Post-Retrieval Semantic Re-Ranking via Zero-shot LLMs for Segmentation-Free Document Image Keyword Spotting

St. Papazis

Master Thesis

— ♦ —

Ioannina, July 2025

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF IOANNINA

Post-Retrieval Semantic Re-Ranking via Zero-shot LLMs for Segmentation-Free Document Image Keyword Spotting

A Thesis

submitted to the designated
by the Assembly
of the Department of Computer Science and Engineering
Examination Committee

by

Stergios Papazis

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN DATA AND COMPUTER SYSTEMS ENGINEERING

WITH SPECIALIZATION
IN DATA SCIENCE AND ENGINEERING

University of Ioannina
School of Engineering
Ioannina 2025

Examining Committee:

- **Christophoros Nikou**, Professor, Department of Computer Science and Engineering, University of Ioannina (Advisor)
- **Aristeidis Lykas**, Professor, Department of Computer Science and Engineering, University of Ioannina
- **Konstantinos Blekas**, Professor, Department of Computer Science and Engineering, University of Ioannina

Dedication

To my family —

for their enduring love, support, and
encouragement throughout the years.

Acknowledgements

Initially, I would like to express my sincere gratitude to my advisor, Dr. Christophoros Nikou, Professor in the Department of Computer Science at the University of Ioannina, for his continuous support, valuable guidance, as well as for entrusting me with this research project. His inspiring lectures sparked my interest in Computer Vision, and I am truly grateful for the opportunity to have conducted this work under his supervision.

I am deeply indebted to Dr. Aggelos P. Giotis, Post-doc Researcher in the Department of Computer Science and Engineering at the University of Ioannina, for his exceptional mentorship throughout this project. His profound expertise in Keyword Spotting, along with his constructive feedback and the time he so generously dedicated to our discussions, was instrumental in shaping both the theoretical and technical aspects of this work.

My heartfelt appreciation goes to my family for their unconditional love and support throughout this academic journey: to my father, Giorgos, whose lifelong pursuit of knowledge has been a constant inspiration; to my mother, Stella, for her unwavering encouragement and care; and to my brother, Alexandros-Vaios, for his steady support and understanding.

Finally, and most importantly, I wish to thank my girlfriend, Sofia. Her enduring love, wholehearted encouragement, comforting presence, and patient understanding have been a beacon of hope, strength and motivation through my darkest hours over the past four years. A special mention goes to our beloved cats, Pixie and Layla, whose mindless antics never fail to brighten our days.

Table of Contents

List of Figures

List of Tables

Glossary

AP	Average Precision. iii, 13–15
BoF	Bag-of-Features. 15
CBIR	Content-Based Image Retrieval. vii, 3
CC	Connected Component. 15
CER	Character Error Rate. 34, 41
CNN	Convolutional Neural Network. 7, 10, 16, 19, 20, 23
CTC	Connectionist Temporal Classification. 7, 25–27, 35
DCT	Discrete Cosine Transformation. 11
DCToW	Discrete Cosine Transform of Words. iii, 10–12, 16, 19, 24, 36
DIoU	Distance-IoU. 24
FPN	Feature Pyramid Network. 16, 24
GW	George Washington. iii, iv, vii, viii, x, 7, 19, 29–37, 39, 41–48, 50, 51
HKS	Heat Kernel Signature. 15
HMM	Hidden Markov Model. 9, 15
HoG	Histogram of Oriented Gradients. 9, 15
HTR	Handwritten Text Recognition. 3, 34
IAM	IAM Handwriting Database. iii, iv, vii, viii, x, 7, 29–32, 34, 36, 38, 40–43, 48, 50, 51
IoU	Intersection over Union. v, vi, 14, 24, 49
IR	Information Retrieval. 13
KWS	Keyword Spotting. iii, vii–x, 3–10, 12–17, 19, 21–27, 29–31, 36, 41, 46, 48, 50, 51

LLM	Large Language Model. iii, 7, 22, 23, 27, 28
mAP	Mean Average Precision. iv, viii, x, 6, 7, 13, 32, 33, 36–40, 44, 45, 48, 51
mAP@25	Mean Average Precision at 25% IoU. iii, 14, 32, 33, 35–45, 47, 48, 51
mAP@50	Mean Average Precision at 50% IoU. iii, 14, 32–45, 47, 48
ML	Machine Learning. 4, 12
NEL	Named Entity Linking. 19
NER	Named Entity Recognition. 19
NLP	Natural Language Processing. 3, 5, 7, 18–20, 48, 50
NMS	Non-Maximum Suppression. 25, 26, 48
OCR	Optical Character Recognition. vi, 3, 7, 28, 34
OOV	Out-of-Vocabulary. 21, 28
PHOC	Pyramidal Histogram of Characters. iii, 10–12, 16, 19, 24, 25
PRF	Pseudo-Relevance Feedback. 20
QbE	Query-by-Example. 4, 10
QbS	Query-by-String. 4, 10, 12, 16, 23, 25, 27, 31
RPN	Region Proposal Network. 16
SD	Standard Deviation. 35, 37, 39, 44
SIFT	Scale Invariant Feature Transform. 9, 15
SPP	Spatial Pyramid Pooling. 10
SRF	Supervised Relevance Feedback. 20
TPP	Temporal Pyramid Pooling. 10
TrOCR	Transformer-based OCR. 7, 28, 34, 36–41, 43, 44, 51
ViT	Vision Transformer. 28
VQA	Visual Question Answering. 19
WS	Word Spotting. 3, 6, 8, 12, 14, 19, 51

Abstract

Stergios Papazis, M.Sc. in Data and Computer Systems Engineering, Department of Computer Science and Engineering, School of Engineering, University of Ioannina, Greece, 2025.
Post-Retrieval Semantic Re-Ranking via Zero-shot LLMs for Segmentation-Free Document Image Keyword Spotting.
Advisor: Christophoros Nikou, Professor.

The digitization of historical handwritten documents plays a crucial role in their preservation. With preservation largely addressed, the focus has shifted toward enhancing accessibility. This has led to the development of Keyword Spotting (KWS), a Content-Based Image Retrieval (CBIR) task that retrieves and ranks word images based on their similarity to a given query, without requiring prior transcription.

Traditional KWS methods typically treat words as visual patterns, relying solely on appearance-based features and often neglecting their underlying semantic content. Even when semantics are considered, it is often within the segmentation-based setting, which assumes prior word-level segmentation, a non-trivial and error-prone requirement, particularly for historical manuscripts.

To address these limitations, we propose a novel unsupervised mechanism for semantic relevance feedback to re-rank the initial output of segmentation-free KWS systems. Our approach operates in three stages: (1) decoding the retrieved word images into text using a neural decoder; (2) projecting the transcriptions into a semantic space using pre-trained transformer-based language models such as RoBERTa, MPNet, and MiniLM, where semantic similarity is measured by cosine distance; (3) re-ranking the retrieved items by combining visual and semantic similarity.

We evaluate our method on the widely used historical George Washington (GW) dataset and the modern IAM Handwriting Database (IAM), using the retrieval ranked lists of two cutting-edge segmentation-free KWS baseline models. We further assess

the performance across two decoder architectures and two naive fusion strategies through an extensive ablative analysis.

Numerical results show consistent improvements in Mean Average Precision (mAP) across all tested configurations, with gains of up to $+2.3\%$ (from 94.31% to 96.59%) on GW and $+3\%$ (from 79.15% to 82.12%) on IAM. Notably, even in scenarios with minimal mAP improvement, we observe significant qualitative gains: semantically relevant but inexact matches are retrieved more frequently. This behavior, known as semantic KWS, is particularly beneficial in real-world scenarios wherein users may not know beforehand the precise query needed to locate relevant content.

These findings demonstrate the effectiveness of incorporating semantic feedback from large language models into visual keyword spotting pipelines. By complementing appearance-based retrieval with NLP-driven semantic re-ranking, our approach enables more flexible and meaningful document search, even in challenging segmentation-free settings. Moreover, it highlights the potential of hybrid vision-language models to advance document image analysis, especially for noisy, heterogeneous, or low-resource historical archives.

Keywords: computer vision; deep learning; keyword spotting; document analysis; segmentation-free retrieval; vision-language models; re-ranking; relevance feedback; semantic embeddings; large language models; NLP-based retrieval; zero-shot learning

Εκτεταμενη Περιληψη

Στέργιος Παπάζης, Δ.Μ.Σ. στη Μηχανική Δεδομένων και Υπολογιστικών Συστημάτων, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πολυτεχνική Σχολή, Πανεπιστήμιο Ιωαννίνων, 2025.
Σημασιολογική Αναδιάταξη Αποτελεσμάτων Εντοπισμού Λέξεων χωρίς Κατάτμηση σε Εικόνες Χειρογράφων με Μεγάλα Γλωσσικά Μοντέλα.
Επιβλέπων: Χριστόφορος Νίκου, Καθηγητής.

Η ραγδαία τεχνολογική πρόοδος του τελευταίου αιώνα έχει δημιουργήσει τις συνθήκες ώστε η ψηφιοποίηση ιστορικών κειμένων να είναι οικονομικά και πρακτικά δυνατή. Καθώς η ύπαρξη αυτών των άυλων ψηφιακών αντιγράφων διασφαλίζει τη διάσωση και τη μακροχρόνια διατήρηση του περιεχομένου των κειμένων, η προσοχή πλέον στρέφεται στην ανάπτυξη τεχνικών που διευκολύνουν την προσβασιμότητα σε αυτά τα ψηφιακά έγγραφα.

Η μέθοδος του Εντοπισμού Λέξεων (Keyword Spotting) έχει αναπτυχθεί για την ανάκτηση πληροφορίας βάσει περιεχομένου σε ψηφιακές συλλογές εικόνων κειμένου. Ένα τέτοιο σύστημα εντοπίζει περιοχές εικόνων εγγράφων όπου εμφανίζεται μια δοθείσα λέξη-κλειδί και τις επιστρέφει ταξινομημένες ως προς την οπτική ομοιότητα που παρουσιάζουν με την είσοδο. Κύριο γνώρισμα της μεθόδου είναι ότι δεν απαιτείται η εκ των προτέρων αναγνώριση του κειμένου, δηλαδή η ανάκτηση βασίζεται αποκλειστικά σε οπτικά χαρακτηριστικά, αποφεύγοντας έτσι τα σφάλματα που συχνά προκύπτουν στην εξαγωγή μεταγραφών από ιστορικά κείμενα.

Οι παραδοσιακές προσεγγίσεις εντοπισμού λέξεων τείνουν να αντιμετωπίζουν τις λέξεις αποκλειστικά ως οπτικά πρότυπα, αγνοώντας το υποκείμενο εννοιολογικό περιεχόμενο, πέρα από ορισμένα επιφανειακά χαρακτηριστικά σε επίπεδο χαρακτήρων. Σε ορισμένες περιπτώσεις, η εννοιολογική διάσταση λαμβάνεται υπόψη, κυρίως στο πλαίσιο κάποιων μεθόδων εντοπισμού λέξεων με κατάτμηση. Ωστόσο, η

διαδικασία κατάτμησης σε επίπεδο λέξεων είναι ιδιαίτερα επιρρεπής σε σφάλματα όταν εφαρμόζεται σε εικόνες ιστορικών εγγράφων.

Για την αντιμετώπιση αυτών των προκλήσεων, προτείνουμε έναν νέο μηχανισμό αναδιάταξης των αποτελεσμάτων εντοπισμού λέξεων, βασισμένο στην ανατροφοδότηση εννοιολογικής σχετικότητας, ο οποίος δεν απαιτεί κατάτμηση των εικόνων σε επίπεδο γραμμής κειμένου ή λέξης. Το προτεινόμενο σύστημα αξιοποιεί σημασιολογική γνώση για την αναδιάταξη της εξόδου ενός συστήματος εντοπισμού λέξεων, δηλαδή της διατεταγμένης λίστας αποτελεσμάτων ανάκτησης που μοιάζουν οπτικά με την προς αναζήτηση λέξη-κλειδί. Η προσέγγισή μας λειτουργεί σε τρία στάδια: (1) αποκωδικοποίηση των ανακτημένων εικόνων λέξεων σε κείμενο μέσω ενός νευρωνικού αποκωδικοποιητή, (2) προβολή της μεταγραφής των αναγνωρισμένων λέξεων σε έναν σημασιολογικό χώρο μέσω προ-εκπαιδευμένων μεγάλων γλωσσικών μοντέλων (όπως τα RoBERTa, MPNet και MiniLM), όπου η σημασιολογική συνάφεια εκφράζεται ως χωρική εγγύτητα, και (3) αναδιάταξη των αποτελεσμάτων μέσω ενός συνδυασμού της οπτικής και εννοιολογικής ομοιότητας.

Για την αξιολόγηση της προτεινόμενης μεθόδου, πραγματοποιήθηκε εκτενής πειραματική μελέτη σε δύο ευρέως χρησιμοποιούμενα σύνολα χειρογράφων της περιοχής: τα George Washington (GW) και IAM Handwriting Database (IAM). Χρησιμοποιήθηκαν δύο σύγχρονα συστήματα εντοπισμού λέξεων χωρίς κατάτμησης, τα οποία παρείχαν τις αρχικές διατεταγμένες λίστες αποτελεσμάτων προς περαιτέρω αναδιάταξη από το σύστημά μας. Επιπλέον, αξιολογήθηκε η απόδοση της μεθόδου σε δύο διαφορετικές αρχιτεκτονικές αποκωδικοποιητών, καθώς και υπό δύο εναλλακτικές στρατηγικές συνδυασμού της οπτικής και της σημασιολογικής ομοιότητας.

Τα αριθμητικά αποτελέσματα παρουσιάζουν σταθερές βελτιώσεις στη μετρική Mean Average Precision (mAP) για όλες τις δοκιμασθείσες πειραματικές διατάξεις, με αυξήσεις έως και $+2.3\%$ (από 94.31% σε 96.59%) στο σύνολο GW και $+3\%$ (από 79.15% σε 82.12%) στη συλλογή IAM. Αξιοσημείωτο είναι ότι, ακόμα και σε περιπτώσεις όπου η αύξηση του mAP είναι περιορισμένη, παρατηρούνται σημαντικά ποιοτικά οφέλη: εννοιολογικά συναφείς αλλά όχι ταυτόσημες λέξεις, ανακτώνται συχνότερα. Η συμπεριφορά αυτή, γνωστή στη βιβλιογραφία ως σημασιολογικός εντοπισμός λέξεων (Semantic Keyword Spotting), είναι ιδιαίτερα χρήσιμη σε πρακτικά σενάρια, όπου οι χρήστες ενδέχεται να μην γνωρίζουν εκ των προτέρων το ακριβές ερώτημα που απαιτείται για την ανάκτηση σχετικού περιεχομένου.

