or more spaces giving several pseudo-words or related components or also sub-words.
- The Arabic characters are mostly consonants, there are only three characters in the Arabic language that represent the vowels. Much of the vocalization is produced by the diacritic.
- Most characters are composed of curves and loops.
- In terms of information density in pixels, The central band is usually the most charged. It corresponds to the location of horizontal ligatures, centered characters and loops, it is called the baseline.
- In the Arabic alphabet, 15 of the 28 letters have one or more points. These diacritical points are located either above or below the form with which they are associated, but never both at once. The maximum number of points a letter can have is three points above the character or two points below.

2.4 Word spotting

In the interest of creating digital libraries for historical documents, word spotting methods are designed to enable the retrieval of all the instances of a query by applying matching techniques in a set of document images. Accordingly, word spotting simplifies indexing and information retrieval in both modern and historical digitized documents that are relatively complex and degraded. In this section, a detailed review of previous research on word retrieval techniques in handwritten and typewritten images of historical documents is provided.

Usually, information is retrieved from documents based on text analysis methods. The ability and performance of optical character recognition (OCR) tools are still inadequate, especially for degraded historical documents. OCR fails to solve the problem completely because of its limitations in dealing with handwriting and degraded ancient collections. On the other hand, OCR techniques cannot be precisely performed given that most character recognition systems are unsuitable for open vocabulary handwritten scripts. For this reason, the idea of word retrieval is presented as an alternative to traditional OCR in various applications to index and retrieve information from digitized document collections.

In recent years, many approaches to word retrieval have been adopted for different scripts,

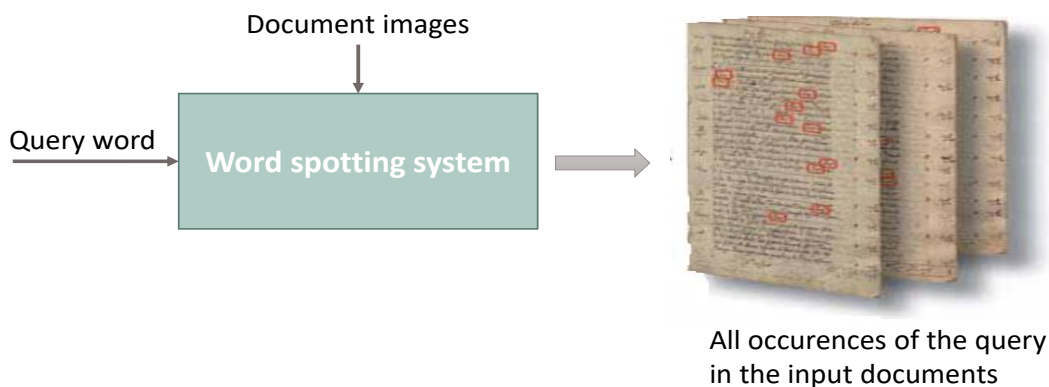


Figure 2.5: An overview of word spotting in historical documents.

such as Latin, Arabic, Greek, etc. These scripts are different and are characterized by different alphabet sizes, writing direction, cursiveness, and some similarities between characters. They appear in the form of handwritten scripts or typewritten scripts. In addition, document analysis researchers have classified word retrieval approaches into several categories. In particular, they can be divided into two main categories according to the matching techniques, e.g. image-based matching techniques and feature-based matching techniques [Rothfeder et al., 2003]. In the first category, there are methods of calculating distances between words directly on the image pixels, such as pattern correlation, while in the second category, various features in word images are compared and then matched. As depicted in Figure 2.5, typically, a word spotting system for historical documents takes the query word and document images as input, and produces all occurrences of the query word in the input documents. There is another classification presented in [Lladós et al., 2012]. There are two main approaches to retrieving words taking into account the query representation. The approaches can be either query-by-string (QbS) or query-by-example (QbE) based. The input for QbS methods [Cao and Govindaraju, 2007] consists of character sequences. Typically, these methods involve a lot of training material as the characters have to be learned a priori, and the model of a query is built at runtime from the models of its constituent characters. For QbE methods [Manmatha et al., 1996], one or more example images of the query word represent the input. Thus, it is necessary to collect one or more examples of the query word. Another common categorization procedure splits the methods into segmentation-based and segmentation-free methods as in [Adamek et al., 2007] [Gatos and Pratikakis, 2009]. Therefore, the main approaches for word retrieval can be classified into two major categories: Holistic analysis techniques and analytical recognition techniques. Furthermore, each of the main categories is divided into two groups: QbE-based techniques and QbS-based techniques. Holistic word-based techniques can be described as segmentation-free techniques. Basically, each word image is considered a unit that is unsegmented. On the other hand, analytical techniques consist of segmentation-based techniques where a word or document image is split into further smaller units that can be identified separately or as a group. In the literature, it is noticed that most of the techniques fall into the holistic analysis class as they have been applied to historical documents presenting difficulties to segment characters in a precise way.

2.4.1 Holistic analysis techniques

Holistic word search techniques treat word images as units. These techniques depend mainly on a word segmentation task. Indeed, the quality of the segmented words in the document images is decisive for the final spotting performance. Figure 2.6 illustrates the different holistic techniques.

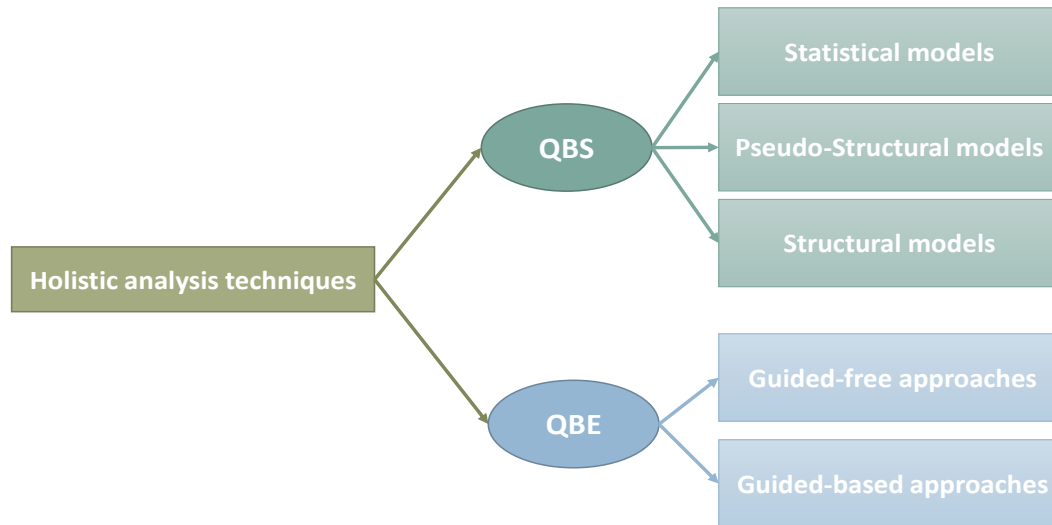


Figure 2.6: An overview of the different holistic methods.

In this section, we split the holistic word-based techniques into two sub-categories according to the type of query representation. Two different approaches to word spotting can be distinguished: QbE and QbS.

2.4.1.1 Query by string-based approach

The holistic query-based string methods may be guided or unguided, i.e. in some cases they require user control to improve the retrieval results. Therefore, such approaches are classified into two categories: QbS-guided and non-QbS-guided approaches.

- **QBS guided-free approaches**

Various holistic word retrieval techniques employed Latex for generating the query word [Marinai et al., 2006]. A word image retrieval system was introduced. It is mainly based on the use of Self Organizing Maps (SOM) allowing the clustering of word images combined with Principal Component Analysis (PCA). The query is a text, i.e. an ASCII word, and then a word image is generated by Latex. The query is formulated as an image having the same typeface as the document to be processed, and then it is correlated with a synthetic noise in order to reproduce the degradation present in the document. The researchers in [Balasubramanian et al., 2006] proposed also a holistic QbS-based system for words. A large collection of images of printed documents is analyzed by the authors by matching image features at the word level. For word representation, they use features based on profile and shape. The authors

consider that some of these features provide sequence information, while others capture structural features. Normalization of the extracted features makes the word representations insensitive to variations in font, style, size and various image degradation. Second, a novel DTW-based matching protocol is applied to support morphologically different words. The proposed system supports cross-linguistic search. Word images are represented by textual and visual representations in the QBS word retrieval method [Aldavert et al., 2013]. At the textual representation level, word transcription is performed using n-grams of character blocks. These transcriptions are divided into a set of bi-grams and tri-gram blocks that constitute overlapping blocks of individual characters. Textual descriptors are obtained by accumulating the occurrences of each n-gram as a histogram and normalizing this histogram by its L2 norm.

- **QBS guided-based approaches**

A tool for keyword-guided retrieval in historical printed documents using synthetic data and user feedback was presented in [Konidaris et al., 2007]. This method is mainly inspired by the one proposed in [Gatos et al., 2005], which developed an approach for keyword retrieval in typed historical documents, combining the technologies of image preprocessing, synthetic data creation, word retrieval and user feedback. The synthetic query word is produced from the manually designated prototype characters in the Greek historical documents. The distance between characters was set to 10% of the average height of the characters in the images as depicted in Figure 2.7.



Figure 2.7: Synthetically generated query word [Konidaris et al., 2007].

The query words are then normalized to fit into a predefined bounding box. In document collection, words are segmented using dynamic parameters. Two types of features are defined for each segmented word. For the first type, the calculation of the areas formed by the upper and lower profiles of the word is done in 30 small areas each. According to the second type, a set of 90 areas where the pixel density of the characters is calculated is divided into the image. The Manhattan distance is applied between the features of the two words to be matched.

In addition, a keyword retrieval system that relies on relevance feedback to improve the accuracy of the document image retrieval system was proposed in [Keyvanpour et al., 2014]. Several strategies to obtain positive and negative feedback, such as "Only Positive Feedback", "Only Negative Feedback" and "Positive and Negative Feedback", were compared. This comparison showed that using a positive feedback strategy such as "Only Positive Feedback" performs better than document image search systems.

To access the content of machine-printed Greek historical documents, the authors of [Kesidis et al., 2011] introduced a framework guided by word retrieval. To generate

the query image, the proposed approach is based on QbS formalism. The feature extraction process consists of two distinct phases. Normalization is performed in order to preserve the scale invariance. Subsequently, two different types of features are extracted, the pixel density of the characters and the top/bottom profile projections. The word-matching step uses a Euclidean distance to rank the matching results.

Regarding QBS approach, the input to the word spotting system is a text string. Different word string embedding models are presented in [Sudholt and Fink, 2017]. Among the extensively used string representations in the embedding approaches for word spotting, there is the Pyramidal Histogram of Characters (PHOC) [Almazán et al., 2014b]. PHOC is represented by a binary histogram of each character's appearance in a particular division of the string. In the same directions, Spatial Pyramid of Characters (SPOC) [Rodríguez-Serrano et al., 2013] based on Bag-of-Characters (BoC) is presented. The idea is to create a histogram of characters that counts occurrences of a character in each split of string and then a BoC is generated for each split. In addition, authors in [Wilkinson and Brun, 2016] proposed an approach based on the Discrete Cosine Transform of Words (DCToW). It consists of three stages: *i*) each character is represented by a one-hot encoded vector according to the alphabet, *ii*) vectors are stacked into a matrix, then applying discrete cosine transform per row and only the most prominent three values are taken and *iii*) the obtained values are combined into a vector to create DCToW descriptor.

2.4.1.2 Query-by-example based approaches

In the literature, the first applications of word retrieval for document image indexing and information retrieval were provided by [Manmatha et al., 1996]. Following this, the authors in [Rath and Manmatha, 2007] presented a word retrieval approach for document image retrieval and historical document image indexing. This approach involves grouping word images as clusters consisting of similar words. Such approaches consider as input one or more sample images representing the query word, therefore they are named QbE-based approaches. The relevance of QbE-based methods depends on the segmentation and word representation process. According to the type of documents processed, each segmented word is represented by features. Generally, holistic word retrieval approaches use three kinds of models to represent each word that can be categorized into statistical, pseudo-structural and structural descriptors.

- **Statistical models**

The main function of statistical descriptors is to give a representation of the image in the form of an n -dimensional feature vector. A distinction can be made between global and local features. In the case of global features, scalar features such as width, height and aspect ratio can be employed, which are calculated from the entire image. However, local features are determined on the basis of local regions of the image or from primitives extracted from the image. For example, such features can be represented by crosses or key points or a region. As a general rule, local features and global features can be considered to provide different information about the image as the medium on which the texture is calculated varies.

Nevertheless, the combination of these two features is advantageous as long as an approximate segmentation of the objects is available [Lisin et al., 2005]. Local features

can represent word pictures in a typical way, but their reliability is not guaranteed. Global features have good reliability but do not typically represent word pictures. The authors [Manmatha and Croft, 1997] proposed a statistical method [El Yacoubi et al., 1995] to identify words. To describe the word images, the authors use statistical features such as area and aspect ratios. Once the word extraction method is completed, two matching algorithms are introduced to classify the matches between words and a particular query image. Basically, the first matching technique was based on Euclidean Distance Mapping (EDM). The second relied on a Scott and Longuet Higgins (SLH) algorithm. Both techniques were evaluated on two pages belonging to two different writers. The first document consists of 192 words and the second of 153 words. The experimental results indicated that in the case of poor handwriting style, EDM performs poorly, as it does not solve the problem of distortions. However, the SLH performed well in the same case. Different characteristics have been investigated by [Rath and Manmatha, 2003]. For the representation phase, only single-valued features including projection profile, top/bottom word profile, background-ink transition, and grey-scale variance were evaluated. During the matching phase, Dynamic Time Warping (DTW) with Euclidean distance was used to match the feature sequences of two words. For the tests, 2381 words from the George Washington manuscript database were used. To determine the performance of each feature type, they used 15 queries. The results indicated that the top word profile feature performed the best.

A Harris corner detector-based statistical description method was employed for every segmented word [Rothfeder et al., 2003]. Computing the relative corner matches was achieved by applying the sum-of-squares-distances (SSD) error metrics in order to compare the grey-level intensity windows around the detected corner points. The correspondences between the feature point pairs provided an effective means of determining the similarity between the local regions. Figure 2.8 illustrates an example of detected corners and recovered matches in two-word images. Indeed, in order to match two words, the size of the words must be the same, which implies that all candidate words must first be reduced to the size of the query word. The Harris detector technique is employed mainly based on its ability to be repeatable, invariant to changes in viewpoint and invariant to changes in illumination.

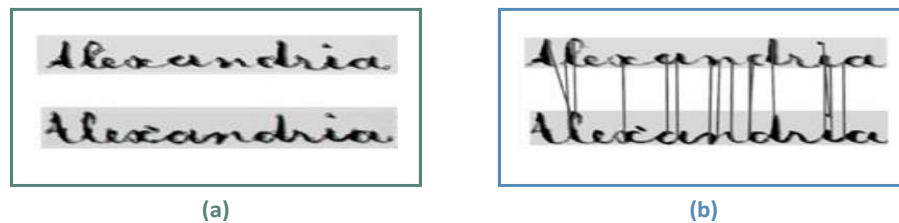


Figure 2.8: (a) Corners detected with the Harris Corner detector on two gray level images. (b) Recovered correspondences in two word images [Rothfeder et al., 2003].

Several works based on the concept of QBE rely on a statistical model based on the Bag of Visual Words (BoW), as presented in [Shekhar and Jawahar, 2012, Yalniz and Manmatha, 2012]. The Scale-invariant Feature Transform (SIFT) features which are obtained by computing the interest points surrounding the key points, constitute the

most widely adopted features to build the representation model. This model is intended to retrieve relevant documents from datasets containing numerous documents [Shekhar and Jawahar, 2012]. Figure 2.9 provides a brief description of the BoW representation model.

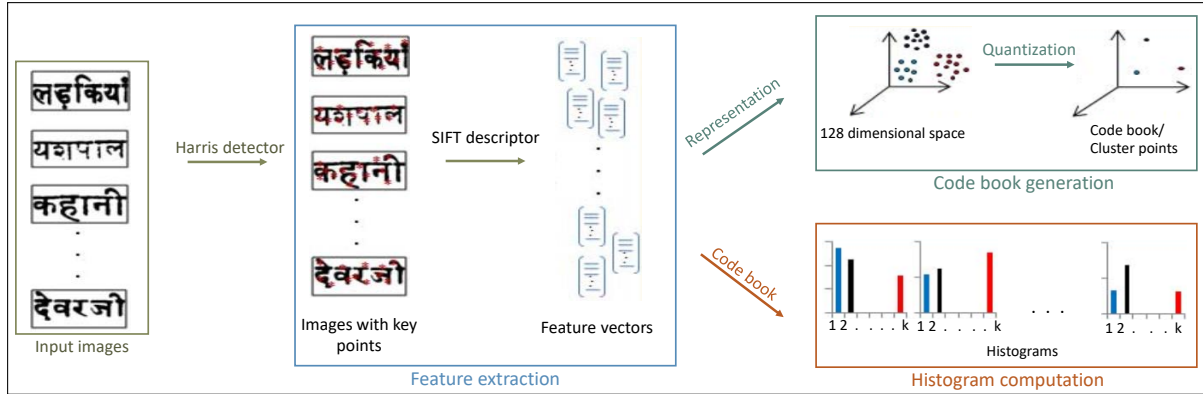


Figure 2.9: An overview of the different holistic methods.

The concept of deep embedding is widely used and it is considered an effective method in terms of their presented results in the state of the art. The proposed method in [Poznanski and Wolf, 2016] is based on a predefined lexicon. The authors used CNN to learn embeddings which take word images as input and output PHOC attributes. A multi-layer perceptron (MLP) is used to predict each level of the PHOC representation. Then, the word spotting task aims to find the nearest neighbors in that space. The authors in [Wicht et al., 2016] proposed a word spotting approach for handwritten documents. The method consists of three steps: *i*) extract small patches from the segmented word images using a horizontal sliding window, *ii*) extract features from patches via Convolutional Deep Belief Networks and *iii*) word spotting using DTW and Hidden Markov Model (HMM). In [Barakat et al., 2018], the authors proposed the use of a Siamese CNN to learn feature representations. First, two images (the query and one of the reference words) were taken as input to the system. After that, they were projected onto the pre-learned embedding space. Second, their novel feature representations were extracted using the Siamese CNN. Finally, the Euclidean distance was computed to check if the two images were associated with the same word.

- **Pseudo-Structural models**

In most cases, word retrieval approaches include pseudo-structural information as part of the descriptors for depicting word images. In this vein, the authors [Fernández et al., 2011] designed a pseudo-structural descriptor based on Loci features and conceived a system for spotting handwritten words in historical documents such as the marriage licenses of the Barcelona Cathedral. They used a pseudo-structural descriptor based on the Loci features in order to classify mixed-character alphabets to represent word images. The number of intersections in four directions (top, bottom, right, and left) is a feature of Loci. For each background pixel in a binary image and for each direction, the number of intersections (a black/white transition between two consecutive pixels) is counted. In this way, every key point produces a code called a Locu number of

length 4. The image skeleton is then determined by iterative thinning. Once the feature code words are organized in a lookup table, Euclidean and Cosine distances are used to match the query word vector feature to all the image vector features. The authors [Fernández-Mota et al., 2014b] calculated pseudo-structural and structural features during the feature extraction phase. In order to extract pseudo-structural features, the authors used the descriptor Loci. As for the structural descriptor, they followed a Shape Context descriptor which allows a coarse distribution of the shape neighborhood of the produced key points.

- **Structural models**

A combination of global and local features in the form of profile signatures and morphological cavities were employed by these authors to characterize a word. In order to encode certain profile features, a DCT is used for encoding cavity features. The profile signatures encoded in DCT were initially matched using a Minkowski distance. Next, the result of the profile matching is introduced into the graph of keyword signatures as an additional feature. Lastly, a probabilistic matching of the graph based on Bayesian proof logic is applied to determine the best match between the keyword signatures and those produced by words located in the candidate regions. This study is considered one of the first keyword retrieval techniques that deal with both cursive and unconstrained handwritten forms.

Authors in [Wang et al., 2014] introduced a coarse to fine graph matching. For the feature extraction step, the authors extract three types of structural points to build graphs. Then, Shape Context labeled graphs are used to generate the different attributes of the graph vertices. In the matching step, a small graph bag technique was used to find candidate words. A graph editing distance based on the DTW alignment algorithm is then applied to produce the true positives.

2.4.1.3 Hybrid methods: QbS & QbE

Recently, the combination of the QbS and QbE methods has been widely employed. To produce feature representation images, the authors in [Krishnan et al., 2016] used the deep CNN which was pre-trained on synthetic data. After that, the obtained features were embedded into a word attribute space using support vector machine classifiers. Finally, word images and textual attributes were projected into a common subspace. Moreover, the authors in [Wilkinson and Brun, 2016] utilized a triplet CNN in order to extract features from word images. These features were embedded into a word attribute space via an MLP. Next, for string embedding, the authors proposed a representation based on DCToW. Finally, to produce word embedding, the representations of images and their corresponding transcription were projected into a common subspace where the spotting task was considered as the closest adjacent search. In the same vein, in [Krishnann et al., 2018] an end-to-end word spotting method was suggested in the interest of learning a common subspace between text and word images. First, the baseline CNN architecture HWNet [Krishnan and Jawahar, 2016] was introduced to learn the feature embedding of word images. Then, recurrent CNN with a spatial transformer layer was employed to perform word spotting and recognition. Regarding the proposed method in [Mhiri et al., 2019], the CNN was used to learn feature embedding of word images. For word text embedding, a recurrent neural network was used to map a

sequence of characters to the subspace. Afterward, an end-to-end DNN architecture was employed to join the embedding of word text and images. Finally, to predict the matching between two embedding vectors, the MLP was used.

2.4.2 Analytical analysis techniques

Analytical processing consists of segmenting word images or even the image of an entire document into smaller units that can be recognized individually or in groups as presented in Figure 2.10. In particular, three categories of analysis techniques are distinguished in the literature. Segmentation-based methods require that each word is segmented into characters. However, the segmentation process is sometimes difficult to perform accurately. To mitigate this challenge, a large number of works consider multiple segmentation assumptions by over-segmenting the images into small features such as connected components, features, etc. Other approaches involve explicit segmentation of words to split them into smaller entities that are expected to be characters to be recognized later. According to the query formulation, such approaches can also be classified into two sub-classes, such as holistic word retrieval approaches.

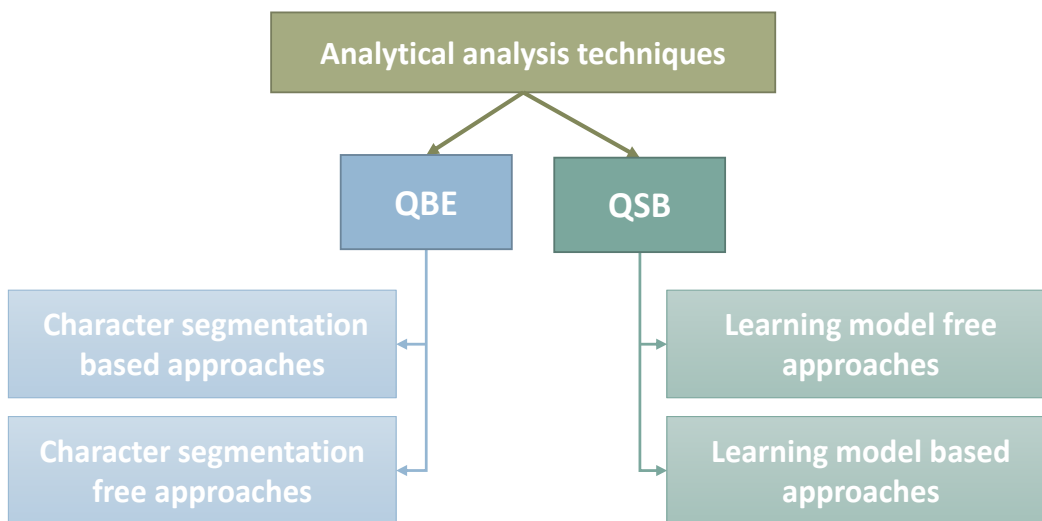


Figure 2.10: An overview of the different analytical methods.

2.4.2.1 Query-by-example based approaches

Analytical approaches relying on QbE concept can be categorized into two main classes: techniques based on character segmentation and techniques without character segmentation. A brief overview of some of the techniques in each category will be given.

- **Character based segmentation-free methods**

Many techniques under this category employ sliding windows for extracting different features. For example, the method based on the slit-like HOG (Histogram of Oriented Gradient) feature was introduced by the authors in [Terasawa and Tanaka, 2009]. They

apply a narrow, rectangular sliding window to each line of text that advances in the direction of writing as shown in Figure 2.11. A HOG feature vector is determined for each sub-picture cut by the window. HOG produces a histogram of the gradient orientations in a certain local region, with orientation bins regularly spaced from 0 to 360 degrees. In general, the best-performing number of bins is 12 or 16.



Figure 2.11: Use of a sliding window for extracting small fragments.

In [Khaissidi et al., 2016], an unsupervised analytical method is introduced. For representing the query and document image, the authors suggested using sliding windows for HOG feature extraction. In order to encode the extracted HOG descriptors, the product quantization method and support vector machines (SVM) are applied.

A similar work using HOG descriptors is reported in [Almazán et al., 2014a]. Images of documents, both hand-written or machine-printed, are partitioned into cells of equal size. HOG descriptors are applied to each cell. More precisely, the different documents are depicted by a grid of HOG descriptors, while a sliding window approach is applied for identifying the regions of the document that are most similar to the query. The Window size is adapted to the size of the query image. In addition, an evolving approach to perform both the word spotting problem and the graph spotting problem was introduced in [En et al., 2016]. In order to perform the feature extraction process, the authors applied the Vectors of Locally Aggregated Descriptors (VLAD) and Fisher Vectors. Then, Both the PQ and the Asymmetric Distance Computation (ADC) techniques are employed to make the approach scalable.

- **Character based segmentation-based methods**

In the same vein, an approach was applied to handwritten manuscripts of Japanese and a few other oriental languages [Terasawa et al., 2005]. Then, a segmentation of the document image into small slits was performed, with narrow rectangular windows scanning the image along the line axis. A low-dimensional descriptor is generated for each slit using the y Eigen space method. The slots have a width of 9 pixels, given that this width should be adequate for the size of a single character, whose average size is about 60 pixels in their study. In order to match the low-dimensional representation of the query word with that of the document images, the process is performed by DTW.

An elastic gradient-oriented histogram (EHOG) for spotting calligraphy words was proposed by the authors [Xia et al., 2014]. The EHOG appears as the modified version of the traditional HOG by taking into account the different characteristics of Chinese calligraphy characters. Finally, in the matching phase, the authors improved the DTW into a Derivative Dynamic Time Warping (DDTW) which considers the feature shape of the Chinese characters.

The authors also proposed an analytical approach on the basis of the QbE applied to Arabic document images [Kassis and EL-Sana, 2014]. A radial descriptor was adopted in this work for extracting the features of word part images. Indeed, it defines the intensity variance of the neighborhood of any point at several levels. The experiment was performed on a set of Arabic historical word parts images consisting of multiple instances of different word parts.

2.4.2.2 Query by string-based approach

In order to solve the word spotting task, most of the approaches in this category use the DTW technique and HMM and models. Therefore, two classes can be distinguished: learning model and Learning model-free.

- **Learning model-free based techniques**

Within this category of approaches, the method proposed by [Moghaddam and Cheriet, 2009a] can be mentioned, which consists in processing cursive Arabic scripts independently of the language, and without segmenting the words or lines. Moreover, this method is applicable to historical document images and enables the extraction of the most relevant information. A Basic Connected Components (BCCs) library is generated from all connected components in the document image, containing the connected components encountered in the text. In this case, this multi-class library is built on the basis of six features. Then, each connected component is matched against all clusters and the closest BCCs are identified as the compatible BCCs of the connected component. Next, the multi-class library is elaborated by matching the normalized horizontal and vertical histograms of a newly connected component to the existing histograms through the DTW method. The matching is performed through the computation of the Euclidean distance.

Similar to the previous work, the authors of [Liang et al., 2012] developed a feature-based representation, designated as a synthesized word from a query word. In the documents, each word image is presented in the form of graphemes that are obtained by segmenting the word images into small units. The synthesized word corresponds basically to a vector of relevant character models assigned a node number related to the labeled graphemes following their similarities in unsupervised learning. The retrieval process consists in determining how probable it is that the K^{th} character of the test word is an instance of the K^{th} character in the query word. Then, the retrieved words are ranked according to their associated probabilities with the synthesized word. In addition, a word retrieval system for historical documents has been introduced [Khurshid et al., 2009]. It is based on the extraction of segmented words and characters in the text. In the query formulation stage, it is possible for the user to either select a word in the document image or enter it as an ASCII word. In the second case, the sequence of features associated with each character is assigned to the word. The method is therefore independent of character size. The matching is performed on two levels, character level, and word level. Once the character extraction stage is finished, each character is labeled with a feature set consisting of six feature sequences and five scalar features.

- **Learning model based techniques**

An alternative method based on gradient features was put forward in [Rodríguez-Serrano and Perronnin, 2009] in which modern fonts are used to produce 34 synthesized queries to be matched with handwritten text. To encode the shape of the words, features based on the local gradient histogram (LGH) are applied. Next, a sliding window is used for splitting the word image into T overlapping windows. Then, each window is divided into a 4×4 cell grid, and the gradient histogram is computed for each of these 16 cells. Therefore, a total of 16×8 features are defined for one window as depicted in Figure 2.12. The same authors [Rodríguez-Serrano and Perronnin,

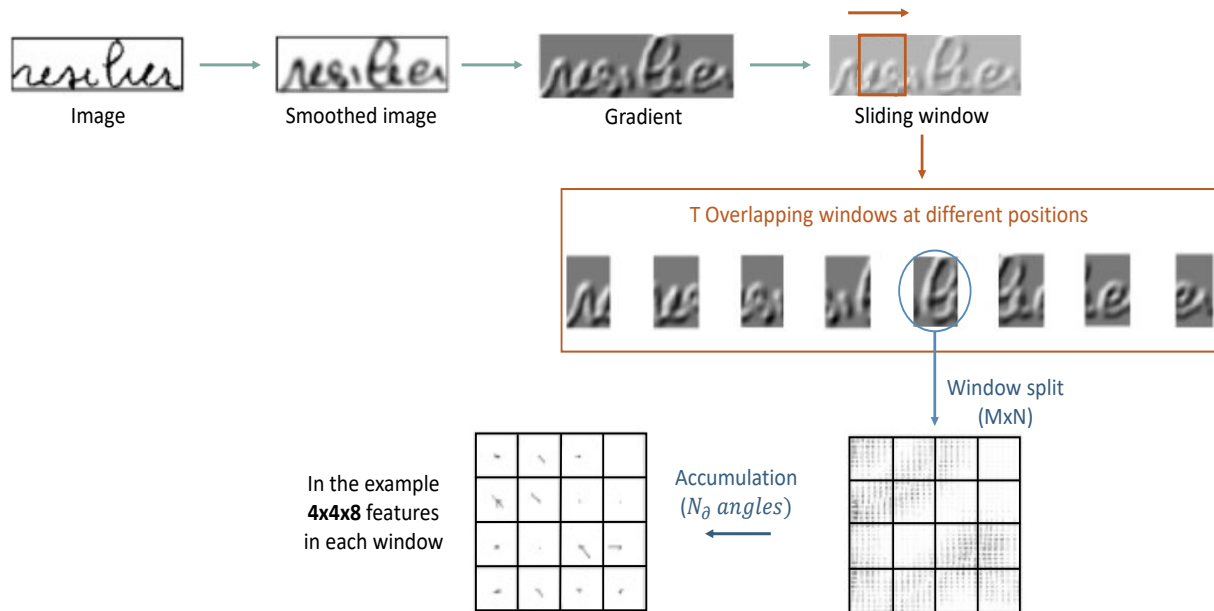


Figure 2.12: The process of extracting LGH features for a small overlapping windows of one word [Rodríguez-Serrano and Perronnin, 2009].

[2012] provided also a novel similarity measure between vector sequences. First, each sequence is modeled by an HMM, and then a similarity measure is computed. In particular, they propose to represent sequences with semi-continuous HMMs (SCHMMs). This approach offers the advantage to simplify the similarity computation between two SC-HMMs into a DTW among their mixture vectors, which allows a relevant reduction of the computing cost. Authors in [Fischer et al., 2012] reported an HMM-based learning-based word spotting method for handwritten text. The input consists of a keyword string and a line image of text. In view of considering different writing styles, normalization preprocessing is applied to each input text line image. Then, the text line images are presented using 9 local features generated through a sliding window. On the basis of the transcribed text line image, for each character of the alphabet, character HMMs are trained. Finally, in the matching step, a likelihood score of the normalized input text line is obtained by matching the trained character HMMs to a keyword text.

2.4.3 Word spotting system

This section discusses the main steps in the word retrieval pipeline. A typical word retrieval system is depicted in Figure 2.13. The entire procedure is split into two phases: an offline phase and an online phase. During the offline phase, features are collected from images of words, lines of text, or even entire pages, which are then interpreted as feature vectors. While in the online phase, a query is formulated by a user who selects a real example from the collection (QbE) or enters an ASCII text word (QbS). A common representation with the offline phase is employed in order to provide a description of the query, and then a matching process is applied between these representations in order to produce a similarity score which, in turn, produces a ranking list of results according to their similarity to the query.

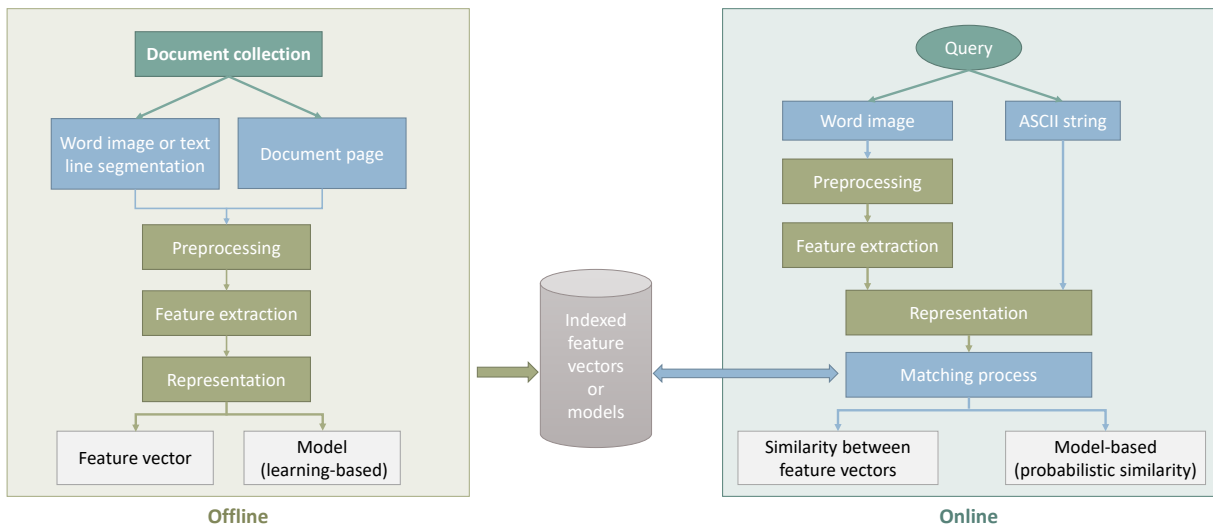


Figure 2.13: General word spotting system architecture.

2.4.3.1 Preprocessing

This step consists of preparing the data from the sensor for the next phase. It is essentially a matter of reducing the noise overlay on the data and trying to keep only the significant information of the represented form. Noise can be due to acquisition conditions (lighting, incorrect document placement, etc.) or to the quality of the original document.

- **Thresholding:** The digitized document is an image that can be binarized or grayscaled, using a thresholding method. The determination of the threshold can be global, i.e. the threshold has the same value for all pixels in the document or local for which the threshold varies from one position to another. The global thresholding methods are not too consuming in terms of consumption calculation time, however, they only give good results if the document is uniformly illuminated. Local thresholding methods are more robust to such degradation but their calculation time is longer.
- **Smoothing:** It allows reducing to a minimum the discontinuities introduced into the image during the different transformations and thus restores the regularity and continuity of the word's outline. Smoothing consists in examining the surroundings of a

pixel and assigning it the value 1 if the number of black pixels in this area is greater than a threshold. The character image may be affected by noise due to acquisition artifacts and document quality, leading to either a missing point or an overload of points. Smoothing techniques allow these problems to be solved by local operations called capping and cleaning operations. The cleaning operation removes small stains and growths from the shape. For the capping, it is a question of equalizing the contours and filling in the inner holes to the shape of the character by adding blackheads.

- Normalization: The normalization method is generally used in word recognition to reduce all types of variations and to obtain normalized data. However, it also distorts the excessive form and eliminates some information useful. The usual methods for normalizing a character are: normalization of the size; correction of line inclination (Skew correction); correction of character inclination (Slant correction).

2.4.3.2 Feature Extraction

The characteristics extraction often called primitives consists in representing the input data (words, characters, graphemes) in a vector of fixed-dimension primitives. This is a crucial step in recognition systems. The purpose of this phase is to select relevant, discriminating and limited information for the classification stage while avoiding the risk of losing important and meaningful information. Indeed, a bad choice of primitives has a negative and clear influence on the results even if a very efficient classifier is used.

The different techniques for extracting characteristics are classified according to the types of primitives :

- Structural characteristics: are generally extracted not from the raw image, but from a representation of the shape by the skeleton or by the contour. It's mainly about: Line segments, Arcs, loops and concavities, slopes, Angularities, extremum points, and endpoints.
- Statistical characteristics: The aim here is to represent the word by statistical measurements of the image. For example, we can use the distribution of pixels in different regions of the image, or histograms (number of black dots per column, per row, or in other directions).
- Global characteristics: are defined when the coding does not involve the specific position of particular elements of the image. The image is considered globally without trying to distinguish the different areas.
- Morphological characteristics : The extraction of morphological characteristics is based on a study of the relative positions of the different black and white components of the image. The word is then described in terms of white and black components, cavities (white parts partially surrounded by black) and loops (white parts entirely surrounded by black). The detection of morphological characteristics can be performed by morphological operators of dilation in all four directions and image intersection.
- Metric characteristics: This category includes features based on physical measurements of the image. In addition, to fairly simple characteristics, such as the height,

width and ratio of these two quantities, more complex characteristics, such as profile coding, can also be used. The profiles can be defined in relation to the four natural axes (left, right, top and bottom).

- Adaptive characteristics: are obtained directly from the image and require a learning phase. In other words, the system operates on a representation close to the original image and must build and optimize the feature extractor.

2.4.3.3 Classification

All pattern recognition methods have been applied to handwriting recognition. However, human capacities to recognize objects, people or the writing of an unknown person, for example, in any context, have led researchers to find flexible and intelligent recognition methods and techniques. Among these classification methods are :

- Statistical methods: Statistical or geometric methods allow a classification decision to be made in an unknown form. It is based on an extensive, rather than a comprehensive, description of the classes. The concept of classification can be expressed in terms of feature space partitioning, where the shape is transformed into a vector of characteristics. The latter presents the measured characteristics. It is based on the statistical study of the measurements made on the shapes to be recognized. The study of their distribution in a metric space and the statistical characterization of the classes, make it possible to make a recognition decision of the type "higher probability of belonging to a class". They benefit from automatic learning methods that are based on sound theoretical foundations, such as Bayesian decision theory.
- Bayesian method: These methods consist in choosing among a set of characters, the one for which the sequence of extracted primitives has the highest probability of being later compared to the previously learned characters. This technique is a basic method for pattern recognition within a probabilistic framework that serves as a reference for other methods. In particular for the evaluation of the error rate.
- Nearest neighbor method: The K Nearest Neighbours KNN algorithm assigns an unknown shape to the class of its nearest neighbor by comparing it to shapes stored in a reference class called prototypes. It returns the K shapes closest to the shape to be recognized according to a similarity criterion. A decision strategy makes it possible to assign confidence values to each of the classes in competition and to assign the most likely class (in the sense of the chosen metric) to the unknown form.
- Artificial neural network: An artificial neural network is a computational model whose design is very schematically inspired by the functioning of biological neurons. The networks of neurons are generally optimized by probabilistic learning methods, especially Bayesian. On the one hand, they are placed in the family of statistical applications, which they enhance with a set of paradigms to create classifications (Kohonen networks in particular), and on the other hand in the family of artificial intelligence to which they provide a perceptual mechanism independent of ideas for implementing it, and providing input information to formal logical reasoning.

2.4.4 Deep Learning in Word Spotting

The task of word spotting in historical documents has been approached using a variety of techniques, as outlined in the preceding section. However, recent advancements in deep learning have demonstrated its suitability as a method for solving this problem. In this section, we present a comprehensive analysis of word spotting in historical documents utilizing deep learning techniques. We will specifically focus on the current study, outlining the methodologies employed and evaluating their effectiveness in comparison to existing solutions.

In the literature, various methods for word spotting have been proposed and can be broadly categorized into two groups: segmentation-free approaches and segmentation-based approaches. The segmentation-free approaches rely on holistic image representations without the need for text-line or word segmentation, while the segmentation-based approaches utilize text-line or word-level segmentation as a pre-processing step before spotting the word. Both categories have their own advantages and limitations, and the choice of approach depends on the specific task and dataset.

2.4.4.1 Segmentation-free approaches

The fundamental concept behind segmentation-free approaches is the selection of patches in the input document. This can be achieved through two primary methods: *i*) a sliding window technique or *ii*) template matching. In the case of sliding window-based approaches, one such method is presented in [Aouadi and Kacem, 2011], where the authors formulate the keyword spotting problem as a search for the parameters of a GHT. The approach consists of two stages: Hash table construction and spotting. The GHT parameters of the reference dataset words are generated and stored in a Hash table, which serves as a dictionary. In the spotting stage, the GHT parameters of the query and reference images are compared using standard distances. This method has been shown to be effective in identifying words in historical documents. In [Khaissidi et al., 2016], the authors proposed a method for word spotting in historical documents by utilizing the Histograms of HOG descriptors to represent the word images. The HOG descriptor is a feature descriptor that captures the local shape information of the word image. The query words and reference words are then compared using a pre-trained SVM model, as outlined in [Vapnik, 1999]. The SVM model is trained on the HOG features of the reference words and is used to classify the query words based on their similarity to the reference words.

In the realm of template matching-based approaches, the authors of [Faisal and Al-Maadeed, 2017] proposed a method for word spotting in historical documents by using Normalized Cross Correlation (NCC) [Lewis, 1995] to identify the location of a query image in the input document. The approach involved image pre-processing and word image matching. In the image pre-processing stage, the image is transformed to improve its quality and reduce noise. In the word image matching stage, the similarity between the query word and reference word image templates is calculated using the NCC algorithm. This approach has been proven to be effective in identifying words in historical documents, particularly when the images are of high quality and contain minimal noise. In [Gatos and Pratikakis, 2009], an approach was proposed for word spotting in historical documents, which consists of three stages: *i*) detection of salient regions using the Run Length Smoothing Algorithm

(RLSA) [Wahl et al., 1982], *ii*) feature extraction based on pixel density, and *iii*) block-based template matching applied only to regions of interest for locating words in documents without segmenting them. In the first stage, salient regions in the image are detected using the RLSA, which is a technique that aims to reduce the noise and preserve meaningful text lines. In the second stage, features are extracted from the salient regions based on pixel density, which describes the number of black pixels in a region. In the final stage, block-based template matching is used only on the regions of interest to locate the words in the document.

2.4.4.2 Segmentation-Based approaches

Segmentation-based approaches for word spotting in historical documents consist of dividing the input document into word parts, which are then compared to the query word image. In [Shahab et al., 2006], the input document is segmented into "sub-words" using the connected component algorithm [Gonzalez et al., 2002]. The word images are then described by angular lines, concentric circles, and geometric features. Finally, the query and reference images are compared using the Euclidean distance. In [Zirari et al., 2013], the authors proposed an approach that involves three main stages: *i*) first, the input document is segmented using the RLSA [Wahl et al., 1982], *ii*) second, word images are described by their vertical/horizontal histogram and upper/lower profile, and *iii*) finally, the Levenshtein Edit distance is used in the matching stage. Additionally, in [Saabni and El-Sana, 2013] an unsupervised segmentation-based approach based on line segmenting and component extraction was proposed. The contour of the connected component is extracted as features from the image word and DTW is used to match the features. Kassis et al. in [Kassis and EL-Sana, 2014], [Kassis and El-Sana, 2016] proposed a method for word spotting in historical documents which involves two primary stages: *i*) feature extraction of word-parts using the radial descriptor, and *ii*) word-parts matching, where the similarity between the word-parts is computed by comparing their occurrence probability histograms. The radial descriptor is a feature descriptor that captures the structural and shape information of the word parts by representing it as a set of radial lines. The occurrence probability histograms are used to represent the probability of the occurrence of a feature in a word part. The distance between the histograms of the query and reference word parts is computed, and the minimum distance is used to identify the matching word.

In [Barakat et al., 2018], a novel approach for word spotting in historical documents was proposed, where the task is formulated as a classification problem. The authors suggested the use of a convolutional Siamese neural network to learn a new feature representation (embedding) for the word images. In the first stage, the query image and one of the reference word images are taken as inputs to the Siamese CNN, which then projects them onto a pre-learned embedding space. In the second stage, the novel feature representations are extracted using the Siamese CNN. Finally, the Euclidean distance is calculated between the feature representations of the query and reference images to determine if they correspond to the same word or not.

Table 2.1 presents a comprehensive summary of the reviewed methods for word spotting in historical documents. The table provides detailed information on the proposed approaches and evaluation protocols for each method. However, it is important to note that the performance of these approaches should not be compared directly, as the authors employed

different evaluation protocols, metrics, and databases in their studies. Therefore, the results should be considered in the context of the specific evaluation conditions and should not be used to make generalizations about the relative effectiveness of the methods.

2.5 Datasets

In this section, we provide a brief overview of publicly available image datasets of historical documents. In particular, we introduce the historical datasets that have been adapted for the word spotting task.

2.5.1 BH2M

The Barcelona Historical Manuscript Marriage Database or BH2M [Fernández-Mota et al., 2014a] is composed of a single volume, written in Old Catalan by the same author from 1617 to 1619. It consists of 174 pages of handwritten marriage records, of which 100 pages are intended for training, 34 for validation and about 40 for testing. The pages included have been kept in the central archives of Barcelona. This database provides the ground truth for layout analysis, text transcription and semantic analysis. The XML format of the annotation files is hierarchized into text blocks, segmented lines and text words for layout analysis. For handwritten text recognition, word retrieval, information extraction and understanding, and contextual algorithms, transcriptions of additional words and semantics regarding license, appearance order, date, and information about the wife and husband can be helpful. Figure 2.14 provides an example of BH2M database structure.



Figure 2.14: An example of BH2M database at each level segmentation.

2.5.2 IFN/ENIT

IFN/ENIT database was developed in 2002, by the Institute of Communications Technology (IFN) at Technical University Braunschweig in Germany and the National School of Engineers of Tunis (ENIT). It consists of 26,549 images of Tunisian town/village names written by 411 writers [4]. It is one of the most widely used databases. Although it has little vocabulary because it contains mainly names of towns and villages of Tunisia.

Table 2.1: Summary of related works to word spotting.

Reference	Year	Query	Features	Segmentation	Dataset	script	Evaluation index
[Shahab et al., 2006]	2006	QbE	Geometric features	Based	Private	Arabic	mAP
[Aouadi and Kacem, 2011]	2011	QbE	GHT	Free	Private	Arabic	-
[Aldavert et al., 2013]	2013	QbS	Character n-grams	Based	Private, GW	Latin	mAP, Recall
[Aouadi and Echi, 2014]	2014	QbE	HGT, Letter morphology	Free	Private	Arabic	-
[Almazán et al., 2014a]	2014	QbE	HOG, SVM	Free	LB, GW	Latin	mAP
[Puigcerver et al., 2014]	2014	QbS	HMMs, character n-grams	Based	Cristo-Salvador	Latin	AP, mAP
[Riba et al., 2015]	2015	QbE	Structural graph	Based	BH2M	Latin	mAP
[Kassis and El-Sana, 2016]	2016	QbE	Graph radial	Based	VML	Arabic	mAP
[Khaissidi et al., 2016]	2016	QbE	HOG	Free	Ibn-Sina	Arabic	mAP
[Faisal and AlMaadeed, 2017]	2017	QbE	Binarisation	Free	HADARA80P	Arabic	TP, FN, Recall
[Sudholt and Fink, 2018]	2018	QbE, QbS	Attribute CNN, PHOC-Net	Based	GW, IFN, IAM	Arabic, Latin	mAP
[Barakat et al., 2018]	2018	QbE	Siamese CNNs	Based	VML, GW	Arabic, Latin	mAP, R-Precision
[Mhiri et al., 2019]	2019	QbE, QbS	CNN, RNN	Based	LB, IFN	Latin, Arabic	mAP
[Wolf and Fink, 2020]	2020	QbS	PHOCNet	Based	GW	Latin	mAP
[Stauffer et al., 2020]	2020	QbE	Graph	Based	GW, Parzival	Latin	AP, mAP
[Riba et al., 2021]	2021	QbE, QbS	Smooth-nDCG	Based	GW	Latin	mAP
[Kundu et al., 2021]	2021	QbE	HT-based features, DTW	Based	IAM, QUWI	Latin	mAP
[Boudraa et al., 2022]	2022	QbE	U-Net, PHOC	Based	IAM	Latin	mAP
[MHIRI et al., 2022]	2022	QbE, QbS	PBCS, CNN, ViT	Based	IAM	Latin	mAP, AP

2.5.3 GRPOLY-DB

The Greek Polytonic Database (GRPOLY-DB) [Gatos et al., 2015] gathers images of printed and handwritten documents from different sources, divided into four subsets. From 1838 to 1977, the documents were written or printed in the old polytonic system. The entire database includes 399 pages, 15,084 lines of text, 102,596 words and 171,511 characters with ground truth. The available annotations include information on the text and word segmentation, as well as transcriptions for text and single character recognition. Experimental results related to layout and content were presented on the dataset using different methods.

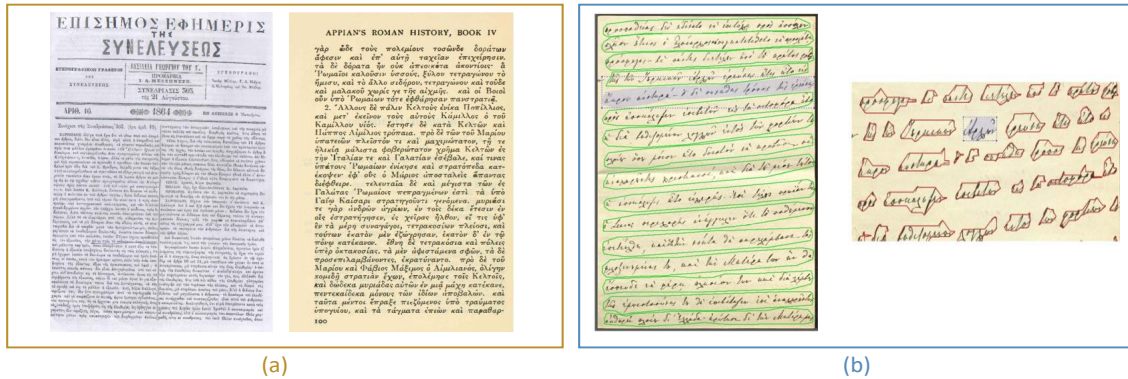


Figure 2.15: (a) Examples of GRPOLY-DB-MachinePrinted grayscale and color page images. (b) Examples of GRPOLY-DB-Handwritten at text line and word level [Gatos et al., 2015].

2.5.4 Hebrew Handwritten dataset

The Hebrew Handwritten dataset (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.

2.5.5 CFRAMUZ

The CFRAMUZ dataset [Arvanitopoulos et al., 2017] includes grayscale image pages from handwritten novels by Charles Ferdinand Ramuz in French between 1910 and 1946. Text and XML annotation files contain the unique word ID, coordinates, width and height of word bounding boxes, word line number, word number in the current line, and word transcription for word spotting without segmentation purposes.

2.5.6 HADARA80P

HADARA80P dataset [Pantke et al., 2014] which consists of high-resolution tiff images of 80 pages of a scanned historical Arabic manuscript. The manuscript is handwritten by one author and was published in the 9th Islamic century or 15th Gregorian century. The data set also provides 25 keyword images segmented out from random pages of the manuscript. The

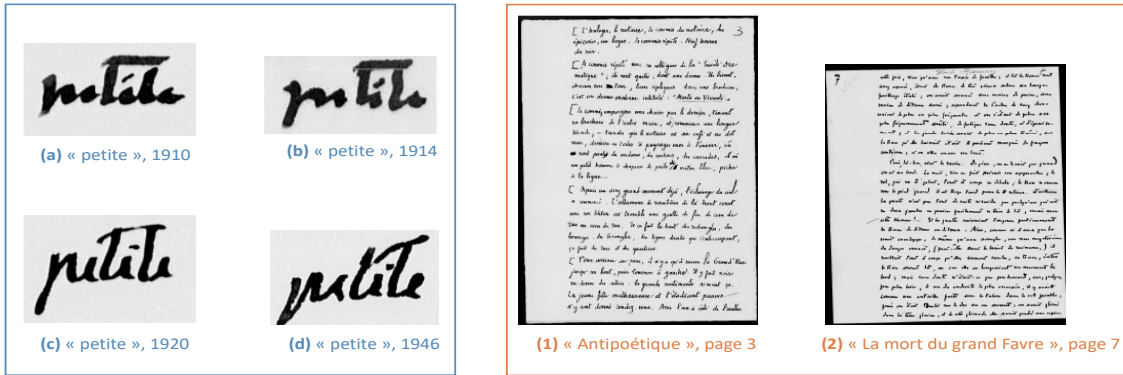


Figure 2.16: Depicting the distinct writing styles in the dataset: (a) and (b) introduce the word ”petite” in a first style while (c) and (d) present the same word in a second style. (1) and (2) present two pages of different novels from the CFramuZ dataset.

page images are 48-bit TIFF images with 16 bits per color channel. Each image is about 50 MB in size.

2.5.7 VML-HD

In [Kassiss et al., 2017] a new database with handwritten Arabic script is presented. It is based on five books written by different writers from the years 1088-1451. 680 pages from these five books. For each page, the authors manually applied bounding boxes on the different sub-words and annotated each bounding box with its corresponding sequence of characters. 121,636 sub-words, containing 244,553 characters out of a vocabulary of 1,731 forms of sub-words.

Each book has its own Hadara XML file which contains the coordinates of all the bounding boxes of all the images annotated of that book, as well as the sequence of characters for each bounding box applied in Arabic text. We also generated a Hadara XML file for each page in addition to the file for each book. This file contains the coordinates of the bounding boxes and the sequence of the characters for the specific page only.

2.5.8 AMADI-LontarSet

The AMADI-LontarSet database [Kesiman et al., 2016] represents the first handwritten dataset of Balinese palm leaf manuscripts. This database consists of three components: binarised images ground truth dataset, the word annotation image dataset and the single character image dataset. The dataset was developed from randomly selected multi-page palm leaf manuscript collections from Bali, Indonesia. This database was part of the ICFHR 2016 challenge. For the word retrieval challenge, 130 training images and 100 test images were used along with 15,022 annotated word patches for training. 36 word-annotated patches were given as a quiz. The challenge involved retrieving similar word image patches from palm leaf manuscripts using a query word image patch.

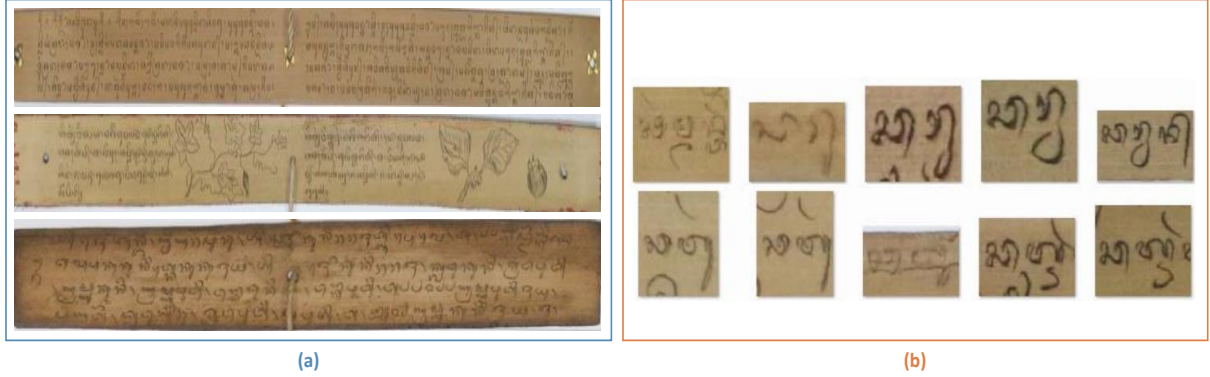


Figure 2.17: (a) Sample images of palm leaf manuscript, (b) Samples of word annotated images [Kesiman et al., 2016].

2.5.9 George Washington

The George Washington dataset (GW) is a collection of historical documents written in English by George Washington and his assistants [Rath and Manmatha, 2007]. It contains 20 pages and is segmented into 4860 words with 1124 different transcriptions, providing ground truth at page, text line and word levels. This dataset is commonly employed for evaluating word retrieval algorithms.

2.5.10 Hebrew Handwritten dataset

The Hebrew Handwritten Dataset (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.

2.6 Evaluation protocols and performance measures

To do an evaluation of a system, certain specific performance parameters are required. In the literature dedicated to word retrieval, several distinct evaluation metrics are available which are defined in the following way.

In the former context, the recall rate and precision rate are determined and often the precision-recall curve is mapped to provide a visual representation of the system's performance [Rodríguez-Serrano and Perronnin, 2009, Srihari and Ball, 2008]. The most common formulas used to evaluate the performance of a word spotting system are as follows.

- **Precision**

The precision is defined as the fraction of retrieved pertinent words to the query word, i.e. the probability that the retrieved image is a target word,

$$P = \frac{|\{\text{relevantinstances}\} \cap \{\text{retrievedinstances}\}|}{|\{\text{retrievedinstances}\}|} \quad (2.1)$$

- **Recall**

Recall rate indicates the fraction of relevant words that are correctly retrieved:

$$R = \frac{|\{relevantinstances\} \cap \{retrievedinstances\}|}{|\{retrievedinstances\}|} \quad (2.2)$$

- **R-Precision**

The R-Precision index is given as the precision at a particular recall value where $P = R$. Since precision is measured for the first k words retrieved, $P@k$ is defined by Eq.(2.3).

$$P@k = \frac{|\{relevantinstances\} \cap \{kretrievedinstances\}|}{|\{kretrievedinstances\}|} \quad (2.3)$$

- **F-measure**

The F-measure refers to the harmonic average of precision and recall. It is defined by Eq.(2.4).

$$F = 2 \cdot \frac{P \cdot R}{P + R} \quad (2.4)$$

- **Average Precision**

The Average Precision score (AP) is calculated as the average of the precision value calculated following the extraction of each relevant word. AP is defined by Eq.(2.5).

$$AP = \frac{\sum_{k=1}^n (P@k \times rel(k))}{|\{relevantinstances\}|} \quad (2.5)$$

where $rel(k)$ denotes an index equal to 1 if the rank k term is appropriate and 0 otherwise.

The average value of the average accuracy over all the queries used in a word retrieval task defines the average accuracy (MAP).

The alternative way of assessing performance is adopted from spoken word spotting [Lavrenko et al., 2004]. This approach is focused on the error rate where the following formulas are used.

- Word Error Rate (WER): the proportion of the words that were not recovered exactly as they were in the manual transcript,
- Out Of Vocabulary words (OOV): words that occur only in the testing pages and not in the training pages or words,
- False Alarm Rate (FAR): an erroneous image target detection decision. The percentage of how many times the word was falsely spotted,

$$FAR = \frac{FP}{FP + TN} \quad (2.6)$$

where TN (True Negative) refers to the total number of OOV image that was not spotted and FP (False Positive) is the total number of spotted samples that are unrecognized.

2.7 Conclusion

This chapter aimed to provide a comprehensive overview of the current state-of-the-art in word spotting techniques for historical documents. We began by highlighting the significance of historical documents and the various challenges associated with their preservation and analysis. We then specifically focused on the challenges of historical Arabic documents, which have unique characteristics that must be taken into account. Subsequently, we discussed different automatic processing systems for historical documents, with a particular emphasis on word retrieval systems. The following section was dedicated to providing an overview of publicly available datasets designed specifically for the word retrieval process. Finally, we discussed the different metrics that are commonly used to evaluate the performance of word retrieval systems in historical documents.

Given the comprehensive state-of-the-art review presented in this chapter, the next chapters will focus on introducing the proposed contributions aimed at addressing the issues and challenges highlighted in the review. These contributions will include novel methods, techniques, and approaches for word spotting in historical documents, with the goal of improving the performance and robustness of these systems. The proposed contributions will be discussed in detail, including the underlying theories, experimental results, and a thorough analysis of their potential impact on the field.

CHAPTER 3

Word Spotting based Triplet-CNN in Historical Arabic Documents

3.1	Introduction	41
3.2	Deep learning for word spotting	41
3.3	Proposed approach for word spotting in historical Arabic document	44
3.3.1	Embedding space construction	46
3.3.2	Word spotting method	48
3.4	Experimental Results	49
3.4.1	Dataset	49
3.4.2	Evaluation Protocol	50
3.4.3	Experimental results	51
3.4.4	Further analysis	53
3.5	Conclusions	56

3.1 Introduction

Word spotting is a crucial task in the field of information retrieval, as it allows for the creation of indexes that aid in understanding and utilizing the content of documents. Specifically, it aims to identify all the instances of a query word within a set of documents, such as a book. The input to the task is a collection of unindexed documents, and the output is a ranked list of words that are similar to the query word. This ranking is based on the similarity measures calculated between the query word and the words in the set of documents. Word spotting is a powerful technique that allows for efficient and effective retrieval of information from large sets of unstructured documents. Due to the complexity of historical documents, including variations in layout and texture, the process of segmenting text into individual words is often prone to errors. This results in degraded quality of the segmented words, making it challenging to extract discriminant hand-crafted features. In recent years, there has been a growing interest in learning feature representations that can effectively extract useful information from images and improve their description and classification. These learned representations, often in the form of deep neural networks, are capable of capturing rich and complex features from the images, overcoming the limitations of traditional hand-crafted features. This shift towards learning feature representations has the potential to significantly improve the performance of word spotting in historical documents, despite the challenges inherent to the task.

This chapter examines the impact of learning feature representations on the performance of word spotting in historical documents. To this end, we propose a novel word spotting method that utilizes a triplet-loss algorithm for feature representation learning. Triplet-loss, introduced in [Schroff et al., 2015a], is a technique for learning similarity or embeddings, and it has been shown to effectively extract discriminative information from data in various domains, such as [Liao et al., 2017, Chen et al., 2017b]. By incorporating this technique, our proposed method aims to improve the performance of word spotting in historical documents by effectively learning more discriminative and robust features from the document images. Our contribution has been formally published in two international conferences [Fathallah et al., 2019, Fathallah et al., 2020].

In the remainder of this chapter, we present an overview of recent word spotting approaches based on deep learning in Section 6.1. Our proposed approach is then described in Section 3.3. The main steps of our approach, which include the construction of an embedding space and the word spotting process, are discussed in more detail in sections 3.3.1 and 3.3.2. The experimental results of our approach are presented in Section 3.4. Finally, conclusions and future research perspectives are provided in Section 3.5.

3.2 Deep learning for word spotting

In the previous chapter, a variety of techniques for word retrieval were discussed. Among them, the use of deep learning techniques has been shown to be particularly promising. In this section, we will specifically focus on the application of deep learning techniques in word spotting, with a particular emphasis on the objectives of the current study. Word retrieval can be performed without the need for recognizing the content of documents, through the use of image-matching tools. The feature extraction process is a crucial step in these tools and an

accurate representation of word images presents a significant challenge in retrieval tasks for historical documents. The use of deep learning techniques, such as representation learning, can help overcome this challenge by effectively extracting discriminative features from the images and improving the performance of word spotting in historical documents.

Handwritten images present a significant challenge in the field of document analysis, as they contain a high degree of variations and complexities compared to printed documents. However, considerable progress has been made in recent years in this area, thanks to the successful application of deep learning methods. For example, the use of matching networks [Krishnan and Jawahar, 2016], semantic feature extraction [Wilkinson and Brun, 2016], and attribute-based approaches [Sudholt and Fink, 2018] have been shown to be effective in addressing the complexities of handwritten images. These methods have been able to overcome the challenges presented by variations in handwriting, such as different writing styles, ink bleed, and degradation of the images over time. The use of learned feature representations for word images simplifies the process of extracting pertinent information and improves image classification performance. The concept of embedding space representations, as illustrated in Figure 3.1, has demonstrated significant success in the task of word spotting in historical documents. This approach allows for the effective capture of rich and complex features from the images, overcoming the limitations of traditional hand-crafted features, and resulting in more accurate and robust word spotting results. The goal of rep-

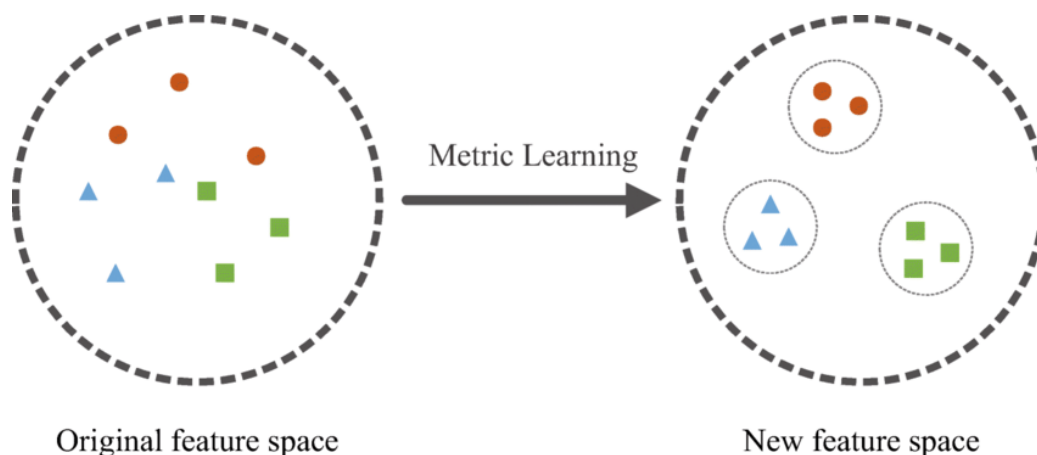


Figure 3.1: Concept of embedding-space in learning-feature representation.

resentation learning is to convert data into low-dimensional feature representations that can be used for classification or retrieval tasks. One of the most effective ways to reduce the dimensionality of the learned vector representations in the embedding space is through the integration of neural networks. This allows the classifier to learn more efficient and discriminative representations. In this regard, authors in [Sharma et al., 2015] have used pre-trained ImageNet to perform word indexing on the IAM database, showing the effectiveness of this approach. Representation learning can significantly improve the performance of word spotting in historical documents by effectively capturing the rich and complex features of the images. In their study, Tang and Lin [Tang and Lin, 2018] proposed a technique for identifying small footnotes in historical documents utilizing deep Residual Networks (ResNets). The depth of the ResNet array was compressed to 12 blocks with only 26 layers, and the authors implemented a Softmax activation function in the output layer of the model. This approach

of using deep ResNets for feature extraction demonstrated effectiveness in identifying small footnotes in historical documents, despite the challenges posed by noise and variations in the images.

Similarity or distance metric learning involves learning a distance function on an input data space that preserves the distance relationship between pairs of similar or dissimilar examples in the training set. As deep neural networks (DNNs) have advanced, the focus of metric learning has shifted from learning distance functions to learning deep feature embeddings that are more adaptable to simple distance functions, such as the Euclidean distance or cosine similarity. DNNs combined with metric learning have proven to be a powerful tool that has been widely used in various computer vision tasks, including object or information retrieval [Oh Song et al., 2016, Ustinova and Lempitsky, 2016, Palangi et al., 2016], question answering [Zhou et al., 2016], machine translation [Zhou et al., 2016, Chen et al., 2017a], object tracking [Tao et al., 2016], face verification [Schroff et al., 2015b] and person re-identification [Shi et al., 2016]. These techniques have been shown to be effective in improving the performance of various applications where the accurate measurement of similarity is crucial.

In the field of word spotting or retrieval, the use of deep embedding techniques has gained widespread acceptance due to its effectiveness in achieving high performance results, as demonstrated in recent studies such as [Mhiri et al., 2019, Barakat et al., 2018, Fathallah et al., 2019]. In terms of the query representation, two main categories of word spotting approaches can be distinguished: Query by Example (QbE) and Query by String (QbS).

With regards to the QbS approach, the input of a word spotting system is a text string. Different string embedding models have been proposed in the literature, such as those presented in [Sudholt and Fink, 2017]. One of the widely used string representations in embedding approaches for word spotting is the Pyramidal Histogram of Characters (PHOC) [Almazán et al., 2014b]. The PHOC is represented by a binary histogram that indicates the presence or absence of each character within a specific segmentation of the string. This representation captures the structural information of the word string and has been shown to be effective in word spotting tasks. textcolorredAuthors in [Rodriguez-Serrano et al., 2013] reformulated the task of recognizing a word from an image, as searching for the nearest match in a common space. The method of creating this space is through the utilization of the Structured SVM (SSVM) framework, which organizes the label-image pairs in a way that the matching pairs are closer together than non-matching ones. In [Wilkinson and Brun, 2016], an approach based on the Discrete Cosine Transform of Words (DCToW) was proposed. This approach consisted of three stages: first, each character was represented by a one-hot encoded vector according to the alphabet. These vectors were then stacked into a matrix. Next, a discrete cosine transform was applied to each row of the matrix, and only the most prominent three values were retained. Finally, these obtained values were combined into a vector to create a DCToW descriptor.