Τέλος, τα ευρήματά μας αναδεικνύουν την αξία της ενσωμάτωσης σημασιολογι-

κής πληροφορίας στα συστήματα εντοπισμού λέξεων, τόσο για την επίτευξη καλύτερης ακρίβειας στην ανάκτηση αποτελεσμάτων, όσο και για μία πιο ευέλικτη και ουσιαστικής αναζήτηση περιεχομένου σε αντίξοες συνθήκες χειρογράφων που δεν έχουν υποστεί προ-επεξεργασία ή κατάτμηση. Παράλληλα, διαφαίνεται η δυναμική υβριδικών μοντέλων που συνδυάζουν οπτικά και γλωσσικά στοιχεία για την αντιμετώπιση προκλήσεων σε συλλογές ιστορικών εγγράφων με χαρακτηριστικά όπως ο θόρυβος, η ετερογένεια και η περιορισμένη διαθεσιμότητα δεδομένων.

Λέξεις κλειδιά: υπολογιστική όραση, μηχανική μάθηση, βαθιά νευρωνικά δίκτυα, εντοπισμός λέξεων εικόνων χειρόγραφων κειμένων χωρίς κατάτμηση, σημασιολογική αναδιάταξη, ανατροφοδότηση σχετικότητας, μεγάλα γλωσσικά μοντέλα

Chapter 1

Thesis Introduction

1.1 Preserving Historical Documents

Writing is one of humanity's most impactful inventions. Emerging from the need to manage practical concerns, such as recording inventories, transactions, and administrative activities like tax collection and census taking, it gradually became integral to every facet of human life. From governance and law to commerce, as well as from science, education, and culture to philosophy and religion, writing has enabled humans to overcome the ephemeral nature of speech and the limitations of memory, codifying both individual and collective knowledge — the foundation upon which knowledge continually expands. Furthermore, it enables the externalization of thought and the recording of ideas, facilitating abstract thinking, refinement, reflection, and revision. It has supported the functioning of complex administrative and governmental systems; underpinned the development of legal codes, treaties, deeds of ownership, and contracts; enabled the recording of historical and environmental events, cultural traditions, folklore, and oral histories; made creative and literary expression possible;

fostered education, and scientific inquiry through the documentation and communication of technical knowledge; and shaped religion through the recording and preservation of sacred texts. Moreover, it has also allowed the ideas and influence of individuals to resonate far beyond their own time and place, constituting one of the earliest forms of communication (typically one-way, although occasionally reciprocal, for instance in written correspondence) in which participants are not bound geographically or temporally.

Recognizing the transformative power of writing, civilizations across the globe developed a wide range of scripts and writing systems, or adapted those of neighboring cultures, to represent their spoken languages. Over the millennia, an immense body of written works has been produced; yet the overwhelming majority of these manuscripts have perished over the centuries. It is telling that, in many cases, the complete works of great ancient writers are known to us only through references, quotations, or critiques by later authors, with the originals themselves long lost.

In the past, preserving documents often required manual duplication and periodic restoration to counteract both natural and accidental forms of age-related degradation, such as fading, tearing, cracking, or damage caused by mold, rot, insects, humidity, or fire. In many cases, restoration efforts were never undertaken because the documents were deemed too insignificant to justify the cost and labor involved. As a result, the number of copies that could be produced and preserved was severely limited. In other cases, documents were lost unexpectedly due to war or natural disasters. Even with the advent of the printing press, which revolutionized the production and distribution of written material, the preservation of documents remained a challenge. For instance, out-of-print texts grew scarcer, and reissuing them was often impractical — especially when the original printing templates were unavailable.

In the 20th century, progress in Computer Science and related fields led to the development of key technologies, including high-resolution scanning, affordable digital storage systems, file compression algorithms, and the Internet, which transformed the way manuscripts are preserved and accessed. The focus has since shifted from merely safeguarding documents to enabling scholars, students, and other interested individuals to engage meaningfully with the contents of these digital libraries. Examples of such efforts in document image analysis include transcribing the manuscripts; modernizing archaic or irregular language (such as outdated scripts, vocabulary, grammar, and spelling); indexing collections to make them searchable; automatic translation;

automatic summarization; and other processes that improve the access to and usability of these digital archives.

Although such work is typically performed by experts, it tends to be costly and time-consuming. Automation is therefore a highly desirable alternative, as it can significantly reduce the resources needed to handle the vast volumes of digitized materials maintained by libraries and cultural institutions.

1.2 Introduction to Keyword Spotting

Keyword Spotting (KWS), also referred to as Word Spotting (WS), is a Content-Based Image Retrieval (CBIR) technique that identifies instances of a specific keyword within a collection of document images [1]. Given a query and a set of document images, the system locates and retrieves all image regions across the collection whose content visually resembles the query. The objective is to enable direct keyword-based search and efficient navigation of digitized manuscript collections.

A naive approach to designing such a CBIR system (or any other document image analysis system) involves two main stages. The first stage employs Handwritten Text Recognition (HTR), also known as *transcription*, in which the location of each word within each document image is identified, and corresponding character sequences are extracted using, for example, Optical Character Recognition (OCR) techniques. The second stage utilizes domain-appropriate Natural Language Processing (NLP) methods to process the textual output.

While HTR approaches can theoretically preserve textual information with high fidelity, practical implementations often struggle with reduced transcription accuracy. This issue is particularly pronounced for handwritten historical documents and degraded machine-printed materials, due to factors such as handwriting variability or physical deterioration (including ink fading, staining, and paper degradation). Most importantly, this process effectively severs the connection between the text and its visual origins, treating it merely as a generic textual resource (e.g., text from a web page).

Sometimes, the nature of the downstream NLP task indicates that exact transcription is not strictly necessary, and enforcing it becomes an avoidable constrain. In such cases, it is often more effective to operate in a *recognition-free* manner, working

directly with image features at the character, word, line, or even document level. These insights have motivated the development of KWS approaches — among the first in document image analysis to adopt this logic — which, in turn, inspired similar methods across other tasks within the field.

There are various ways to categorize KWS systems, each reflecting different underlying assumptions and design choices made by researchers. A predominant classification distinguishes between *segmentation-based* and *segmentation-free* approaches. Segmentation-based methods [2–9] assume that a preliminary segmentation step is performed prior to retrieval, and as such, they operate explicitly on pre-segmented word or line images. On the contrary, segmentation-free methods [10–18] process the entire document image directly, without relying on any prior segmentation.

Another common classification is based on the nature of the query input. *Query-by-Example* (QbE) systems [2, 4, 5, 19] accept a sample word image as input, requiring the user to manually identify at least one instance of the target word within the document collection. Conversely, *Query-by-String* (QbS) systems [15–18] accept textual queries This behavior is often preferable, making QbS systems more user-friendly. Nevertheless, QbE approaches can be adapted to assist manual transcription workflows: once a single word instance is transcribed, all other occurrences are automatically identified and transcribed, thereby reducing the overall transcription effort.

Furthermore, KWS systems can be categorized based on whether they employ Machine Learning (ML) techniques to learn features or rely on manually-engineered features, distinguishing between *learning-based* [5, 12, 15–18, 20] and *learning-free* [6, 8, 19] approaches. In recent years, learning-based methods have increasingly dominated the field due to their superior adaptability and performance, albeit at the cost of requiring extensive annotated training data.

Finally, a more recent distinction among KWS systems classifies approaches by their retrieval objective, dividing them into *Verbatim Keyword Spotting* and *Semantic Keyword Spotting*, two terms coined by Wilkinson et al. [21]. *Verbatim* KWS refers to the conventional approach discussed previously, where retrieval is primarily guided by visual similarity between the query and document content. On the other hand, Semantic KWS [21–24] retrieves word instances that are conceptually related to the query, even when they bear little or no visual resemblance to it. Notably, these two approaches are not mutually exclusive: a KWS system may employ a two-stage retrieval process, first identifying verbatim matches (exact or near-exact) before supplementing

them with semantically relevant candidates.

The material and terminology presented in this section are primarily adapted from the seminal survey by Giotis et al. [1]. For a more comprehensive treatment of the topic, we refer the interested reader to the original work.

1.3 Motivation and Objectives

Notwithstanding substantial advances in Keyword Spotting, current systems appear to be approaching a fundamental performance ceiling. Conventional KWS methods rely predominantly on visual characteristics, matching queries to image regions or word images based solely on character-level features. While computationally efficient, its representational capacity remains limited to shallow semantic relationships, such as shared character subsequences. Nevertheless, since false positives (i.e., visually similar inexact retrievals ranked above true matches) typically exhibit little semantic relevance to the query, the integration of a deeper semantic understanding within the KWS pipeline could aid in the detection and elimination of these errors, thereby improving retrieval accuracy.

Modern NLP techniques, such as pre-trained word embeddings (e.g., Word2Vec [25], GloVe [26], FastText [27], BERT [28], RoBERTa [29]), have revolutionized the modeling of semantics through their ability to capture deep linguistic patterns. Recent advances in document image analysis combining visual and textual features [23, 24, 30–34] demonstrate a viable alternative to conventional vision-only methodologies. Yet their application to KWS remains limited, since even when adopted they are typically confined to the segmentation-based scenario.

Segmentation-based approaches face inherent limitations, particularly when applied to historical manuscripts, where reliable segmentation remains an open challenge [1]. Issues such as inconsistent spacing, overlapping strokes, decorative lettering, and local skew significantly hinder accurate word segmentation. Dey et al. [35] demonstrated that the effectiveness of conventional segmentation-based KWS methods is highly sensitive to the quality of the preceding word segmentation: suboptimal segmentation leads to substantial degradation in retrieval accuracy.

Taken together, the need to address these limitations *motivates* the present work. To this end, we propose a novel *post-processing* mechanism that leverages *semantic*

relevance feedback to *re-rank* the output of a *segmentation-free* KWS system. This mechanism exploits modern pre-trained transformer based language models to map the retrieved instances into a latent word space, where semantic relevance is assessed. The resulting semantic similarity is then combined with the initial visual similarity to produce a hybrid similarity score, which is used to re-rank the original retrieval list. Furthermore, our design pursues three primary objectives: (1) improving Mean Average Precision (mAP) via false positive suppression and true positive boosting; (2) enhancing retrieval quality by elevating semantically relevant inexact matches without compromising exact-match recall — a core aim of semantic KWS; and (3) ensuring a flexible, modular, lightweight framework, which bridges the gap between vision and language without requiring end-to-end retraining.

1.4 Thesis Outline and Contributions

We conclude this introduction with a brief overview of the thesis structure and a summary of the key contributions.

The remainder of this thesis is structured as follows:

- Chapter 2 expands upon the introduction to Keyword Spotting in Section 1.2, establishing the theoretical framework and surveying relevant methodologies. It examines the verbatim and semantic Word Spotting paradigms, along with re-ranking techniques and relevance feedback.
- Chapter 3 introduces the *main contribution* of this thesis: a semantic-aware re-ranking framework. It outlines the framework's architecture and design rationale, and describes the two baseline models (namely WordRetrievalNet [17] and KWS-Simplified [18]) whose outputs form the basis for the re-ranking.
- Chapter 4 presents a comprehensive experimental evaluation of the proposed approach, detailing the datasets, evaluation protocol, and implementation setup. It also analyzes component-level contributions and qualitative behaviors, concluding with a discussion of the main findings.
- Chapter 5 concludes the thesis by summarizing the main findings and outlining potential directions for future research.

In addition to the structure outlined above, the main contributions of this thesis are summarized below:

- We propose a novel framework for semantic-aware re-ranking in handwritten document image Keyword Spotting, integrating cutting-edge NLP techniques (e.g, RoBERTa [36], MPNet [36, 37], MiniLM [36, 38]) into the retrieval pipeline. To the best of our knowledge, this is the first work to: (1) leverage Large Language Models (LLMs) for semantic re-ranking in handwritten document image KWS; and (2) address semantic modeling in the segmentation-free formulation.

- We explore two distinct neural decoding strategies for converting retrieved word snippets into text: (1) Transformer-based OCR (TrOCR) [39], a transformer-based vision-to-text model, as well as (2) a compact CNN-based decoder adapted from KWS-Simplified [18], that combines character-counting heuristics with Connectionist Temporal Classification (CTC) [40] re-scoring.

- We conduct extensive experiments on the George Washington (GW) dataset [41] and the IAM Handwriting Database (IAM) [42] using two state-of-the-art baseline systems. Our ablation studies assess the impact of key pipeline components on Mean Average Precision (mAP), including the transcription decoder, semantic embedding method, and fusion strategy.

- Our results show that semantic re-ranking consistently improves retrieval performance with low variance across cross-validation iterations, Beyond quantitative gains, the method enables semantically aware, recognition-free retrieval effectively bridging the gap between verbatim and semantic KWS.

Chapter 2

An Overview of Keyword Spotting

2.1 Introduction

Building upon the basic concepts and taxonomy introduced in Section 1.2, this chapter examines the core technical foundations of KWS systems. It is organized as follows. Section 2.2 introduces the fundamentals of KWS. Section 2.3 reviews related work across traditional and modern approaches. Section 2.4 explores recent developments in semantic Word Spotting, while Section 2.5 discusses methods based on relevance feedback and re-ranking.

2.2 Keyword Spotting Fundamentals

2.2.1 A Typical KWS System

A typical KWS system operates in two main stages: *offline processing* and *online querying*, as illustrated in Figure 2.1.

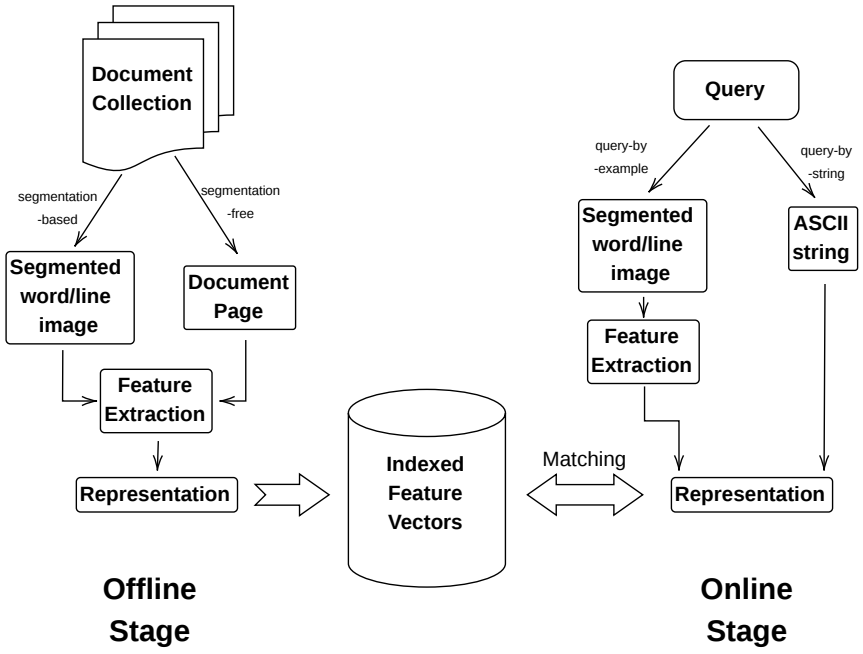


Figure 2.1: The architecture of a general KWS system

During the *offline processing* stage, the system extracts feature representations from document images (at word, line, or page level). The feature vectors are indexed to enable efficient retrieval during the *online* stage, when they are matched against query-derived representations. Over time, the feature extraction pipeline has evolved significantly.