Regarding the QbE approaches, the input of a word spotting system is an image of a word. One of the proposed methods in [Poznanski and Wolf, 2016] utilizes a pre-defined lexicon and a Convolutional Neural Network (CNN) to learn an embedding space for word images. The authors used the PHOC representation as the output, and a Multilayer Perceptron (MLP) was used to predict each level of the PHOC representation. The goal of the word spotting task is then to find the nearest neighbors in this embedding space. In [Wicht et al.,

2016], a word spotting approach for handwritten documents was proposed. The method was composed of three steps: first, extracting small patches from the segmented word images using a horizontal sliding window technique; second, extracting features from patches using convolutional deep belief networks; and, finally, word spotting utilizing dynamic time warping and hidden Markov models to match the extracted features with those from a pre-defined lexicon.

In [Barakat et al., 2018], the authors proposed a word spotting approach that utilizes a CNN to learn a novel feature representation, also known as embedding. The system takes in two images, the query and one of the reference words, as input and projects them onto a pre-learned embedding space. The feature representations are then extracted using the Siamese CNN and the Euclidean distance is computed to determine if the two images are associated with the same word. This approach has been shown to effectively improve performance in word spotting tasks for historical documents.

Recently, the combination of the Query-by-Example (QbE) and Query-by-String (QbS) methods have been widely used to improve the performance of word spotting in historical documents. In [Krishnan et al., 2016], the authors proposed a method that utilizes a deep CNN pre-trained on synthetic data to extract features from word images. These features are then embedded into a word attribute space using SVM classifiers. Finally, the word image and textual attributes are projected into a common subspace, allowing for more accurate comparison and retrieval of words. Additionally, the authors in [Wilkinson and Brun, 2016] employed a triplet CNN to extract features from word images. These features were then embedded into a word attribute space using an MLP. For string embedding, the authors proposed a representation based on the Discrete Cosine Transform of Words (DCToW). Lastly, to generate word embedding, the representations of images and their corresponding transcriptions were projected into a common subspace, where the word spotting task was treated as a nearest neighbor search. In the same vein, in [Krishnann et al., 2018], an approach to word spotting in historical documents was proposed that aimed to learn a common feature representation between text and word images. The method consisted of two main steps: first, feature extraction using a baseline CNN architecture called HWNet; second, word spotting and recognition using a recurrent CNN with a spatial transformer layer to project the features into a common subspace for comparison. This approach aimed to improve the performance of word spotting by learning a shared representation between text and images. In [Mhiri et al., 2019], a word spotting method was proposed that utilizes a combination of deep learning techniques to learn feature embeddings of both word images and text. A CNN was used to extract features from the word images, while a recurrent neural network (RNN) was employed to map the text sequence of characters to a common subspace. An end-to-end architecture was then utilized to join the embeddings of the word images and text. Lastly, a MLP was employed to predict the matching between the two embedding vectors.

3.3 Proposed approach for word spotting in historical Arabic document

In this study, we propose a supervised approach for word spotting in historical documents that utilizes learned feature representations of word images to simplify the process of ex-

tracting pertinent information and enhance image representation. It is important to note that the effectiveness of a word spotting algorithm is contingent upon both the accuracy of the segmentation algorithm and the performance of the word search algorithm. Our approach focuses on evaluating the performance of the word search algorithm and as such, we have chosen to use a pre-segmented dataset. In contrast to the approach presented in [Barakat et al., 2018], our methodology for constructing the embedding space utilizes a triplet loss based CNN. This CNN is designed to take in three distinct word images as inputs: an anchor, a positive sample, and a negative sample, with the anchor serving as the reference input. By minimizing the distance between the anchor and the positive sample while simultaneously maximizing the distance between the anchor and the negative sample, the CNN is able to learn a novel feature space. This feature space, in turn, is utilized as the embedding space for our proposed method. In their study, Barakat et al. [Barakat et al., 2018] utilized a contrastive loss function to learn the embedding space. Specifically, their neural network takes as input a set of pairs of images, comprising positive and negative examples, and learns a novel space where semantically similar examples are embedded in close proximity to one another. In contrast, our proposed method for word spotting is based on the use of a triplet-loss function for representation learning which allows us to overcome the threshold tuning issue that is commonly encountered in Siamese networks and provides a promising solution for HAD recognition. Triplet-loss, as first introduced in [Schroff et al., 2015a], is a powerful learning algorithm that has been used to extract discriminative information from data in various domains [Liao et al., 2017, Chen et al., 2017b].

As illustrated in Figure 3.2, our proposed approach for word spotting consists of two primary steps: first, construction of an embedding space using a triplet-loss function, and, second, word spotting based on the resulting embedding features. The first step is focused on creating a feature representation for the input images through the construction of an embedding space. To accomplish this, a set of triplets is generated from a pre-segmented training dataset, where each triplet comprises three word images: an anchor, a positive sample, and a negative sample. A CNN is then trained on these triplets using a triplet-loss function. The goal of this training phase is to minimize the triplet-loss by creating an embedding space that maximizes the distance between word images associated with different classes and minimizes the distance between word images associated with the same class. In the second step, the embedding features obtained from the first step are used for word spotting. The embedding features are used to represent the input images in a compact feature space, which can be used to perform word spotting tasks with high accuracy. In the word spotting step, the process of inputting query and reference words is undertaken in order to project them onto a pre-established embedding space. This embedding space has been previously learned through a process of training and optimization. Once the query and reference words have been projected onto this space, novel embedding features are extracted for further analysis. These features are then matched utilizing the Euclidean distance metric, which calculates the distance between two points in a multidimensional space. Finally, the output of this process is a list of retrieved words, which have been sorted in accordance with their distance to the query word. This allows for a more efficient and accurate retrieval of relevant words from the reference set.

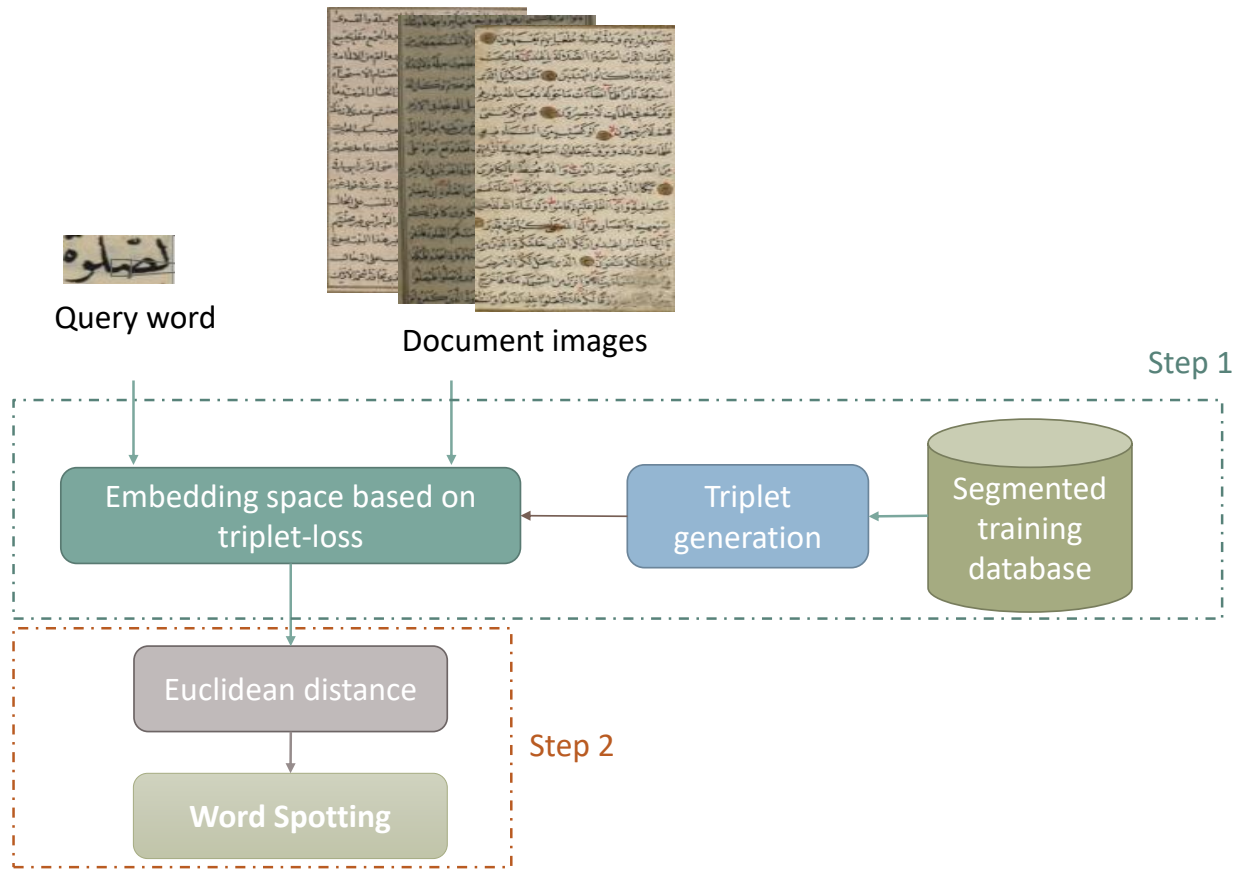


Figure 3.2: Principal steps the proposed approach.

3.3.1 Embedding space construction

The process of creating an embedding space involves the transformation of an original feature space into a new, reduced representation. This is achieved through two key steps: the generation of triplets and the construction of the embedding space utilizing CNNs. Specifically, the generation of triplets involves identifying sets of three related data points, with a focus on preserving the relative relationships between them. These triplets are then used as input for the CNN, which learns to map the data into a new, lower-dimensional feature space. This new representation captures the essential characteristics of the original data while discarding unnecessary information, making it well-suited for a variety of machine-learning tasks.

3.3.1.1 Triplet CNN architecture

CNNs were first proposed in [LeCun et al., 1999] as a method for recognizing objects in images. Since then, they have been successfully applied to a wide range of computer vision tasks, including pedestrian detection and face recognition. A CNN is composed of multiple layers, including convolutional layers, non-linear processing units, and subsampling layers. In our approach, we utilize a network architecture consisting of three CNN instances with shared parameters. The input to the network is a triplet of images, each with a resolution of 60x110 pixels. Each of the three instances processes one of the samples from the triplet.

The architecture of our CNN includes five convolutional layers, each followed by a ReLU activation function and a pooling layer, with the exception of the fourth layer. Additionally, the CNN contains a fully-connected (FC) layer with 1024 neurons, followed by a dropout layer. The output of the FC layer is the final embedding of the input. As depicted in Figure 3.3, the detailed architecture of our CNN is presented. The figure provides a comprehensive illustration of the various components that make up the network, including the number and types of layers, as well as their interconnections. The figure also highlights the specific configurations of each layer, including the number of kernel sizes in the convolutional layers and the number of neurons in the fully-connected layer. The figure serves as an important visual aid in understanding the design of the CNN and how it processes the input data to produce the final embedding.

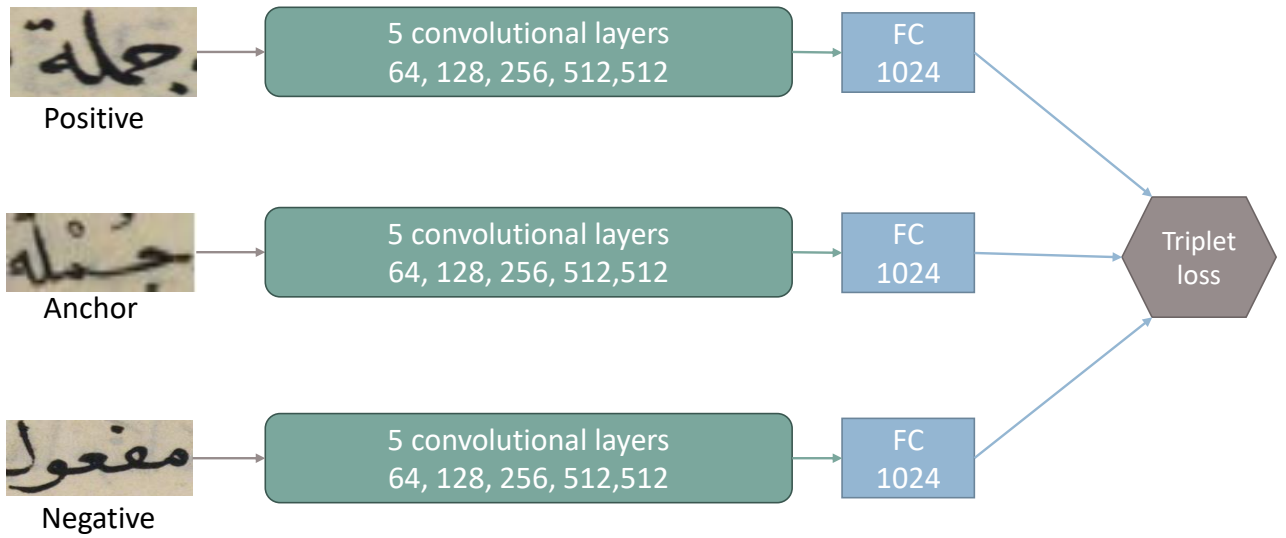


Figure 3.3: Triplet-CNN architecture proposed in our framework.

3.3.1.2 Triplet CNN Learning

In order to effectively train the CNN, a dataset of triplets must be generated. Each triplet comprises three images, an anchor, a positive sample, and a negative sample, where the anchor and the positive sample belong to the same class, and the anchor and the negative sample belong to different classes. To generate this triplet dataset, given a training dataset, a set of image triplets (a, p, n) is created where a is the anchor image, p is a positive sample and n is a negative sample. The set of all possible combinations of generated triplets (a, p, n) is defined as $T = (a^{(i)}, p^{(i)}, n^{(i)})$. To train the CNN using these generated triplets, a triplet-loss function is employed. This loss function, initially proposed in [Schroff et al., 2015b], calculates the distance between the anchor and positive sample, and the distance between the anchor and negative sample. The goal is to minimize the distance between the anchor and positive sample while maximizing the distance between the anchor and negative sample, thus allowing the CNN to learn the underlying relationships between the images in the triplets and produce a more accurate embedding of the data.

During the training stage, the objective is to optimize the CNN's ability to discriminate between different classes of images. To accomplish this, a global triplet-loss function $L(T)$

is used, which calculates the loss over all the triplets in the dataset. The goal is to minimize this loss function over the entire dataset, as represented in (Eq.3.1).

$$L(T) = \sum_{(a,p,n) \in T} \max(0, loss + \alpha) \quad (3.1)$$

where $loss$ is the triplet-loss calculated on one triplet represented in (Eq.3.2).

$$loss = \|Net(a) - Net(n)\|_2^2 - \|Net(a) - Net(p)\|_2^2 \quad (3.2)$$

where the neural network function, represented as $Net(\cdot)$, is responsible for converting an image into its embedding representation and α is an empirical value that represents the margin enforced between positive and negative pairs. The triplet-loss function compares the distances between the anchor image, positive image and negative image. The idea behind it is to ensure that the distance between the anchor image and the positive image is smaller than the distance between the anchor image and the negative image by a margin α . This margin is added to make sure that the embeddings of the anchor and positive images are closer than the embeddings of the anchor and negative images. The function enforces this constraint by computing the difference between the distances and adding the margin, if the difference is less than the margin the function penalizes this by adding the difference to the final loss function. α is chosen through experimentation, and it is used to fine-tune the performance of the CNN. In other words, it allows adjusting the trade-off between allowing some negative images to be closer to the anchor image than the positive image and the risk of missing some positive images that are farther away from the anchor image.

3.3.2 Word spotting method

In this subsection, we provide an overview of the process of word spotting. As illustrated in

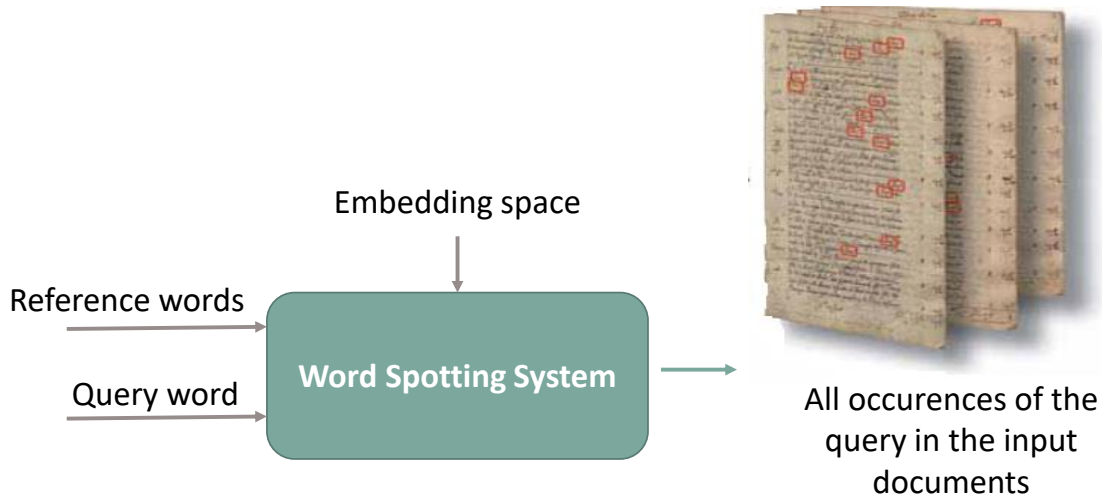


Figure 3.4: Word spotting flowchart.

Figure 3.4, our proposed word spotting process is described as follows:

First, the technique takes as input a query word image and a set of reference word images, which have been pre-segmented from the dataset. The reference words are used as a set of

predefined templates. The input images are then projected onto the embedding space, which is created using the CNN and the triplet-loss function previously described. The embedding space is a reduced representation of the original feature space, and it captures the essential characteristics of the input images while discarding unnecessary information.

In the next step, the query word image and the reference word images are compared in the embedding space. The comparison is performed by calculating the Euclidean distance, defined by (Eq.3.3), between the embedding representations of the query word and the reference words. Finally, the technique outputs all reference words that are similar to the query word. The output is a list of reference words that are closest to the query word in the embedding space.

$$d = \text{Euclidean}(Net(q), Net(r)) \tag{3.3}$$

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

3.4 Experimental Results

3.4.1 Dataset

The VML-HD dataset, proposed in [Kassis et al., 2017], is a recently developed dataset that consists of five books with handwritten Arabic scripts from the years 1088-1451. It is composed of 680 pages, containing a total of 121,636 sub-words. Each page of the dataset is associated with bounding boxes that correspond to the different sub-words. Figure 3.7 provides examples from the dataset.



(a) Example of pages from each book.

(b) Example of segmented data.

Figure 3.7: Examples of images extracted from the VML-HD database.

In this dataset, a word is defined as an entity that has meaning in a sentence, and it can be composed of one or more sub-words as depicted in Figure 3.7 (b). Each bounding box is annotated with the corresponding sequence of characters, which allows for the recognition of the sub-words. However, it is important to note that the distinction between a word and a sub-word is not considered, and they are simply referred to as "words" in the dataset.

3.4.2 Evaluation Protocol

3.4.2.1 Dataset

In order to conduct a fair and accurate comparison with the results presented in [Barakat et al., 2018], the same evaluation protocol was employed in this study. Specifically, only the first book of the VML-HD dataset was used for training, and all five books were used for testing. To further evaluate the performance of the proposed method, the first book was divided into three sets: training, validation, and test. The test set was composed of words that were completely different from the words used in the training set. This partitioning of the first book is illustrated in Table 3.1. The purpose of the validation set is to provide an unbiased estimate of the model’s performance during the training stage, this set is used to tune the hyperparameters of the model. It is important to note that for all stages of the evaluation, words and samples were randomly selected to ensure that the results were not biased. The random selection of the samples and words is important to avoid any bias in the results and to make sure that the performance of the model is generalizable.

Table 3.1: First book dataset partition for model training.

Dataset	#words	#Samples/word
Training	144	10
Validation	80	10
Test	21	100

3.4.2.2 Training process

The training dataset used in this study consists of 144 word classes, and for each class, 10 samples are randomly selected. Triplets are generated from this dataset to train the CNN. A triplet consists of an anchor image, a positive image, and a negative image. The anchor and positive images belong to the same class, while the negative image belongs to a different class.

To generate the triplets, the training dataset is first divided into positive pairs by selecting all possible combinations of two images from the 10 samples of each class. In total, there are C_{10}^2 possible combinations of positive pairs, of which only 42 are selected. Next, for each positive pair, 42 triplets are formed by adding to this pair a negative sample randomly selected from the other word class samples. This random addition is repeated 42 times to generate a total of 42 triplets per positive pair. Finally, the training dataset is composed of $(144 \times 42 \times 42)$ triplets. This large number of triplets ensures that the CNN is exposed to a wide variety of training examples, which helps the network to learn the underlying relationships between the images and produce a more accurate embedding of the data.

In the training stage of the proposed method, an Adam optimization algorithm was employed to adjust the parameters of the CNN. The Adam algorithm is a widely used optimization algorithm that is well suited for large datasets and deep neural networks. It is a combination of two other optimization algorithms, namely, AdaGrad and RMSprop. It adapts the learning rate of each parameter based on the historical gradient information, which helps to

improve the stability of the optimization process. The learning rate of the Adam algorithm, which determines the step size during the optimization process, was set to 10^{-3} . A batch size of 512 and 100 epochs were used. The batch size is the number of samples processed before the model’s internal parameters are updated. The number of epochs is the number of times the learning algorithm will work through the entire training dataset. All parameter values were empirically selected based on the training dataset. Empirical selection means that the values were chosen through experimentation, by evaluating the performance of the model on the training dataset. This process is commonly used to fine-tune the performance of the model and ensures that it generalizes well to new data.

3.4.2.3 Evaluation metrics

The proposed approach is evaluated using two performance metrics: Precision at the top- K -retrievals ($P@K$) and the mean Average Precision (mAP). These metrics are commonly used to evaluate the performance of information retrieval and image retrieval systems. The $P@K$ metric, proposed in [Deng et al., 2011], is a measure of the proportion of relevant items in the top- K retrieved items. An item is considered relevant if it belongs to the same word class as the query word. The $P@K$ metric is commonly used to evaluate the effectiveness of retrieval systems in terms of their ability to retrieve relevant items in the top- K retrieved items.

The mAP metric, proposed in [Everingham et al., 2010], is a measure of the mean Average Precision over all queries and all ranks K . The Average Precision is the average of the precision values at the point where a relevant item is retrieved. The mAP metric is commonly used to evaluate the effectiveness of retrieval systems in terms of the average performance across all queries and all ranks.

3.4.3 Experimental results

Table 3.2 and Figure 3.8 present the results of our proposed approach in terms of $P@K$, compared to the Siamese-based embedding space proposed in [Barakat et al., 2018]. To the best of our knowledge, [Barakat et al., 2018] is the only work that has been evaluated on the VML-HD dataset. The results presented in Table 3.3 and Figure 3.9 show the performance of our proposed approach in terms of mAP on the VML-HD dataset, and they also are compared to the results obtained by the Siamese-based embedding space proposed in [Barakat et al., 2018]. The comparison of our results with the results of the previous work allows to evaluate the effectiveness of our approach and to identify its strengths and weaknesses.

The results for $P@K$ metrics also illustrated in Figure 3.8 are limited to the first five ranks (from $P@1$ to $P@5$). Furthermore, the mean of the $P@K$ metric is reported for all five books included in the dataset.

Several observations could be drawn from the obtained results. The first observation concerns the results according to the mAP metric. Our approach achieves an average improvement of 7% over all the 5 books compared to [Barakat et al., 2018]. This improvement varies from 0% to 17%. It is significant given the large size of the test dataset.

Another observation that can be made from the results is the variation in mAP across the different books in the dataset. For example, the mAP for Book1 was found to be 77%, which is close to the highest mAP achieved (79%). This can be attributed to the fact that our

Table 3.2: Results on VML-HD dataset according to $P@K$.

$P@K$	Siamese [Barakat et al., 2018]	Our approach
$P@1$	0.89	0.90
$P@2$	0.85	0.89
$P@3$	0.86	0.89
$P@4$	0.89	0.88
$P@5$	0.89	0.89
$mP@K$	0.87	0.89

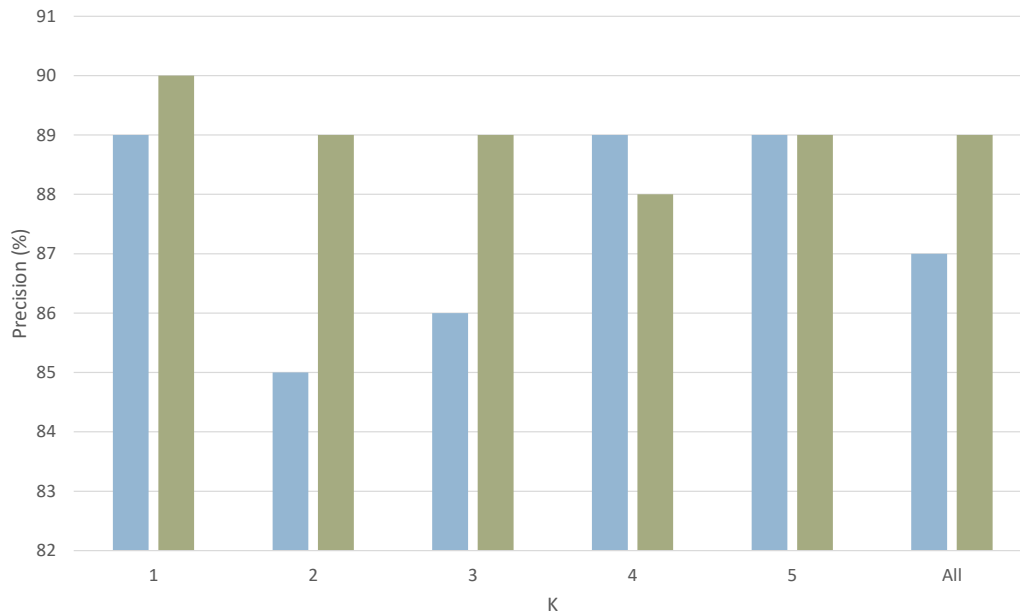


Figure 3.8: Comparison of the performances of the Siamese and Triplet networks based on $P@K$ metric.

Table 3.3: Results on VML-HD dataset according to mAP .

Book	Siamese [Barakat et al., 2018]	Our approach
Book1	0.65	0.77
Book2	0.68	0.72
Book3	0.66	0.69
Book4	0.69	0.69
Book5	0.62	0.79
All	0.66	0.73

triplet-CNN model was trained specifically using document images from Book1. Similarly, at Book5, a mAP of 79% was achieved, which can be explained by the characteristics of the words in the book. They are relatively simple and composed mostly of only two letters with a clear background. These characteristics may have helped the triplet-CNN to better

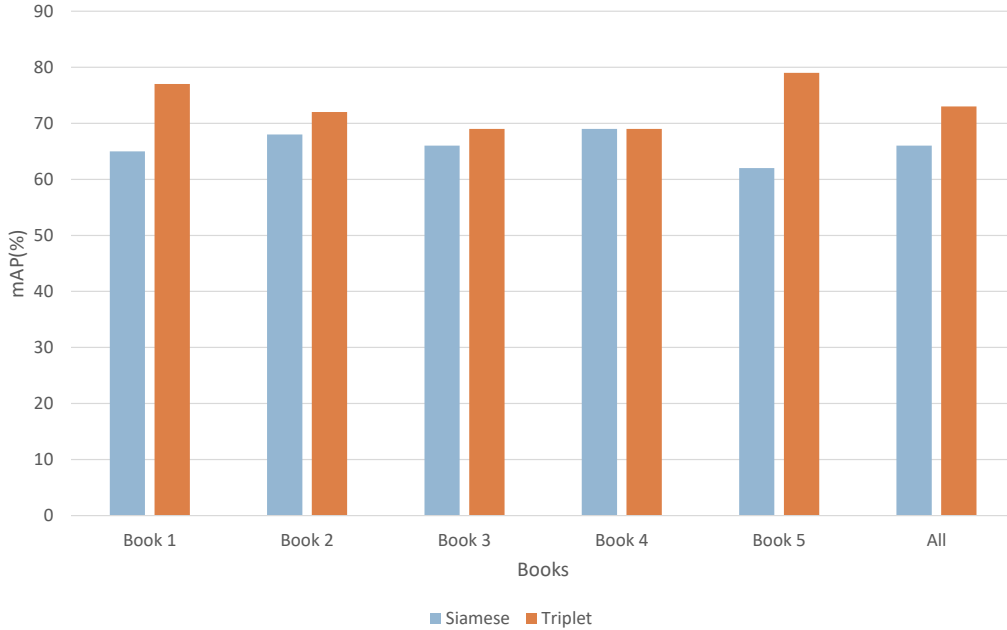


Figure 3.9: Comparison of the performances of the Siamese and Triplet networks based on mAP metric.

distinguish the words of this book from those of the other books in the dataset.

A further observation that can be drawn from the results is related to the precision at K ($P@K$) metric. Our proposed approach was found to outperform the Siamese CNN-based method by 1% at rank 1 across all test datasets. Furthermore, at ranks 2 and 3, our approach achieved improvements of 4% and 3% respectively. However, at ranks 4 and 5, there was no improvement observed. This discrepancy may be attributed to the small number of words used in the training phase. As stated in [Wang et al., 2017], the triplet architecture tends to perform better when a large number of words are used in the training set, with a smaller number of samples per word. This suggests that increasing the number of words used in the training phase may improve the performance at higher ranks.

Despite utilizing a relatively large number of words in our work, specifically 144 word classes with 10 samples per word class, equivalent to 254016 training triplets, this number may still be considered insufficient to fully exploit the capabilities of the triplet architecture. However, it should be noted that increasing the number of words used in the training phase would have been constrained by the limitations of our GPU performance. This highlights the importance of the computational resources available when utilizing such architectures in order to achieve optimal performance.

3.4.4 Further analysis

In this section, we investigate the impact of different embedding models on the performance of word spotting in historical Arabic documents. Our goal in this study is to evaluate and compare the performance of our proposed model against various techniques used for representing word image embeddings in word spotting systems. Our aim is to provide a comprehensive analysis of the effectiveness of these different embedding models in word spotting,

with the ultimate objective of identifying the best model for this task. In order to achieve this, we employ dimensionality reduction techniques to reduce the high-dimensional data to a smaller set of features. There are several dimensionality reduction and feature extraction techniques that have been proposed for image representations, which create new embedding spaces to represent the data. These techniques can be broadly categorized into two groups: linear methods, such as Principal Component Analysis (PCA) [Pearson, 1901] and Linear Discriminant Analysis (LDA) [Balakrishnama and Ganapathiraju, 1998], and non-linear techniques such as the Siamese CNN [Barakat et al., 2018] and our proposed triplet-CNN model.

3.4.4.1 Training process

As outlined in Table 3.1, the training set is composed of 143 classes, each containing 10 images. To train the Principal Component Analysis (PCA) model, 1,430 images with a size of 60x110 were used. The eigenvalues and eigenvectors are then determined from this dataset. The eigenvalues are sorted in decreasing order and the explained variance is used to identify the number of principal components that should be selected for the new feature subspace. A variance of approximately 61% is considered as a sufficient percentage for PCA, which results in a dimension reduction from the original 6600 dimensions to 29 principal components.

In the case of the LDA technique, the same dataset partition is utilized. An LDA model is trained using a set of 1,430 images and subsequently tested on a separate set of 2,100 images. The LDA technique is a dimensionality reduction method that aims to maximize the separation between classes in a feature set. In this case, the number of classes is 143, and thus the number of LDA components will be equal to $143-1 = 142$. This reduction of dimensionality allows for a more efficient representation of the data while preserving important information.

Regarding the Siamese method, to train the model, a set of image pairs is created. These pairs are defined as either positive or negative, depending on their class association. A positive pair is composed of two images from the same class and is associated with the label "0". Conversely, a negative pair is composed of two images from different classes, and is associated with the label "1". The model is trained using an equal number of positive and negative pairs, which allows the model to learn the similarity and dissimilarity between the images in the dataset. The Siamese model is trained using the same CNN architecture that is employed in the training of the triplet model. This allows for a fair comparison between the two models in terms of the underlying architecture, and any differences in performance can be attributed to the specific training method used.

3.4.4.2 Results

The results of our proposed model evaluated on the VML-HD dataset, according to the mAP metric, are presented in Table 3.5 and illustrated in Figure 3.11. On the other hand, the results according to P@K metric are presented in Table 3.4 and illustrated in Figure 3.10. These tables and figures provide a comprehensive evaluation of our proposed model for a clear comparison with other methods.

Our results, as presented, provide a detailed analysis of the performance of our proposed

Table 3.4: Results according to $P@K$.

P@K	Siamese [Barakat et al., 2018]	Triplet	PCA	LDA
P@1	0.89	0.90	0.88	0.77
P@2	0.85	0.89	0.89	0.75
P@3	0.86	0.89	0.87	0.74
P@4	0.89	0.88	0.87	0.70
P@5	0.89	0.89	0.85	0.68

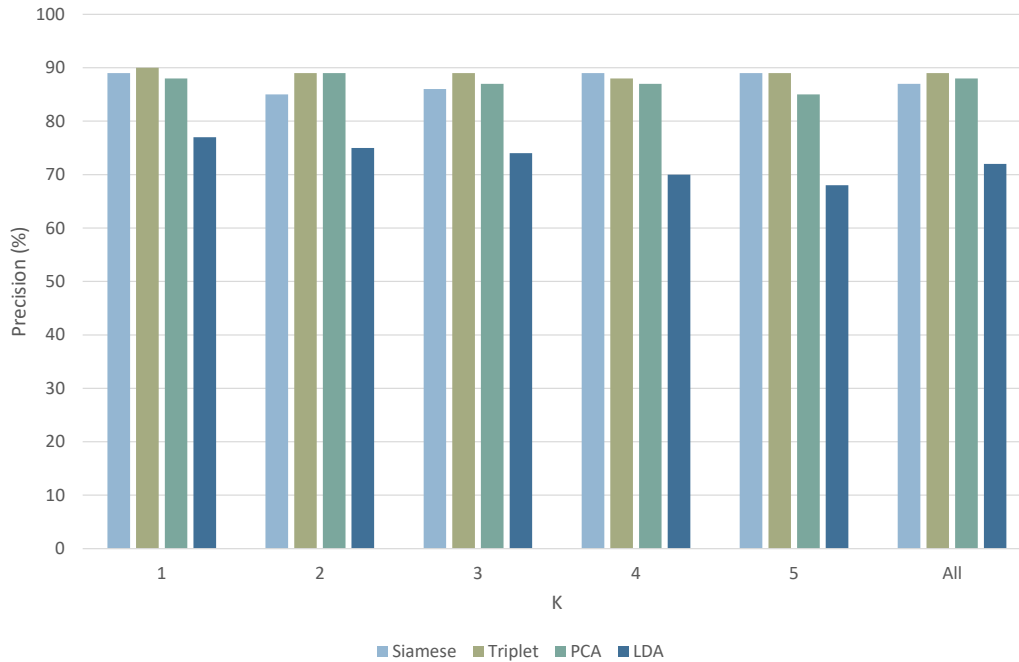


Figure 3.10: Comparison of performances of embedding methods based on $P@K$ metric.

Table 3.5: Results according to mAP .

Book	Siamese [Barakat et al., 2018]	Triplet	PCA	LDA
Book1	0.65	0.77	0.54	0.42
Book2	0.68	0.72	0.57	0.38
Book3	0.66	0.69	0.66	0.46
Book4	0.69	0.69	0.47	0.33
Book5	0.62	0.79	0.48	0.32
All	0.66	0.73	0.54	0.38

model on the VML-HD dataset, by evaluating the mAP metric on a per-book basis. Additionally, the results of the $P@K$ metric are presented specifically for the top five ranks. The results of the experiments indicate several noteworthy observations. The first observation

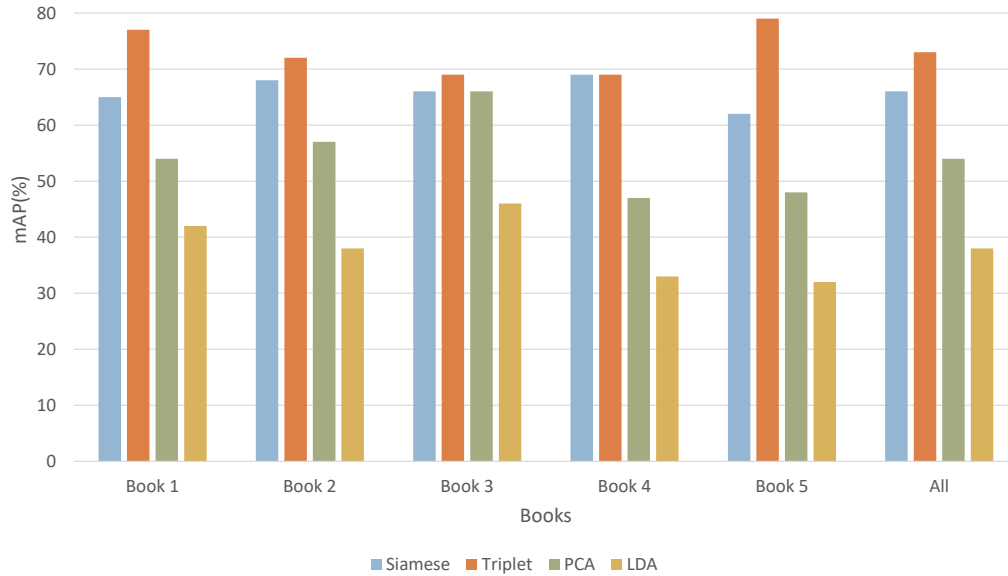


Figure 3.11: Comparison of performances of embedding methods based on mAP metric.

pertains to the performance of the models as measured by the P@K metric. At rank 1, the PCA and Siamese model achieve rate of 88% and 89% respectively, while the LDA method reached a rate of 77%, and our proposed Triplet model exhibits the highest rate of 90%. This can be attributed to the CNN architecture used in the Siamese and Triplet methods, which effectively extract features from word images. Furthermore, as seen in Table 3.4, PCA, one of the top dimensionality reduction algorithms, outperforms LDA. This can be explained by the small number of samples per class (10 images/class for the training stage) used in the experiment, as LDA is known to perform better with larger datasets and multiple classes.

Another significant observation is related to the performance of the models as evaluated by the mAP metric. The triplet-CNN model was found to be the most effective in separating words across all datasets, achieving a mean mAP rate of 73%. The highest mAP rate was observed on the Book5 dataset, with a rate of 79%. Similarly, the Book2 and Book1 datasets also exhibited high mAP rates of 72% and 77%, respectively. Additionally, both the Siamese and PCA techniques demonstrated good performance across all datasets, with mAP rates ranging from 66% to 54%, respectively. It is worth noting that PCA is commonly employed for dimensionality reduction, and is often utilized in conjunction with other methods to enhance its main advantage [Xu and Gowen, 2019].

3.5 Conclusions

In this chapter, we examined the efficacy of learning feature representations to enhance word spotting performance in the context of historical Arabic documents. We proposed an approach that involves the construction of an embedding space in order to achieve more effective feature extraction for describing word images. The proposed method consists of two main steps. Firstly, based on a pre-segmented training dataset, we constructed an embedding space utilizing the triplet-loss within the framework of a convolutional neural network. The objective of this step was to maximize the distance between word images associated

with different classes and minimize the distance between word images associated with the same class. Secondly, to spot a query word, images are projected on the previously learned embedding space and then matched using the Euclidean distance. Despite the complexity of historical Arabic documents in terms of variations in backgrounds and writing styles, the experiments, performed on the VML-HD dataset, demonstrate that the Triplet CNN provides superior results compared to its closest competitor, the Siamese CNN. Furthermore, we investigated different dimensionality reduction techniques to identify the best image representations in the context of historical Arabic documents. We presented a comparative study of four embedding methods: Triplet CNN, Siamese CNN, PCA and LDA.

In the following chapter, we will present new strategies aimed at improving the performance of the proposed Triplet CNN model. The ultimate goal of these efforts is to further optimize the performance of the Triplet CNN model for word spotting in historical Arabic documents.

CHAPTER 4

Enhancement Strategies for Word Spotting in Historical Documents

4.1	Introduction	59
4.2	Related Work	60
4.2.1	Word spotting	60
4.2.2	Transfer Learning	61
4.2.3	Triplet loss	62
4.3	Enhancement strategies for word spotting in historical documents	63
4.3.1	Word spotting process	64
4.4	Experiments	67
4.4.1	Experimental setup	67
4.4.2	Results	68
4.5	Qualitative Evaluation Analysis	70
4.6	Ablation Study	71
4.7	Conclusion	73

4.1 Introduction

In recent years, significant technological advances have led to a proliferation of digitized historical documents, both handwritten and printed. The process of digitization has made it possible to preserve and access a vast amount of historical information. However, it is widely recognized that digital historical documents, in their raw form, pose significant challenges for computer vision-based processing and analysis. This is due to variations in writing styles, document degradation, and other factors that make it difficult for computer vision tools to automatically interpret and understand the content of these documents. Therefore, there is a need to transform these digital historical documents into a more usable format to facilitate their processing and analysis by computer vision tools.

The analysis and processing of historical documents is a complex endeavor, often hindered by the poor physical condition of the manuscripts. The degradation of these documents over time, due to factors such as wrinkles, dust damage, nutritional stains, and discoloration from exposure to sunlight, can significantly impair their readability and comprehensibility [Farghaly and Shaalan, 2009]. Therefore, the implementation of automated pre-processing techniques is crucial in order to effectively extract and interpret the valuable information contained within these documents. This is particularly important for documents that are of significant cultural or historical significance, as it allows for their preservation and accessibility to both human and computer vision systems.

The processing of Historical Arabic Documents (HADs) poses a significant challenge due to the intricacies of the Arabic script and the diversity of its forms. Over time, these documents undergo various forms of degradation which further complicates the task. Additionally, a significant number of digitized HADs are unlabelled, particularly at the word level. In light of these challenges, we propose a pre-processing step based on Generative Adversarial Networks (GANs) for degraded document enhancement. The use of GANs in this manner aims to overcome the complexity of historical documents and improve their legibility when compared to existing methods.

The ability to learn effective representations from input data is crucial for addressing any pattern recognition problem. The extraction of useful information and construction of a suitable representation for word images in historical documents are essential for achieving improved performance and gaining insights into the specific challenges of these documents. Deep embedding approaches aim to create an embedding space that transforms the input image into new representations by selecting relevant features. However, these approaches can be hindered by the need for regular updating of offline mined triplets. To circumvent these limitations, an online learning approach can be applied that selects the most appropriate triplets during the training process. This approach utilizes mini-batches of data, generating more triplets per batch of inputs, and eliminates the need for offline extraction. In this context, we propose a process that focuses on learning features from word images through a triplet-CNN, incorporating an online learning procedure and a semi-hard triplet selection strategy. This approach allows for the selection of the most appropriate triplets during the training process, thus providing a more effective representation of the input data and improving the performance of the historical document analysis.

A significant challenge facing research in HADs is the limited availability of annotated public datasets for evaluating word spotting approaches. Despite the availability of various public datasets for document analysis, there is a scarcity of annotated historical documents

for assessing the performance of these approaches. To mitigate this limitation, the transfer learning technique has been proposed as a viable solution [Mohammed et al., 2022, Pramanik and Bag, 2021]. This technique involves utilizing knowledge acquired from similar historical scripts or different scripts to improve the word spotting process. In light of this, we adopt a transfer learning technique by incorporating knowledge from a similar historical script and a different script in order to investigate the effect of this technique on improving the word spotting process. This approach allows the model to leverage pre-existing knowledge and improve the efficiency of the model training. Our contributions have been submitted to a regarded journal and published in an international conference [Fathallah et al., 2023b].

The remainder of this chapter is organized as follows. In section 4.2, a comprehensive review of the relevant literature on the topic is presented. Section 4.3 provides a detailed description of the proposed enhancement strategies for word spotting in HADs. The experimental study and results are presented in Section 4.4. A thorough analysis and discussion of the improvement strategies are provided in Section 4.5. Finally, the conclusions and future perspectives of the research are outlined in Section 4.7.

4.2 Related Work

In this section, we present a comprehensive examination of the various techniques utilized to improve the performance of word spotting systems in historical documents. We begin by examining the current state-of-the-art in the field of enhancing word spotting systems. Subsequently, we provide an overview of transfer learning tools and their applications. Finally, we delve into the concept of triplet loss and its various applications.

4.2.1 Word spotting

Word spotting in historical document images can be utilized to exploit document content in digital form. It delineates different query word occurrences in such document sets. 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 SVM and Regularized Discriminant Analysis (RDA) classifiers to increase the performance of closed lexicon word spotting systems. They achieved an enhancement in the precision rate of 4%. Additional methods have been used to boost the performance of word retrieval approaches presented in [Westphal et al., 2020, Sudholt and Fink, 2018, Gurjar et al., 2018]. In [Sudholt and Fink, 2018], a novel method was proposed to improve word spotting performance through the application of data augmentation techniques. The authors demonstrated that the proposed method was able to effectively enhance the performance of word spotting models. On the other hand, Gurjar et al. in [Gurjar et al., 2018] proposed the use of pre-training CNN architecture using a synthetic dataset [Krishnan and Jawahar, 2016]. The authors showed that this approach was able to achieve improved word spotting performance, despite the limited amount of training samples available. This study highlights the utility of pre-training techniques and synthetic data in enhancing the performance of deep learning models for word spotting tasks. On the other hand, Westphal et al. in [Westphal et al., 2020] demonstrated the effectiveness of sample selection approaches in the context of word spotting. Specifically,

they reduced the amount of training data required in the word spotting training stage by utilizing PHOC representation. Moreover, other studies [Tushar et al., 2018, Can and Kabadayi, 2020, Khayyat and Elrefaei, 2020] in the field of historical document processing have employed transfer learning techniques [Weiss et al., 2016] as a means of leveraging acquired knowledge from a source task to improve performance on a target task. These studies have shown that transfer learning can be a powerful tool in improving the performance of deep learning models for historical document processing tasks.

4.2.2 Transfer Learning

In order to improve the extraction and analysis of information contained within historical documents, a number of approaches have been proposed that focus on developing more advanced word spotting models through the utilization of transfer learning. This is done by leveraging pre-existing knowledge acquired from similar or different scripts, which can improve the performance of the word spotting process in historical documents.

In the field of document similarity detection, the HWNet model, as described in the study by [Krishnan and Jawahar, 2016], utilizes a convolutional network architecture to effectively compare two distinct documents written by different authors. The model was initially trained on a synthetic dataset, known as HW-SYNTH, which consists of 750 publicly available handwritten characters. To improve the performance of the model, the weights were fine-tuned using additional handwritten datasets. A more advanced version of the HWNet, referred to as HWNet v2 and reported in [Krishnan and Jawahar, 2019], was later developed. This variant features an adaptation of the ResNet-34 architecture, including the incorporation of region-of-interest pooling layers, which allows for the efficient analysis of images of varying sizes. In this study, a synthetic dataset was created using publicly available handwritten character images, comprising approximately one million word images. This extensive dataset eliminates the need for any data augmentation techniques. In the study by [Lladós et al., 2012], the authors sought to address the challenge of handwriting recognition in historical documents, where little or no training data is available. They recognized that the variations in handwriting style, which can be influenced by factors such as the time period and geographical location, pose significant difficulties for recognition. As an alternative, other researchers have proposed the use of transfer learning (TL) as a means of addressing this issue. For example, authors in [Can and Kabadayi, 2020, Khayyat and Elrefaei, 2020] have proposed the use of TL as a method for processing historical documents. The approach involves leveraging information acquired from a source task to improve performance on a target task. This can be particularly useful in cases where data is limited or not available, such as in the case of historical documents. Transfer learning, as a technique in deep learning, has gained significant attention in recent years due to its ability to leverage knowledge learned from one task to improve the performance of another related task. One particularly interesting variant of transfer learning is transductive transfer learning, as proposed in [Pan and Yang, 2009]. This approach focuses on adapting a model trained on one set of data, referred to as the source domain, to perform well on a different set of data, referred to as the target domain, for the same task.

In this chapter, we aim to expand upon this method by incorporating an increased number of annotated data sources and introducing the concept of parameter transfer. By utilizing a larger number of sources, we anticipate a further improvement in performance on the target

domain. Additionally, by introducing parameter transfer, we aim to optimize the adaptation process and enhance the overall performance of the model on the target domain.

4.2.3 Triplet loss

The triplet loss, introduced in the previous chapter, is a loss function that includes three distinct instances within the same network and shared parameters, as depicted in Figure 4.2. Formally, assuming that the three inputs are denoted by x , x^+ and x^- , and the embedded representation of the network by $f(x)$, the loss function for the triplet is given by Eq.(4.1):

$$L(x, x^-, x^+) = \max[\|f(x) - f(x^+)\|_2^2 - \|f(x) - f(x^-)\|_2^2 + \alpha] \quad (4.1)$$

where α is a margin that is applied between positive and negative pairs. The triplet loss encourages the network to learn features that can differentiate between the positive and negative samples, while also preserving the relative ordering of the embeddings within the same class.

In offline triplet mining, a pre-selection process is employed where the triplets are chosen and stored in a dataset prior to the training of the model. This is typically accomplished by initially selecting a substantial number of data points, followed by the utilization of clustering algorithms to group similar data points together. Once these clusters have been formed, triplets can be chosen by randomly selecting an anchor from one cluster, a positive sample from the same cluster, and a negative sample from a different cluster. As an illustration, authors in [Ustinova and Lempitsky, 2016] proposed a modified version of the triplet loss known as the histogram loss. The aim of this modification was to enhance the diversity of the learned feature representations. The loss is calculated by estimating two distributions of similarities for positive and negative sample pairs. Then, the probability of a positive pair having a lower similarity score than a negative pair is computed based on the estimated similarity distributions. This modification allows for a more comprehensive and diverse feature representation to be learned. Additionally, in [Ma et al., 2021], the authors presented a comprehensive examination of recent advancements in deep similarity learning, with a focus on triplet loss and its variants and modifications. The authors thoroughly analyzed the literature, and discussed the usage of both offline methods as a means of selecting triplets for training. They also provided an in-depth analysis of the strengths and limitations of each method, providing valuable insight for practitioners in the field.

In online triplet mining, the triplets are generated on the fly during training. This approach is typically used when the dataset is too large to be pre-selected or when new data is constantly being added to the dataset. The model is trained on a batch of data, and a set of triplets are selected from that batch, with the anchor and positive sample coming from the same class and the negative sample coming from a different class. Authors in [Schroff et al., 2015b] proposed the use of triplet loss in a deep CNN architecture to learn feature representations that are robust to variations in pose, lighting, and other factors. To achieve this, they use an online mining method to select the triplets during training. This process starts by first training the model on a large dataset of labeled faces, then using the trained model to generate embeddings for each face. Then, during training, for each image in a batch of images, the model generates an embedding and using a function to compare the embedding to every other embedding, the closest and most distant embeddings are selected as positive

and negative samples, respectively. The triplets are then used to compute the triplet loss and update the model’s parameters. In [Wang et al., 2022], authors aimed to establish a learnable margin for each class, in order to better maintain the intra-class variance in the final embedding space. Additionally, a loss function between proxies was also introduced, to enhance discrimination between classes and further preserve the intra-class distribution.

4.3 Enhancement strategies for word spotting in historical documents

In this section, we present our proposed approach for addressing the task of word spotting in historical document images using deep learning techniques. Our approach is an end-to-end framework that incorporates various stages, as illustrated in Figure 4.1. The initial step in our proposed framework involves the utilization of a conditional GANs based image enhancement model to enhance the quality of the input document images. This enhancement process is accomplished by training the GANs on a dataset of degraded document images, and then using the trained GANs to improve the quality of new input images. The technical details and implementation of this enhancement process will be discussed in greater depth in the subsequent chapter. The second step is a segmentation process, where we follow previous research such as [Almazán et al., 2014b, Fathallah et al., 2019, Barakat et al., 2018] to pre-segment the document images into individual words by utilizing their annotation XML files. We assume in this work that document images have been segmented into separate words, which are considered as candidates for matching the query word.

In the third step, we apply a novel triplet-CNN scheme for feature extraction and embedding space construction. Finally, in the last step, we match the embedding of the query word with each reference word using similarity distances.

To evaluate the effectiveness of our proposed framework, we conduct experiments on three publicly available historical image datasets: VML-HD and HADARA80P for the Arabic script, and George Washington for the Latin script. By analyzing the impact of our proposed approach on the performance of the word spotting model, we aim to demonstrate its effectiveness and potential for future applications.

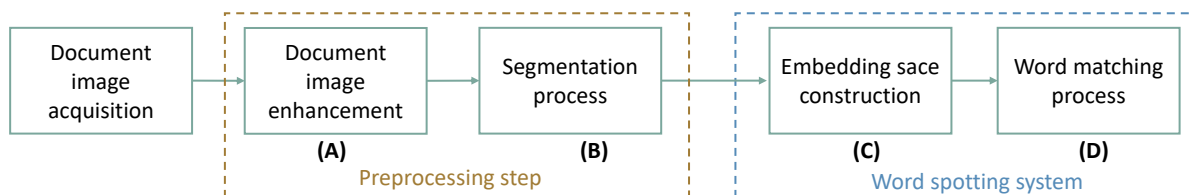


Figure 4.1: Illustration of general steps presented in our proposed approach: (A) An enhancement method based on conditional GANs, (B) Document images pre-segmented using their annotation XML files, (C) Triplet-CNN for feature extraction and embedding space construction, (D) Similarity distances to match embeddings of query word and each reference word.

4.3.1 Word spotting process

In recent years, various updated versions of the triplet loss have been proposed to improve the efficiency and extend the generalization capability of deep learning models. In this chapter, we investigate the role of triplet mining and transfer learning in enhancing the performance of the word spotting process in historical documents. Specifically, we propose a novel triplet mining strategy, based on the triplet loss, and explore how transfer learning can improve the learning model.

It is important to note that the performance of a word spotting algorithm is highly dependent on the performance of the segmentation algorithm. Therefore, in our proposed approach, we focus on the performance of the word spotting algorithm only and choose a pre-segmented dataset to eliminate the effect of segmentation on our results. For evaluation, we aim to investigate the impact of the proposed triplet mining strategy and transfer learning on the performance of the word spotting algorithm, by conducting a series of experiments and analyzing the results.

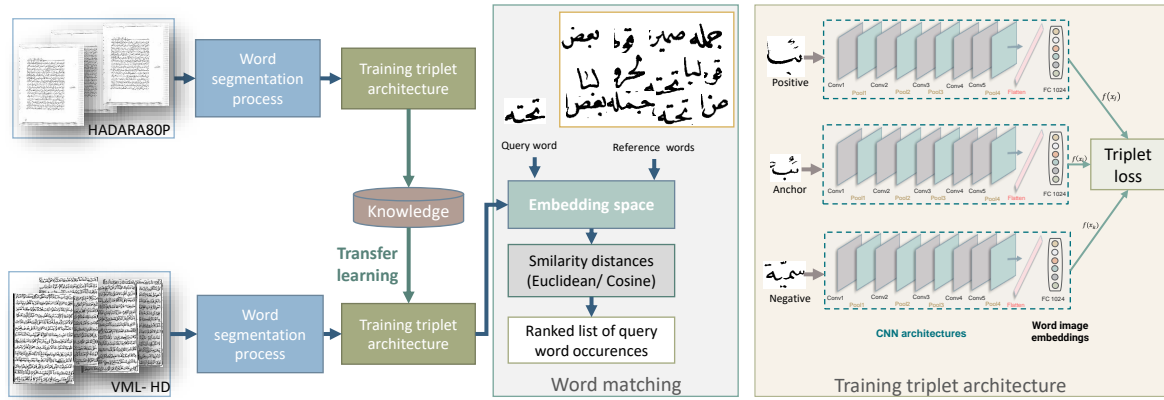


Figure 4.2: Proposed word spotting steps for historical documents.

4.3.1.1 Strategies of triplets mining

Triplet mining is a technique utilized to select a set of triplets, which consist of an anchor, a positive and a negative sample, for training a deep learning model with the triplet loss. There are two main approaches to triplet mining: online and offline. Online triplet mining, also referred to as dynamic triplet mining, selects triplets during the training process in real-time, based on the current state of the model. This approach allows the model to adapt to the current training data and improve the quality of the triplets. On the other hand, offline triplet mining, also referred to as static triplet mining, selects triplets before the training process begins, based on the entire training dataset. The pre-selected triplets are then used during the training process.

The construction of a triplet for a given anchor image x_a in a deep learning model requires selecting a positive image x_p that belongs to the same class as the anchor, and a negative image x_n that belongs to a different class. For a dataset with N training images, the number of possible triplets that can be constructed is $O(N^3)$. However, it is important to note that a large number of these triplets may not be useful for the learning convergence. For example, triplets where $d_{an} \gg d_{ap}$, where d_{an} is the distance between the anchor and

negative samples and d_{ap} refers to the distance between the anchor and positive samples, may not be beneficial for the learning process. Therefore, it is crucial to select only the most useful triplets to ensure the rapid convergence of triplet-based networks.

In literature, there are several online triplet mining strategies available that are based on the distances of the data instances in mini-batches. These strategies are designed to select triplets during the training process in real-time.

- **Easy triplet mining:** This strategy is based on the selection of easy negative examples, which are defined as the least similar images that have a different label from the anchor image. However, it is important to note that this condition is not useful in triplet construction, as it will not produce useful gradients for updating the model. The reason being that the model is already able to correctly classify the easy negative examples, thus the triplet will not contribute to the model’s learning. This can lead to a suboptimal training process as the model will not be able to improve its ability to distinguish between similar images of different classes.
- **Hard triplets:** Hard triplets are defined as the triplets where the negative example is the closest sample to the anchor image in feature space among all samples that have a different class label from the anchor as defined by Eq.(4.2). Thus, the distance between the anchor and the negative sample is smaller than the distance between the anchor and the positive sample. This strategy is beneficial as it allows the model to focus on the most challenging negative examples, which are closest to the anchor in feature space, and can lead to better generalization of the model.

$$\|f(x) - f(x^+)\|_2^2 < \|f(x) - f(x^-)\|_2^2 \quad (4.2)$$

- **Semi-hard Triplet:** Semi-hard triplets are defined as the triplets where the negative example is further away from the anchor image in feature space compared to the positive example but still within a predefined margin as described in Eq.(4.3). This is achieved by selecting an anchor-negative pair that is further away than the anchor-positive pair but still within the margin, which leads to a positive loss. This strategy allows the model to focus on the samples that are challenging but not too difficult and can lead to better generalization of the model. By training on semi-hard triplets, the model is forced to learn representations that can separate similar images of different classes, which is important for the performance of the model.

$$\|f(x) - f(x^+)\|_2^2 < \|f(x) - f(x^-)\|_2^2 < \|f(x) - f(x^+)\|_2^2 + \alpha \quad (4.3)$$

where α is a margin that is applied between positive and negative pairs.

While the use of hard negative triplets can accelerate the convergence of the training process, it can also lead to poor learning due to the problem of mislabeling, which is unavoidable and critical in word retrieval tasks, particularly in the case of subjective historical document annotations. To overcome this problem, hard negative triplets need to be removed from the triplet network training. In the study by Schroff et al. [Schroff et al., 2015b], it was demonstrated that the use of semi-hard negatives can result in better performance than networks trained with either random or hard negatives. Based on this idea, we propose to investigate

an online semi-hard triplet strategy for the word retrieval task in historical documents. This strategy aims to select semi-hard negatives during the training process in real-time, based on the current state of the model, to achieve better performance while avoiding the problem of mislabeling.

4.3.1.2 Transfer learning

Transfer learning approaches have been shown to be effective in a wide range of contexts, as demonstrated in recent studies such as the one by Agarwal et al. in 2021 [Agarwal et al., 2021]. In the context of historical document analysis, the wide variety of scripts and writing styles present in HAD can make it challenging to train a word spotting model to be generalized. To address this problem, we propose to employ transfer learning techniques. Transfer learning involves reusing pre-trained networks that have been trained on more complex HADs and adapting them to the task of word spotting in HADs.

One of the main advantages of such a transfer learning approach is that it allows for more convergent learning, where supplementary knowledge and features are easily transferred to the word image representation task. This can improve the performance of the word spotting model, as it can leverage the knowledge of feature representations learned from the source task τ_s to the target task τ_t . In this study, we investigate the problems of word spotting enhancement using the transfer learning technique, with the goal of exploiting and leveraging the knowledge of feature representations from source task to target task.

Our contribution to the field of word spotting in historical documents consists of two key steps. Firstly, we train a triplet-CNN on a dataset of historical documents, specifically on the HADARA80P dataset which comprises 80 pages of historical Arabic handwriting. By training the model on this dataset, we aim to learn robust feature representations of historical Arabic handwriting. Secondly, we train the same triplet-CNN architecture on a different dataset of historical Arabic documents and apply transfer learning based on the models generated in the first step. By reusing the knowledge learned from the first step, we aim to improve the performance of the model on the second dataset, while also making it more generalizable to other historical Arabic handwriting datasets.

4.3.1.3 Matching process

In order to find the closest word images to a given query word, a word spotting system must compute a similarity score between all the word images in the test set (referred to as reference words) and the query word. Subsequently, it ranks all those images based on their similarity to the query, resulting in the top retrieved word images. A commonly employed strategy for learning these similarities is to learn representations, often referred to as embeddings, of images and queries in a shared vectorial space, known as the embedding space.

The embedding space construction involves projecting the images and queries onto a dense vector representation that captures the underlying characteristics of the image. This vector representation, also known as an embedding, can then be used to compute the similarity between the images and queries. By computing the similarity in the embedding space, the word spotting system can effectively compare and rank images based on their similarity to the query word.

In order to ensure that two word images are matched correctly, we investigate the use of

three similarity distances, namely Euclidean, Cosine and Manhattan, which are computed between the embedding representations of each image pair (query, reference word) in the embedding space. These similarity distances are commonly used in image retrieval tasks as they provide a measure of the similarity between two images based on their embedding representations. Euclidean distance, also known as L2 distance, measures the straight-line distance between two points in the embedding space. Cosine similarity, on the other hand, measures the cosine of the angle between two vectors in the embedding space, with a value of 1 indicating that the vectors are identical, and a value of 0 indicating that they are orthogonal. Manhattan distance, also known as L1 distance, measures the distance between two points in the embedding space as the sum of the absolute differences of their coordinates.

By evaluating the performance of the word spotting system using these different similarity distances, we aim to determine the most suitable similarity measure for the task at hand and improve the performance of the word spotting system.

4.4 Experiments

In this section, we present the main experimental evaluations to evaluate the effectiveness of our proposed approaches. To this end, we have conducted extensive experiments on various publicly available datasets to assess the performance of our proposed methodologies under different conditions and settings.

4.4.1 Experimental setup

In this study, we present an analysis of the results obtained from the three datasets: VML-HD [Kassis et al., 2017], HADARA80P [Pantke et al., 2014] and GW [Rath and Manmatha, 2007], which have been previously described in prior chapters. In the context of the word spotting process, Table 4.1 illustrates the division of the three datasets employed in our experiments. The GW dataset is exclusively utilized in the evaluation phase and is not incorporated in the training or validation stages. This design enables us to evaluate the generalizability of the proposed approach to various datasets. The table displays the number of images and the number of words in each dataset, providing a means to evaluate the performance of the proposed approach.

Table 4.1: Partitioning of datasets according to the level of the word classes and number of images per class.

Dataset	Sub-set	#words	#Samples/word
VML-HD	Train	141	10
	Val	20	10
	Test	105	100
HADARAP80	Train	40	30
	Val	40	10
	Test	50	10
GW	Test	85	10

4.4.1.1 Evaluation protocol

In order to evaluate the performance of our proposed word spotting system, we have selected two performance assessments, namely $P@K$ (Precision at the top- K -retrievals) [Deng et al., 2011] and mAP (mean Average Precision) [Everingham et al., 2010]. Both of these metrics are commonly used to evaluate the performance of image retrieval systems, as they provide a comprehensive evaluation of the system’s performance. They allow us to determine the trade-off between precision and recall and evaluate the effectiveness of our proposed approach.

4.4.1.2 Implementation details

The learning process of our proposed approach is performed using the stochastic gradient descent algorithm for optimization, with a learning rate of 10^{-3} and a batch size of 512. To prevent overfitting, an early stopping function with a patience of 50 is employed. Additionally, all parameter values are selected empirically through a process of experimentation and analysis. For all experiments, we use the PyTorch framework and the model is trained on an NVIDIA Quadro RTX 6000 GPU with 24 GB of RAM. This allows for efficient and fast training of the model, while also leveraging the power of the GPU to accelerate the computations. It should be noted that the choice of the batch size and learning rate, as well as the use of early stopping, are commonly used techniques to ensure the stability and the generalization of the model. The use of the Pytorch framework allows for a more efficient implementation of the proposed approach and the use of a powerful GPU allows for faster training of the model.

4.4.1.3 Word spotting model training

In online triplet mining, a batch of φ embeddings is computed from a batch of φ inputs. These embeddings are then used to generate triplets. However, it is important to note that most of these triplets are not valid, i.e. they do not contain two positives and one negative. To select valid triplets, we consider the labels of the input images, and for each three indices $i, j, k \in [1, \varphi]$, if examples i and j share the same label but are distinct and example k has a different label, then (i, j, k) is considered a valid triplet. A batch of input word images of size $\varphi = \omega\sigma$ is composed of ω different word classes with σ examples per word class. This gives a total $\omega\sigma(\sigma - 1)(\omega\sigma - \sigma)$ of triplets where $\omega\sigma$ represents anchors, $(\sigma - 1)$ denotes possible positives per anchor, and $(\omega\sigma - \sigma)$ represents possible negatives. Once the valid triplets are selected, the semi-hard strategy is applied to select triplets that are used for computing the loss. This involves combining the largest distance between the anchor and the positive samples, and the smallest distance between the anchor and the negative samples in the final triplet loss. The final loss is then computed as the average loss on the semi-hard triplets.

4.4.2 Results

In this section, we report and analyze the performance of the proposed strategies for different parameters applied to the word retrieval task. The results of the word retrieval task are evaluated after applying the preprocessing stage, online triplet mining, selection strategies, and transfer learning techniques.

Table 4.2 presents the results of our experiments on the VML-HD dataset in terms of $P@K$ and mAP metrics. To better highlight the impact of our suggested strategies on embedded-feature representation and matching, we have evaluated the framework on five books of the VML-HD dataset. These results provide an in-depth analysis of the performance of our proposed approach and demonstrate its effectiveness in enhancing the word retrieval task in historical documents.

Table 4.2: Results of our proposed model on the VML-HD dataset according to $P@K$ and mAP metrics.

Books	P@1	P@2	P@3	P@4	P@5	mAP
Book 1	1	1	1	1	1	0.82
Book 2	1	0.95	0.95	0.95	0.95	0.83
Book 3	1	1	1	1	1	0.79
Book 4	1	1	0.98	0.97	0.96	0.73
Book 5	0.95	0.98	0.97	0.98	0.98	0.77

To further demonstrate the performance of the proposed framework, additional results are presented in Table 4.3. These results are presented in terms of $P@K$ and mAP metrics, and were obtained using different datasets: VML-HD and HADARA80P for the Arabic script, and GW for the Latin script. The results reported in Table 4.3 provide a comprehensive evaluation of the proposed framework, by considering different datasets, script types and similarity distances. The results demonstrate the effectiveness of the proposed approach in enhancing the word retrieval task.

Table 4.3: Results of our proposed model on different datasets according to $P@K$ and mAP metrics.

Database	Distance	P@1	P@2	P@3	P@4	P@5	mAP
VML-HD	Euclidean	0.99	0.99	0.98	0.98	0.98	0.79
	Cosine	0.98	0.98	0.98	0.98	0.98	0.78
	Manhattan	0.98	0.98	0.98	0.98	0.98	0.77
HADARA80P	Euclidean	0.89	0.88	0.91	0.91	0.90	0.73
	Cosine	0.88	0.87	0.90	0.90	0.89	0.70
	Manhattan	0.88	0.87	0.90	0.90	0.89	0.70
GW	Euclidean	0.87	0.88	0.88	0.89	0.87	0.63
	Cosine	0.86	0.88	0.88	0.88	0.86	0.61
	Manhattan	0.86	0.85	0.85	0.86	0.84	0.59

In Table 4.4, we present a comparison of our proposed approach with other state-of-the-art methods designed for word retrieval, specifically those that are based on feature representations. The comparison is performed on the same test dataset from the VML-HD dataset. As previously demonstrated in Table 4.3, the Euclidean distance is the most appropriate similarity distance for matching embeddings, thus, the comparison results are based on this distance.

The results of the comparison clearly demonstrate that our proposed method outperforms existing methods by a significant margin, with a considerable improvement in the performance of the word retrieval task. According to the P@K metric, at rank 1, our proposed approach outperforms the methods proposed by [Barakat et al., 2018] and [Fathallah et al., 2019] by a margin of 10% and 9%, respectively. Additionally, for other ranks, an improvement range from 9% to 14% is achieved. The results obtained in terms of mAP also show a significant improvement, with our proposed approach achieving the best results compared to [Barakat et al., 2018] and [Fathallah et al., 2019] with an improvement of 13% and 6%, respectively. These results clearly demonstrate the effectiveness of our proposed approach in enhancing the word retrieval task in historical documents.

Table 4.4: Results on VML-HD according to P@K and mAP metrics using Euclidean distance: Comparison with state-of-the-art methods.

Methods	P@1	P@2	P@3	P@4	P@5	mAP
[Barakat et al., 2018]	0.88	0.85	0.86	0.89	0.89	0.66
[Fathallah et al., 2019]	0.90	0.89	0.89	0.88	0.89	0.73
Our	0.99	0.99	0.98	0.98	0.98	0.79

To thoroughly evaluate the performance of our proposed word spotting framework, we propose to conduct an error analysis on the VML-HD validation set in the following section. This will provide insights into the areas where the proposed approach may need improvement and help to identify potential areas for future research.