Early systems employed extensive preprocessing (smoothing, filtering, Otsu's binarization [43]) and normalization prior to extracting handcrafted features. These approaches utilized manually engineered feature representations — including, but not limited to: local descriptors (e.g., stroke contours, pixel transitions, region-based attributes); gradient-based features such as Histogram of Oriented Gradients (HoG) and Scale Invariant Feature Transform (SIFT); geometric and statistical measures (including black pixel distributions, contour inclinations); structural features (e.g., graphemes, contours, or skeletons); and semantic attribute features (i.e., appearance-independent properties that bridge visual and textual domains). Finally, these features were often modeled with statistical frameworks such as Hidden Markov Models (HMMs).

With the recent advent of deep learning-based KWS methods, many of these steps

have been simplified or replaced. A common practice is to normalize input images to a fixed size before feeding them directly into a Convolutional Neural Network (CNN). This reflects a broader trend in pattern recognition: the transition from handcrafted pipelines to end-to-end learned representations, marking a paradigm shift from explicit *feature engineering* to implicit *feature learning* [5, 17, 18, 20, 21].

During the *online querying* stage, users may submit queries in either of the modalities: QbE, using word image samples, and QbS, using ASCII text inputs. The system projects the query into the offline feature representation space, computes pairwise similarities against all indexed entries, and returns the ranked retrieval results. These modalities differ fundamentally: QbE requires an actual instance of the query word to exist in the dataset limiting lexical coverage, while QbS enables arbitrary queries through character-level semantic models at the cost of requiring robust cross-modal alignment.

KWS systems must strike a balance between accuracy and efficiency — delivering reliable matching while remaining scalable to large document collections. A common strategy for improving retrieval speed is the use of fixed-length feature representations, which enable rapid comparisons using standard similarity measures such as the cosine distance, along with algorithms like the nearest neighbor search. These representations typically encode each word image as a compact, discriminative vector, supporting both efficiency and effective generalization. Some fixed-length vectors are obtained directly from image features, while others are derived through encoding or pooling mechanisms — such as Spatial Pyramid Pooling (SPP) and Temporal Pyramid Pooling (TPP) [5] — which aggregate variable-length representations and variable-size inputs.

Beyond purely visual approaches, a significant research direction in KWS leverages semantic attribute-based embeddings [1, 2]. These techniques encode both the visual characteristics of word images and their corresponding textual labels into a unified representation space. By capturing linguistically meaningful properties, they create discriminative features that are largely invariant to visual appearance. This dual capability facilitates not only efficient retrieval but also cross-modal matching between image and text representations. Additionally, such attribute-based methods demonstrate robustness to variations in writing style while maintaining computational efficiency for large-scale comparisons. Prominent examples of this approach include the Pyramidal Histogram of Characters (PHOC) and Discrete Cosine Transform of

Words (DCToW) representations.

The *Pyramidal Histogram of Characters (PHOC)* [2] represents words as fixed-dimensional binary vectors by hierarchically encoding character distribution patterns. Through multi-level spatial decomposition — partitioning words into progressively finer segments (typically halves, thirds, quarters, and fifths) — it records character and frequent bigram occurrences across these regions. This approach preserves approximate character positioning while maintaining robustness to visual appearance variations. The resulting binary embedding enables efficient storage and comparison, with spatial proximity in the subspace reflecting linguistic similarity (e.g., "letter" and "better" clustering nearby due to shared characters).

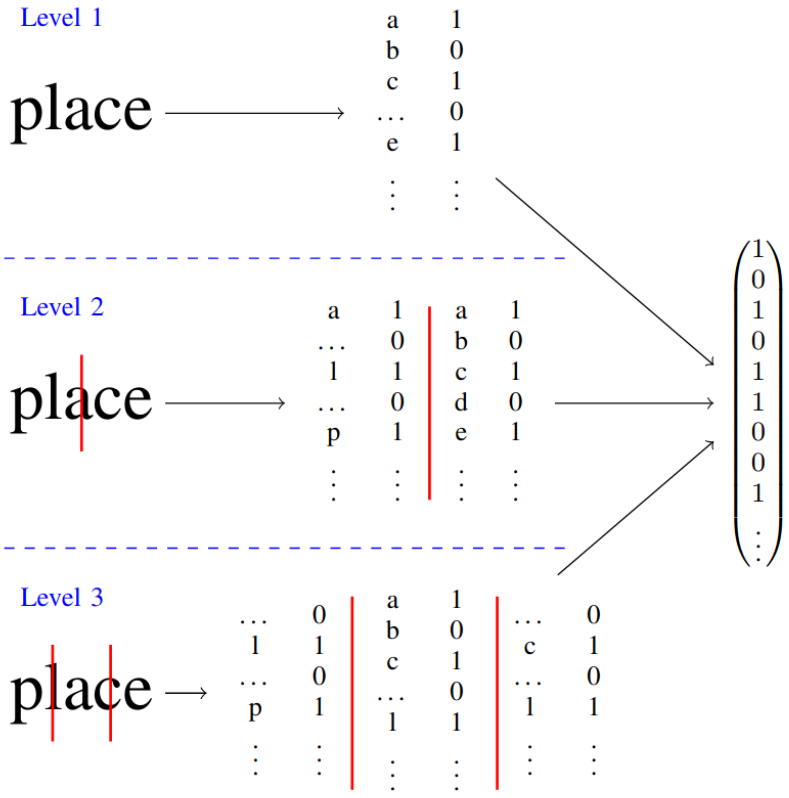


Figure 2.2: This figure demonstrates the generation of a three-level PHOC representation for the word "place". Figure reproduced from [44].

The *Discrete Cosine Transform of Words (DCToW)* [21] represents words as compact real-valued vectors through spectral analysis of character distributions. Unlike binary approaches like PHOC, DCToW first encodes words as $K \times m$ matrices (where K=36 represents the English alphabet size and m the word length) using one-hot character vectors. It then applies a DCT-II transform row-wise, retaining only the

first R=3 low-frequency coefficients per character dimension. This process yields a 108-dimensional embedding ($K \times R$) that implicitly captures both global character ordering and distribution patterns through its spectral signature.

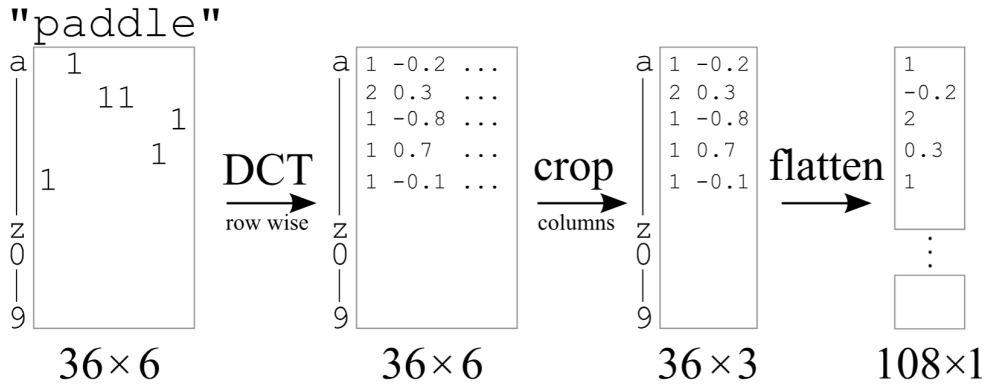


Figure 2.3: A visualization of the DCToW representation for the string “paddle”. Figure reproduced from [15].

Notably, the attribute-based design of both PHOC and DCToW facilitates strong generalization capabilities, enabling effective zero-shot Word Spotting even for queries absent from training data. By encoding fundamental linguistic properties — whether through character distributions or spectral signatures — these representations achieve robust retrieval without requiring prior exposure to specific word forms. Their fixed-dimensional embeddings preserve structural information while maintaining computational efficiency, making them particularly suitable for the cross-modal retrieval of QbS KWS, when integrated with ML techniques.

2.2.2 Applications of Word Spotting

KWS serves as a powerful tool across diverse domains, enabling efficient interaction with unstructured or historical documents [1]. In cultural heritage preservation, libraries and archives employ KWS to index historical manuscripts, newspapers, and rare books, making them searchable without costly manual annotation. Additionally, it serves as an assistive technology for human transcribers, particularly when working with degraded or historical documents. These systems enhances transcription efficiency by suggesting probable keywords, even for previously unseen terms.

In corporate environments, KWS systems can automate the sorting of handwritten mail into predefined categories, such as “urgent”, “complaint”, and “cancelation”,

streamlining workflow efficiency. Beyond mailrooms, corporations leverage KWS to index archival records, including invoices, contracts, and legal documents allowing rapid retrieval without the need for full-text transcription.

Furthermore, the healthcare sector can benefit significantly from KWS, particularly in processing handwritten medical records. For example, Patient Care Reports (PCRs) and prescriptions often contain critical but unstructured notes, including patient and doctor information, symptoms, medication names, dosages, and care instructions. KWS extracts this key data while compensating for handwriting variations, improving accessibility for patients, doctors, and pharmacists. By automating prescription processing, it prevents dangerous medical errors caused by handwriting misinterpretations, thereby improving patient safety through more accurate treatment administration.

The application scope of KWS extends well beyond conventional documents, finding utility in a range of visually diverse and unstructured sources. Notable examples include mobile-captured handwritten notes [45], as well as graphical materials such as maps, posters, and technical diagrams. In mobile note-taking scenarios, KWS can facilitate real-time transcription and digitization. Alternatively, when applied to graphical documents, it enhances both searchability and contextual understanding by detecting and leveraging embedded textual elements. For instance, it can link keywords to figures in technical manuals or associate place names with their corresponding locations on maps.

Overall, KWS offers a scalable and cost-effective alternative to full transcription for both large-scale and focused applications.

2.2.3 Evaluation Metrics

One of the most widely adopted performance metrics in KWS, as well as in broader Information Retrieval (IR) tasks, is the Mean Average Precision (mAP), valued for its objectiveness and reliability [1]. As the name suggests, mAP is the mean of the Average Precision (AP) across all queries.

For a given query, *precision* is defined as the fraction of retrieved instances that are relevant to that query:

$$\text{Precision} = \frac{\text{number of } \textbf{relevant retrieved} \text{ instances}}{\text{number of } \textbf{retrieved} \text{ instances}},$$

while *recall* is the fraction of relevant instances that the system was able to retrieve

successfully:

$$\text{Recall} = \frac{\text{number of } \textbf{relevant retrieved} \text{ instances}}{\text{number of } \textbf{relevant} \text{ instances}}.$$

In other words, precision quantifies the accuracy of the predictions (fraction of true positives over true positives and false positives), whereas recall assesses how many true positives were actually identified (fraction of true positives over true positives and false negatives). *Precision at rank* k ($P@k$) is defined as the precision computed solely over the top k retrieved results, reflecting the reliability of the system's highest-ranked outputs.

For a given query, the Average Precision (AP) is defined as:

$$AP = \frac{1}{R} \sum_{k=1}^{n} P@k \cdot rel(k),$$

where $rel(k)$ is an indicator function that returns 1 if the retrieved instance at index k is considered relevant and 0 otherwise; and R denotes the total number of relevant instances for the query. This metric approximates the area under the precision-recall curve (see Figure 2.4) for a single query, providing a comprehensive evaluation of the precision-recall characteristics across the system's output ranking.

Finally, a precise definition of relevance is critical to prevent evaluation bias. For example, if an entire document is counted as relevant simply because it contains one matching snippet, this will artificially inflate performance metrics. Therefore, in the context of segmentation-free KWS, a retrieved image region is deemed relevant if it sufficiently overlaps with a ground-truth bounding box annotated with the same transcription as the query. Overlap is measured using the Intersection over Union (IoU) criterion, which is satisfied when the IoU exceeds a predefined threshold. Commonly used thresholds, which we also employ in our study, are 25% (mAP@25) and 50% (mAP@50). If multiple retrieved regions overlap with a single ground truth region, only one is considered relevant — typically, the one with the greatest overlap.

2.3 Related Works

Originally, Word Spotting was proposed in the speech recognition community [46] and was later adapted for printed [47] and handwritten [48] document indexing.

One of the major issues of the preprocessing stage is that possible segmentation errors are regularly conveyed in the spotting phase. Particularly, accurate word seg-

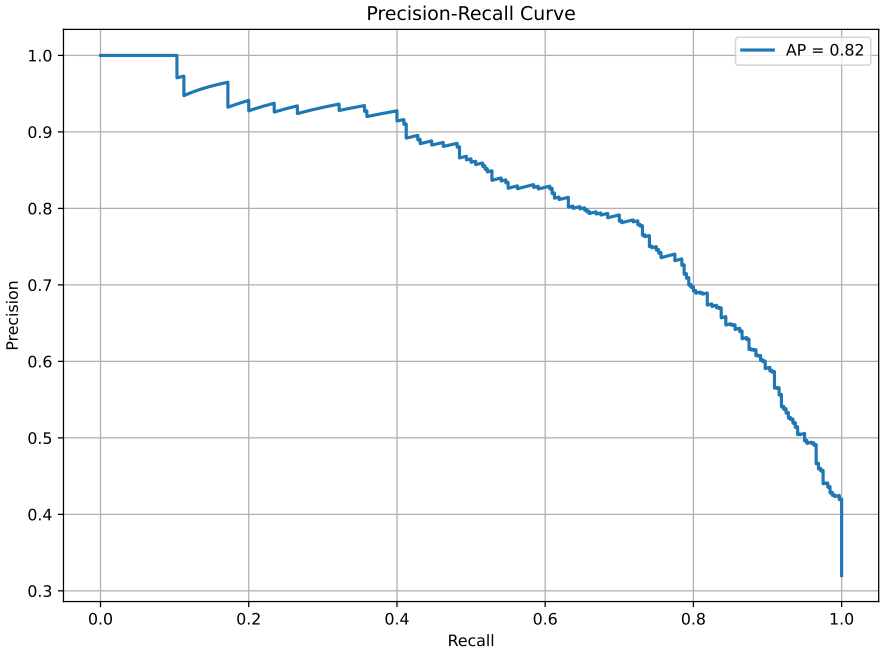


Figure 2.4: The precision-recall curve of a ranked list and the corresponding AP.

mentations are difficult to obtain in handwritten and degraded documents. For this reason, several *segmentation-free* word spotting techniques have emerged.

Early segmentation-free KWS methods addressed the problem of avoiding explicit word or line segmentation by analyzing entire document pages directly. This direction was initially dominated by hand-crafted feature extraction and region-of-interest proposals. Leydier et al. [49, 50] and Zhang et al. [51, 52] used local keypoints and gradient-based descriptors or Heat Kernel Signatures (HKSs), matched through elastic or manifold-based similarity metrics. However, these techniques incurred high computational costs and did not scale well.

A more scalable direction emerged with patch-based sliding-window frameworks [10, 12, 19, 53], where descriptors like SIFT, HoG, or pixel densities are extracted over image regions. In this respect, Rusiñol et al. [10] enhanced retrieval effectiveness via a Latent Semantic Indexing (LSI) projection, while Rothacker et al. [19] used a BoF-HMM formulation for robust query modeling.

Graph- and component-based techniques [13, 54, 55] modeled spatial or structural properties of documents using Connected Components (CCs) or grapheme graphs, typically matched using graph edit distances or geometric constraints. These methods, while segmentation-free, were often sensitive to noise, image binarization, and relied on heuristics to construct valid word proposals.

More recent advances leverage deep learning for end-to-end segmentation-free retrieval. Ghosh and Valveny [56] combined region proposal CNNs with attribute-based deep embeddings (PHOCNet representations [5]) to aggregate features across word-like regions. Wilkinson et al. [15] introduced Ctrl-F-Net, an end-to-end architecture that employs a ResNet34 backbone [57], Region Proposal Networks (RPNs), and word string embeddings such as the PHOC [2] and DCToW [21], to enable robust QbS retrieval. Rothacker et al. [58] further enhanced region detection under uncertainty by incorporating extremal region proposals and class activation maps into the word spotting pipeline.

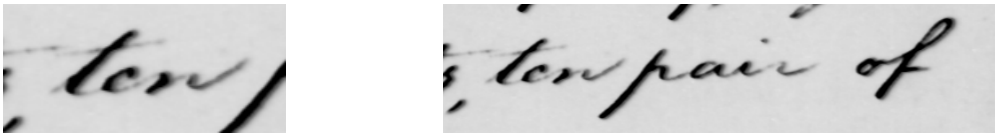
At the multi-task and multi-scale learning frontier, Zhao et al. [17] integrated a Feature Pyramid Network (FPN) into a CNN architecture, jointly training for pixel classification, bounding box regression, and visual-to-textual embedding learning (via DCToW). These contributions have advanced the segmentation-free paradigm towards dense, discriminative, and scalable retrieval pipelines.

While most prior work in this area focus on visual representations and similarity, recent trends point toward bridging the gap between visual and semantic domains using embeddings that encode language-aware properties [33]. However, such semantic alignment has remained under-explored in segmentation-free KWS settings. This motivates our work, which aims to enhance KWS effectiveness by introducing relevance-aware re-ranking based on *language models* and *semantic embeddings* of word image transcriptions.

2.4 Semantic Keyword Spotting

In a traditional KWS system, document regions are represented using descriptors such as PHOC or DCToW and matched against queries based solely on visual appearance and character-level similarity. However, words are more than sequences of characters; they also carry *semantic meaning*, which is often overlooked. On the other hand, when a person struggles to read a word in an illegible portion of a text, they often rely on the broader semantic context to disambiguate visually similar words. This process involves determining whether the interpreted letters form a valid word and whether that word fits logically within the surrounding sentence and passage. Figure 2.5 exemplifies this peculiarity. This observation motivates the integration of semantic

reasoning into word spotting systems.



(a) (b)

Figure 2.5: An example where relying solely on visual information can lead to ambiguity. The word shown in sub-figure (a) may be interpreted as “tcw”, “ton”, or “ten” based on visual cues alone. Word-level semantic context rules out “tcw” as it does not correspond to a valid English word. In sub-figure (b), the sentence-level semantic context enables the correct identification of the word as “ten”.

Semantic KWS was first introduced by Wilkinson et al. [21] as a method to enrich word image retrieval with language-level knowledge. The authors also coined the terms *Semantic KWS* and *Verbatim KWS* for the new and conventional approaches, respectively. In the semantic approach, retrieved results are ranked based on their semantic similarity to the query. When exact matches — results with the exact same transcription — are also included in the retrieved results, semantic KWS can be regarded as an extension of verbatim KWS.

The inexact — but conceptually relevant — results introduced by semantic KWS, can be especially useful when the query provided by the user does not accurately reflect their intended search targets. This may occur, for instance, when a user explores a document collection that is unfamiliar to them, when they misspell a query (either accidentally or because modern spelling has evolved away from historical forms), or when dealing with hyphenation and word splits across line breaks.

For instance, consider a scenario in which a user searches a document collection for instances of the word “book”. A typical verbatim system might retrieve visually similar but semantically irrelevant words such as “look” or “cook”, simply because they resemble the query when written. However, this fails to take into account the user’s underlying intent, as the user might not only be interested in exact matches of “book”, but could instead be referring to a specific book. Deeper semantic cues could help identify relevant content — even when the term “book” is not explicitly mentioned.

In other words, even if all instances of “book” are correctly retrieved at the top of the ranked list, the presence of visually similar yet irrelevant words in high-ranking

positions diminishes the effectiveness of the search. These positions could instead be occupied by semantically related terms such as "writer" (if it refers to the book's author), "library" or "bookshelf" (if the text mentions where the book is located), or "publication" (or another synonymous term for "book"). A search system that retrieves not only exact matches but also semantically related terms could significantly enhance the quality of both the search and the browsing experience.

The notion of semantic similarity can be subjective and context-dependent, which makes it challenging to formalize. For example, a word may be considered semantically related to another word if they are synonyms, morphological variants sharing the same stem (e.g., singular and plural forms of nouns, different tenses and participles of verbs, adjectival variations), or categorically related terms (e.g., "cat", "dog", "mammal", "animal").

In the literature, semantic retrieval methods can be broadly categorized based on how they address this challenge: (1) *ontology-based* techniques and (2) *context-based* techniques [24]. Ontology-based approaches [22, 59] use lexical resources such as the WordNet [60] to identify categorical or lexical similarities. Yet, as noted by Krishnan et al. [23], such manually constructed databases are limited in both the quality and quantity of their semantic relationships, as they depend on human annotations, and typically support only a handful of languages.

Context-based approaches are grounded in the distributional hypothesis [61], which posits that words appearing in *similar contexts* tend to share *similar meanings*. Recent advancements in NLP have produced a wide range of text embedding techniques, including Word2Vec, GloVe, FastText, and transformer-based models like BERT, which generate dense vector representations that encode semantic similarity beyond surface-level string matching. These embeddings have opened new opportunities for mapping visual content, such as handwritten words, into semantically meaningful spaces.

To bring this semantic capability into keyword spotting and obtain semantic representations, two general strategies have emerged: (1) learning visual-to-semantic mappings directly in an *end-to-end*, *recognition-free* manner, or (2) *transcribing* word images first, followed by transforming those transcriptions through a text-based embedding.

End-to-end approaches are particularly common in segmentation-based settings, where they circumvent explicit recognition — a process known to yield irrecoverable errors when mapped directly to embedding spaces. This paradigm was introduced

by Wilkinson et al. [21], who trained a two-stage CNN with cosine embedding loss to project word images into a pre-trained semantic space. Subsequent work by Krishnan et al. [23] extended this approach using the HWNet architecture [20] to jointly learn syntactic (PHOC or DCToW) and semantic (e.g., FastText) representations. Tüselmann et al. [33] further evaluated the impact of different embeddings — including FastText, GloVe, and BERT — using the same architecture, and explored combinations thereof for document-level semantic understanding. Finally, end-to-end approaches are particularly valuable because they enable the tackling of downstream NLP tasks — such as Named Entity Recognition (NER) [30, 62], Visual Question Answering (VQA) [32, 45], Named Entity Linking (NEL) [31] — directly on image documents, without explicit recognition.

Although effective, direct embedding methods remain limited in practice: they require large amounts of annotated training data, and no publicly available pre-trained models currently exist for generic semantic KWS. A case in point is the recommendation by Wilkinson et al. [21] to train their system on all 40 volumes of *The Writings of George Washington from the Original Manuscript Sources, 1745–1799*[1] to effectively capture corpus-level semantics of the George Washington (GW) database. Similarly, for model pre-training, other works [20, 33] employ a subset of HW-SYNTH [63] comprising approximately 600,000 word images and covering the 12,000 most frequent English words.

Despite these efforts, recent analysis [24] reveals that visual-semantic embeddings often retain mostly syntactic characteristics. This suggests a gap between visual and semantic domains, and highlights the underutilized potential of modern pre-trained language models in Word Spotting.

Last but not least, to the best of our knowledge, all existing semantic KWS methods are limited to the segmentation-based setting — a gap we bridge in this work.

2.5 Retrieval Enhancement via Relevance Feedback and Re-Ranking

A prominent challenge in KWS systems is the presence of false positives within the top-ranked retrieval results. To address this issue, several techniques for refining

[1] Available online: `https://archive.org/details/writingsofgeorge01wash`

the ranked list have been proposed In what follows, we review relevant literature in retrieval enhancement, particularly focusing on *relevance feedback* and *re-ranking* schemes that aim to improve the final ranked list beyond raw visual similarity. This includes both supervised and unsupervised paradigms and sets the foundation for our proposed *embedding-based re-ranking* method that leverages *NLP-driven semantic proximity* in the re-ranking process.

Supervised Relevance Feedback

In Supervised Relevance Feedback (SRF), the user manually labels several results from the initial retrieval as relevant or irrelevant. This information is then used to either reformulate the query vector (query refinement) or adjust the ordering of the existing ranked list (re-ranking). Approaches based on Rocchio's algorithm [64], as employed by Bhardwaj et al. [65] and Cao et al. [66], optimize the position of the query in the embedding space by incorporating positive and negative instances. Rusiñol et al. [67] explore both query reformulation and re-ranking using relevance scores. Konidaris et al. [68] and Kesidis et al. [69] involve user-selected positives from synthetic queries to improve real-word image retrieval. Additionally, Wolf et al. [70] investigate CNN-based confidence metrics to identify and suppress unreliable predictions, employing dropout, surrogate models, and sigmoid-activated meta-classifiers to evaluate prediction trustworthiness and prune false positives accordingly.

Unsupervised Feedback and Re-Ranking

The obvious benefits of Supervised Relevance Feedback lie in leveraging user judgments to guide ranking refinement. However, this process can be costly and subjective, particularly in degraded or historical documents. This limitation has motivated the adoption of *unsupervised* alternatives, such as Pseudo-Relevance Feedback (PRF) [71], where top-N ranked results are automatically assumed to be relevant and used for re-ranking or query expansion. Almazán et al. [12] introduced a two-stage ranking, combining fast approximate ranking with Fisher vector-based re-ranking and iterative query expansion. Similar approaches have been employed by Ghosh and Valveny [72], as well as Shekhar and Jawahar [71], who incorporate spatial pyramids for refinement. Vats and Fornés [73] propose a local query expansion technique based on confidence thresholds and keypoint matching, repeated across document pages,

yielding consistent performance gains.

While these methods have advanced re-ranking pipelines in the visual domain, they typically rely on low-level image similarity and neglect deeper semantic relationships between words; particularly useful in the presence of visual ambiguity or Out-of-Vocabulary (OOV) terms. This limitation further motivates the incorporation of language-based semantic reasoning into KWS systems, as we explore in the following chapters.

Chapter 3

Semantically-Informed Relevance Feedback

3.1 Introduction

In this chapter, we introduce a post-processing semantically-aware relevance feedback mechanism, designed to enhance the retrieval performance of KWS systems — the main contribution of this thesis. We first analyze two state-of-the-art KWS approaches (Sections 3.2 and 3.3), each embodying a distinct approach to traditional segmentation-free KWS. Their ranked output lists are utilized as inputs in the testing of our re-ranking process. Finally, we conclude the chapter by presenting our proposed re-ranking methodology in Section 3.4.

An overview of the proposed architecture is illustrated in Figure 3.1. Given a query, the KWS system retrieves the top-k most visually similar image regions from the document collection. Each retrieved region is transcribed, and an LLM projects both the transcription and query into a shared semantic subspace to determine their

semantic similarity. The final ranking combines the verbatim (visual) and semantic similarity scores to re-order the ranking and optimize retrieval accuracy.

Additionally, the re-ranking capability of the proposed framework extends beyond conventional KWS objectives. As demonstrated in Figure 3.1, visually similar but irrelevant terms (e.g., “letter”, “understand”, “better”) are replaced with semantically relevant military terms (e.g.,“sergeant”, “regiment”), providing users with more meaningful results — a qualitative improvement that traditional KWS metrics fail to capture.