4.5 Qualitative Evaluation Analysis

The error analysis process is conducted on miss-retrieved word images in a validation set from the VML-HD dataset, where the P@K metric is considered to evaluate the model. The validation set is composed of 20 different word classes, each of which contains 10 samples.

Figure 4.3 presents examples of word classes that were miss-retrieved by the model in [Fathallah et al., 2019]. The figure displays the query word and the first five retrieved samples (from P@1 to P@5) and then the correct ranked occurrences of the query word. Some examples of mismatches predicted by the [Fathallah et al., 2019] model are depicted at the top of the figure, and the same examples predicted by the proposed model are shown at the bottom. This comparison allows us to visually evaluate the improvement of the proposed model over the existing method and provide insights into the model’s ability to correctly retrieve word images.

As shown in Figure 4.3, the proposed approach exhibits a significant improvement in the accuracy of word spotting in historical documents compared to our first model presented in [Fathallah et al., 2019]. Through an error analysis process conducted on a validation set from the VML-HD dataset, it can be observed that the proposed approach is able to correctly retrieve a larger number of words that were previously miss-retrieved by the [Fathallah et al., 2019] model. This is a result of the implementation of relevant strategies, namely the enhancement model, online triplet mining, selection strategies, and transfer learning tech-

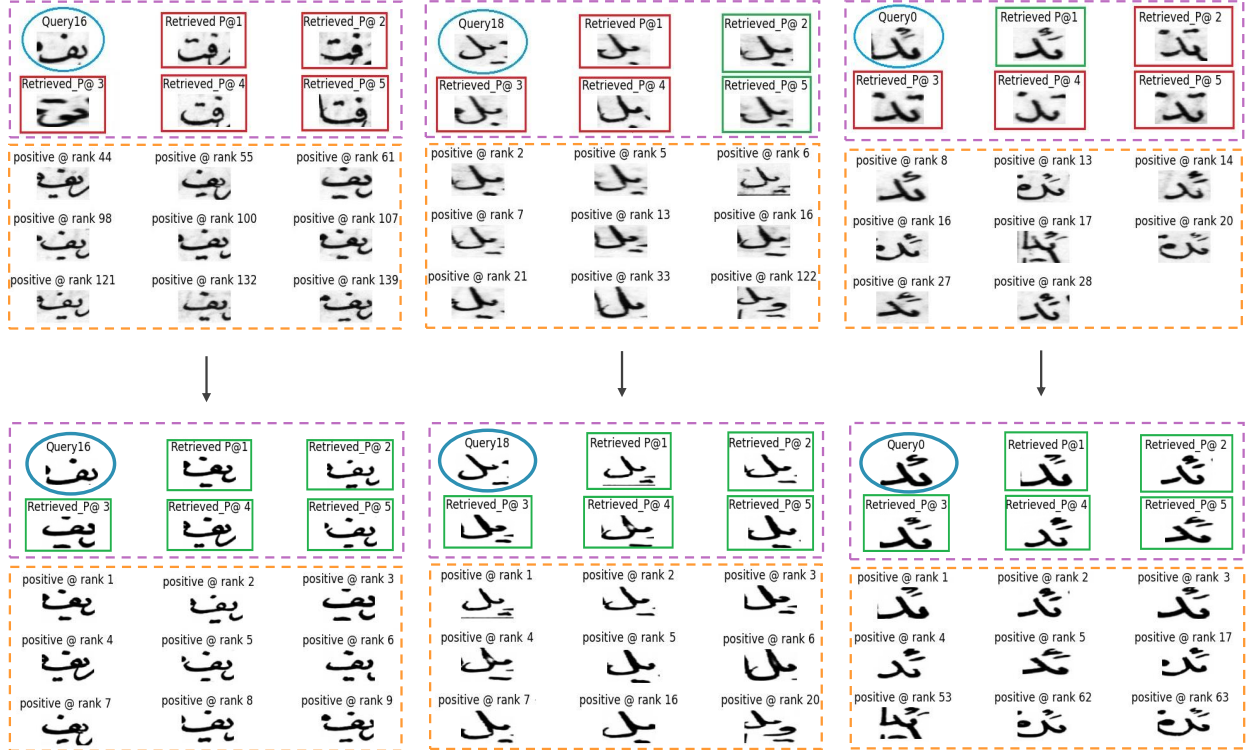


Figure 4.3: Some examples of miss-retrieved words: Error Analysis with displaying the first five ranks spotted (from P@1 to P@5) and the right occurrences ranks. On the top, the images without enhancement while on the bottom, the images with enhancement.

niques, aimed at improving the performance of the word spotting process in historical documents.

4.6 Ablation Study

Although our experiments show that the proposed method achieves competitive performance against existing state-of-the-art methods, we would like experimentally to prove the benefits brought by transfer learning (TL), online learning-based semi-hard triplets (OL) and document enhancement(DE). Here, we perform our ablation study to verify the advantage of each enhancement strategy for the task of word spotting on three popular databases, namely VML-HD, HADARA80P and GW.

First, the model is trained without any improvement strategy and it is observed that there is a significant decrease in the performance of the model. Then, to investigate the impact of each improvement strategy, all possible combinations of the three strategies (TL, OL and DE) are considered. 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 4.5.

We notice a considerable improvement in mAP (6% additional), using all three proposed strategies. We assume that this can be explained by the following phenomena. The first

Table 4.5: Results on VML-HD according mAP metric using Euclidean distance.

Model	TL	OL	DE	VML-HD	HADARA80P	GW
	X	X	X	0.73	0.58	0.56
	✓	X	X	0.75	0.61	0.59
	✓	✓	X	0.78	0.67	0.61
Ours	✓	X	✓	0.76	0.65	0.60
	X	✓	X	0.75	0.64	0.59
	X	✓	✓	0.76	0.66	0.61
	X	X	✓	0.74	0.63	0.58
	✓	✓	✓	0.79	0.73	0.63

phenomenon consists of the difference between online triplet learning and offline triplet learning for training a model using triplet loss. In online triplet learning, the model is trained on individual triplets that are dynamically selected from the training dataset. In contrast, offline triplet learning involves pre-computing all possible triplets in the training dataset and then training the model on the entire set of triplets at once. Thus, online triplet learning is more efficient in terms of computational resources, because it only uses a small number of triplets at a time, rather than processing the entire set of triplets at once. This can be particularly important for large datasets, where pre-computing all possible triplets can be computationally expensive. Another advantage of online triplet learning is that it is more flexible than offline triplet learning. This is because the triplets used for training are selected during the training process, rather than being pre-computed, which means that the model can adapt to changes in the dataset and incorporate new information as it becomes available. This can help the model learn more effectively and improve its performance.

Second, transfer learning is particularly useful for historical Arabic documents, given the limited data available for training a word spotting model from scratch. By using a pre-trained model that was trained on a similar dataset, the word spotting model can use its pre-existing knowledge as a starting point, where the pre-trained model can provide a strong foundation for the word spotting model, and can help it learn the features and patterns of Arabic text more effectively based on additional knowledge and expertise that can help it to better understand the structure and layout of the document. This can make the word spotting model more accurate and reliable and can help it to accurately recognize words in historical documents.

Lastly, document enhancement can be useful for word spotting in historical documents that have been processed or altered in some way to make the text or other features more visible or easier to read and improve the overall quality of the document images. These enhancements can include techniques such as image sharpening, color correction, or noise reduction, which can help to improve the legibility of the text in historical documents. This can make it easier for the word spotting algorithm to accurately identify the words in the document, which can improve the overall performance of the word spotting system.

4.7 Conclusion

In this chapter, we have presented an approach to enhance the performance of word spotting in historical documents by introducing a conditional GAN to generate clean images from degraded inputs. Furthermore, we proposed several strategies to improve the learning of feature embeddings for the word spotting task, including the use of transfer learning, online triplet mining, and semi-hard triplet selection techniques. Our experimental results on several datasets have shown that our approach outperforms state-of-the-art methods for word spotting in historical documents. Additionally, we have conducted an error analysis to evaluate the effectiveness of our proposed method, which confirmed the improved performance of our approach.

In the subsequent chapter, we will present and thoroughly examine our proposed GAN-based model for degraded document image restoration. This model utilizes a conditional GAN architecture to generate high-quality images from degraded inputs. The goal of this model is to produce images with improved fine-detail recovery and hyper-clean results for use in the word spotting task in historical documents.

CHAPTER 5

Enhancement of Historical Document Images via Generative Adversarial Networks

5.1	Introduction	75
5.2	Research related to enhancing degraded document images	75
5.2.1	Degraded document enhancement	76
5.2.2	Generative adversarial networks for image-to image transform	78
5.3	Proposed Method	80
5.3.1	Generator architecture	80
5.3.2	Discriminator architecture	81
5.3.3	Loss functions of proposed GAN	81
5.4	Experiments	83
5.4.1	Datasets	83
5.4.2	Experimental setup	83
5.4.3	Results	85
5.5	Conclusion	93

5.1 Introduction

The process of document processing can be accomplished through a combination of computer vision tools and human analysis. With the advent of various publicly available databases, the scope and scale of document processing have grown significantly in recent years. However, despite these advancements, the process may still prove ineffective when applied to heavily degraded documents due to their poor quality and the various forms of deterioration they have undergone over time [Schreiber et al., 2017, Chaieb et al., 2015]. The analysis of historical documents presents significant challenges due to the degraded state of the manuscripts. These documents can be adversely affected by a wide range of deteriorations, including wrinkles, dust damage, nutritional stains, and discolored sunspots [Zamora-Martínez et al., 2007]. Additionally, the process of digitizing historical documents through scanning can also introduce further degradation, such as poor image quality due to the use of smartphone cameras (e.g. shadows [Finlayson et al., 2002], blurring [Chen et al., 2011], variations in lighting, and warping) which can lead to inefficiency in the processing of these documents. Furthermore, historical documents may also be encumbered with additional features such as stamps, watermarks, and annotations, which can further complicate the analysis process. In light of these challenges, this study presents a novel document enhancement model, aimed at improving the visual quality of degraded historical documents. The proposed model utilizes a Generative Adversarial Networks (GANs) based approach to treat the document enhancement task as an image-to-image conversion process. The proposed GAN architecture is designed to be robust and complex, specifically tailored to address the unique challenges associated with historical documents, with the goal of producing cleaner document images compared to existing methods.

The remainder of this chapter is organized as follows. A comprehensive review of related work in the field of document enhancement is presented in Section 5.2. The proposed enhancement approach for degraded historical documents is described in detail in Section 5.3. The results of the experimental study conducted to evaluate the effectiveness of the proposed approach are presented in Section 5.4. Finally, the conclusions and future perspectives of this research are outlined in Section 5.5.

5.2 Research related to enhancing degraded document images

In this section, we first identify the most challenging aspects of the document enhancement process. Subsequently, we provide a comprehensive overview of existing document enhancement approaches and their corresponding strengths and limitations. The purpose of this review is to provide a clear understanding of the state of the art in the field and to identify potential areas for improvement in the enhancement of historical documents.

Access to valuable cultural heritage in the form of Historical Arabic Documents (HADs) is often hindered due to inadequate storage conditions. These documents, in their original form, may not be amenable to automatic processing by machine vision algorithms and thus require transcription into a more readable format. However, the process of transcription and digitization can be further complicated by the presence of various types of degradation, such as wrinkles, dust damage, nutritional stains, and discolored sunspots. Additionally, the

presence of watermarks, stamps, or annotations on the documents further exacerbates the difficulties in enhancing and restoring them. The task of document enhancement becomes even more challenging when such forms of degradation occur in the vicinity of text, particularly when the stain color matches or is more intense than the font color of the document, as illustrated in Figure 5.1.

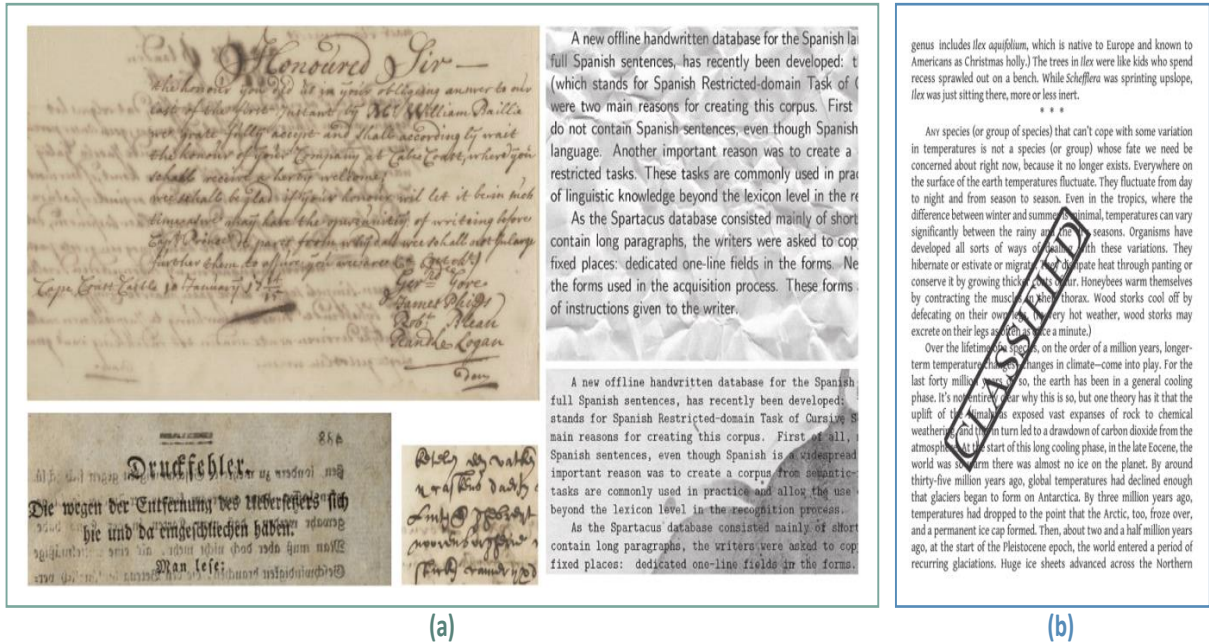


Figure 5.1: Examples of documents: (a): Degraded documents, (b): A document including a big stamp.

An optimal system for document enhancement should be able to effectively perform two tasks simultaneously: removing noise and watermarks while preserving the quality of the text in the document images. Recently, deep neural networks have demonstrated remarkable success in natural image generation and restoration, particularly through the use of deep convolutional neural networks (such as auto-encoders and Variational Auto-Encoders (VAE)) [Mao et al., 2016, Dong et al., 2015, Kingma and Welling, 2013], and Generative Adversarial Networks (GANs) [Isola et al., 2017, Kingma and Welling, 2013]. These methods have shown promising results in the context of natural image enhancement and restoration, and it is believed that they could be adapted to address the specific challenges of historical document enhancement.

5.2.1 Degraded document enhancement

The process of document enhancement aims to address issues pertaining to the visual quality of document images and the removal of degradation and artifacts, with the ultimate goal of restoring the original appearance of the documents. This process encompasses a wide range of techniques, including image restoration and enhancement methods [Moghaddam and Cheriet, 2009b], and the application of various algorithms and models to improve the overall quality of document images and make them more legible [Hedjam and Cheriet, 2013].

This process is crucial for preserving and making accessible the valuable cultural heritage contained in historical documents.

A commonly employed technique for enhancing the quality of historical documents is document binarization, which involves separating text pixels from the background [Sauvola and Pietikäinen, 2000, Ntirogiannis et al., 2012]. The goal of this technique is to eliminate noise and other unwanted information from the document images and to improve the overall legibility and readability of the documents.

Traditionally, document binarization methods [Otsu, 1979, Sauvola and Pietikäinen, 2000, Niblack, 1985, Phansalkar et al., 2011, Chutani et al., 2015, Cheriet et al., 1998] utilize thresholding techniques as the main method for separating text pixels from the background. This process involves the use of a threshold value, or multiple threshold values, to classify pixels as part of the text or degradation. The selection of the optimal threshold value(s) is critical for the success of the binarization process, and various approaches have been developed over the years to determine the best threshold(s) for a given document image. The choice of threshold(s) can significantly impact the quality of the final binarized image and can affect the overall performance of the document enhancement process.

An approach for locating text in degraded document images that relies on edge detection is presented in [Lelore and Bouchara, 2013]. Furthermore, in [Annabestani and Saadatmand-Tarzjan, 2019], an algorithm based on global threshold selection is proposed. The image contrast is enhanced using fuzzy expert systems (FESs). Then, different FESs and a pixel counting algorithm are applied to adjust the range of the threshold which is finally fixed at the average value of this range. Another method for determining the binarization threshold is proposed in [Chou et al., 2010], which is based on machine learning techniques. Each region of the image is defined as a three-dimensional feature vector obtained from the gray-level pixel value distribution. Then, a support vector machine (SVM) was employed to classify each region into one of four different threshold values.

Traditional methods for document image processing have been widely used for various tasks such as text recognition, document analysis, and image restoration. However, these methods often suffer from limitations due to the dependence of their results on the conditions of the document image. This can lead to difficulties in the presence of complex image backgrounds or atypical intensity levels. To address these limitations, various approaches have been proposed in the literature. One such approach is the use of variational models, as demonstrated in [Moghaddam and Cheriet, 2009b], where a method for eliminating transparencies from degraded two-sided document images was introduced. The model was based on the minimization of an energy functional, which combines data fidelity and regularization terms. Additionally, a modified version of this model was developed to address situations where the reverse side of the document is missing. Another approach, proposed in [Hedjam et al., 2014], treated ink as a target and aimed to detect it by maximizing an energy function. The authors proposed a graph-cut-based optimization method that uses local and global constraints. This technique was later extended to the task of binarizing scene text in [Milyaev et al., 2015]. In [Xiong et al., 2018], a mathematical morphology-based approach was used to estimate and extract the background from the image. The background was estimated using a combination of morphological opening and closing operations. Subsequently, a Laplacian energy-based segmentation method was applied to the enhanced document image for pixel classification. The Laplacian energy was defined as a combination of image gradient and regional information.

Over recent years, deep learning techniques have garnered significant attention for their potential to enhance document images. Several authors have proposed the use of deep architectures for document binarization, a process of converting a grayscale image into a binary image by separating text from the background. For instance, in the study conducted by authors in [Tensmeyer and Martinez, 2017], the binarization problem was formulated as a learning task for pixel classification and a fully convolutional neural network (FCN) architecture was applied, which was capable of operating at various image scales as well as at full resolution. In another study by [Afzal et al., 2015], a method based on a fully convolutional neural network was proposed, where the binarization of document images was treated as a sequence learning task utilizing Long Short-Term Memory (LSTM). Each pixel was then classified as text or background. Additionally, authors in [Westphal et al., 2018] proposed an approach based on recurrent neural networks, specifically using LSTM and Pseudo-F-Measure for image binarization.

A deep-network-based approach for removing shadows in document images was proposed in [Lin et al., 2020]. The authors introduced a background estimation module to extract the global background color of the document image. This module learns the spatial distribution of the background as well as the pixels that are not part of the background during the color estimation process. By creating an attention map based on this encoded information, the authors were able to produce a document image without shadows by estimating the global background color and applying the attention map.

5.2.2 Generative adversarial networks for image-to image transform

Recently, GANs have demonstrated impressive results in both image generation and translation tasks. In the field of document processing and enhancement, GANs have also shown potential. One example is the application of GANs in image segmentation, as demonstrated by Ledig et al. in their study of SRGAN [Ledig et al., 2017], which uses a generative adversarial network for super-resolution of images. The authors employed conditional GANs for document enhancement tasks based on image-to-image translation. Another example is the use of Dual GAN generator algorithms as presented by [Yi et al., 2017] which were specifically designed for underwater image enhancement. These algorithms exploit two or more generators for predicting the enhanced image. The intention behind using two generators along with one discriminator or two generators with two discriminators is to either share features between the generators or to consider the prediction of one generator as an input to the other generator.

In their study, Li et al. [Li et al., 2018] proposed a model called UWGAN (UnderWater Generative Adversarial Network) for color correction of underwater images, which is inspired by the principles of GANs as introduced in [Goodfellow et al., 2020]. The model is weakly supervised, meaning it does not require the use of matched underwater images for training. This is achieved by considering underwater images in unknown locations, which enables adversarial learning. This approach allows for the model to adapt to different underwater environments and lighting conditions, making it more versatile in correcting color distortion in underwater images. This study has demonstrated the potential of GANs-based models in addressing the specific challenges faced in color correction of underwater images.

Isola et al. [Isola et al., 2017] developed a GANs named Pix2Pix, which is designed for image-to-image translation using Conditional GAN (CGAN). The model utilizes an adver-

serial loss to train the generator, promoting plausible image generation in the target domain while minimizing the computed L1 loss between the generated image and the expected output image. The discriminator in the model assesses whether the generated image is a real transformation of the source image. By using the CGAN architecture, the Pix2Pix model is able to leverage additional information in the form of image labels, making it more effective in image-to-image translation tasks. While there are various document improvement methods available, many of them focus on a specific issue. For instance, in a study by Souibgui et al. [Souibgui and Kessentini, 2020], the authors examined the problem of degraded documents caused by dense watermarks or stamps, and proposed a document enhancement approach based on conditional GANs to recover a clean version of historical documents. Similarly, in a study by authors in [Jemni et al., 2022], the authors presented an approach that integrated a recognition stage in a document binarization model.

In a study by Gangeh et al. [Gangeh et al., 2021], the authors focused their approach on specific problems such as salt and pepper noises, blurred text, and watermarks in document images. They proposed a unified architecture by integrating a deep network with a cycle-consistent GANs as the basic network for the denoising problem of document images. This approach aims to address the specific issues faced in document image denoising, such as noise reduction and text preservation. Additionally, the authors in [Kang et al., 2021] presented an auto-encoder architecture that executed a cascade of pre-trained U-Net models [Ronneberger et al., 2015] in order to learn binarization using fewer data. This approach aims to make the binarization process more efficient and effective by utilizing pre-trained models and reducing the amount of data required for training.

Recent advancements in document enhancement have led to the emergence of various techniques that utilize deep learning tools, specifically CNNs and GANs, to produce a pristine binary version of degraded document images. These techniques have been demonstrated in a number of studies, such as [Zhao et al., 2019, Souibgui and Kessentini, 2020, Tamrin et al., 2021, Kang et al., 2021], and have shown promising results. However, many of these studies have employed basic architectures and optimization techniques that may not be fully suited to address the intricacies and complexities of historical document images. Therefore, there is a need for more sophisticated and adaptable approaches to enhance these types of document images.

The ultimate goal of document enhancement is to create a high-quality, polished version of degraded document images, which can be used for a variety of downstream processing tasks. In order to achieve this, we propose a robust and versatile GAN architecture that is trained and optimized using multiple loss functions. This approach is specifically designed to tackle the complexity of historical documents and to produce a cleaner and clearer image regardless of the type of degradation. By implementing this advanced GAN architecture, we aim to overcome the limitations of current document enhancement techniques and generate a significantly improved version of degraded document images, which can be applied to a wide range of use cases. Our contribution has been formally submitted to a regarded journal and published in an international conference [Fathallah et al., 2023a].

5.3 Proposed Method

In this section, we present the key steps of our proposed approach for enhancing degraded document images. The primary objective of GANs as outlined in recent research [Souibgui and Kessentini, 2020, Marnissi et al., 2021, Jemni et al., 2022, Qins and El-Yacoubi, 2022] is to train generative models that can learn the distribution of real data and produce output images from random noise.

Our Enhancement Historical Document Images (EHDI) model is specifically designed to generate a clean version of degraded historical documents. We treat this task as an image-to-image conversion process, where our model aims to learn a mapping from the degraded document image x to the clean document image y . During the training process, the proposed GAN takes a degraded document image as input and attempts to generate a clean version of it. On the other hand, the discriminator takes two inputs: the generated image and the ground truth, which refers to the cleaned version of the degraded image. The discriminator then determines whether the generated image is realistic or not based on the ground truth. As depicted in Figure 5.2, our proposed model is composed of a generator (G) and a discriminator (D) that work in tandem to enhance degraded document images. The generator G is trained to transform a degraded document image into a clean version, while the discriminator D helps G to produce a more realistic image by distinguishing between generated and real images. In the following section, we will delve deeper into the architecture of our proposed model, providing a more comprehensive explanation of its various components and how they work together to achieve the goal of document enhancement.

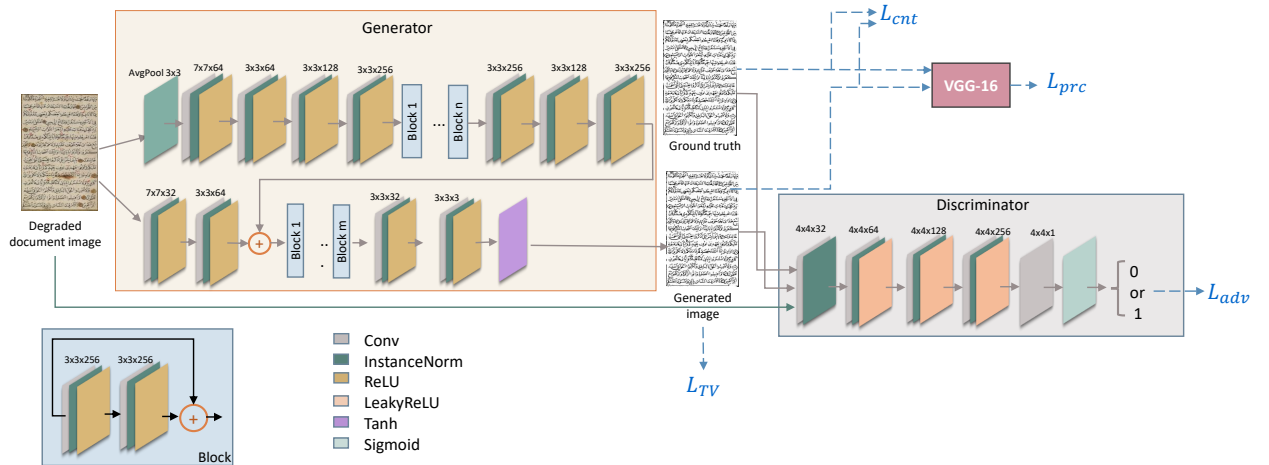


Figure 5.2: Proposed GAN architecture for document enhancement process.

5.3.1 Generator architecture

The proposed generator in this study is an image transformation network that is designed to generate transformed images by utilizing input images. This design is consistent with the concept of an auto-encoder model, which comprises both an encoder and a decoder. The input image is typically encoded through a series of convolutional layers with decreasing spatial resolution, resulting in a condensed representation of the input image at a particular

layer. Subsequently, the image is decoded through a series of layers with increasing spatial resolution, comprising up-sampling and convolutional layers. The architecture of the generator network is illustrated in Figure 5.2, which is similar to the architecture proposed in [Kuang et al., 2020] for the colorization of thermal infrared images. Furthermore, the generator network is divided into two sub-networks, with each sub-network utilizing the architecture proposed in [Johnson et al., 2016].

5.3.2 Discriminator architecture

The discriminator in this study is an FCN that receives two input images, namely the generated image and its corresponding ground truth image. Its primary function is to distinguish between the authenticity of the generated image and a fake image. The architecture of the proposed discriminator is depicted in Figure 5.2. It consists of five convolutional layers designed to extract features from the input images. The first four layers are followed by a normalization layer, which helps to stabilize the training of the model and prevent overfitting. Additionally, a LeakyReLU activation function, which allows for small negative input values, is applied after the second, third and fourth layers to improve the gradient flow and enhance the performance of the discriminator. This architecture is inspired by PatchGAN [Isola et al., 2017] which utilizes a 70×70 patch as input to the discriminative network, aiming to distinguish whether local image patches are real or fake. The overall objective of the discriminator is to determine if the input patch in an image is authentic or synthetic.

5.3.3 Loss functions of proposed GAN

In order to effectively train our proposed GANs, we incorporate a content loss based on an L_1 term, which penalizes the distance between the generated and ground-truth images. The adversarial loss from the discriminator helps the generator to synthesize fine and specific details in the generated image. Additionally, we incorporate a combination of perceptual and Total Variation (TV) losses to further enhance the quality of the generated details. This results in an objective loss function comprising four different losses: the adversarial loss, the content loss, the perceptual loss and the TV loss. These losses are defined as follows:

- Adversarial loss:

In order to enhance the ability of the generator to produce high-fidelity images with accurate details, an adversarial loss is utilized. This loss function is designed to ensure that the generated clean images $G(x)$ are indistinguishable from true clean images indicated by y . The adversarial loss function is defined by Eq.(5.1) and is optimized during the training process of the model. It is based on the binary cross-entropy between the discriminator predicted probability that the generated image is real and an array of ones (real labels). This loss helps the generator to produce images that are similar to the ground-truth images, and it is also a measure of how well the generator is able to fool the discriminator. As the generator produces more realistic images, the discriminator will be less able to distinguish between the generated images and the real images, and the adversarial loss will decrease.

$$\mathcal{L}_{adv} = \mathbb{E}_y[-\log(D(G(x), y))] \quad (5.1)$$

- The proposed GAN is improved by incorporating a content loss, which aims to ensure that the content information present in the ground-truth image y is also present in the generated image $G(x)$. To achieve this goal, we use a pixel-wise mean squared error loss to minimize low-level content errors between the generated cleaned images and their corresponding ground-truth images, as described in Eq.(5.2). The content loss function compares the generated image with the ground-truth image on a pixel-by-pixel basis, ensuring that the generated image preserves the same features and details as the ground-truth image. The mean squared error is a commonly used loss function for this purpose; it provides a good balance between the ability to preserve fine details and the ability to handle noise and outliers.

$$\mathcal{L}_{content} = \frac{1}{WH} \sum_{i=1}^W \sum_{j=1}^H \| y_{i,j} - G(x)_{i,j} \|_1 \quad (5.2)$$

where W and H respectively represent the height and width of the degraded image, $\| \cdot \|_1$ refers to the L_1 norm, $y_{i,j}$ represents the pixel values of the ground truth image, and $G(x)_{i,j}$ are the pixel values of the generated image.

- Perceptual loss: In order to generate results with improved perceptual quality and to address distorted textures that may be introduced by the adversarial loss, we propose to utilize the perceptual loss introduced in [Johnson et al., 2016]. This loss function is used to compute the distance between the generated image and its ground-truth, based on the high-level representations obtained from a pre-trained VGG-16 model. The perceptual loss function, defined by Eq.(5.3), compares the features of the generated image with those of the ground-truth image at a certain layer of the VGG-16 model. By comparing high-level representations, the perceptual loss can better capture the overall quality of the image and texture, rather than just the individual pixels. The use of a pre-trained model allows the GANs to learn from a vast dataset and generalize well to unseen images. This loss function helps the generator to produce images that are more similar to the ground-truth images, not just in terms of pixel values, but also in terms of their high-level representations, resulting in a generated image with improved perceptual quality.

$$\mathcal{L}_{prc} = \sum_k \frac{1}{C_k H_k W_k} \sum_{i=1}^{H_k} \sum_{j=1}^{W_k} \| \Phi_k(y)_{i,j} - \Phi_k(G(x))_{i,j} \|_1 \quad (5.3)$$

where Φ_k represents the feature representations of the k^{th} maxpooling layer in the VGG-16 network, and $C_k H_k W_k$ represents the size of these feature representations.

- Total variation loss: In order to prevent over-pixelization and enhance the spatial smoothness of the cleaned document images, we adopt the Total Variation (TV) loss introduced in [Aly and Dubois, 2005]. The TV loss is defined by Eq.(5.4). It is based on the total variation of the generated image, which measures the smoothness of the image. By minimizing the TV loss, the generator is encouraged to produce images that are smooth and have a low level of noise. This is particularly useful in image denoising

and deblurring where the goal is to generate a smooth image that is close to the original image. The TV loss helps to prevent the generator from producing images with high frequency noise, which can result in over-pixelization and a loss of fine details in the generated image.

$$\mathcal{L}_{tv} = \frac{1}{WH} \sum |\nabla_x G(\tilde{y}) + \nabla_y G(\tilde{y})| \quad (5.4)$$

where $|\cdot|$ refers to the absolute value per element of the indicated input.

The optimization of network parameters in G is achieved through the estimation of a global loss function, denoted as \mathcal{L} , as outlined in Eq.(5.5). This loss function serves as a metric for evaluating the performance of the network and is used to guide the adjustment of network parameters in order to minimize the loss and improve overall network performance.

$$\mathcal{L} = \mathcal{L}_{cnt} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{prc}\mathcal{L}_{prc} + \lambda_{tv}\mathcal{L}_{tv} \quad (5.5)$$

where λ_{adv} , λ_{prc} and λ_{tv} represent the weights that control the share of different losses in the full objective function. The setting of weights is based on preliminary experiments on the training dataset.

5.4 Experiments

In this section, we present a thorough evaluation of our proposed approaches in document enhancement tasks through a series of experimental studies. Our experiments have been conducted using a variety of publicly available datasets, in order to provide a comprehensive assessment of the effectiveness of our proposed EHDI.

5.4.1 Datasets

In order to evaluate the effectiveness of our proposed document cleaning method, we have employed the Noisy Office database [Zamora-Martínez et al., 2007] as the primary training dataset. This database comprises 144 images, which encompasses a wide range of document degradation types. For the evaluation phase, we have considered several benchmark datasets, namely DIBCO 2013 [Pratikakis et al., 2013], DIBCO 2017 [Pratikakis et al., 2017], H-DIBCO 2018 [Pratikakis et al., 2018] and [Pantke et al., 2014], to provide a comprehensive assessment of our proposed approach’s performance in comparison with state-of-the-art methods.

5.4.2 Experimental setup

5.4.2.1 Evaluation protocol

In order to thoroughly evaluate the effectiveness and quality of our proposed EHDI model, we have conducted a series of experimental studies on various datasets. We have employed a combination of qualitative and quantitative metrics to provide a comprehensive assessment of

our model’s performance. In particular, we have utilized four widely accepted performance metrics [Pratikakis et al., 2013] in the field of document enhancement and image processing to evaluate the performance of our EHDI model in comparison to state-of-the-art methods. The used metrics are:

- Peak signal-to-noise ratio (PSNR): It is a measure of the quality of image reconstruction. It compares the original image with the reconstructed image. It is defined as the ratio of the maximum possible power of a signal to the power of corrupting noise that affects the quality of its representation. PSNR is commonly used to measure the quality of the lossy image and video compression.
- Pseudo-F-measure (F_{ps}): It is a measure of the quality of the result of a binary segmentation. It is a combination of precision and recall.
- Distance reciprocal distortion metric (DRD): It is a measure of the quality of the result of a binary segmentation. It is defined as the harmonic mean of the inverse of the distances between the true positive pixels and the corresponding pixels in the obtained segmentation.
- F-measure: It is a measure of a test’s accuracy. F-measure is the harmonic mean of precision and recall. It is used to compare different binary classifiers.

5.4.2.2 Implementation details

In the proposed approach, the optimization of the model parameters is carried out using the stochastic gradient descent (SGD) algorithm with a learning rate of 10^{-3} and a batch size of dimension 512. To mitigate the risk of overfitting, an early stopping function with a patience value of 50 is employed during the training process. Additionally, all parameter values were determined through an empirical process. In order to facilitate the implementation of the proposed approach, the PyTorch framework was utilized, and the EHDI model was trained on an NVIDIA Quadro RTX 6000 GPU with 24 GB of RAM to ensure efficient and expedient training and evaluation.

5.4.2.3 Enhancement architecture training

To evaluate the effectiveness of our proposed document cleaning method, we utilized the dataset presented in [Zamora-Martínez et al., 2007], which includes several different types of degradation. Specifically, we employed the entire Noisy Office dataset for the training phase, which comprises 144 images. To improve the performance of our enhancement architecture, each image was resized to 1024×1024 pixels and a set of stacked patches of size 256×256 pixels were mined. This resulted in a total of 2,304 patch pairs that were introduced into our model for training. To balance the different losses in the full objective function, we set the weights of the adversarial loss λ_{adv} , the perceptual loss λ_{prc} and the total variation loss λ_{tv} to 0.3, 1 and 1 respectively. Figure 5.3 illustrates some of the stacked patches employed in model training, and demonstrates how our EHDI model is able to restore a document as closely as possible to the ground truth.

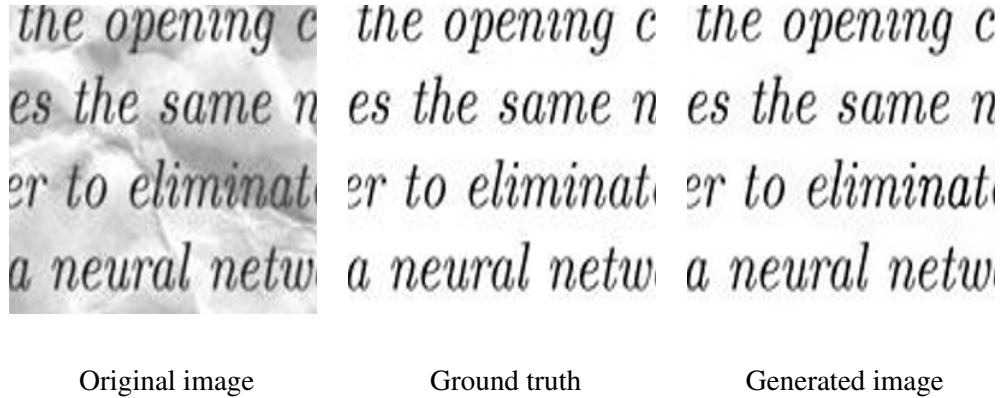


Figure 5.3: Example of stacked patches of size 256×256 pixels during training phase of our proposed model.

5.4.3 Results

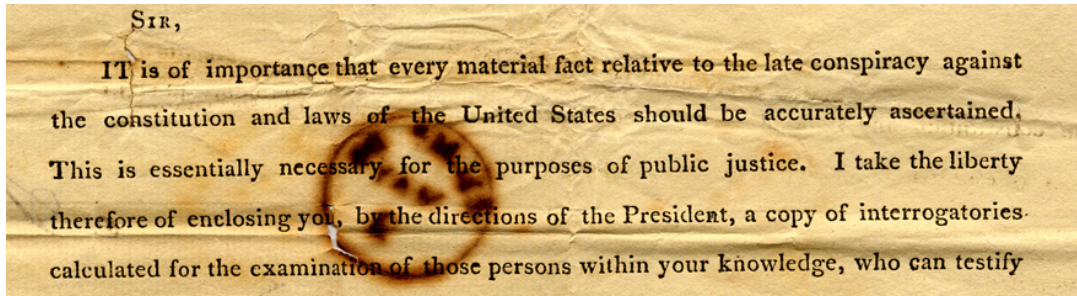
In this section, we present and analyze the performance of the proposed EHDI model. It is important to note that the model was trained exclusively using the Noisy Office dataset [Zamora-Martínez et al., 2007], and the assessment of its performance was conducted on external datasets that were not included in the training process. This approach allows us to evaluate the generalization capabilities of our model. To provide a fair comparison of our proposed method, we have compared the results of our EHDI with the state-of-the-art methods in document binarization. A comparison of the results obtained on the DIBCO 2013 dataset against different approaches is provided in Table 5.1. Additionally, a qualitative comparison of the results obtained on the DIBCO 2013 dataset is illustrated in Figure 5.4. The results demonstrate that our EHDI model produces a cleaner image quality than DE-GAN, especially when the degradation occurs very densely. This is likely due to the fact that in such cases, DE-GAN fails to remove the degradation from the document background.

Table 5.1: Results of our proposed EHDI on DIBCO 2013 dataset.

Model	PSNR	F-measure	F_{ps}	DRD
[Otsu, 1979]	16.6	83.9	86.5	11.0
[Niblack, 1985]	13.6	72.8	72.2	13.6
[Sauvola and Pietikäinen, 2000]	16.9	85.0	89.8	7.6
[Gatos et al., 2004]	17.1	83.4	87.0	9.5
[Su et al., 2012]	19.6	87.7	88.3	4.2
[Tensmeyer and Martinez, 2017]	20.7	93.1	96.8	2.2
[Xiong et al., 2018]	21.3	93.5	94.4	2.7
[Vo et al., 2018]	21.4	94.4	96.0	1.8
[Howe, 2013]	21.3	91.3	91.7	3.2
[Souibgui and Kessentini, 2020]	24.9	99.5	99.7	1.1
Our	26.8	99.9	99.9	0.97

In this section, we report and analyze the performance of the proposed EHDI.

It is important to note that the training of our model was performed only with the Noisy Office [Zamora-Martínez et al., 2007] database while the assessment is conducted on external databases that were not included in the model training. The results of our EHDI are compared to the current state of the art in document binarization. A comparison of the re-



Original image

SIR,
IT is of importance that every material fact relative to the late conspiracy against the constitution and laws of the United States should be accurately ascertained. This is essentially necessary for the purposes of public justice. I take the liberty therefore of enclosing you, by the directions of the President, a copy of interrogatories calculated for the examination of those persons within your knowledge, who can testify

Ground truth

SIR,
IT is of importance that every material fact relative to the late conspiracy against the constitution and laws of the United States should be accurately ascertained. This is essentially necessary for the purposes of public justice. I take the liberty therefore of enclosing you, by the directions of the President, a copy of interrogatories calculated for the examination of those persons within your knowledge, who can testify

DE-GAN [Souibgui and Kessentini, 2020]

SIR,
IT is of importance that every material fact relative to the late conspiracy against the constitution and laws of the United States should be accurately ascertained. This is essentially necessary for the purposes of public justice. I take the liberty therefore of enclosing you, by the directions of the President, a copy of interrogatories calculated for the examination of those persons within your knowledge, who can testify

Ours

Figure 5.4: Example of degraded documents enhancement by our EHDI and DE-GAN on sample PR08 from DIBCO-2013 [Souibgui and Kessentini, 2020].

sults according to DIBCO 2013 dataset against different approaches is provided in Table 5.1. Moreover, qualitative results from DIBCO 2013 are depicted in Figure 5.4; EHDI produces

a cleaner image quality than DE-GAN [Souibgui and Kessentini, 2020]. The failure of DE-GAN to effectively remove dense degradation from document backgrounds can be attributed to its limitations in accurately distinguishing between degraded and non-degraded pixels. In contrast, traditional methods are reliant on threshold-based approaches to classify degraded pixels as text or text pixels as removable degradation, which may also contribute to their limitations in effectively addressing dense degradation.

Our proposed EHDI method demonstrates a significant improvement in document image generation, as evidenced by the results presented in Figure 5.5. The example depicted in the figure illustrates the ability of the EHDI method to generate document images that are not only highly similar to, but also surpass the quality of the ground truth images. This serves as a clear demonstration of the effectiveness and efficiency of the EHDI method.

Table 5.2: A comparative review of competitor approaches of DIBCO 2018 on DIBCO 2017 and DIBCO 2018 Datasets.

Model	DIBCO 2018				DIBCO2017			
	PSNR	F-measure	F_{ps}	DRD	PSNR	F-measure	F_{ps}	DRD
1 [Pratikakis et al., 2018]	19.11	88.34	90.24	4.92	17.99	89.37	90.17	5.51
7 [Pratikakis et al., 2018]	14.62	73.45	75.94	26.24	15.72	84.36	87.34	7.56
2 [Pratikakis et al., 2018]	13.58	70.04	74.68	17.45	14.04	79.41	82.62	10.70
3b [Pratikakis et al., 2018]	13.57	64.52	68.29	16.67	15.28	82.43	86.74	6.97
6 [Pratikakis et al., 2018]	11.79	46.35	51.39	24.56	15.38	80.75	87.24	6.22
[Souibgui and Kessentini, 2020]	16.16	77.59	85.74	7.93	18.74	97.91	98.23	3.01
Our	20.31	92.69	90.83	3.94	19.15	98.56	99.44	2.87

In addition to the quantitative evaluation of our proposed EHDI method, a qualitative comparison with state-of-the-art results in the document enhancement task was also conducted. The comparison revealed the clear superiority of the EHDI method over the DE-GAN model as reported in [Souibgui and Kessentini, 2020]. This superiority was consistently observed across all experiments conducted. This highlights the potential of EHDI as a powerful tool for document enhancement.

As demonstrated in Table 5.2, our proposed EHDI method achieves the highest performance on the 2017 DIBCO test set and the 2018 H-DIBCO test set. This is further illustrated in Figure 5.7, which presents an example of the visual results of the EHDI method applied to a sample document image (sample 16) from the DIBCO 2017 dataset. The proposed EHDI method is benchmarked against the winner’s method [Pratikakis et al., 2017] and the DE-GAN model. The superior performance of the EHDI method is evident when compared with the winner’s method, which utilized a U-net architecture and data augmentation techniques, as well as the DE-GAN model, which used a simple GANs to generate a clean document image. This can be attributed to the use of several loss functions in the EHDI method that aid in optimizing the generator to produce images that are closer to the ground truth.

In assessing the visual quality of enhanced document images, a qualitative examination was conducted utilizing sample images from the HADARA80P dataset [Pantke et al., 2014] as presented in Figure 5.8. Our proposed model is compared against DE-GAN, as it represents the current state-of-the-art method utilizing GANs for document enhancement. The results of this comparison demonstrate that EHDI consistently produces superior image quality when compared to DE-GAN.

Inhalt.	
Ein wunderlicher Traum.	Seite 251
Palingenese und Wiederauflebung.	ibid.
Von der Wünschelruthe.	254
Von der verschiedenen Anwendung der Wünschelruthe.	255
Meynungen von der Wünschelruthe.	263
Meine Erfahrung über eine Art von Wünschel-	

Inhalt.	
Ein wunderlicher Traum.	Seite 251
Palingenese und Wiederauflebung.	ibid.
Von der Wünschelruthe.	254
Von der verschiedenen Anwendung der Wünschelruthe.	255
Meynungen von der Wünschelruthe.	263
Meine Erfahrung über eine Art von Wünschel-	

Inhalt.	
Ein wunderlicher Traum.	Seite 251
Palingenese und Wiederauflebung.	ibid.
Von der Wünschelruthe.	254
Von der verschiedenen Anwendung der Wünschelruthe.	255
Meynungen von der Wünschelruthe.	263
Meine Erfahrung über eine Art von Wünschel-	

Handwritten text, very faded and difficult to read.

Handwritten text, clearly legible and well-structured.

Handwritten text, clearly legible and well-structured.

Personne n'avait aperçu le jeune homme.

— Que faire? se dit Pajou en regardant en vain de tous côtés; puis se rappelant tout-à-coup le plaisir qu'éprouvait toujours Augustin à feuilleter les cartons de gravures exposées près de l'Institut, il prit rapidement cette direction, tout en explorant des yeux les différents quartiers qu'il trouvait sur son passage.

Personne n'avait aperçu le jeune homme.

— Que faire? se dit Pajou en regardant en vain de tous côtés; puis se rappelant tout-à-coup le plaisir qu'éprouvait toujours Augustin à feuilleter les cartons de gravures exposées près de l'Institut, il prit rapidement cette direction, tout en explorant des yeux les différents quartiers qu'il trouvait sur son passage.

Personne n'avait aperçu le jeune homme.

— Que faire? se dit Pajou en regardant en vain de tous côtés; puis se rappelant tout-à-coup le plaisir qu'éprouvait toujours Augustin à feuilleter les cartons de gravures exposées près de l'Institut, il prit rapidement cette direction, tout en explorant des yeux les différents quartiers qu'il trouvait sur son passage.

Handwritten text, very faded and difficult to read.

Handwritten text, clearly legible and well-structured.

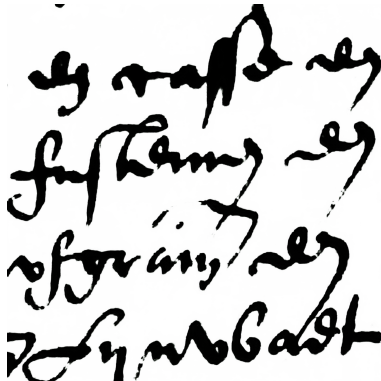
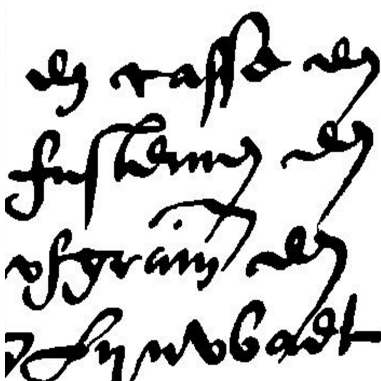
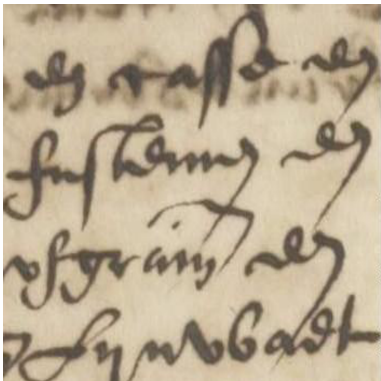
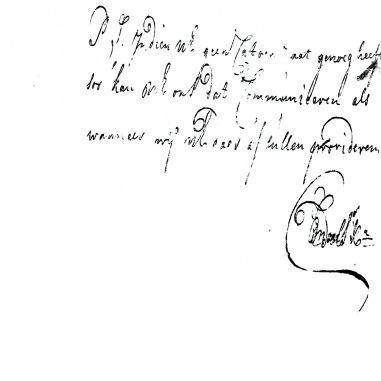
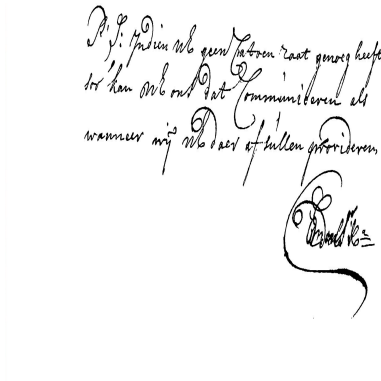
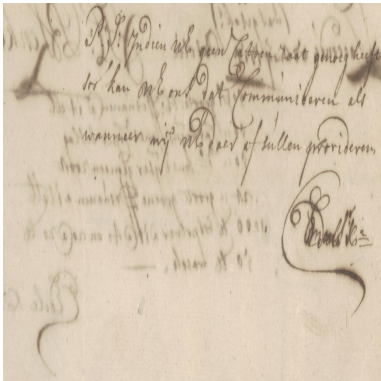
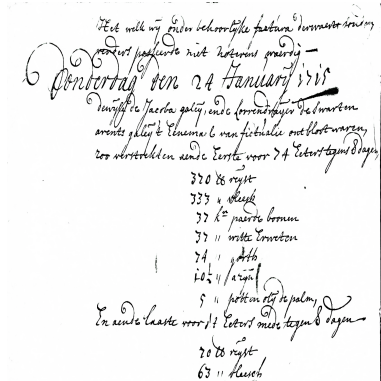
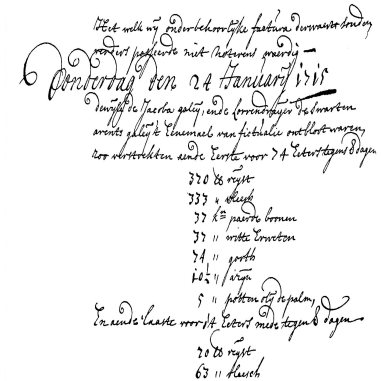
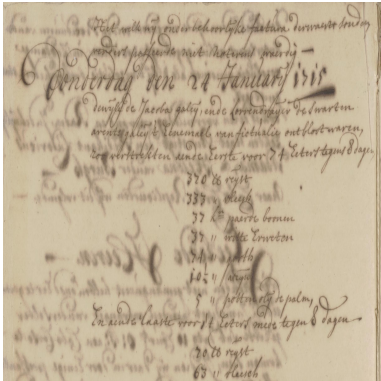
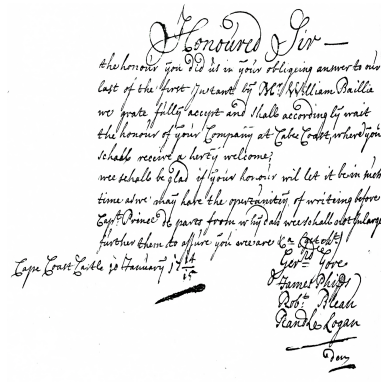
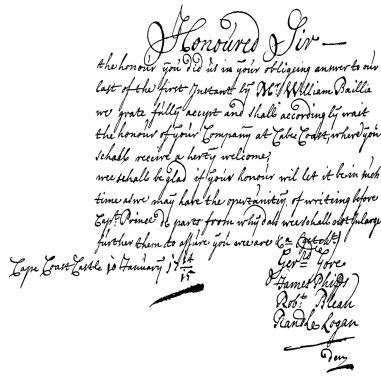
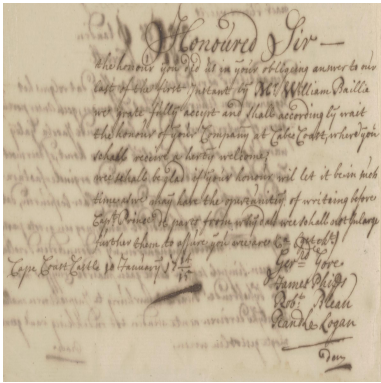
Handwritten text, clearly legible and well-structured.

Original image

Ground truth

Generated image

Figure 5.5: Example of enhancing degraded documents by our EHDI.

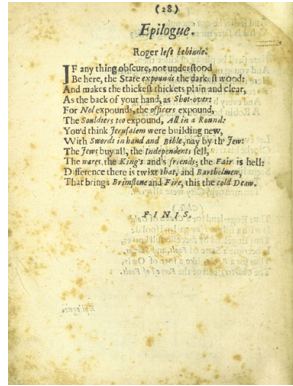


Original image

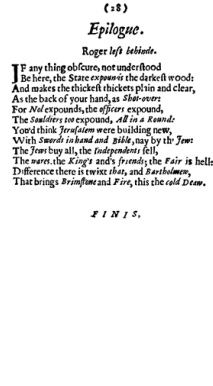
Ground truth

Generated image

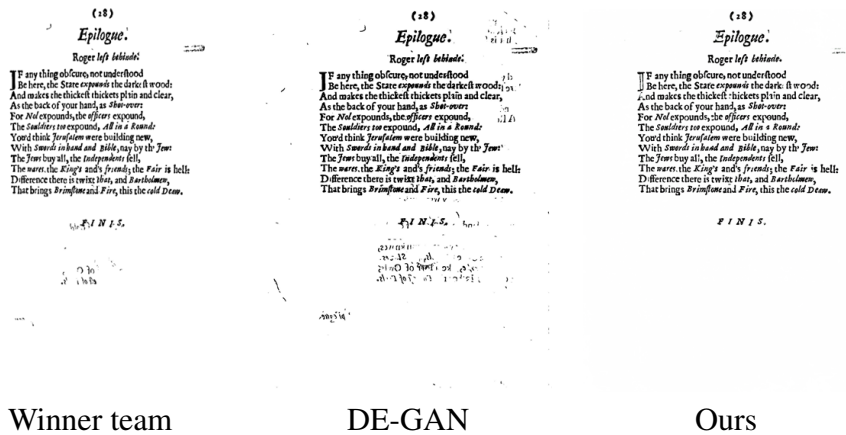
Figure 5.6: Example of enhancing degraded documents by our proposed model.



Original image



Ground truth



Winner team

DE-GAN

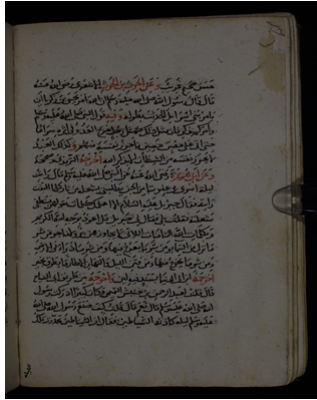
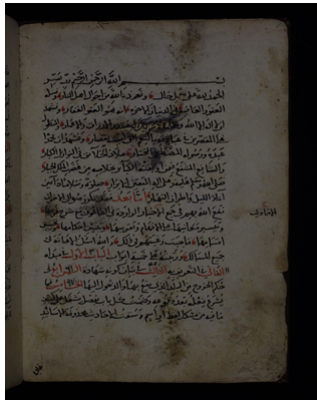
Ours

Figure 5.7: Qualitative binarization results on sample 16 in DIBCO 2017 dataset. Here, we compare the results of our proposed model with the winner’s approach [Pratikakis et al., 2017] and DE-GAN [Souibgui and Kessentini, 2020].

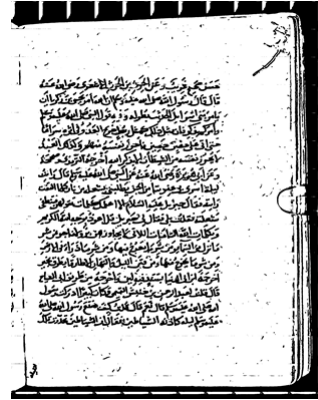
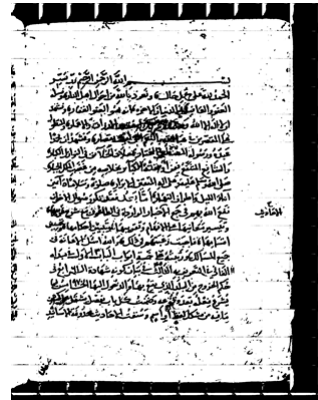
Table 5.3: Results of our proposed model on DIBCO 2018 dataset.

Model	PSNR	F-measure	F_{ps}	DRD
cycleGAN	11.00	56.33	58.07	30.07
pix2pix-HD	14.42	72.79	76.28	15.13
DE-GAN	16.16	77.59	85.74	7.93
Our	20.31	92.69	90.83	3.94

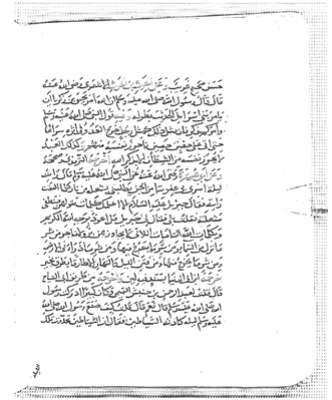
In order to conduct a comprehensive evaluation of our proposed model, EHDI, against state-of-the-art methods, a consistent set of sample images were utilized for testing by all methods. In alignment with the assessment methodology presented in the DE-GAN paper [Souibgui and Kessentini, 2020], our comparison includes the models pix2pix-HD [Wang et al., 2018a] and CycleGan [Zhu et al., 2017]. The results of this comparison, as illustrated in Figure 5.9 and quantitatively summarized in Table 5.3, demonstrate the superior performance of EHDI in achieving the highest visual quality results when compared to cycleGAN, pix2pix-HD, and DE-GAN.



Original image



DE-GAN [Souibgui and Kessentini, 2020]



Ours

Figure 5.8: Example of qualitative results on HADARA dataset compared to DE-GAN enhancement approach [Souibgui and Kessentini, 2020].

Previous research in [Souibgui and Kessentini, 2020] proposed the development of three separate cGAN-based models, each specifically tailored to address a distinct task of document image processing, namely binarization, watermarking, and deblurring. In contrast, the current study presents a unique approach by implementing a single model capable of effectively addressing all of these tasks simultaneously.

5.5 Conclusion

In this chapter, we proposed a conditional Generative Adversarial Network (cGAN) approach, referred to as EHDI, for generating high-quality document images from highly degraded inputs. The proposed EHDI model was designed to address various degradation tasks, including watermark removal and chemical degradation, with the ultimate goal of producing hyper-clean document images and effectively recovering fine details.

Extensive experimental evaluations were conducted to demonstrate the effectiveness of the proposed EHDI in cleaning extremely degraded documents. The results of these experiments indicate that the proposed approach outperforms recent state-of-the-art methods on reference datasets, particularly in the context of historical documents, representing thereby an interesting improvement in this domain.

CHAPTER 6

A Triplet Vision Transformers for Word Spotting in Historical Documents

6.1	Introduction	95
6.2	Related work	96
6.2.1	Vision transformer	96
6.3	Proposed approach	97
6.3.1	Pre-processing	98
6.3.2	Enhancement based Transfer Learning	98
6.3.3	Transformer architecture	99
6.3.4	Triplet loss	101
6.3.5	Embeddings matching	101
6.4	Experiments	102
6.4.1	Experimental setup	102
6.4.2	Results and discussions	103
6.4.3	Error Analysis	105
6.4.4	Ablation Study	107
6.5	Conclusion	108

6.1 Introduction

Processing historical Arabic documents (HADs) is extremely challenging, first due to the nature of the Arabic script which is a cursive text containing many diacritics, and second, to the poor quality of these documents which have undergone several types of degradation including distortions and noisy pixels.

There are several types of neural networks available for processing different types of data. For example, recurrent neural networks (RNN) [Mikolov et al., 2010] can process sequential data, while CNNs [Krizhevsky et al., 2012] are employed for image data. Many different CNN models have been designed for feature learning in the word spotting task [Mohammed et al., 2022, Serdouk et al., 2019, Pramanik and Bag, 2021, Mhiri et al., 2019]. Despite the remarkable performance of CNNs in image processing tasks, it has been acknowledged that a deeper network architecture is necessary for achieving a comprehensive representation of an image in the feature space and for handling long-term dependencies. While CNNs employ local receptive fields and max pooling layers to extract features, this technique can lead to a loss of valuable information during the downsampling process. On the other hand, RNNs are able to utilize the entire dataset at each stage of feature representation, which enables them to effectively capture temporal dependencies and contextual information. However, RNNs are not parallelizable since an input at a particular time step relies on the output of the preceding time step. Transformer networks [Tay et al., 2020, Lin et al., 2021] have emerged as a solution to the limitations of CNNs and RNNs for image processing tasks. The transformer uses self-attention mechanisms to weigh the importance of different parts of an image, allowing it to effectively capture both local and global features. Additionally, the transformer’s ability to process input in parallel, rather than sequentially as in RNNs, allows for faster training and inference. These characteristics make transformer models a promising solution for image processing tasks requiring both local and global information.

Recently, inspired by the success of vision transformer approaches in many applications, their encoding-decoding architectures have begun to be effectively applied to data representation. In view of the above facts, we have developed in this chapter a new word spotting technique for historical document images based on a triplet transformer. Particularly, we aim to build an embedding space for word image representation using triplet transformer architectures.

The analysis of HADs poses a significant challenge, particularly when using deep neural networks. This is due to the complexity of the task and the large amount of data required to train models to address it. The data scarcity problem is further exacerbated by the time and effort required to label and annotate the data, making it a significant undertaking. To overcome these challenges, transfer learning (TL) has become a popular approach. TL allows for the transfer of knowledge from pre-trained models to new tasks and domains, enabling researchers to leverage the knowledge of pre-trained models to solve new issues. This is achieved by fine-tuning pre-trained models on the target task, which can lead to improved performance compared to training models from scratch. The ability of TL to look beyond specific tasks and domains, and draw on the knowledge of pre-trained models, makes it a valuable approach for solving new issues in HADs.

In this chapter, we investigate the potential of TL as a means to leverage learned features from previous document datasets to train triplet transformers for word spotting in HADs. Our goal is to improve the representation of embedding features of Arabic word images. Specif-

ically, we propose a triplet transformer-based approach for word spotting in HADs, which is based on TL. Our model features a triplet architecture, with a backbone of transformers. We provide an in-depth analysis of the adaptation of TL techniques and their interaction with transformers. In particular, we investigate the transfer of knowledge from two domains, which have different characteristics: historical documents written in English and handwritten documents written in Hebrew. The results of our study demonstrate the potential of TL to improve the performance of word spotting in HADs.

This chapter is organized as follows. Section 6.2 provides an overview of the related work. Section 6.3 describes our approach to building an embedding space using the image word representation. It also illustrates the different steps in the entire process. Section 6.4 depicts the datasets and parameters used for our experiments. The conclusion is provided in 6.5.

6.2 Related work

In this section, we outline existing methods for the most recent vision transformer-based approaches.

6.2.1 Vision transformer

The transformer architecture was originally introduced in [Vaswani et al., 2017] in the field of automatic machine translation. It simply relies on fully connected layers and self-attention, achieving an interesting trade-off between performance and efficiency. Accordingly, it provides the best possible performance for various natural language processing tasks [Lee and Toutanova, 2018, Radford et al., 2018].

Several efforts have been carried out in computer vision in order to include different forms of attention, whether in conjunction [Wang et al., 2018b] or as an alternative to convolution [Ramachandran et al., 2019]. Other methods for detection have employed transformer layers over convolutional trunks [Carion et al., 2020]. Recently, some convolution-free models, relying only on the transformer layers, have shown interesting performance [Chen et al., 2020, Dosovitskiy et al., 2020, Touvron et al., 2021], allowing them to be a possible alternative to convolutional architectures. Image classification using the Vision Transformer (ViT) model, developed by [Dosovitskiy et al., 2020], constitutes the first example which is able to outperform, or even beat, the most advanced convolutional models for image classification. The authors in [Touvron et al., 2021] later refined the optimization procedure, leading to competitive results using ImageNet training only [Deng et al., 2009].

Transformers have also brought significant impact in computer vision challenges [Khan et al., 2021, Liu et al., 2021]. The ViT [Dosovitskiy et al., 2020] represents the current state of the art in using transformers for scalable image recognition. It assumes the 16×16 image patches as the sequential input to the transformer encoder. The ViT significantly performs better than CNNs for visual recognition. Recently, ViT architectures have achieved wide popularity for various vision tasks, including visual tracking [Wang et al., 2021, Chen et al., 2021], density prediction [Ranftl et al., 2021], medical image segmentation [Li et al., 2021a], person re-identification [Li et al., 2021b], text extraction from visual data [Miech et al., 2021], human-object interaction detection [Kim et al., 2021], among others. There

are specific designs of transformer architectures adopted to document processing [Xu et al., 2020, Beltagy et al., 2020, Appalaraju et al., 2021, Li et al., 2022]. For instance, an end-to-end transformer-based approach performing at the paragraph level was suggested in [Rouhou et al., 2022]. The authors proposed a transformer model adapted to recognize named entities in handwritten documents. Very few attempts have been undertaken to use the transformer for word spotting in historical documents. The authors in [MHIRI et al., 2022] put forward a word retrieval system based on QbE and QbS approaches. They adopted the strengths of convolutions and transformer layers for constructing a representation to encode both forms of words (texts and images). The word retrieval task is processed as a binary classification task in the space varying between representations.

In view of the great success of transformers in many domains, in this chapter, we propose a Triplet loss-based vision Transformer (TripTran) model for word retrieval in historical documents. We employ the transformer architecture and we detail how it can be adapted to build an embedding space basing on TL for image representations for performing word retrieval in a particular historical document. Section 6.3 provides more details about the proposed approach.

6.3 Proposed approach

In this chapter, we present a novel approach for embedding representations dedicated to word retrieval in historical documents.

More precisely, we propose a vision transformer-based triplet loss. To address the different aspects of document degradation, we suggest performing a pre-processing step. The main objective is to eliminate as many as possible forms of degradation in order to improve the visual quality of the documents while keeping the same textual content. A set of triplets is generated during the training process from a pre-segmented dataset. These triplets are formed by three word images: an anchor, a positive sample and a negative sample. Taking the enhanced triplets as input, a transformer architecture is trained using a triplet loss function. This training phase is designed to minimize the triplet loss and create an embedding space to minimize the distance between images of words associated with different classes and to minimize the distance between images of words associated with the same class. We investigate the problem of word spotting enhancement using the TL 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.

Figure.6.1 shows the flowchart of our proposed approach. More specifically, the triplet transformer takes as input a triplet of word images.

The idea involves training transformer architectures based on triplet 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). The leveraged features contributed in extracting a better embedding

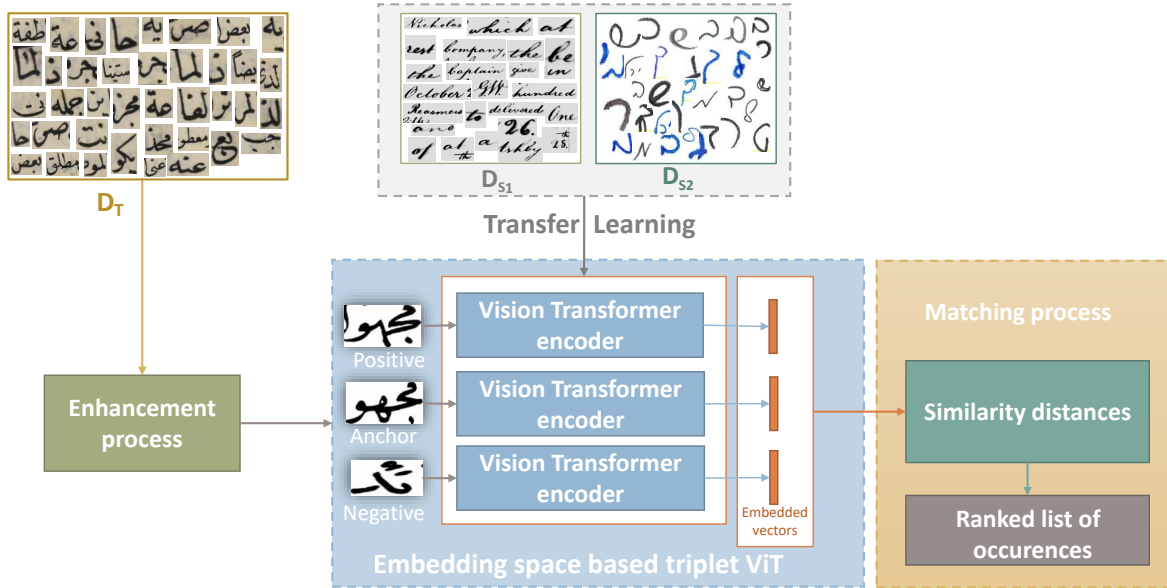


Figure 6.1: Proposed approach based on triplet transformer: TL is applied from historical English (D_{S1}) and handwritten Hebrew documents (D_{S2}) to Arabic documents (D_T).

features representation for word images.