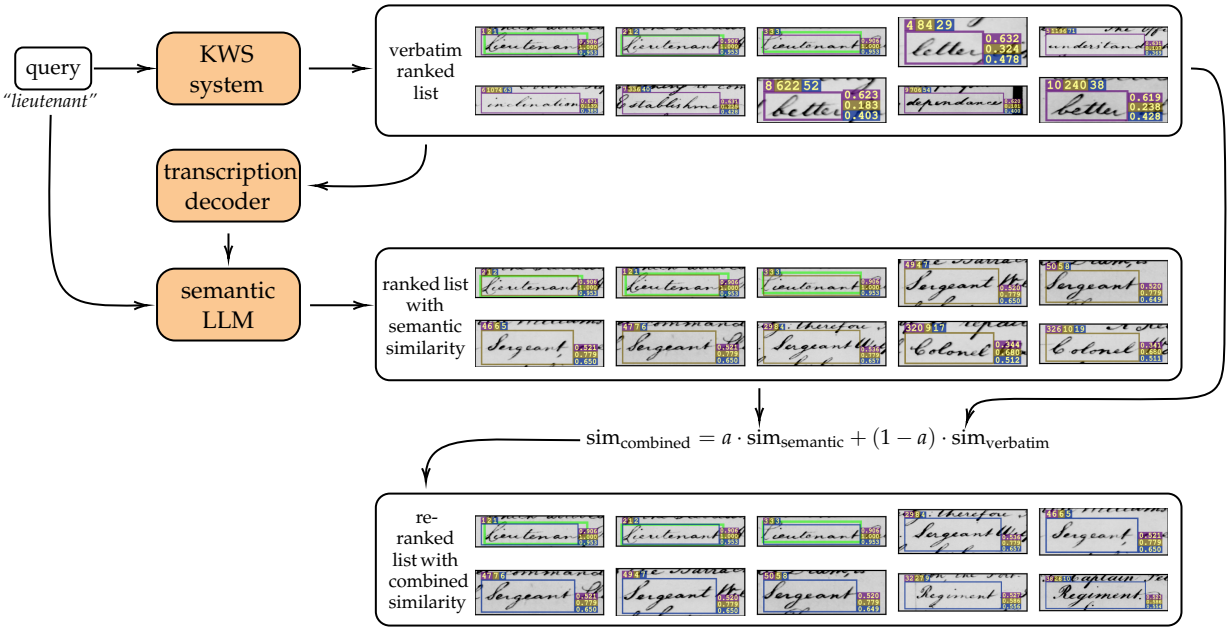


Figure 3.1: Proposed semantic relevance framework integrating LLM-based contextual similarities into the re-ranking of candidate word instances in segmentation-free KWS ranking lists.

3.2 The WordRetrievalNet

The WordRetrievalNet is a state-of-the-art segmentation-free QbS KWS system introduced by Zhao et al. [17]. It operates in two stages:

i. Offline stage: A CNN is trained to generate a database of candidate bounding boxes along with their representations in a latent space.

ii. Online stage: The descriptor of a query is matched against the database and a

ranked list of bounding boxes having the highest cosine similarity is returned.

Its end-to-end design eliminates the need for complex pre- and post-processing steps commonly required in other traditional KWS approaches. Additionally, it is scale-insensitive, since its FPN-based architecture [57, 74, 75] enables the extraction of multi-scale features directly from document images, which are then processed by three prediction heads:

1. a classification head that identifies pixels belonging to a positive word region,
2. a regression head that predicts the offsets between each pixel in a positive word region and the boundaries of its corresponding bounding box,
3. an embedding head that maps word regions into the latent space (e.g., DCToW, PHOC, etc).

The network is trained in a supervised manner minimizing the loss function:

$$\mathcal{L}_{all} = \mathcal{L}_{cls} + \mathcal{L}_{bbox} + \mathcal{L}_{embed}.$$

For the classification task, a loss based on the Dice coefficient [75, 76] is used:

$$\mathcal{L}_{cls} = 1 - \frac{2\sum_{i,j} \hat{y}_{cls}^{i,j} \cdot y_{cls}^{i,j}}{\sum_{i,j} \left(\hat{y}_{cls}^{i,j}\right)^2 + \sum_{i,j} \left(y_{cls}^{i,j}\right)^2},$$

where $\hat{y}_{cls}^{i,j}$, $y_{cls}^{i,j}$ denote the values of pixel (i, j) in the word classification prediction $\hat{y}_{cls}$, and the ground truth y_{cls}, respectively. This loss function counteracts the bias introduced by the imbalance between word pixels and background pixels.

For the bounding box regression, the Distance-IoU (DIoU) loss [77] is utilized — a variant of the IoU loss that augments it by incorporating information about the distance between the centers of predicted and ground-truth boxes. This formulation leads to faster convergence and improved localization accuracy compared to the conventional IoU. The loss function is defined as:

$$\mathcal{L}_{bbox} = \frac{1}{|C|} \sum_{i \in C} \mathrm{DIoU}\left(\hat{y}_{bbox}, y_{bbox}\right),$$

where C denotes the set of positively classified word pixels, $\hat{y}_{bbox}$ is the predicted bounding box, and y_{bbox} is the ground truth box.

For the word embedding, the cosine loss $\mathcal{L}_{embed} = 1 - \cos(\hat{y}_{embed}, y_{embed})$ is used in order to penalizes dissimilarity between the predicted representation $\hat{y}_{embed}$ and actual representation y_{embed}.

During inference, the set of positive word pixels and their offsets are combined to construct bounding boxes for the candidate word regions, and a Non-Maximum Suppression (NMS) filter is applied to reduce the density of predictions. The embedding of each bounding box is computed as the mean embedding of the pixels it contains. Figure 3.2 summarizes the above-mentioned key model components.

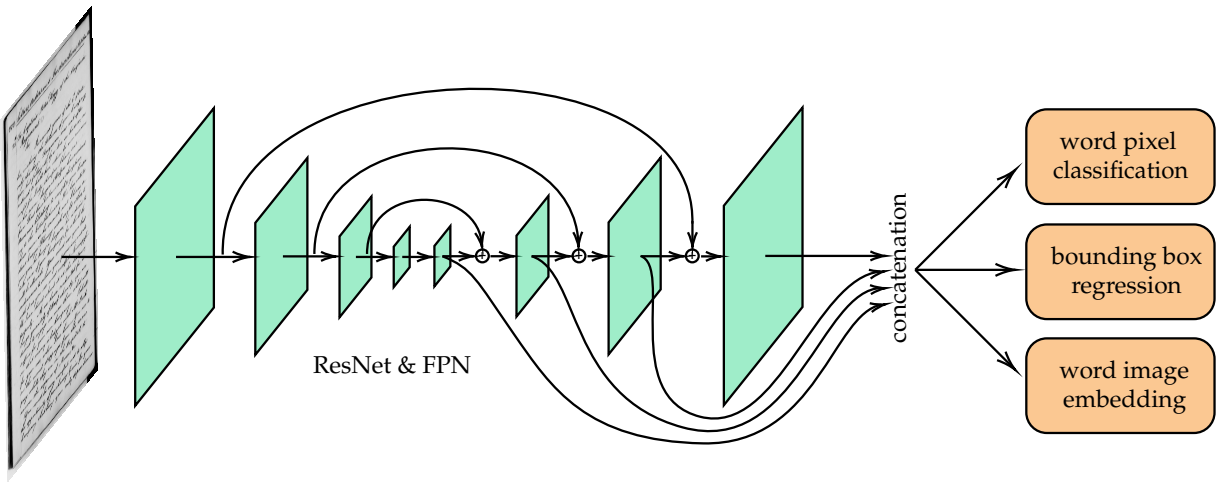


Figure 3.2: The WordRetrievalNet architecture

3.3 Segmentation-free KWS Simplified

Retsinas et al. [18] introduced a segmentation-free QbS KWS system, *KWS-Simplified*, that formulates KWS as a character counting problem: the goal is to identify image regions containing the same character histograms as the query. Unlike WordRetrievalNet, the lightweight system architecture eliminates synthetic training data requirements. However, to refine the initial predictions and compute their similarity with the query, several post-processing steps are applied. These steps include: (1) *Pyramidal Counting*, where a descriptor similar to PHOC is constructed and compared against the query descriptor; and (2) *CTC-based re-scoring* via force alignment. This alignment approach is used to improve the bounding box estimation and refine the ranking of candidates, in line with recent developments in CTC alignment and

scoring [78]. NMS is further applied to reduce overlapping predictions.

The system consists of a ResNet [57] backbone with two prediction heads: (1) a decoder head that estimates the probability distribution of each character occurrence across a document image; and (2) a scaler head that predicts the character scale at each image location. Figure 3.3 depicts the overall pipeline of the reference segmentation-free KWS system.

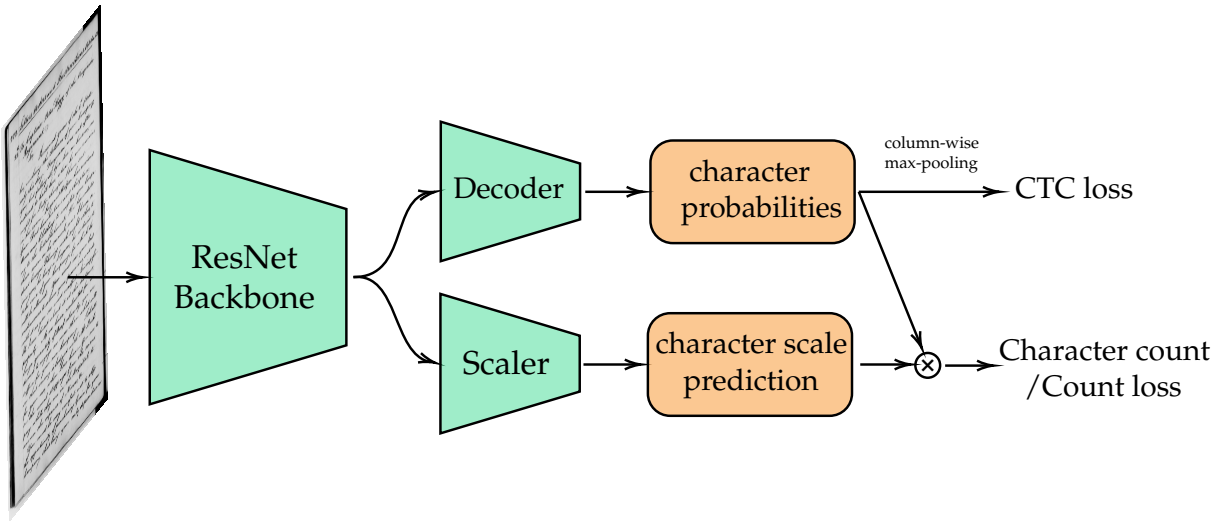


Figure 3.3: An overview of the KWS-Simplified network

For a given character c, let $F(i, j, c)$ denote the feature probability distribution output by the decoder head at image coordinates (i, j), and let $S(i, j)$ represent the predicted scale factor at the same coordinates. The number of occurrences of character c within a bounding box spanning from $(s1, s2)$ to $(e1, e2)$ is given by:

$$y_c = \sum_{i=s_1}^{e_1} \sum_{j=s_2}^{e_2} F(i, j, c) \cdot S(i, j)$$

This formulation gives rise to the counting loss function: $\mathcal{L}_{count} = \|y_c - t_c\|_2$, where t_c is the target character count.

During the training of the network, the loss function optimized combines the counting loss with the CTC loss [40] — a sequence alignment objective, commonly used in handwriting and speech recognition tasks, bridging continuous input signals (e.g., audio or image columns) with discrete output sequences (e.g., character labels) via its unique mechanism of blank token insertion and repetition collapsing. Following the approach of Retsinas et al. [18], a weighting factor of 10 is applied to the

counting loss to balance its contribution relative to the CTC loss:

$$\mathcal{L} = \mathcal{L}_{\text{CTC}} + 10 \cdot \mathcal{L}_{\text{count}}.$$

Herein, the feature map produced by the decoder head undergoes column-wise max-pooling before being fed into the CTC loss. Figure 3.4 overviews these post-processing scoring steps.

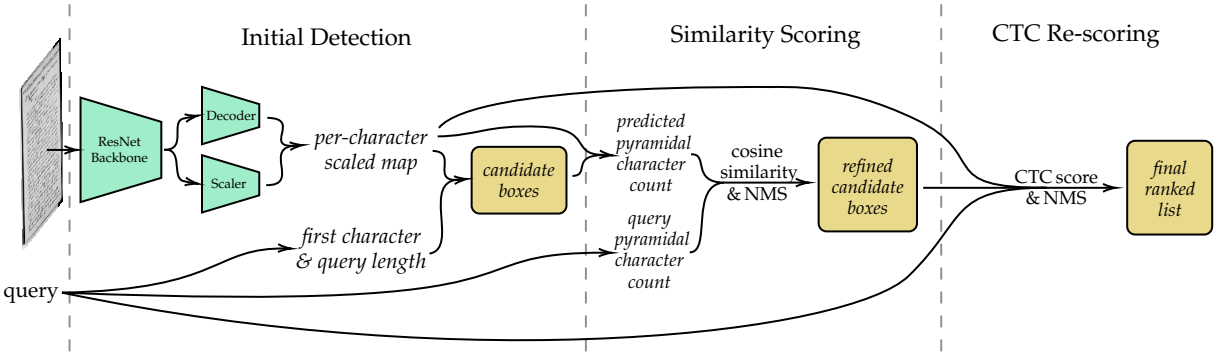


Figure 3.4: The post-processing stages of the KWS-Simplified pipeline

3.4 Proposed Re-ranking Pipeline

We conclude this chapter with the presentation of our proposed pipeline: an unsupervised relevance feedback mechanism enabled by the incorporation of semantic information during the re-ranking process. Its goal is to enhance retrieval performance through the suppression of false positive results that appear high in the initial ranking, while simultaneously promoting instances whose original rank underestimated their actual relevance.

The proposed framework operates in three stages (see Figure 3.1). Initially, a segmentation-free QbS KWS system performs verbatim retrieval, generating a list of candidate image regions that are ranked by their visual resemblance to a given query. Next, each candidate image region in the list is transcribed by a decoder, and a semantically aware LLM embeds the resulting transcriptions into a semantic space [36], where spatial proximity reflects semantic relatedness. Finally, the candidates are re-ranked based on a combination of their verbatim and semantic similarities.

We examine multiple variations of the framework:

i. two distinct decoder architectures,

ii. three alternative state-of-the-art semantic LLMs,

iii. two late fusion strategies.

The first decoder is adapted from the KWS-Simplified network. Notably, its character probability prediction head can operate as an independent module that generates transcriptions when a softmax function is applied to its output. For the second decoder, we utilize Transformer-based OCR (TrOCR), a state-of-the-art text recognition model. It integrates a Vision Transformer (ViT) encoder [79] with a Transformer-based text generator (typically initialized using either RoBERTa or MiniLM). The combined model has been pre-trained on large-scale synthetic textual data and can be fine-tuned on both machine-printed and handwritten document collections.

For the generation of semantic embeddings, we use three state-of-the-art BERT-like LLMs derived from RoBERTa, MPNet, and MiniLM, respectively. Each of these models has been specifically adapted for the task of semantic search through supervised contrastive learning [36], a training paradigm that teaches models to distinguish between correct sentence pairs and randomly sampled negative pairs. In order to handle Out-of-Vocabulary (OOV) terms (i.e., words for which no instances or transcriptions are available during training), while maintaining embeddings of a fixed dimension, the WordPiece [80] subword tokenization technique is used. The final embedding is obtained by mean pooling all token embeddings, and semantic similarity is quantified as the cosine similarity between the generated embedding vectors. Mean pooling is employed as it offers greater resilience to errors in the transcription of individual subtokens.