To perform the word retrieval task, both images of the query words and of all the words in the document are projected into the embedding space previously built based on transformer and triplet loss in order to encode them with new representations (embedding vectors). Then, the embedding vectors are matched according to different similarity distances. Finally, the output is a list of retrieved words ordered by their distance to the query word. The proposed approach for building an embedding space is presented in Figure 6.3. Specifically, the pre-processing step is given in Section 6.3.1. The triplet loss function is then presented in Section 6.3.3. The details concerning how the proposed transformer is used to extract the relevant features from the word images are explained in Section 6.3.4, followed by the corresponding loss functions. A word spotting process is detailed in Section 6.3.5 where the embeddings matching process is introduced. The effectiveness of our proposed framework is highlighted by studying its impact in improving the performance of the word retrieval model on different historical documents, including Latin and Arabic.

6.3.1 Pre-processing

The present study utilizes a document enhancement preprocessing technique to mitigate the effects of image degradation in document images. The specific model employed in this endeavor is thoroughly described and discussed in the preceding chapter, with a focus on its ability to generate high-quality legible document images from degraded inputs.

6.3.2 Enhancement based Transfer Learning

TL [Pan and Yang, 2009] 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 $\mathcal{X} = \{x_1, \dots, x_n\} \in \mathcal{X}$. In this way, a mathematical description of the domain is given as $\mathcal{D} = \{\mathcal{X}, \mathcal{P}(\mathcal{X})\}$. TL includes 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 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}_{S_1} refers to the first source domain, historical English documents, and \mathcal{D}_{S_2} corresponds to the second source domain, handwritten Hebrew documents, while \mathcal{D}_T represents the target domain historical Arabic documents. All domains are assigned the same task \mathcal{T} .

Given a source domains \mathcal{D}_{S_1} and \mathcal{D}_{S_2} , corresponding source tasks \mathcal{T}_{S_1} and \mathcal{T}_{S_2} , as well as a target domain \mathcal{D}_T and a target task \mathcal{T}_T , the objective of TL now is to enable us to learn the target conditional probability distribution $\mathcal{P}(\mathcal{Y}_T|\mathcal{X}_T)$ in \mathcal{D}_T with the information gained from \mathcal{D}_S and \mathcal{T}_S where $\mathcal{D}_{S_1} \neq \mathcal{D}_{S_2} \neq \mathcal{D}_T$ and $\mathcal{T}_{S_1} = \mathcal{T}_{S_2} = \mathcal{T}_T$.

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

6.3.3 Transformer architecture

The proposed vision transformer architecture is depicted in Figure 6.2. It includes a patch embedding generation and a transformer encoder for image representation. The input image

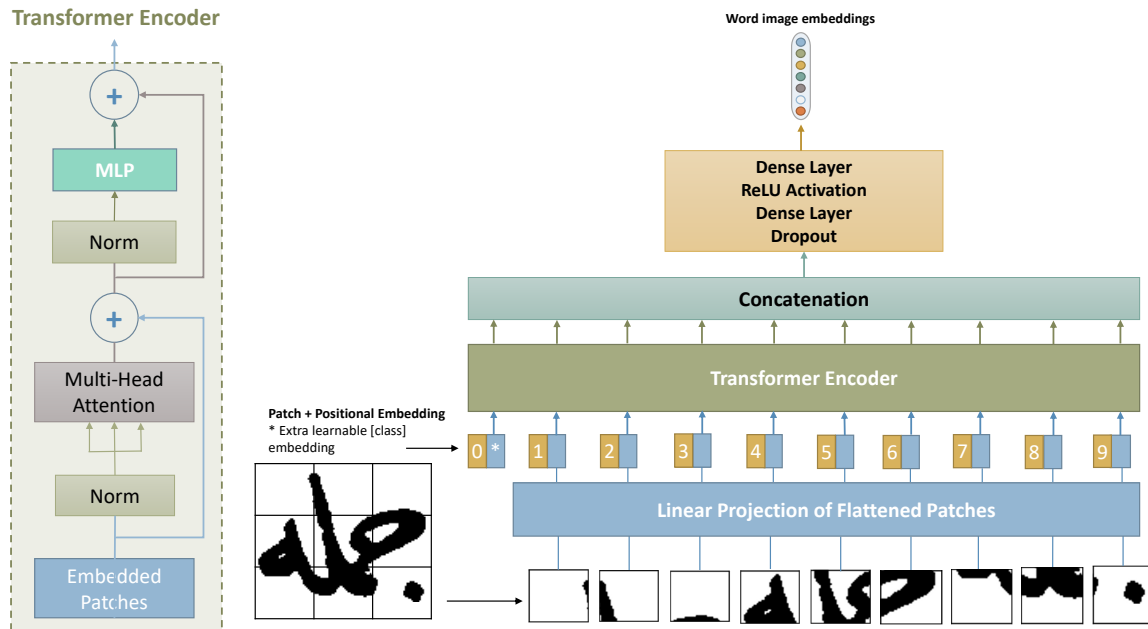


Figure 6.2: Vision Transformer architecture for word image representations.

$I \in \mathbb{R}^{m,m,c}$ is split into N non-overlapping patches $I_i \in \mathbb{R}^{k,k,c}$ with $i = 1, 2, \dots, N$, $m^2 = Nk^2$ and $c = 3$ the number of color channels. The input patches I_i are converted to vectors

$V_i \in \mathbb{R}^{1,d}$ with $d = ck^2$ and $i = 1, 2, \dots, N$.

Patch embeddings \mathcal{PE} will be generated based on a linear projection applied to each flattened vector. \mathcal{PE} are defined by Eq.(6.1) which will be optimized during the model training.

$$\mathcal{PE}_i = V_i \times \mathcal{W}_{\mathcal{PE}} \quad (6.1)$$

where $V_i \in \mathbb{R}^{1,d}$ denotes the flattened vector associated with the i^{th} patch, $\mathcal{W}_{\mathcal{PE}} \in \mathbb{R}^{d,de}$ represents the parameter matrix and $\mathcal{PE}_i \in \mathbb{R}^{1,de}$ is the projected vector associated with the i^{th} patch. In this case, de stands for the embeddings dimension (or hidden size) of the projection. Consequently, the projected embeddings associated with the entire input image are written in the form of $\mathcal{PE} \in \mathbb{R}^{N,de}$. We introduce an embedding of class $\mathcal{CT} \in \mathbb{R}^{1,de}$ as learnable parameters initialized to zero and concatenated with the projected embeddings in order to generate the extended embeddings ($\mathcal{EE} \in \mathbb{R}^{N+1,de}$) in the form of $\mathcal{EE} = [\mathcal{CT}, \mathcal{PE}]$ in the 1^{st} dimension. The position parameters ($\mathcal{PoE} \in \mathbb{R}^{N+1,de}$) are learnable parameters initialized by zeros and fed to the extended embeddings for incorporating spatial information. As a result, the final projection embedding (\mathcal{FE}) is given by Eq.(6.2).

$$\mathcal{FE} = dropout(\mathcal{EE} + \mathcal{PoE}, 0.1) \quad (6.2)$$

considering that *dropout* is a dropout layer with 0.1 as the dropout factor while $\mathcal{FE} \in \mathbb{R}^{N+1,de}$ represents the final projection of the final projection embeddings that will be provided as input to the transformer encoder.

6.3.3.1 Transformer encoder

The transformer encoder is based on a stack of L transformer blocks. By means of a self-attention mechanism, each transformer block converts the input hidden state into the same dimension output.

Here, the input of a transformer block of $\mathcal{T}_{j|j=1,2,\dots,L}$ is the hidden state h_{j-1} which corresponds to the output of the preceding transformer block \mathcal{T}_{j-1} , given that $h_{j-1} \in \mathbb{R}^{N+1,de}$. The input of the first transformer block \mathcal{T}_1 is the final projection integration (i.e., $h_0 = \mathcal{FE}$). The output of the last transformer block \mathcal{T}_L is the hidden state h_L . Typically, the output and input of the j^{th} transformer block are respectively notated as h_{j-1} and h_j .

The layer normalization constitutes the first layer of the j^{th} transformer block which receives as input h_{j-1} and generates $l_{j-1} \in \mathbb{R}^{N+1,de}$ as output. The input of the self-attention layer corresponds to the output of the layer normalization. It is formed by three configured projections according to the Query, Key and Value vectors employing the following learnable weight matrices: \mathcal{W}_Q , \mathcal{W}_K and \mathcal{W}_V .

We can define the output of the linear projections by Eq.(6.3).

$$\mathcal{S}_t = l_{j-1} \times \mathcal{W}_t \quad (6.3)$$

where $\mathcal{S}_t \in \mathbb{R}^{N+1,de}$ for $t = \{Q, K, V\}$.

Since the attention layer implements multiple attention heads, the *query* (\mathcal{S}_Q) and *value* vectors (\mathcal{S}_V) are reshaped to a dimension of $(\mathcal{A}_h, N + 1, \mathcal{A}_S)$. Moreover, the *Key* factor is reshaped to the dimension of $(\mathcal{A}_h, \mathcal{A}_S, N + 1)$ in which \mathcal{A}_h refers to the number of attention heads while \mathcal{A}_S refers to the size of the attention head derived as $\mathcal{A}_S = de/\mathcal{A}_h$.

Then, we perform a matrix multiplication of \mathcal{S}_Q and \mathcal{S}_K to generate an output matrix $\mathcal{S}_{QK} \in \mathcal{R}^{\mathcal{A}_h, N+1, N+1}$ which will be normalized to the square root of the attention head size as depicted by Eq.(6.4).

$$\mathcal{S}_{QK} = \frac{\mathcal{S}_Q \times \mathcal{S}_K}{\sqrt{\mathcal{A}_S}} \quad (6.4)$$

The attention weights labeled \mathcal{A}_w are subsequently computed by applying the softmax function on \mathcal{S}_{QK} , as shown in Eq.(6.5).

$$\mathcal{A}_w = \text{softmax}(\mathcal{S}_{QK}) \quad (6.5)$$

The attention-weighted features ($\mathcal{F}_A \in \mathbb{R}^{\mathcal{A}_h, N+1, \mathcal{A}_S}$) are then generated in terms of incorporating the attention weights \mathcal{A}_w with the value vector \mathcal{S}_V as defined by Eq.(6.6).

$$\mathcal{F}_A = \mathcal{A}_w \times \mathcal{S}_V \quad (6.6)$$

The attention-weighted features \mathcal{F}_A is then reshaped in the dimension of $(N + 1, de)$ where $\mathcal{F}_A \in \mathbb{R}^{N+1, de}$.

A linear projection is lastly performed on \mathcal{F}_A using a learnable parameter \mathcal{W}_A for generating the output of the self-attention module designed by (\mathcal{F}_S), where $\mathcal{F}_S = \mathcal{F}_A \times \mathcal{W}_A$. For a better learning process, the transformer blocks consist of the residual connection depicted by Eq.(6.7).

$$\mathcal{F}_R = h_{j-1} + \mathcal{F}_S \quad (6.7)$$

where h_{j-1} stands for the input of the transformer block, \mathcal{F}_S denotes the output of the self-attention module and \mathcal{F}_R stands for the output of the residual connection.

6.3.4 Triplet loss

The last layer of the transformer encoder is defined as a fully connected layer of size 1024. By means of L2 normalization, the embedding vectors are obtained and the triplet loss is calculated based on this feature representation. In our proposed approach, the output of the transformer encoder network will be the input of the triplet loss function, which is an embedding mapping function represented as $\mathcal{G}(x)$. The triplet loss maps an image x into a d -dimensional Euclidean space as depicted in Figure 6.3. Considering that there is no target to reach during the training process, the selection of the triples has a significant impact on the model convergence as well as on the experimental results. Following the authors of [Schroff et al., 2015b], we focus on a semi-hard triplet selection strategy given by Eq.(6.8).

$$\|\mathcal{G}(a) - \mathcal{G}(p)\|_2^2 < \|\mathcal{G}(a) - \mathcal{G}(n)\|_2^2 < \|\mathcal{G}(a) - \mathcal{G}(p)\|_2^2 + \alpha \quad (6.8)$$

where the embedded representation of the network is designed by $\mathcal{G}()$ and α represents a margin used between positive and negative pairs.

6.3.5 Embeddings matching

To determine the images of the words most similar to the query word, it is sufficient to perform a comparison between the image of the query word and all the word images identified in the document. Otherwise, the similarity between all images in the document and the query

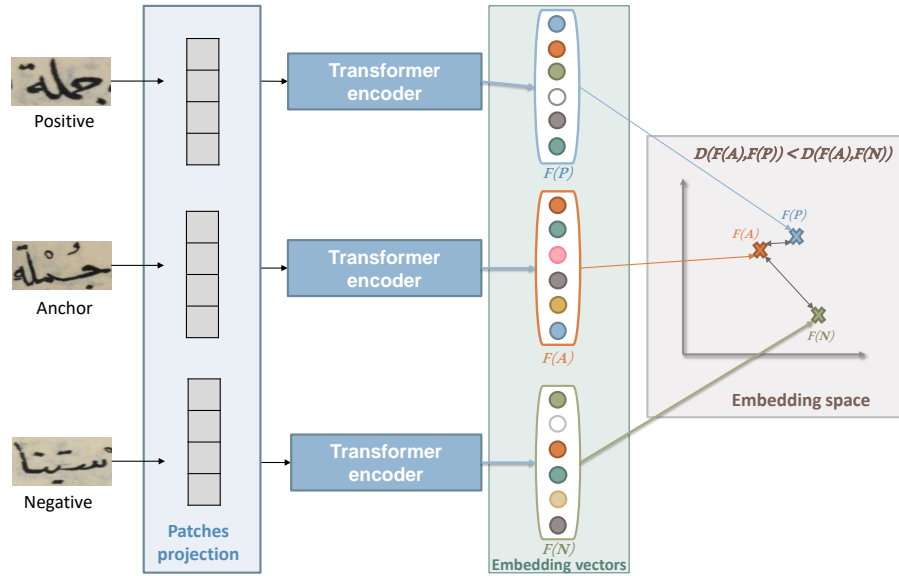


Figure 6.3: Flowchart of the proposed TripTran architecture.

word will be calculated. It is relying on a basic concept: The more the feature values are close, the more the similarity is high. Each word or sub-word will be presented by a vector, i.e. an embedding vector following the TripTran model application. Subsequently, the spotting system quantifies the similarity in the embedding space between two embedding vectors and ranks all these images according to their similarity with the query, which provides the occurrences of the query word. Considering that all feature vectors are normalized, the most basic and commonly adopted similarity measure is the Euclidean distance in the embedding space.

6.4 Experiments

As previously mentioned, the main objective of this chapter is to provide a word retrieval system in historical document images based on the vision transformer approach. For the assessment, a brief description of the employed datasets is included. Then, the experimental protocol and the evaluation measures used are described. Finally, the obtained results are analyzed and discussed in order to obtain an objective evaluation against the latest state-of-the-art systems.

6.4.1 Experimental setup

We conduct an analysis of the data obtained from three datasets: VML-HD [Kassis et al., 2017], HADARA80P [Pantke et al., 2014] and GW [Rath and Manmatha, 2007]. These datasets have been previously detailed in preceding chapters. For transfer learning, both GW and Hebrew Handwritten dataset (HHD) [Rabaev et al., 2020] are employed. Table 6.1 provides the partition of the employed datasets. “Train” refers to the training set for each data set. Meanwhile “Val” is the validation set used for model learning. Then “Test” corresponds to the test set that is introduced in the evaluation phase.

Table 6.1: Partitioning of datasets according to the level of word classes and the number of images per class.

Dataset	Sub-set	#words	#Samples/word
VML-HD	Train	141	10
	Val	20	10
	Test	105	100
HDD	Train	27	20
	Val	27	10
HADARAP80	Test	50	10
GW	Test	85	10

6.4.1.1 Evaluation protocol

Following the same evaluation criteria as most word spotting techniques [Barakat et al., 2018, Mhiri et al., 2019], the performance of the proposed TripTran model is assessed in terms of Precision at the top- K -retrievals ($P@K$) [Deng et al., 2011] and the mean Average Precision (mAP) [Everingham et al., 2010].

6.4.1.2 Implementation details

The learning process is optimized by the stochastic gradient descent algorithm with a learning rate of 10^{-3} and a batch size of 512. The early stopping function with *patience* = 50 is employed to prevent overfitting. All parameter values are empirically chosen. The model is trained on an NVIDIA Quadro RTX 6000 GPU with 24 GB of RAM.

6.4.2 Results and discussions

The experimental results in terms of $P@K$ and mAP are summarized in Table 6.2 for word image retrieval on the VML-HD dataset. We assess the model on five books of the VML-HD dataset to emphasize the impact of our proposed TripTran on the embedded feature representation and matching. The results presented in Table 6.2 are based on a matching process involving the Euclidean distance.

Table 6.2: Results of our proposed TripTran model on VML-HD dataset according to $P@K$ and mAP metrics.

Books	P@1	P@2	P@3	P@4	P@5	mAP
Book 1	1	1	1	1	1	0.88
Book 2	1	1	1	1	0.97	0.90
Book 3	1	1	1	1	1	0.86
Book 4	1	1	1	0.99	0.98	0.82
Book 5	1	1	1	0.98	0.98	0.84

Supplementary results are presented in Table 6.4 to highlight the performance of the proposed system. The historical datasets used are derived from different scripts. VML-HD and HADARA80P present the Arabic script, while GW is the Latin script. We provide the matching results based on the different similarity distances calculated between each pair of embedding vectors.

Table 6.3: Word spotting results in terms of P@k metric for some query word images used in our experiment.








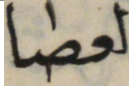




















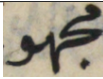





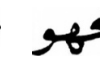
Query image	Enhanced image	Top 5 retrieved				
		P@1	P@2	P@3	P@4	P@5
						
						
						
						
						

Table 6.3 displays some query words randomly selected from the VML-HD dataset in original, enhanced forms and their top five occurrences retrieved by TripTran. It can be noticed that the proposed model succeeds in finding the occurrences of all the query words correctly except the 3rd word. This word is wrongly retrieved at the 5th rank. This is explained by model confusion and the high similarity between the two matched images.

Table 6.4: Results of our proposed model on different datasets according to P@K and mAP metrics.

Database	Distance	P@1	P@2	P@3	P@4	P@5	mAP
VML-HD	Euclidean	1	1	1	1	0.98	0.86
	Cosine	1	1	1	0.99	0.98	0.85
	Manhattan	1	1	1	0.98	0.98	0.85
HADARA80P	Euclidean	0.95	0.96	0.93	0.93	0.92	0.82
	Cosine	0.95	0.96	0.94	0.93	0.92	0.81
	Manhattan	0.88	0.87	0.90	0.90	0.89	0.80
GW	Euclidean	1	1	0.98	0.99	0.98	0.99
	Cosine	1	0.99	0.98	0.98	0.99	0.98
	Manhattan	0.99	0.98	0.98	0.98	0.99	0.99

In Table 6.5, our results are compared with other methods designed for word retrieval, specifically the methods that are based on feature representations. The comparison is performed on the same test part from the VML-HD dataset. As it has been proved in Table

6.2, the Euclidean distance is the most appropriate similarity distance for matching embeddings, so we perform the comparison results based on this distance. Our method clearly outperforms the existing methods by a significant margin.

Table 6.5: Results on VML-HD according to P@K and mAP metrics using Euclidean distance: Comparison with state-of-the-art methods.

Methods	P@1	P@2	P@3	P@4	P@5	mAP
[Barakat et al., 2018]	0.88	0.85	0.86	0.89	0.89	0.66
[Fathallah et al., 2019]	0.90	0.89	0.89	0.88	0.89	0.73
TripTran	1	1	1	0.99	0.98	0.86

Several different observations are proposed in order to investigate the impact of ViT technique on the word retrieval system. First, according to $P@1$, our proposed TripTran system provides an average improvement of 12% and 10% on all books compared to Siamese and Triplet. Furthermore, at ranks 2, 3, 4 and 5, improvements ranging from 11% to 15% are obtained. We can note that the transformer architectures significantly contribute to training the model on VML-HD for better feature embedding representations. Secondly, according to mAP values, The TripTran performance is improved by 19% and 12% compared to [Barakat et al., 2018] and [Fathallah et al., 2019] respectively.

Finally, from a different point of view, the combination of the vision transformer and the triplet loss improves the word retrieval system by constructing an appropriate embedding space for word image representations. This can be explained by the fact that the strong transformer architecture and the preprocessing step are applied to all datasets to improve the visual image quality.

Additionally, in terms of mAP , our TripTran model is compared to different methods from the state of the art according to the VML-HD and GW datasets. The same evaluation protocol is applied. As depicted in Figure 6.4, the result is again satisfactory: TripTran presents the best results obtained compared to [Barakat et al., 2018], [Kassis and El-Sana, 2016] and [Fathallah et al., 2019] according to the VML-HD dataset. Clearly, this figure confirms the advantage of the vision transformer representation over CNN and graph techniques.

Our proposed TripTran model considerably exceeds the performance of much prominent work regarding the GW dataset. Notably, the same matching algorithm is applied. Figure 6.4 gives an overview of the superiority of our approach over GW in terms of the mAP metric. These results can be explained by the high ability of the vision transformer to extract meaningful features from the word images for encoding into the embedding vectors. Our approach beats the HWNet v2 [Krishnan and Jawahar, 2019], PHOCNet [Sudholt and Fink, 2016], Attribute SVM [Almazán et al., 2014b] and SC-HMM [Rodríguez-Serrano and Perronnin, 2012] methods by an improvement range from 3.29% to 50%. However, there is no improvement compared to WSNet [Mohammed et al., 2022].

6.4.3 Error Analysis

Our proposed Triplet transformer-based TL has demonstrated better performance on word retrieval in HADs. Despite its high performance, the word spotting process has shown miss-

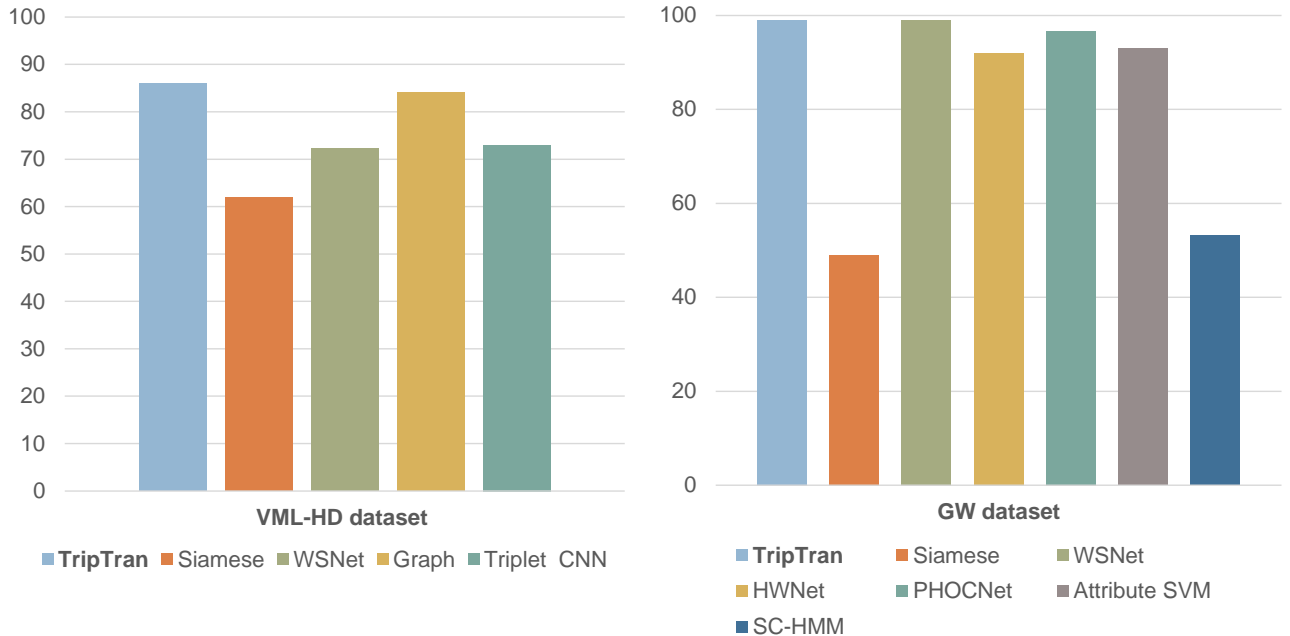


Figure 6.4: Results on VML-HD and GW datasets in terms of mAP metric using Euclidean distance: Comparison with state-of-the-art methods.

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.6.5 introduces some examples of word classes that were 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, ambiguity between similar word classes is the first source. It is commonly known that the transformer architecture performs well in representing data when there is a clear separation between word classes but this 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 letter parts of another preceding or 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 triplet transformer model: in general, data labeling is a subjective task because it is provided by human judgments.

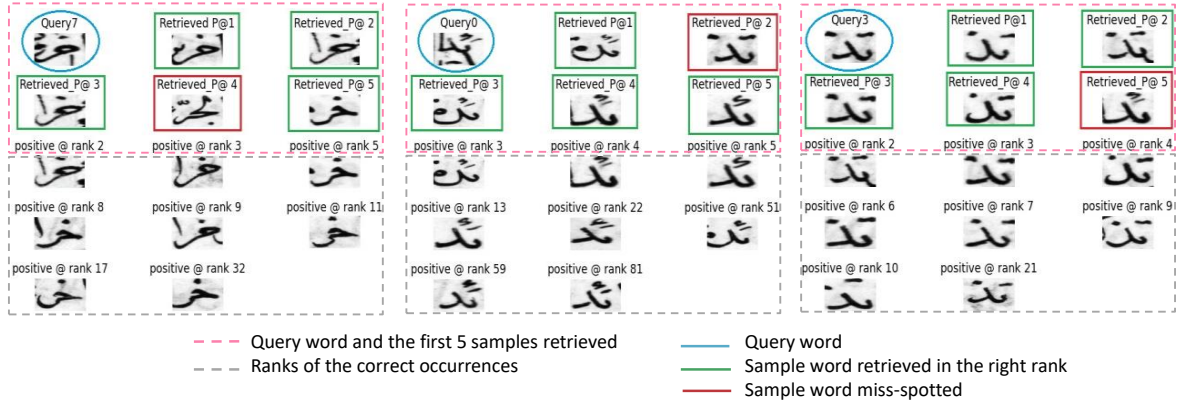


Figure 6.5: Some examples of miss-retrieved words: Error analysis with displaying the first five ranks spotted (from P@1 to P@5) and the right occurrences ranks.

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 missegmented word images.

6.4.4 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 transfer learning via GW (TL-GW) and transfer learning via HHD (TL-HDD), we conduct an ablation study on three widely used datasets: VML-HD, HADARA80P, 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 6.6: 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	✓	✗	0.82	0.78	0.86
	✗	✓	0.84	0.80	0.96
	✓	✓	0.86	0.82	0.99

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 transfer learning based on the combination of the two domains, as shown in the last row of Table 6.6.

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 7%, 6% and 16% 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.

6.5 Conclusion

In this chapter, we introduced a revolutionary vision transformer called TripTran, which is based on a triplet network architecture for word retrieval in historical documents. The primary goal of this model is to generate more discriminative, efficient, and accurate feature embeddings for the word spotting task. Through the implementation of transfer learning from different domains, specifically TL-HDD and TL-GW, we have seen a substantial improvement in enhancing a word spotting model. Our results demonstrate that TripTran surpasses previous methods for word retrieval in a diverse set of historical documents. The combination of the triplet loss function and the feature extraction-based vision transformer enhances the generalization ability of the model. Therefore, it is clear that the vision transformer for feature representation significantly improves the performance of word spotting due to its capability to learn discriminative features.

CHAPTER 7

Conclusion and perspectives

7.1	Conclusion	110
7.2	Perspectives	112

7.1 Conclusion

A huge collection of historical documents is available in libraries and national archives across the world. These documents suffer from different forms of degradation, which makes them difficult to interpret and significantly affects the performance of any word retrieval process. The cultural heritage of historical Arabic documents is not always accessible. Storage conditions are still inadequate to the requirements of these documents. To safeguard and preserve these precious resources, countless of pages are scanned and deposited at different servers. With the enormous technological advances of recent years, the amount of digitized historical documents, both handwritten and printed, has increased. It is well known that digital historical documents are not easily processed in their original form, but they need to be transformed into a readable form in order to be automatically understood by computer vision tools.

A recent trend in the image processing of HADs is word spotting which has gained attention in the document analysis field over the last two decades. It always involves many challenges owing to the complexity of these documents. The ability to locate the occurrences of a particular word in a large set of historical document images is a daunting task due to the intricacies of the Arabic script. Hence, an approach to spot a query word in such a document is needed. The above challenges have been addressed in this thesis based on Deep Learning approaches. In particular, the thesis work has been outlined as follows:

- The first chapter provided the motivation and objectives of this PhD thesis. Firstly, we gave a brief introduction to the research field of document analysis and indexing, including the analysis of historical documents in particular. Then, an overview of the word retrieval architecture was discussed, and its applicability to historical documents was depicted. Next, the general context of the PhD thesis was introduced and the main encountered difficulties and challenges were highlighted. Finally, we outlined the objectives and the contributions of this work.
- **Chapter 2** was designed to present the state of the art of indexing historical Arabic documents where the importance of historical documents and its several challenges were introduced. Then, the same presentation was made for the historical Arabic documents. Following this, different automatic processing systems for historical documents were discussed, in particular word retrieval systems. The next section was devoted to a set of public datasets designed for the word retrieval process. Finally, some metrics for evaluating the word retrieval process were highlighted.
- The contribution, introduced in **chapter 3**, relates to our first application for word spotting in historical Arabic documents which is to perform building embedding space for features representation. We investigated the efficiency of learning feature representations to ensure the ability of a more effective features extraction to describe word images. The proposed approach consisted of two major steps. First, based on a pre-segmented training dataset, we constructed an embedding space using the triplet-loss in the context of a convolutional neural network. Its construction aims to maximize the distance between word images associated with different classes and minimize the distance between word images associated with the same class. Second, to spot a query word, images are projected on the previously learned embedding

space, and then matched via the Euclidean distance. Despite the complexity of HAD in terms of backgrounds, writing styles variation, etc., the experiments, performed on the VML-HD dataset, show that the Triplet CNN provided better results than its immediate competitor, the Siamese CNN. Furthermore, different dimensionality reduction techniques have been investigated in order to highlight the best image representations in the context of historical Arabic documents. we presented a comparative study of four embedding methods: triplet CNN, Siamese CNN, PCA and LDA.

- **Chapter 4** presented an improved version to build an embedding space for a word spotting model by following many enhancement strategies. In particular, it has investigated a solution that enables to learn more discriminative, reliable and efficient feature embeddings for the word spotting task. As a first step, we have put forward a conditional GAN as a means to generate clean document images from highly degraded images. Our enhancement model has been designed to handle different degradation tasks such as watermark removal and chemical degradation with the goal to producing hyper-clean document images and fine detail recovery performances. Moreover, we employed transfer learning, which involves reusing pre-trained networks based on more complex HADs. One of the main advantages of such a transfer learning approach is that it allows for more convergent learning where supplementary knowledge and features are easily transferred to the word image representation task. To do this, we firstly trained a triplet-CNN on historical documents and precisely on the HADARA80P, which is an 80-page HADs. Secondly, we train the same triplet-CNN architecture and use historical Arabic documents, while applying transfer learning based on the previous models generated in the first step. In addition, the training process is enhanced by following an online triplet mining and semi-hard triplet selection techniques. Extensive experiments have shown that our employed strategies have resulted in an interesting improvement of the word spotting task in historical documents compared to many recent state-of-the-art methods. We also performed an evaluation of our solution on reference datasets such as HADARA80P, VML-HD and GW.
- In **Chapter 5**, we have presented a conditional GAN as a means to generate clean document images from highly degraded images. Our enhancement model has been designed to handle different degradation tasks such as watermark removal and chemical degradation with the goal to producing hyper-clean document images and fine detail recovery performances. The task of our GAN model is related to the ability to generate a clean version of a degraded historical document. Therefore, we considered the task as an image-to-image conversion process. Specifically, our proposed model aimed at learning a mapping from the degraded document image x to the clean document image y . During the training process, the proposed GAN took a degraded document image as input and tried to generate a clean version of this image. On the other hand, the discriminator took two inputs: the generated image and the ground truth which referred to the cleaned version of the degraded images. Then, it discriminated whether the generated image was real or not according to the ground truth. Our model network involved generator G and discriminator D . G was trained to transform a degraded document image into a clean document image, while D helped G to produce a more realistic image by distinguishing between generated and real images.

- **Chapter 6** presented a novel approach for embedding representations dedicated to word retrieval in historical documents. More precisely, we proposed a vision transformer based triplet loss. To address the different aspects of document degradation, we suggested to perform a pre-processing step. The main objective was to eliminate as many as possible forms of degradation in order to improve the visual quality of the documents while keeping the same textual content. Taking the enhanced triplets as input, a transformer architecture was trained using a triplet loss function. This training phase was designed to minimize the triplet loss and create an embedding space to minimize the distance between images of words associated with different classes and to minimize the distance between images of words associated with the same class. To perform the word retrieval task, both images of the query words and of all the words in the document were projected into the embedding space previously built based on transformer and triplet loss in order to encode them with new representations (embedding vectors). Then, the embedding vectors were matched according to different similarity distances. Finally, the output was a list of retrieved words ordered by their distance to the query word.

7.2 Perspectives

In the present thesis, we illustrated the effectiveness of deep learning models in improving historical document indexing tasks. However, the research is not over yet, there are still a lot of exciting issues to investigate. In the following, we sketch some upcoming work.

- Basically, whenever a user wants to spot for a particular word in a given digital document, the default approach is to use a query by string, so the user simply types the letters of the word to be found in the corresponding box. For historical documents, this can only be done through the transcription phase. Automatic transcription of a historical document is challenging due to the unavailability of labeled training data and requires that each model has to be trained on the particular alphabet to be recognized, and whenever the alphabet changes, the system has to be re-trained from scratch with samples of the new script. The goal is to provide word spotting system based on query by string approach in historical documents. In order to accomplish this task, i.e., to locate words in historical documents, we can develop a model that takes a text word as input and outputs the occurrences of this word in historical document images. From a text word, a capture of that word is generated as a printed word image. The goal is to design a general historical image synthesis framework capable of rendering various printed word images with controllable historical styles. The challenge of this task lies in the difficulty to generate high quality results due to significant geometric changes in the script and variation in the texture style. Thus, unlike predefined styles that can be learned directly by training on large-scale databases, we will propose a GAN model to transfer the historical style. In particular, we will intend to train a common historical translator which can render printed word images as historical as possible. It is challenging due to the complexity of the task and the absence of paired data. The core idea is to introduce gated cycle mapping, that employs a novel gated mapping unit to produce the category-specific style code and embeds this code into cycle networks to

control the translation process. We will use only a historical collection of mixed styles with the aim to render printed word images in a historical style with the trained model.