Finally, we evaluate two strategies that fuse visual and semantic relevance: *weighted combination of similarities* and *semantic pruning*. In the weighted combination strategy, we compute a weighted average of the verbatim and semantic similarity scores:

$$\text{sim}_{\text{combined}} = a \cdot \text{sim}_{\text{semantic}} + (1 - a) \cdot \text{sim}_{\text{verbatim}},$$

where $a \in [0, 1]$ is a hyperparameter controlling the relative importance of semantic similarity in the final score. The resulting values are used to re-rank the candidate list, incorporating both semantic and verbatim relevance. On the other hand, semantic pruning refers to the process of filtering candidate items by discarding those with a semantic similarity below a predefined threshold. The remaining candidates are then re-ranked based on their verbatim similarity.

Chapter 4

Experimental Evaluation

4.1 Introduction

This chapter is devoted to the experimental evaluation of the proposed re-ranking pipeline detailed in Section 3.4. To this end, we conduct experiments on two standard KWS benchmarks: the George Washington dataset and the IAM Handwriting Database, presented in Section 4.2. The employed evaluation protocol, outlined in Section 4.3, follows widely adopted practices for segmentation-free KWS systems. To ensure transparency and reproducibility, key implementation details are provided in Section 4.4. Quantitative results are then presented and analyzed in Section 4.5, assessing both the overall performance of the method and the contributions of its individual components. Section 4.6 complements this with a qualitative analysis of retrieval behaviors that are not fully captured by standard evaluation metrics. The chapter concludes with a discussion and summary of the main findings in Section 4.7.

4.2 Datasets

We evaluated our approach on two widely adopted KWS benchmarks: the *George Washington (GW)* dataset [41] and the *IAM Handwriting Database (IAM)* [42]. These datasets serve as standard references for performance comparison in the field [1].

4.2.1 The George Washington Dataset

The George Washington dataset[1] comprises 20 handwritten letters from George Washington's Papers at the Library of Congress [81]. These 18th-century documents, written in historical English by Washington and his aides, contain 4,860 annotated words with corresponding bounding boxes. Due to the minimal variation in the writing style, it can be characterized as a single-writer dataset.

Unlike segmentation-based KWS approaches that employ the standardized partition of Almazán et al. [2] for the GW database, no official partition exists for the evaluation of segmentation-free methods. Instead, we follow the established experimental practices used in prior works [14–18]. Given its limited size, it is customary to adopt a 4-fold cross-validation scheme, where each fold consists of 5 pages. Thus, four experimental iterations are conducted. During each iteration, one fold is reserved as the test set, while the remaining three serve as the training set. Additionally, one page from the training set is set aside for validation. Test queries are extracted from the unique transcriptions of the test pages by removing punctuation and lowercasing, whereas stopwords are retained as queries. Table 4.1 presents the exact partition of the dataset used in our experiments (the partition was obtained by shuffling the page indices 0–19 using NumPy with seed 0.)

Table 4.1: The partition of GW used in our experiments.

Fold No.	Pages across each fold
1	274, 309, 276, 272, 303
2	306, 273, 301, 300, 278
3	270, 302, 277, 275, 308
4	307, 304, 279, 271, 305

[1] Available online: `https://fki.tic.heia-fr.ch/databases/washington-database`

4.2.2 The IAM Handwriting Database

The IAM Handwriting Database[2,3] contains 1,539 pages of modern cursive handwritten English text produced by 657 writers. The pages are segmented and annotated, comprising a total of 115,320 words. The variability introduced by the multi-writer setting is a principal factor contributing to the difficulty of the dataset. We use the official partition of the database, as is common practice in the literature [15, 18]; however, unlike GW, no cross-validation is performed. The test queries comprise all unique, lowercased transcriptions from the test set, excluding words that contain non-alphanumeric characters, punctuation marks, erroneous annotations, or those words appearing in the official stopword list [2, 5].

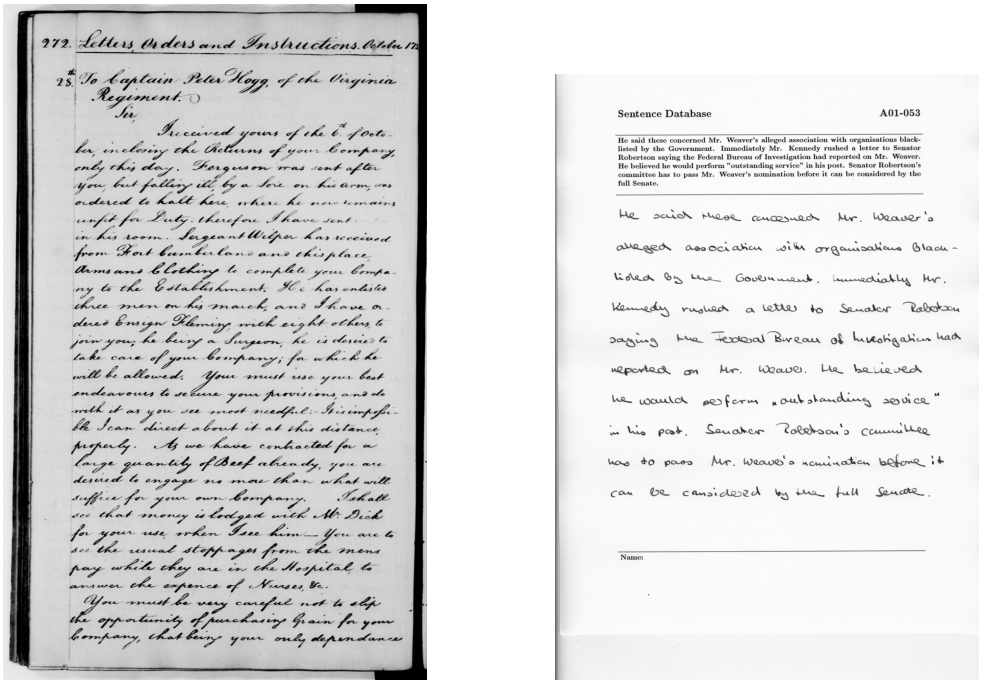


Figure 4.1: Examples of document images from the GW and IAM datasets.

4.3 Evaluation Protocol

In recent years, a standard evaluation protocol has emerged for segmentation-free QbS KWS systems introduced by Rothacker et al. [14] as an extension of the Almazán protocol [2]. We adhere to this established procedure, ensuring direct comparability

[2] Available online: `https://fki.tic.heia-fr.ch/databases/iam-handwriting-database`

[3] IAM = Institut für Informatik und Angewandte Mathematik, University of Bern, Bern, Switzerland

with prior work [15–18].

Hence, the test queries are derived from all unique transcriptions in the test pages. Additionally, these queries undergo normalization, including lowercasing all words, removing non-alphanumeric characters and punctuation, and optionally filtering stopwords (see Section 4.2 for dataset-specific details). Following standard practice, we report mAP@25 and mAP@50 as our primary evaluation metrics.

4.4 Implementation Details

This section describes the implementation[4] details of the proposed framework, the training and evaluation of the baseline models, and the modifications made to ensure fair and consistent comparisons. Implementation-specific adjustments are documented to support reproducibility and to explain observed deviations from previously reported results.

Both WordRetrievalNet and KWS-Simplified offer open-source implementations, as well as pre-trained models. However, while KWS-Simplified includes training on the IAM dataset, neither system has been pre-trained on GW.

In order to enable a direct comparison between these architectures and to establish a rigorous baseline against which we evaluate our framework's relative improvements, we trained both systems on GW using the typical 4-fold cross-validation scheme described in Section 4.2 and the partition shown in Table 4.1. For each cross-validation iteration, we trained a WordRetrievalNet model for 120 epochs and a KWS-Simplified model for 200 epochs, evaluating mAP@25 performance on the validation set within every 10 epochs and retaining the best-performing model as the final baseline. The training configuration of each system (e.g., optimizer selection, hyperparameter values) followed the the settings reported by Zhao et al. [17] and Retsinas et al. [18] as optimal — those yielding the highest mAP performance in their evaluations. The reproduced and reported mAP scores are recorded in Table 4.2.

Similarly, WordRetrievalNet was trained on the official IAM split for 80 epochs, achieving mAP@25 of 79.15% and mAP@50 of 72.85%. It is worth noting that Zhao et al. [17] do not report results for this dataset.

[4]The source code, along with the pre-trained models, is publicly available at `https://github.com/stevepapazis/semantically-reranked-kws`.

Table 4.2: Comparison of reproduced and reported mAP scores for the two baseline models, WordRetrievalNet and KWS-Simplified, on the GW dataset.

Model	**mAP@25**	**mAP@50**
WordRetrievalNet (Reproduced)	94.31 ± 1.8	88.29 ± 4.0
WordRetrievalNet (Reported)	96.46	94.06
KWS-Simplified (Reproduced)	89.74 ± 0.7	72.29 ± 3.0
KWS-Simplified (Reported)	91.6	66.4

While our reproduced results deviate from prior work, our goal is to establish a baseline reference system, and therefore, such observed divergences are to be expected and can be attributed to several factors.

First and foremost, the original implementations used different partitions of GW. In the case of WordRetrievalNet, an additional source of variability arises from the use of randomly generated synthetic training data.

Second, we note minor implementation-specific differences in the training data preparation pipelines. Since both systems are segmentation-free, they employ a detection phase in which the model identifies image regions that are likely to contain words. These candidate regions are then refined to match the query and produce the final retrieval results. To train such a detector effectively, it is not sufficient to use only the tight bounding boxes of the ground truth word images. Instead, it is beneficial to enlarge these regions so it can learn to segment words more reliably. Overestimating the bounding box is generally preferable to underestimating it, as missing a word entirely would hinder recall.

In the training of KWS-Simplified on GW, the authors generate training images by extending the ground truth bounding boxes by a constant factor. In our implementation, we enlarged the word image areas by a factor of 2.5. We observed that this larger context window helped the model learn more robust representations corresponding to the segmented word regions, which in turn led to a substantial improvement in mAP@50 performance.

Furthermore, the implementation of WordRetrievalNet uses training data extracted as image crops that cover regions larger than the ground truth word bounding boxes, often including multiple word instances. We note here that recovering the exact parameterization which yielded the optional result in the original work is not possible.

Therefore, we modified the patch cropping algorithm used for extracting positive examples during the training of WordRetrievalNet, aiming to improve the computational efficiency of the training loop. The original algorithm repeatedly sampled image patches until no word within a patch crossed its boundary, or until a limit of 1,000 iterations was reached. In practice, this limit was frequently exhausted, creating an artificial bottleneck that significantly slowed down training and rendered the computation CPU-bound. Our modified approach addresses this issue by retaining only the words for which at least 70% of the bounding box falls within the sampled patch. While the trade-off improved computational efficiency, it came at the expense of retrieval accuracy, particularly visible in the mAP@50 scores.

Finally, some variability is inevitably introduced by the stochastic nature of neural network training, such as the random weight initialization and the random sampling during the minimization of the loss function.

As discussed earlier in Section 3.4, we evaluate two distinct decoder architectures for transcribing the retrieved image regions. Each *KWS-Simplified-based decoder* inherently shares weights with the corresponding KWS-Simplified backbone trained on the same partition, thereby requiring no additional training before deployment. For the *TrOCR-based* alternative, we initialized the model in each cross-validation iteration of GW using weights from a HTR model[5] pre-trained on the IAM database. Next, the model was fine-tuned for 20 epochs using the set of word images from the training set of the current iteration, while the queries from one page were used for validation. We employed the AdamW optimization algorithm [82] with an initial learning rate of $5e{-}5$, which increased linearly during a short warm-up period and then decayed linearly for the remaining epochs. The fine-tuning reduced the average Character Error Rate (CER) on the validation sets of GW from 26.76% to 11.05% — a standard metric in OCR and HTR, which measures the number of character-level errors in a predicted transcription compared to the ground truth. Although the strong 3.42% CER performance reported for TrOCR on IAM [39] suggests that there is potential room for improvement on GW, accurate retrieval does not necessarily require explicit transcription. We report CER solely to support reproducibility. Moreover, given this performance, we used the decoder in our experiments on IAM without further training.

To embed the decoded transcriptions into a semantic space, we employed three

[5]Model card: `https://huggingface.co/microsoft/trocr-base-handwritten`

pre-trained models from the SentenceTransformer[6] Python library: (1) stsb-roberta-base[7] (the RoBERTa architecture fine-tuned on the Semantic Textual Similarity benchmark [83]), (2) all-mpnet-base-v2[8] (the MPNet architecture fine-tuned on a corpus of diverse datasets comprising over one billion sentence pairs), and (3) all-MiniLM-L12-v2[9] (the MiniLM architecture fine-tuned on the same diverse corpus).

Ultimately, both WordRetrievalNet and the three semantic embedding models utilize the cosine distance to compute the verbatim and semantic similarities, respectively. In comparison, the KWS-Simplified system relies on a similarity measure derived from CTC scores. To ensure that these values are on a comparable scale when combined, we normalize the CTC-based scores to the range $[-1, 1]$ using the minimum and maximum values observed in the training set.

4.5 Ablation Experiments

This Section presents the ablation experiments conducted to evaluate the impact of different configurations on the proposed pipeline. These experiments aim to isolate the effects of the re-ranking strategy and assess the generalization ability of our method across different initial word spotters. Tables 4.3 to 4.6 present the numerical results obtained after re-ranking the initial WordRetrievalNet and KWS-Simplified ranking lists using the proposed weighted combination strategy. These results confirm our intuition that the combination of verbatim and semantic information actually enhances retrieval accuracy. On the contrary, Tables 4.7 and 4.8 show the results for the alternative strategy based on semantic pruning. Despite its aim to alleviate the retrieval by identifying semantically irrelevant instances, the strategy is overly simplistic and effectively hampers system performance. Performance on GW is reported as the average mAP@25 and mAP@50, along with their corresponding Standard Deviations (SD), computed across four experimental trials. The contribution of each pipeline component is discussed in detail below.

[6] Available online: `https://sbert.net/`

[7] Model card: `https://huggingface.co/sentence-transformers/stsb-roberta-base`

[8] Model card: `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

[9] Model card: `https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2`

4.5.1 Impact of Baseline KWS Model

The effect of our semantic re-ranking pipeline varies depending on the baseline KWS model. On one hand, WordRetrievalNet exhibits stronger improvements compared to the KWS-Simplified variant. As shown in Tables 4.3 and 4.4, as well as Figure 4.2, the gains on the GW dataset reach $\sim 2.3\%$ in mAP@25 (from 94.31% to 96.59%) and $\sim 0.9\%$ in mAP@50 (from 88.29% to 90.17%). Even more noticeable are the improvements on the IAM dataset, where mAP@25 rises by $\sim 3\%$ (from 79.15% to 82.12%) and mAP@50 by $\sim 2.6\%$ (from 72.85% to 75.43%).

On the other hand, KWS-Simplified achieves smaller gains, as reported in Tables 4.5 and 4.6, as well as Figure 4.3. On GW, the re-ranking yields an increase of $\sim 1.2\%$ in mAP@25 (from 89.74% to 90.94%) and $\sim 0.7\%$ in mAP@50 (from 72.29% to 72.97%). On IAM, the respective gains are $\sim 1.9\%$ (mAP@25: 86.40% to 88.25%) and $\sim 1.2\%$ (mAP@50: 63.73% to 64.88%).

These differences can be attributed to the design of each backbone. WordRetrievalNet produces a larger and richer set of candidate detections for each query, since it performs query-independent spotting by precomputing visual features (e.g., DCToW) across the document. This allows the re-ranking module to refine the initial results more effectively, as it has access to more potential matches, including those ranked low in the initial retrieval.

Conversely, KWS-Simplified retrieves a very limited set of candidates, typically just a handful per query. This restricts the impact of re-ranking, as relevant instances not retrieved initially cannot be recovered in later stages. Therefore, KWS-Simplified may benefit from integrating query expansion techniques aimed at enlarging the initial candidate pool, making semantic re-ranking more effective.

4.5.2 Impact of Decoder

The decoder module directly influences the effectiveness of the re-ranking process, as reflected in the performance variations across datasets and architectures. On GW, the KWS-Simplified decoder yields higher mAP gains than TrOCR, improving mAP@25 by $\sim 2.3\%$ (from 94.31% to 96.59%) and mAP@50 by $\sim 0.9\%$ (from 88.29% to 90.17%). Unlike the first case, TrOCR shows smaller gains: $\sim 0.6\%$ for mAP@25 (from 89.74% to 90.38%) and $\sim 0.3\%$ for mAP@50 (from 72.29% to 72.57%).

This pattern reverses on IAM, where the TrOCR decoder, when paired with Wor-

Table 4.3: mAP performance on GW. *WordRetrievalNet* is the employed backbone paired with the *weighted combination* strategy across semantic importance thresholds and embeddings.

Relative semantic importance	Semantic embeddings	KWS-Simplified decoder		TrOCR decoder	
		mAP@25±SD	mAP@50±SD	mAP@25±SD	mAP@50±SD
0.0*	all-MiniLM -L12-v2	94.31 ± 1.8	88.29 ± 4.0	94.31 ± 1.8	88.29 ± 4.0
0.1		95.80 ± 1.5	89.51 ± 3.8	**94.51 ± 1.8**	**88.43 ± 4.0**
0.2		96.10 ± 1.5	89.75 ± 3.8	93.96 ± 2.0	87.84 ± 4.0
0.3		**96.30 ± 1.3**	**89.79 ± 3.7**	93.20 ± 2.1	87.07 ± 3.8
0.4		96.04 ± 1.5	89.64 ± 3.6	91.71 ± 2.5	85.73 ± 3.9
0.5		95.34 ± 1.4	89.01 ± 3.5	89.05 ± 2.9	83.19 ± 3.8
0.6		94.39 ± 1.5	88.20 ± 3.5	84.82 ± 4.2	79.41 ± 4.2
0.7		93.25 ± 1.6	87.37 ± 3.5	79.04 ± 4.8	74.03 ± 3.7
0.8		91.92 ± 1.9	86.26 ± 3.5	73.12 ± 6.1	68.42 ± 4.5
0.9		90.93 ± 2.0	85.42 ± 3.6	69.68 ± 6.6	65.20 ± 4.8
1.0		89.83 ± 2.2	84.28 ± 3.7	64.58 ± 7.5	60.07 ± 6.0
0.0*	all-mpnet -base-v2	94.31 ± 1.8	88.29 ± 4.0	94.31 ± 1.8	88.29 ± 4.0
0.1		95.90 ± 1.4	89.67 ± 3.8	**94.68 ± 1.6**	**88.62 ± 3.9**
0.2		96.27 ± 1.3	89.86 ± 3.8	94.09 ± 1.7	87.98 ± 3.8
0.3		96.50 ± 1.4	89.97 ± 4.0	93.07 ± 2.1	87.02 ± 3.7
0.4		**96.56 ± 1.5**	**90.08 ± 3.9**	91.78 ± 2.4	85.85 ± 3.7
0.5		96.00 ± 1.3	89.62 ± 3.6	89.42 ± 2.9	83.85 ± 3.8
0.6		95.04 ± 1.2	88.82 ± 3.5	85.13 ± 3.6	79.84 ± 3.8
0.7		93.83 ± 1.6	87.86 ± 3.4	79.10 ± 5.3	74.10 ± 4.1
0.8		92.57 ± 1.7	86.71 ± 3.4	74.06 ± 6.2	69.30 ± 4.6
0.9		91.53 ± 1.8	85.84 ± 3.6	70.50 ± 6.4	65.90 ± 4.6
1.0		90.21 ± 2.1	84.45 ± 3.7	64.91 ± 7.5	60.31 ± 6.0
0.0*	stsb-roberta -base	94.31 ± 1.8	88.29 ± 4.0	94.31 ± 1.8	88.29 ± 4.0
0.1		95.83 ± 1.6	89.60 ± 3.9	**94.58 ± 2.0**	**88.55 ± 4.2**
0.2		96.19 ± 1.3	89.82 ± 3.7	94.01 ± 2.1	87.91 ± 4.2
0.3		**96.59 ± 1.3**	**90.17 ± 3.7**	92.87 ± 2.9	86.92 ± 4.5
0.4		96.44 ± 1.6	89.99 ± 3.7	90.75 ± 3.2	84.96 ± 4.4
0.5		96.32 ± 1.6	89.90 ± 3.7	87.91 ± 3.5	82.32 ± 4.3
0.6		96.04 ± 1.7	89.57 ± 3.8	84.15 ± 4.3	78.83 ± 4.9
0.7		95.27 ± 1.7	88.91 ± 4.0	80.37 ± 4.4	75.25 ± 4.7
0.8		94.37 ± 1.8	88.08 ± 3.9	77.09 ± 4.4	72.16 ± 4.3
0.9		93.54 ± 2.0	87.46 ± 4.0	74.05 ± 4.8	69.22 ± 4.1
1.0		92.32 ± 2.2	86.19 ± 3.9	67.98 ± 5.4	63.34 ± 4.3

*This is essentially the baseline model. It does not use a decoder.

Table 4.4: mAP performance on IAM. *WordRetrievalNet* is the employed backbone paired with the *weighted combination* strategy across semantic importance thresholds and embeddings.

Relative semantic importance	Semantic embeddings	KWS-Simplified decoder		TrOCR decoder	
		mAP@25	mAP@50	mAP@25	mAP@50
0.0*	all-MiniLM -L12-v2	79.15	72.85	79.15	72.85
0.1		**80.60**	**73.98**	**82.04**	**75.40**
0.2		79.29	72.62	80.59	73.77
0.3		75.87	69.30	77.68	71.05
0.4		71.36	65.13	74.04	67.66
0.5		65.95	60.19	70.68	64.57
0.6		60.26	54.99	67.17	61.33
0.7		55.95	50.99	64.47	58.88
0.8		53.12	48.38	62.88	57.42
0.9		51.46	46.88	62.22	56.83
1.0		47.96	43.71	54.34	49.48
0.0*	all-mpnet -base-v2	79.15	72.85	79.15	72.85
0.1		**80.84**	**74.18**	**82.12**	**75.43**
0.2		80.01	73.21	80.77	74.02
0.3		77.55	70.82	77.87	71.29
0.4		73.86	67.22	74.73	68.33
0.5		69.24	62.95	71.23	65.12
0.6		63.70	57.97	67.84	61.99
0.7		58.18	52.97	65.43	59.78
0.8		53.92	49.10	63.80	58.25
0.9		51.43	46.89	62.78	57.36
1.0		49.26	44.84	54.99	50.11
0.0*	stsb-roberta -base	79.15	72.85	79.15	72.85
0.1		**81.16**	**74.49**	**81.88**	**75.18**
0.2		80.39	73.55	80.97	74.22
0.3		77.40	70.68	77.51	70.90
0.4		73.54	66.94	74.29	67.87
0.5		69.45	63.25	71.28	65.11
0.6		64.83	59.05	68.70	62.85
0.7		61.28	55.73	66.71	61.03
0.8		58.23	53.05	65.16	59.66
0.9		56.22	51.22	64.13	58.74
1.0		53.99	49.11	56.03	51.24

*This is essentially the baseline model. It does not use a decoder.

Table 4.5: mAP performance on GW. *KWS-Simplified* is the employed backbone paired with the *weighted combination* strategy across semantic importance thresholds and embeddings.

Relative semantic importance	Semantic embeddings	KWS-Simplified decoder		TrOCR decoder	
		mAP@25±SD	mAP@50±SD	mAP@25±SD	mAP@50±SD
0.0*	all-MiniLM -L12-v2	89.74 ± 0.7	72.29 ± 3.0	89.74 ± 0.7	72.29 ± 3.0
0.1		90.62 ± 0.5	72.77 ± 3.0	90.27 ± 0.8	72.53 ± 3.1
0.2		90.68 ± 0.4	72.82 ± 3.0	**90.38 ± 0.7**	**72.57 ± 3.1**
0.3		90.70 ± 0.5	72.84 ± 3.0	90.35 ± 0.7	72.55 ± 3.2
0.4		90.72 ± 0.5	72.84 ± 3.0	90.30 ± 0.7	72.51 ± 3.2
0.5		90.79 ± 0.4	72.90 ± 3.0	90.32 ± 0.6	72.55 ± 3.2
0.6		90.85 ± 0.5	72.93 ± 3.0	90.32 ± 0.6	72.54 ± 3.2
0.7		**90.91 ± 0.5**	**72.97 ± 3.0**	90.26 ± 0.7	72.51 ± 3.2
0.8		90.83 ± 0.4	72.90 ± 2.9	90.12 ± 0.6	72.42 ± 3.1
0.9		90.82 ± 0.4	72.89 ± 2.9	89.93 ± 0.6	72.29 ± 3.0
1.0		90.43 ± 0.7	72.73 ± 2.7	87.16 ± 0.6	70.39 ± 2.1
0.0*	all-mpnet -base-v2	89.74 ± 0.7	72.29 ± 3.0	89.74 ± 0.7	72.29 ± 3.0
0.1		90.63 ± 0.5	72.78 ± 3.0	90.22 ± 0.7	72.49 ± 3.1
0.2		90.69 ± 0.5	72.83 ± 3.0	90.31 ± 0.7	72.51 ± 3.1
0.3		90.72 ± 0.5	72.84 ± 3.0	90.29 ± 0.6	72.49 ± 3.1
0.4		90.81 ± 0.5	72.91 ± 3.1	**90.32 ± 0.7**	**72.54 ± 3.1**
0.5		90.86 ± 0.5	72.96 ± 3.1	90.29 ± 0.7	72.53 ± 3.2
0.6		90.87 ± 0.4	72.94 ± 3.0	90.26 ± 0.6	72.49 ± 3.1
0.7		**90.94 ± 0.5**	**72.97 ± 3.1**	90.10 ± 0.7	72.39 ± 3.1
0.8		90.91 ± 0.5	72.96 ± 3.0	90.06 ± 0.6	72.35 ± 3.0
0.9		90.85 ± 0.4	72.92 ± 3.0	89.74 ± 0.4	72.09 ± 2.9
1.0		90.47 ± 0.5	72.78 ± 2.8	87.06 ± 0.7	70.23 ± 2.1
0.0*	stsb-roberta -base	89.74 ± 0.7	72.29 ± 3.0	89.74 ± 0.7	72.29 ± 3.0
0.1		90.62 ± 0.5	72.78 ± 3.0	90.21 ± 0.6	72.46 ± 3.0
0.2		90.63 ± 0.5	72.79 ± 3.0	90.27 ± 0.6	72.47 ± 3.0
0.3		90.66 ± 0.5	72.80 ± 3.0	90.27 ± 0.6	72.47 ± 3.1
0.4		90.76 ± 0.5	72.88 ± 3.0	90.32 ± 0.6	72.52 ± 3.1
0.5		90.89 ± 0.5	**72.97 ± 3.1**	**90.33 ± 0.5**	**72.54 ± 3.1**
0.6		90.90 ± 0.4	**72.97 ± 3.1**	90.30 ± 0.4	72.50 ± 3.0
0.7		**90.92 ± 0.4**	**72.97 ± 3.1**	90.13 ± 0.4	72.41 ± 2.9
0.8		90.86 ± 0.4	72.92 ± 3.0	89.96 ± 0.5	72.28 ± 2.9
0.9		90.74 ± 0.4	72.86 ± 3.0	89.75 ± 0.5	72.12 ± 2.8
1.0		90.33 ± 0.5	72.71 ± 2.8	87.08 ± 0.9	70.24 ± 2.0

*This is essentially the baseline model. It does not use a decoder.

Table 4.6: mAP performance on IAM. *KWS-Simplified* is the employed backbone paired with the *weighted combination* strategy across semantic importance thresholds and embeddings.

Relative semantic importance	Semantic embeddings	KWS-Simplified decoder		TrOCR decoder	
		mAP@25	mAP@50	mAP@25	mAP@50
0.0*	all-MiniLM -L12-v2	86.40	63.73	86.40	63.73
0.1		86.71	63.86	87.49	64.22
0.2		**86.79**	**63.94**	87.84	64.47
0.3		86.57	63.72	88.01	64.58
0.4		86.11	63.36	**88.06**	**64.60**
0.5		85.65	63.01	87.74	64.43
0.6		84.66	62.38	87.28	64.17
0.7		83.32	61.47	86.30	63.56
0.8		81.31	60.20	84.66	62.49
0.9		77.61	57.85	81.51	60.47
1.0		71.97	54.39	76.74	57.24
0.0*	all-mpnet -base-v2	86.40	63.73	86.40	63.73
0.1		86.69	**63.88**	87.53	64.25
0.2		**86.70**	63.87	87.89	64.51
0.3		86.56	63.75	88.16	64.80
0.4		86.32	63.59	**88.25**	**64.88**
0.5		85.91	63.33	87.97	64.69
0.6		85.03	62.71	87.52	64.37
0.7		84.15	62.10	86.60	63.77
0.8		82.56	61.09	85.07	62.78
0.9		79.47	59.19	81.97	60.86
1.0		74.04	55.89	77.30	57.78
0.0*	stsb-roberta -base	86.40	63.73	86.40	63.73
0.1		86.58	63.82	87.35	64.13
0.2		**86.78**	**63.85**	87.78	64.52
0.3		86.72	63.74	88.04	64.72
0.4		86.45	63.59	**88.12**	**64.84**
0.5		85.87	63.29	87.62	64.60
0.6		85.09	62.79	86.78	64.16
0.7		83.82	61.99	85.44	63.32
0.8		81.96	60.79	83.06	61.79
0.9		79.45	59.12	80.17	59.88
1.0		76.10	57.18	76.49	57.40

*This is essentially the baseline model. It does not use a decoder.