- Inspired by the success of vision transformers approaches in many different applications, their encoding-decoding architectures have actually started to be applied to data representation. Their mechanism allows to highlight the global interactions among contextual features. Combining the use of local information with the knowledge of the global long-range spatial arrangement provides an interesting contribution to an efficient image restoration model. Such local information content is typically encoded at the patch level of an image, while the large-scale organization is contained in the redundancy of this information across patches in the image. Unlike CNNs, which deal with arrays of pixels, vision transformers (ViTs) allow an image to be divided into fixed-size patches, properly integrate each of them as a latent representation, and provide positional embedding information as input to the transformer's encoder. This allows the relative location of patches to be encoded, as well as local (spatial) and global (semantic) long-term dependencies. We will propose a global base model based on the use of ViTs. A missing or degraded patch in the distorted document image can be recovered from information about neighboring patches through the power of multi-headed self-attention of ViTs, which quantifies the global pairwise reasoning between them. Furthermore, the ViTs were incorporated into the global model pipeline in an encoder-decoder based framework, inspired by the concept of a denoising autoencoder used in the reconstruction of corrupted input data. The encoder maps degraded image patches into latent representations while the decoder generates the restored image.

Bibliography

- [Adamek et al., 2007] Adamek, T., O'Connor, N. E., and Smeaton, A. F. (2007). Word matching using single closed contours for indexing handwritten historical documents. *International Journal of Document Analysis and Recognition (IJ DAR)*, 9(2):153–165.
- [Afzal et al., 2015] Afzal, M. Z., Pastor-Pellicer, J., Shafait, F., Breuel, T. M., Dengel, A., and Liwicki, M. (2015). Document image binarization using lstm: A sequence learning approach. In *Proceedings of the 3rd international workshop on historical document imaging and processing*, pages 79–84.
- [Agarwal et al., 2021] Agarwal, N., Sondhi, A., Chopra, K., and Singh, G. (2021). Transfer learning: Survey and classification. In *Smart Innovations in Communication and Computational Sciences*, pages 145–155. Springer.
- [Aldavert et al., 2013] Aldavert, D., Rusinol, M., Toledo, R., and Lladós, J. (2013). Integrating visual and textual cues for query-by-string word spotting. In *2013 12th International conference on document analysis and recognition*, pages 511–515. IEEE.
- [Almazán et al., 2014a] Almazán, J., Gordo, A., Fornés, A., and Valveny, E. (2014a). Segmentation-free word spotting with exemplar svms. *Pattern recognition*, 47(12):3967–3978.
- [Almazán et al., 2014b] Almazán, J., Gordo, A., Fornés, A., and Valveny, E. (2014b). Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2552–2566.
- [Aly and Dubois, 2005] Aly, H. A. and Dubois, E. (2005). Image up-sampling using total-variation regularization with a new observation model. *IEEE Transactions on Image Processing*, 14(10):1647–1659.
- [Annabestani and Saadatmand-Tarzjan, 2019] Annabestani, M. and Saadatmand-Tarzjan, M. (2019). A new threshold selection method based on fuzzy expert systems for separating text from the background of document images. *Iranian journal of science and technology, transactions of electrical engineering*, 43(1):219–231.

- [Aouadi and Echi, 2014] Aouadi, N. and Echi, A. K. (2014). Prior segmentation of old arabic manuscripts by separator word spotting. In *2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, pages 31–36. IEEE.
- [Aouadi and Kacem, 2011] Aouadi, N. and Kacem, A. (2011). Word spotting for arabic handwritten historical document retrieval using generalized hough transform. In *Proceedings of the Third International Conference on Pervasive Patterns and Applications, Rome, Italy*, pages 67–71.
- [Appalaraju et al., 2021] Appalaraju, S., Jasani, B., Kota, B. U., Xie, Y., and Manmatha, R. (2021). Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 993–1003.
- [Arvanitopoulos et al., 2017] Arvanitopoulos, N., Chevassus, G., Maggetti, D., and Süssstrunk, S. (2017). A handwritten french dataset for word spotting: Cframuz. In *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing*, pages 25–30.
- [Balakrishnama and Ganapathiraju, 1998] Balakrishnama, S. and Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18:1–8.
- [Balasubramanian et al., 2006] Balasubramanian, A., Meshesha, M., and Jawahar, C. (2006). Retrieval from document image collections. In *International Workshop on Document Analysis Systems*, pages 1–12. Springer.
- [Barakat et al., 2018] 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.
- [Beltagy et al., 2020] Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- [Boudraa et al., 2022] Boudraa, O., Michelucci, D., and Hidouci, W. K. (2022). Punet: Novel and efficient deep neural network architecture for handwritten documents word spotting. *Pattern Recognition Letters*, 155:19–26.
- [Can and Kabadayı, 2020] 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.
- [Cao and Govindaraju, 2007] Cao, H. and Govindaraju, V. (2007). Template-free word spotting in low-quality manuscripts. In *Advances In Pattern Recognition*, pages 135–139. World Scientific.
- [Carion et al., 2020] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer.

- [Chaieb et al., 2015] Chaieb, R., Kalti, K., and Essoukri Ben Amara, N. (2015). Interactive content-based document retrieval using fuzzy attributed relational graph matching. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 921–925. IEEE.
- [Chen et al., 2017a] Chen, K., Zhao, T., Yang, M., Liu, L., Tamura, A., Wang, R., Utiyama, M., and Sumita, E. (2017a). A neural approach to source dependence based context model for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2):266–280.
- [Chen et al., 2020] Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. (2020). Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR.
- [Chen et al., 2011] Chen, X., He, X., Yang, J., and Wu, Q. (2011). An effective document image deblurring algorithm. In *CVPR 2011*, pages 369–376. IEEE.
- [Chen et al., 2021] Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., and Lu, H. (2021). Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8126–8135.
- [Chen et al., 2017b] Chen, Y., Duffner, S., Stoian, A., Dufour, J.-Y., and Baskurt, A. (2017b). Triplet cnn and pedestrian attribute recognition for improved person re-identification. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE.
- [Cheriet et al., 1998] Cheriet, M., Said, J. N., and Suen, C. Y. (1998). A recursive thresholding technique for image segmentation. *IEEE transactions on image processing*, 7(6):918–921.
- [Chou et al., 2010] Chou, C.-H., Lin, W.-H., and Chang, F. (2010). A binarization method with learning-built rules for document images produced by cameras. *Pattern Recognition*, 43(4):1518–1530.
- [Chutani et al., 2015] Chutani, G., Patnaik, T., and Dwivedi, V. (2015). An improved approach for automatic denoising and binarization of degraded document images based on region localization. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2272–2278. IEEE.
- [Deng et al., 2011] 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.
- [Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- [Dong et al., 2015] Dong, C., Loy, C. C., He, K., and Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307.

- [Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- [El Yacoubi et al., 1995] El Yacoubi, A., Bertille, J.-M., and Gilloux, M. (1995). Conjoined location and recognition of street names within a postal address delivery line. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pages 1024–1027. IEEE.
- [En et al., 2016] En, S., Petitjean, C., Nicolas, S., and Heutte, L. (2016). A scalable pattern spotting system for historical documents. *Pattern Recognition*, 54:149–161.
- [Everingham et al., 2010] 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.
- [Faisal and AlMaadeed, 2017] Faisal, T. and AlMaadeed, S. (2017). Enabling indexing and retrieval of historical arabic manuscripts through template matching based word spotting. In *Arabic Script Analysis and Recognition (ASAR), 2017 1st International Workshop on*, pages 57–63. IEEE.
- [Farghaly and Shaalan, 2009] Farghaly, A. and Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4):1–22.
- [Fathallah et al., 2023a] Fathallah, A., El-Yacoubi, M. A., and Amara, N. E. B. (2023a). Ehdi: Enhancement of historical document images via generative adversarial network. In *VISAPP 2023: 18th International Conference on Computer Vision Theory and Applications*. VISAPP.
- [Fathallah et al., 2023b] Fathallah, A., El-Yacoubi, M. A., and Amara, N. E. B. (2023b). Transfer learning for word spotting in historical arabic documents based triplet-cnn. In *VISAPP 2023: 18th International Conference on Computer Vision Theory and Applications*. VISAPP.
- [Fathallah et al., 2020] 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 et al., 2019] 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.
- [Fernández et al., 2011] Fernández, D., Lladós, J., and Fornés, A. (2011). Handwritten word spotting in old manuscript images using a pseudo-structural descriptor organized in a hash structure. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 628–635. Springer.

- [Fernández-Mota et al., 2014a] Fernández-Mota, D., Almazán, J., Cirera, N., Fornés, A., and Lladós, J. (2014a). Bh2m: The barcelona historical, handwritten marriages database. In *2014 22nd International Conference on Pattern Recognition*, pages 256–261. IEEE.
- [Fernández-Mota et al., 2014b] Fernández-Mota, D., Riba, P., Fornés, A., and Lladós, J. (2014b). On the influence of key point encoding for handwritten word spotting. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 476–481. IEEE.
- [Finlayson et al., 2002] Finlayson, G. D., Hordley, S. D., and Drew, M. S. (2002). Removing shadows from images. In *European conference on computer vision*, pages 823–836. Springer.
- [Fischer et al., 2012] Fischer, A., Keller, A., Frinken, V., and Bunke, H. (2012). Lexicon-free handwritten word spotting using character hmms. *Pattern recognition letters*, 33(7):934–942.
- [Gangeh et al., 2021] Gangeh, M. J., Plata, M., Motahari, H., and Duffy, N. P. (2021). End-to-end unsupervised document image blind denoising. *arXiv preprint arXiv:2105.09437*.
- [Gatos et al., 2005] Gatos, B., Konidakis, T., Ntzios, K., Pratikakis, I., and Perantonis, S. J. (2005). A segmentation-free approach for keyword search in historical typewritten documents. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 54–58. IEEE.
- [Gatos and Pratikakis, 2009] Gatos, B. and Pratikakis, I. (2009). Segmentation-free word spotting in historical printed documents. In *2009 10th international conference on document analysis and recognition*, pages 271–275. IEEE.
- [Gatos et al., 2004] Gatos, B., Pratikakis, I., and Perantonis, S. J. (2004). An adaptive binarization technique for low quality historical documents. In *International Workshop on Document Analysis Systems*, pages 102–113. Springer.
- [Gatos et al., 2015] Gatos, B., Stamatopoulos, N., Louloudis, G., Sfikas, G., Retsinas, G., Papavassiliou, V., Sunistira, F., and Katsouros, V. (2015). Grpoly-db: An old greek polytonic document image database. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 646–650. IEEE.
- [Gonzalez et al., 2002] Gonzalez, R. C., Woods, R. E., et al. (2002). Digital image processing. *Publishing house of electronics industry*, 141(7).
- [Goodfellow et al., 2020] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.
- [Gurjar et al., 2018] 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.

- [Hedjam and Cheriet, 2013] Hedjam, R. and Cheriet, M. (2013). Historical document image restoration using multispectral imaging system. *Pattern Recognition*, 46(8):2297–2312.
- [Hedjam et al., 2014] Hedjam, R., Cheriet, M., and Kalacska, M. (2014). Constrained energy maximization and self-referencing method for invisible ink detection from multispectral historical document images. In *2014 22nd International Conference on Pattern Recognition*, pages 3026–3031. IEEE.
- [Howe, 2013] Howe, N. R. (2013). Document binarization with automatic parameter tuning. *International journal on document analysis and recognition (ijdar)*, 16(3):247–258.
- [Isola et al., 2017] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.
- [Jemni et al., 2022] Jemni, S. K., Souibgui, M. A., Kessentini, Y., and Fornés, A. (2022). Enhance to read better: A multi-task adversarial network for handwritten document image enhancement. *Pattern Recognition*, 123:108370.
- [Johnson et al., 2016] Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer.
- [Kang et al., 2021] Kang, S., Iwana, B. K., and Uchida, S. (2021). Complex image processing with less data—document image binarization by integrating multiple pre-trained u-net modules. *Pattern Recognition*, 109:107577.
- [Kassis et al., 2017] Kassis, M., Abdalhaleem, A., Droby, A., Alaasam, R., and El-Sana, J. (2017). Vml-hd: The historical arabic documents dataset for recognition systems. In *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, pages 11–14. IEEE.
- [Kassis and EL-Sana, 2014] Kassis, M. and EL-Sana, J. (2014). Word spotting using radial descriptor. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 387–392. IEEE.
- [Kassis and El-Sana, 2016] Kassis, M. and El-Sana, J. (2016). Word spotting using radial descriptor graph. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 31–35. IEEE.
- [Kesidis et al., 2011] Kesidis, A. L., Galiotou, E., Gatos, B., and Pratikakis, I. (2011). A word spotting framework for historical machine-printed documents. *International Journal on Document Analysis and Recognition (IJDR)*, 14(2):131–144.
- [Kesiman et al., 2016] Kesiman, M. W. A., Burie, J.-C., Wibawantara, G. N. M. A., Sunarya, I. M. G., and Ogier, J.-M. (2016). Amadi_lontarset: The first handwritten balinese palm leaf manuscripts dataset. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 168–173. IEEE.

- [Keyvanpour et al., 2014] Keyvanpour, M., Tavoli, R., and Mozafari, S. (2014). Document image retrieval based on keyword spotting using relevance feedback. *INTERNATIONAL JOURNAL OF ENGINEERING*.
- [Khaissidi et al., 2016] Khaissidi, G., Elfakir, Y., Mrabti, M., Lakhliai, Z., Chenouni, D., et al. (2016). Segmentation-free word spotting for handwritten arabic documents. *International Journal of Interactive Multimedia & Artificial Intelligence*, 4(1).
- [Khan et al., 2021] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2021). Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*.
- [Khayyat and Suen, 2018] 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 and Elrefaei, 2020] 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.
- [Khurshid et al., 2009] Khurshid, K., Faure, C., and Vincent, N. (2009). A novel approach for word spotting using merge-split edit distance. In *International Conference on Computer Analysis of Images and Patterns*, pages 213–220. Springer.
- [Kim et al., 2021] Kim, B., Lee, J., Kang, J., Kim, E.-S., and Kim, H. J. (2021). Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83.
- [Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- [Konidaris et al., 2007] Konidaris, T., Gatos, B., Ntzios, K., Pratikakis, I., Theodoridis, S., and Perantonis, S. J. (2007). Keyword-guided word spotting in historical printed documents using synthetic data and user feedback. *International Journal of Document Analysis and Recognition (IJDAR)*, 9(2):167–177.
- [Krishnan et al., 2016] Krishnan, P., Dutta, K., and Jawahar, C. (2016). Deep feature embedding for accurate recognition and retrieval of handwritten text. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 289–294. IEEE.
- [Krishnan and Jawahar, 2016] Krishnan, P. and Jawahar, C. (2016). Matching handwritten document images. In *European Conference on Computer Vision*, pages 766–782. Springer.
- [Krishnan and Jawahar, 2019] Krishnan, P. and Jawahar, C. (2019). Hwnet v2: An efficient word image representation for handwritten documents. *International Journal on Document Analysis and Recognition (IJDAR)*, 22(4):387–405.
- [Krishnann et al., 2018] Krishnann, P., Dutta, K., and Jawahar, C. (2018). Word spotting and recognition using deep embedding. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 1–6. IEEE.

- [Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- [Kuang et al., 2020] Kuang, X., Zhu, J., Sui, X., Liu, Y., Liu, C., Chen, Q., and Gu, G. (2020). Thermal infrared colorization via conditional generative adversarial network. *Infrared Physics & Technology*, 107:103338.
- [Kundu et al., 2021] Kundu, S., Malakar, S., Geem, Z. W., Moon, Y. Y., Singh, P. K., and Sarkar, R. (2021). Hough transform-based angular features for learning-free handwritten keyword spotting. *Sensors*, 21(14):4648.
- [Lavrenko et al., 2004] Lavrenko, V., Rath, T. M., and Manmatha, R. (2004). Holistic word recognition for handwritten historical documents. In *IEEE*, pages 278–287.
- [LeCun et al., 1999] LeCun, Y., Haffner, P., Bottou, L., and Bengio, Y. (1999). Object recognition with gradient-based learning. In *Shape, Contour and Grouping in Computer Vision*, pages 319–, London, UK, UK. Springer-Verlag.
- [Ledig et al., 2017] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690.
- [Lee and Toutanova, 2018] Lee, J. D. M. C. K. and Toutanova, K. (2018). Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [Lelore and Bouchara, 2013] Lelore, T. and Bouchara, F. (2013). Fair: a fast algorithm for document image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):2039–2048.
- [Lewis, 1995] Lewis, J. P. (1995). Fast template matching. In *Vision interface*, pages 15–19.
- [Li et al., 2018] Li, C., Guo, J., and Guo, C. (2018). Emerging from water: Underwater image color correction based on weakly supervised color transfer. *IEEE Signal processing letters*, 25(3):323–327.
- [Li et al., 2022] Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., and Wei, F. (2022). Dit: Self-supervised pre-training for document image transformer. *arXiv preprint arXiv:2203.02378*.
- [Li et al., 2021a] Li, S., Sui, X., Luo, X., Xu, X., Liu, Y., and Goh, R. (2021a). Medical image segmentation using squeeze-and-expansion transformers. *arXiv preprint arXiv:2105.09511*.
- [Li et al., 2021b] Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., and Wu, F. (2021b). Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2907.

- [Liang et al., 2012] Liang, Y., Fairhurst, M. C., and Guest, R. M. (2012). A synthesised word approach to word retrieval in handwritten documents. *Pattern Recognition*, 45(12):4225–4236.
- [Liao et al., 2017] Liao, W., Ying Yang, M., Zhan, N., and Rosenhahn, B. (2017). Triplet-based deep similarity learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 385–393.
- [Lin et al., 2021] Lin, T., Wang, Y., Liu, X., and Qiu, X. (2021). A survey of transformers. *arXiv preprint arXiv:2106.04554*.
- [Lin et al., 2020] Lin, Y.-H., Chen, W.-C., and Chuang, Y.-Y. (2020). Bedsr-net: A deep shadow removal network from a single document image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12905–12914.
- [Lisin et al., 2005] Lisin, D. A., Mattar, M. A., Blaschko, M. B., Learned-Miller, E. G., and Benfield, M. C. (2005). Combining local and global image features for object class recognition. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)-Workshops*, pages 47–47. IEEE.
- [Liu et al., 2021] Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J., and He, Z. (2021). A survey of visual transformers. *arXiv preprint arXiv:2111.06091*.
- [Lladós et al., 2012] Lladós, J., Rusinol, M., Fornés, A., Fernández, D., and Dutta, A. (2012). On the influence of word representations for handwritten word spotting in historical documents. *International journal of pattern recognition and artificial intelligence*, 26(05):1263002.
- [Ma et al., 2021] Ma, G., Ahmed, N. K., Willke, T. L., and Yu, P. S. (2021). Deep graph similarity learning: A survey. *Data Mining and Knowledge Discovery*, 35(3):688–725.
- [Manmatha and Croft, 1997] Manmatha, R. and Croft, W. (1997). Word spotting: Indexing handwritten archives. *Intelligent Multimedia Information Retrieval Collection*, pages 43–64.
- [Manmatha et al., 1996] Manmatha, R., Han, C., and Riseman, E. M. (1996). Word spotting: A new approach to indexing handwriting. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 631–637. IEEE.
- [Mao et al., 2016] Mao, X., Shen, C., and Yang, Y.-B. (2016). Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Advances in neural information processing systems*, 29.
- [Marinai et al., 2006] Marinai, S., Faini, S., Marino, E., and Soda, G. (2006). Efficient word retrieval by means of som clustering and pca. In *International Workshop on Document Analysis Systems*, pages 336–347. Springer.

- [Marnissi et al., 2021] Marnissi, M. A., Fradi, H., Sahbani, A., and Essoukri Ben Amara, N. (2021). Thermal image enhancement using generative adversarial network for pedestrian detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6509–6516. IEEE.
- [Mhiri et al., 2019] 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.
- [MHIRI et al., 2022] MHIRI, M., Hamdan, M., and Cheriet, M. (2022). Handwriting word spotting in the space of difference between representations using vision transformers. *Available at SSRN 4113859*.
- [Miech et al., 2021] Miech, A., Alayrac, J.-B., Laptev, I., Sivic, J., and Zisserman, A. (2021). Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836.
- [Mikolov et al., 2010] Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Interspeech*, pages 1045–1048. Makuhari.
- [Milyaev et al., 2015] Milyaev, S., Barinova, O., Novikova, T., Kohli, P., and Lempitsky, V. (2015). Fast and accurate scene text understanding with image binarization and off-the-shelf ocr. *International Journal on Document Analysis and Recognition (IJ DAR)*, 18(2):169–182.
- [Moghaddam and Cheriet, 2009a] Moghaddam, R. F. and Cheriet, M. (2009a). Application of multi-level classifiers and clustering for automatic word spotting in historical document images. In *2009 10th International Conference on Document Analysis and Recognition*, pages 511–515. IEEE.
- [Moghaddam and Cheriet, 2009b] Moghaddam, R. F. and Cheriet, M. (2009b). A variational approach to degraded document enhancement. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1347–1361.
- [Mohammed et al., 2022] 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*.
- [Niblack, 1985] Niblack, W. (1985). *An introduction to digital image processing*. Strandberg Publishing Company.
- [Ntirogiannis et al., 2012] Ntirogiannis, K., Gatos, B., and Pratikakis, I. (2012). Performance evaluation methodology for historical document image binarization. *IEEE Transactions on Image Processing*, 22(2):595–609.
- [Oh Song et al., 2016] Oh Song, H., Xiang, Y., Jegelka, S., and Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4004–4012.

- [Otsu, 1979] Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66.
- [Palangi et al., 2016] Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., and Ward, R. (2016). Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(4):694–707.
- [Pan and Yang, 2009] Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.
- [Pantke et al., 2014] Pantke, W., Dennhardt, M., Fecker, D., Märgner, V., and Fingscheidt, T. (2014). An historical handwritten arabic dataset for segmentation-free word spotting-hadara80p. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 15–20. IEEE.
- [Pearson, 1901] Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.
- [Phansalkar et al., 2011] Phansalkar, N., More, S., Sabale, A., and Joshi, M. (2011). Adaptive local thresholding for detection of nuclei in diversity stained cytology images. In *2011 International conference on communications and signal processing*, pages 218–220. IEEE.
- [Poznanski and Wolf, 2016] Poznanski, A. and Wolf, L. (2016). Cnn-n-gram for handwriting word recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2305–2314.
- [Pramanik and Bag, 2021] Pramanik, R. and Bag, S. (2021). Handwritten bangla city name word recognition using cnn-based transfer learning and fcn. *Neural Computing and Applications*, 33(15):9329–9341.
- [Pratikakis et al., 2013] Pratikakis, I., Gatos, B., and Ntirogiannis, K. (2013). Icdar 2013 document image binarization contest (dibco 2013). In *2013 12th International Conference on Document Analysis and Recognition*, pages 1471–1476. IEEE.
- [Pratikakis et al., 2017] Pratikakis, I., Zagoris, K., Barlas, G., and Gatos, B. (2017). Icdar2017 competition on document image binarization (dibco 2017). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1395–1403. IEEE.
- [Pratikakis et al., 2018] Pratikakis, I., Zagoris, K., Kaddas, P., and Gatos, B. (2018). Icfhr 2018 competition on handwritten document image binarization (h-dibco 2018). *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 489–493.
- [Puigcerver et al., 2014] Puigcerver, J., Toselli, A. H., and Vidal, E. (2014). Word-graph-based handwriting keyword spotting of out-of-vocabulary queries. In *2014 22nd International Conference on Pattern Recognition*, pages 2035–2040. IEEE.

- [Qins and El-Yacoubi, 2022] Qins, H. and El-Yacoubi, M. A. (2022). End-to-end generative adversarial network for hand-vein recognition. In *Advances in Pattern Recognition and Artificial Intelligence*, pages 47–60. World Scientific.
- [Rabaev et al., 2020] 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.
- [Radford et al., 2018] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding with unsupervised learning. *Technical report, OpenAI*.
- [Ramachandran et al., 2019] Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., and Shlens, J. (2019). Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, 32.
- [Ranftl et al., 2021] Ranftl, R., Bochkovskiy, A., and Koltun, V. (2021). Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188.
- [Rath and Manmatha, 2003] Rath, T. M. and Manmatha, R. (2003). Word image matching using dynamic time warping. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–II. IEEE.
- [Rath and Manmatha, 2007] Rath, T. M. and Manmatha, R. (2007). Word spotting for historical documents. *International Journal of Document Analysis and Recognition (IJ DAR)*, 9(2):139–152.
- [Riba et al., 2015] Riba, P., Lladás, J., and Fornés, A. (2015). Handwritten word spotting by inexact matching of grapheme graphs. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 781–785. IEEE.
- [Riba et al., 2021] Riba, P., Molina, A., Gomez, L., Ramos-Terrades, O., and Lladós, J. (2021). Learning to rank words: optimizing ranking metrics for word spotting. In *International Conference on Document Analysis and Recognition*, pages 381–395. Springer.
- [Rodríguez-Serrano and Perronnin, 2009] Rodríguez-Serrano, J. A. and Perronnin, F. (2009). Handwritten word image retrieval with synthesized typed queries. In *2009 10th International Conference on Document Analysis and Recognition*, pages 351–355. IEEE.
- [Rodríguez-Serrano and Perronnin, 2012] Rodríguez-Serrano, J. A. and Perronnin, F. (2012). A model-based sequence similarity with application to handwritten word spotting. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2108–2120.
- [Rodriguez-Serrano et al., 2013] Rodriguez-Serrano, J. A., Perronnin, F., and Meylan, F. (2013). Label embedding for text recognition. In *BMVC*, pages 5–1. Citeseer.
- [Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

- [Rothfeder et al., 2003] Rothfeder, J. L., Feng, S., and Rath, T. M. (2003). Using corner feature correspondences to rank word images by similarity. In *2003 Conference on Computer Vision and Pattern Recognition Workshop*, volume 3, pages 30–30. IEEE.
- [Rouhou et al., 2022] Rouhou, A. C., Dhiaf, M., Kessentini, Y., and Salem, S. B. (2022). Transformer-based approach for joint handwriting and named entity recognition in historical document. *Pattern Recognition Letters*, 155:128–134.
- [Saabni and El-Sana, 2013] Saabni, R. M. and El-Sana, J. A. (2013). Keywords image retrieval in historical handwritten arabic documents. *Journal of Electronic Imaging*, 22(1):013016.
- [Sauvola and Pietikäinen, 2000] Sauvola, J. and Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern recognition*, 33(2):225–236.
- [Schreiber et al., 2017] Schreiber, S., Agne, S., Wolf, I., Dengel, A., and Ahmed, S. (2017). Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 1162–1167. IEEE.
- [Schroff et al., 2015a] Schroff, F., Kalenichenko, D., and Philbin, J. (2015a). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.
- [Schroff et al., 2015b] Schroff, F., Kalenichenko, D., and Philbin, J. (2015b). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.
- [Serdouk et al., 2019] Serdouk, Y., Eglin, V., Bres, S., and Pardoën, M. (2019). Keyword spotting using siamese triplet deep neural networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1157–1162. IEEE.
- [Shahab et al., 2006] Shahab, S., Al-Khatib, W. G., and Mahmoud, S. A. (2006). Computer aided indexing of historical manuscripts. In *Computer Graphics, Imaging and Visualisation, 2006 International Conference on*, pages 287–295. IEEE.
- [Sharma et al., 2015] Sharma, A. et al. (2015). Adapting off-the-shelf cnns for word spotting & recognition. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 986–990. IEEE.
- [Shekhar and Jawahar, 2012] Shekhar, R. and Jawahar, C. (2012). Word image retrieval using bag of visual words. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 297–301. IEEE.
- [Shi et al., 2016] Shi, H., Yang, Y., Zhu, X., Liao, S., Lei, Z., Zheng, W., and Li, S. Z. (2016). Embedding deep metric for person re-identification: A study against large variations. In *European conference on computer vision*, pages 732–748. Springer.
- [Souibgui and Kessentini, 2020] Souibgui, M. A. and Kessentini, Y. (2020). De-gan: A conditional generative adversarial network for document enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

- [Srihari and Ball, 2008] Srihari, S. N. and Ball, G. R. (2008). Language independent word spotting in scanned documents. In *International Conference on Asian Digital Libraries*, pages 134–143. Springer.
- [Stauffer et al., 2020] Stauffer, M., Fischer, A., and Riesen, K. (2020). Filters for graph-based keyword spotting in historical handwritten documents. *Pattern Recognition Letters*, 134:125–134.
- [Su et al., 2012] Su, B., Lu, S., and Tan, C. L. (2012). Robust document image binarization technique for degraded document images. *IEEE transactions on image processing*, 22(4):1408–1417.
- [Sudholt and Fink, 2016] Sudholt, S. and Fink, G. A. (2016). Phocnet: A deep convolutional neural network for word spotting in handwritten documents. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 277–282. IEEE.
- [Sudholt and Fink, 2017] Sudholt, S. and Fink, G. A. (2017). Evaluating word string embeddings and loss functions for cnn-based word spotting. In *2017 14th iapr international conference on document analysis and recognition (icdar)*, volume 1, pages 493–498. IEEE.
- [Sudholt and Fink, 2018] 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.
- [Tamrin et al., 2021] Tamrin, M. O., El-Amine Ech-Cherif, M., and Cheriet, M. (2021). A two-stage unsupervised deep learning framework for degradation removal in ancient documents. In *International Conference on Pattern Recognition*, pages 292–303. Springer.
- [Tang and Lin, 2018] Tang, R. and Lin, J. (2018). Deep residual learning for small-footprint keyword spotting. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5484–5488. IEEE.
- [Tao et al., 2016] Tao, R., Gavves, E., and Smeulders, A. W. (2016). Siamese instance search for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1420–1429.
- [Tay et al., 2020] Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. (2020). Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*.
- [Tensmeyer and Martinez, 2017] Tensmeyer, C. and Martinez, T. (2017). Document image binarization with fully convolutional neural networks. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 99–104. IEEE.
- [Terasawa et al., 2005] Terasawa, K., Nagasaki, T., and Kawashima, T. (2005). Eigenspace method for text retrieval in historical document images. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 437–441. IEEE.
- [Terasawa and Tanaka, 2009] Terasawa, K. and Tanaka, Y. (2009). Slit style hog feature for document image word spotting. In *2009 10th international conference on document analysis and recognition*, pages 116–120. IEEE.

- [Touvron et al., 2021] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR.
- [Tushar et al., 2018] Tushar, A. K., Ashiquzzaman, A., Afrin, A., and Islam, M. R. (2018). A novel transfer learning approach upon hindi, arabic, and bangla numerals using convolutional neural networks. In *Computational Vision and Bio Inspired Computing*, pages 972–981. Springer.
- [Ustinova and Lempitsky, 2016] Ustinova, E. and Lempitsky, V. (2016). Learning deep embeddings with histogram loss. *Advances in Neural Information Processing Systems*, 29.
- [Vapnik, 1999] Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999.
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- [Vo et al., 2018] Vo, Q. N., Kim, S. H., Yang, H. J., and Lee, G. (2018). Binarization of degraded document images based on hierarchical deep supervised network. *Pattern Recognition*, 74:568–586.
- [Wahl et al., 1982] Wahl, F. M., Wong, K. Y., and Casey, R. G. (1982). Block segmentation and text extraction in mixed text/image documents. *Computer graphics and image processing*, 20(4):375–390.
- [Wang et al., 2017] Wang, C., Zhang, X., and Lan, X. (2017). How to train triplet networks with 100k identities? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1907–1915.
- [Wang et al., 2021] Wang, N., Zhou, W., Wang, J., and Li, H. (2021). Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1571–1580.
- [Wang et al., 2014] Wang, P., Eglin, V., Garcia, C., Largeron, C., Lladós, J., and Fornés, A. (2014). A coarse-to-fine word spotting approach for historical handwritten documents based on graph embedding and graph edit distance. In *2014 22nd International Conference on Pattern Recognition*, pages 3074–3079. IEEE.
- [Wang et al., 2018a] Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018a). High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807.
- [Wang et al., 2018b] Wang, X., Girshick, R., Gupta, A., and He, K. (2018b). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803.

- [Wang et al., 2022] Wang, Y., Liu, P., Lang, Y., Zhou, Q., and Shan, X. (2022). Learnable dynamic margin in deep metric learning. *Pattern Recognition*, 132:108961.
- [Weiss et al., 2016] Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1):1–40.
- [Westphal et al., 2020] 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.
- [Westphal et al., 2018] Westphal, F., Lavesson, N., and Grahn, H. (2018). Document image binarization using recurrent neural networks. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 263–268. IEEE.
- [Wicht et al., 2016] Wicht, B., Fischer, A., and Hennebert, J. (2016). Deep learning features for handwritten keyword spotting. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3434–3439. IEEE.
- [Wilkinson and Brun, 2016] Wilkinson, T. and Brun, A. (2016). Semantic and verbatim word spotting using deep neural networks. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 307–312. IEEE.
- [Wolf and Fink, 2020] Wolf, F. and Fink, G. A. (2020). Annotation-free learning of deep representations for word spotting using synthetic data and self labeling. In *International Workshop on Document Analysis Systems*, pages 293–308. Springer.
- [Xia et al., 2014] Xia, Y., Yang, Z.-B., and Wang, K.-Q. (2014). Chinese calligraphy word spotting using elastic hog features and derivatives dynamic time warping. *Journal Of Harbin Institute of Technology*, pages 21–27.
- [Xiong et al., 2018] Xiong, W., Jia, X., Xu, J., Xiong, Z., Liu, M., and Wang, J. (2018). Historical document image binarization using background estimation and energy minimization. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3716–3721. IEEE.
- [Xu and Gowen, 2019] Xu, J.-L. and Gowen, A. A. (2019). Spatial-spectral analysis method using texture features combined with pca for information extraction in hyperspectral images. *Journal of Chemometrics*, page e3132.
- [Xu et al., 2020] Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., et al. (2020). Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.
- [Yalniz and Manmatha, 2012] Yalniz, I. Z. and Manmatha, R. (2012). An efficient framework for searching text in noisy document images. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 48–52. IEEE.
- [Yi et al., 2017] Yi, Z., Zhang, H., Tan, P., and Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857.

- [Zamora-Martínez et al., 2007] Zamora-Martínez, F., España-Boquera, S., and Castro-Bleda, M. (2007). Behaviour-based clustering of neural networks applied to document enhancement. In *International Work-Conference on Artificial Neural Networks*, pages 144–151. Springer.
- [Zhao et al., 2019] Zhao, J., Shi, C., Jia, F., Wang, Y., and Xiao, B. (2019). Document image binarization with cascaded generators of conditional generative adversarial networks. *Pattern Recognition*, 96:106968.
- [Zhou et al., 2016] Zhou, G., Xie, Z., He, T., Zhao, J., and Hu, X. T. (2016). Learning the multilingual translation representations for question retrieval in community question answering via non-negative matrix factorization. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(7):1305–1314.
- [Zhu et al., 2017] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.
- [Zirari et al., 2013] Zirari, F., Ennaji, A., Nicolas, S., and Mammass, D. (2013). A methodology to spot words in historical arabic documents. In *Computer Systems and Applications (AICCSA), 2013 ACS International Conference on*, pages 1–4. IEEE.