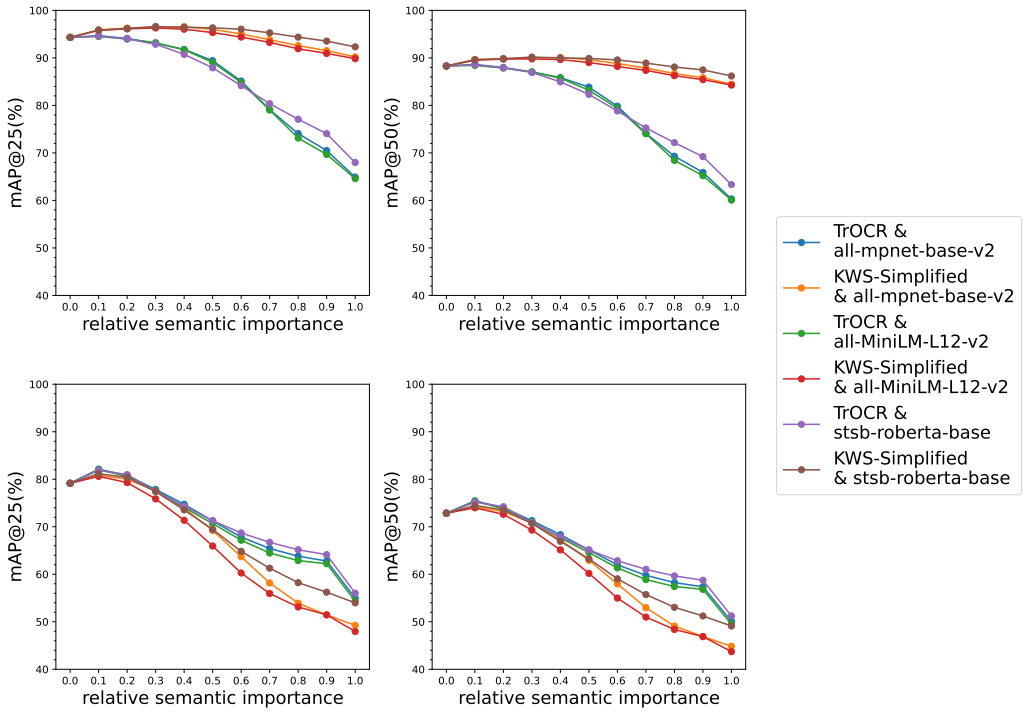


Figure 4.2: mAP@25 and mAP@50 curves for the *WordRetrievalNet* baseline paired with the *weighted combination* strategy reflecting the results from Tables 4.3 and 4.4 on GW (top) and IAM (bottom) benchmarks.

dRetrievalNet, achieves stronger improvements: $\sim 3\%$ (mAP@25: from 79.15% to 82.16%) and $\sim 2.6\%$ (mAP@50: from 72.85% to 75.43%). The KWS-Simplified decoder on the same dataset shows slightly lower gains: $\sim 2\%$ for mAP@25 and $\sim 1.6\%$ for mAP@50. These trends are summarized in Tables 4.4 and 4.6.

Such results are consistent with prior work in KWS: the final performance is shaped not only by the quality of the decoder but also by the accuracy of the initial bounding boxes [35]. Even a strong decoder like TrOCR may struggle on GW due to limited candidate diversity, whereas it benefits more on IAM, where segmentation is more fine-grained and the linguistic space is richer. This highlights the interplay between decoder expressiveness and the underlying retrieval quality.

Ultimately, these findings emphasize the importance of holistic system design. Decoder selection should not be based solely on CER or transcription quality, but also on how well it complements the retrieval front-end. While our current strategy employs a simple late-fusion scheme, it still provides meaningful gains with no additional supervision and minimal computational overhead.

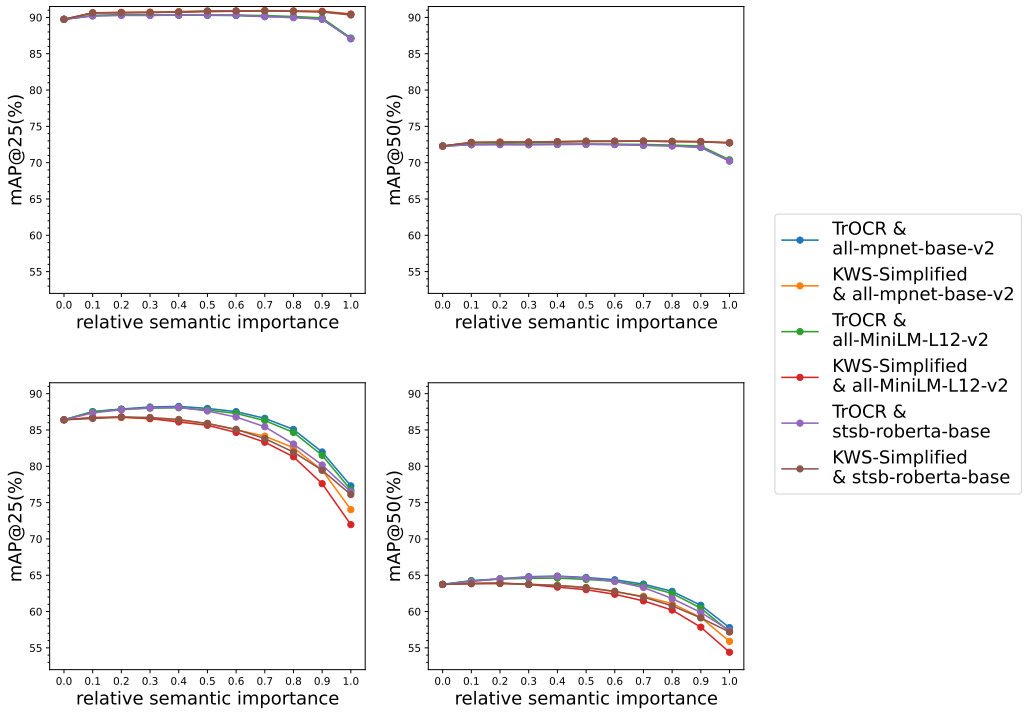


Figure 4.3: mAP@25 and mAP@50 curves for the *KWS-Simplified* baseline paired with the *weighted combination* strategy reflecting the results from Tables 4.5 and 4.6 on GW (top) and IAM (bottom) benchmarks.

4.5.3 Impact of Semantic Embedding Model

We expected that the more diversely trained semantic embedding models (all-MiniLM-L12-v2, all-mpnet-base-v2) would outperform stsb-roberta-base. However, across all model and decoder combinations, we observed at most a 0.5% difference in both mAP@25 and mAP@50, with performance typically being indistinguishable across the GW and IAM benchmarks. This can be attributed to factors such transcription errors, the reliance on word-level context alone and dataset limitations (i.e., the small size of GW), all of which prevent the more expressive models from demonstrating their full semantic representational power.

4.5.4 Impact of Fusion Strategy: Weighted Combination

The weighted combination of verbatim and semantic relevance scores leads to consistent performance gains, as showcased in Tables 4.3 to 4.6, along with Figures 4.2 and 4.3. Nonetheless, the extent of the improvement and the importance of the semantic relevance level depend on the performance of the underlying keyword spotter

and the decoder architecture.

On the GW database, the already strong performance of WordRetrievalNet shifts the optimal semantic relevance weight toward lower values, favoring verbatim relevance, for both decoders. This effect is less pronounced for the KWS-Simplified decoder, where optimal performance is achieved when the semantic weight lies within the range [0.2, 0.5]. On the contrary, for the TrOCR-based decoder, as well as the WordRetrievalNet on IAM, the best performance occurs when the semantic influence is minimal, typically within [0.1, 0.2]. Outside of these ranges, and particularly when relying solely on semantic similarity, performance deteriorates. This decline indicates that recognition errors introduced during transcription propagate into the semantic space as well. Additionally, it highlights the importance of leveraging both verbatim and semantic signals for effective re-ranking. In each of these cases, the performance drops sharply as the level of semantic importance increases. For instance in the TrOCR-decoded pipeline, mAP@25 decreases from 94.31% to approximately 65% and mAP@50 drops from 88.29% to about 62% on average across all choices of semantic embeddings.

In comparison, as shown in the re-ranking results of KWS-Simplified in Table 4.5, the less accurate initial ranking leads to a stronger reliance on the semantic component for achieving the best results. Furthermore, the re-ranking consistently outperforms the baseline for the KWS-Simplified decoder while remains on par with its TrOCR counterpart. Notably, there appears to be a synergy between the KWS-Simplified decoder and the baseline model. Since the decoder is an integral part of this baseline model, the initially predicted bounding boxes are optimally aligned for the decoder to use. Consequently, the incorporation of semantic relevance offers clear advantages for this model.

4.5.5 Impact of Fusion Strategy: Semantic Pruning

Thus far, we have presented results based on the weighted combination strategy, which consistently delivers the best performance among the tested fusion methods. On the contrary, semantic pruning, a simpler alternative, performs noticeably worse across all configurations, as shown in Tables 4.7 and 4.8, as well as Figure 4.4.

The key limitation of semantic pruning lies in its rigid filtering rule, which only accepts candidates exceeding a predefined similarity threshold. While this may boost

Table 4.7: mAP performance on GW. *WordRetrievalNet* is the employed backbone paired with the *semantic pruning* strategy across filtering thresholds and embeddings.

Filtering threshold	Semantic embeddings	KWS-Simplified decoder		TrOCR decoder	
		mAP@25±SD	mAP@50±SD	mAP@25±SD	mAP@50±SD
0.1	all-MiniLM -L12-v2	**94.25 ± 1.7**	**88.24 ± 3.9**	**94.16 ± 1.7**	**88.20 ± 4.1**
0.3		92.02 ± 2.3	86.32 ± 3.7	77.32 ± 6.3	71.77 ± 4.3
0.5		89.87 ± 2.3	84.52 ± 3.7	67.59 ± 6.7	63.15 ± 4.8
0.7		88.58 ± 2.6	83.68 ± 3.8	63.25 ± 6.3	59.34 ± 4.7
0.9		87.81 ± 2.7	83.06 ± 3.8	60.10 ± 6.6	56.45 ± 5.3
0.1	all-mpnet -base-v2	**94.29 ± 1.8**	**88.28 ± 4.0**	**93.96 ± 2.1**	**87.97 ± 4.2**
0.3		91.74 ± 2.0	86.21 ± 3.4	76.40 ± 5.8	71.46 ± 4.0
0.5		89.51 ± 2.2	84.27 ± 3.7	66.96 ± 7.0	62.76 ± 5.4
0.7		88.50 ± 2.3	83.51 ± 3.8	63.11 ± 6.2	59.20 ± 4.7
0.9		87.50 ± 2.9	82.76 ± 4.0	59.70 ± 6.8	56.05 ± 5.4
0.1	stsb-roberta -base	**94.33 ± 1.8**	88.32 ± 4.0	**93.49 ± 1.9**	**87.45 ± 4.2**
0.3		94.32 ± 2.1	**88.33 ± 4.2**	85.81 ± 3.4	80.42 ± 4.3
0.5		92.95 ± 2.0	87.08 ± 3.7	73.83 ± 4.6	69.09 ± 4.0
0.7		90.72 ± 2.5	85.36 ± 3.7	65.99 ± 5.3	62.04 ± 4.0
0.9		87.82 ± 2.9	83.02 ± 3.9	60.61 ± 6.2	56.93 ± 4.9

Table 4.8: mAP performance on GW. *KWS-Simplified* is the employed backbone paired with the *semantic pruning* strategy across filtering thresholds and embeddings.

Filtering threshold	Semantic embeddings	KWS-Simplified decoder		TrOCR decoder	
		mAP@25±SD	mAP@50±SD	mAP@25±SD	mAP@50±SD
0.1	all-MiniLM -L12-v2	**89.63 ± 0.8**	**72.18 ± 3.1**	**89.50 ± 0.9**	**72.05 ± 3.0**
0.3		87.52 ± 1.2	70.26 ± 3.7	71.46 ± 5.1	56.03 ± 3.9
0.5		86.53 ± 1.6	69.69 ± 4.1	63.25 ± 6.1	49.08 ± 4.3
0.7		85.62 ± 2.2	69.01 ± 4.7	59.35 ± 5.9	45.98 ± 4.3
0.9		85.24 ± 2.3	68.71 ± 4.7	56.53 ± 5.9	43.64 ± 4.4
0.1	all-mpnet -base-v2	**89.72 ± 0.7**	**72.27 ± 3.0**	**89.34 ± 0.9**	**71.90 ± 2.9**
0.3		87.91 ± 1.4	70.75 ± 3.8	71.68 ± 4.9	56.63 ± 3.4
0.5		86.46 ± 2.0	69.72 ± 4.5	62.64 ± 5.7	48.51 ± 4.4
0.7		85.72 ± 1.9	69.04 ± 4.5	59.04 ± 5.8	45.68 ± 4.0
0.9		85.24 ± 2.5	68.71 ± 4.8	56.15 ± 6.0	43.26 ± 4.4
0.1	stsb-roberta -base	**89.74 ± 0.7**	**72.29 ± 3.0**	**88.96 ± 1.1**	**71.52 ± 3.2**
0.3		89.05 ± 1.2	71.73 ± 3.4	81.33 ± 2.7	64.75 ± 3.0
0.5		88.00 ± 1.6	70.90 ± 4.1	69.24 ± 3.8	54.77 ± 3.5
0.7		86.28 ± 2.4	69.62 ± 4.7	62.33 ± 4.6	48.99 ± 3.3
0.9		85.35 ± 2.4	68.83 ± 4.8	57.18 ± 5.4	44.19 ± 3.9

semantic purity, it often leads to valid instances being discarded due to minor recognition or decoding errors. As a result, retrieval performance deteriorates, especially under stricter thresholds.

Despite its suboptimal results in verbatim keyword spotting, this strategy serves an illustrative role: it highlights the value of combining semantic and lexical signals rather than treating them in isolation. In future applications, semantic pruning might be better suited for tasks where conceptual alignment or query expansion is more critical than exact string matches.

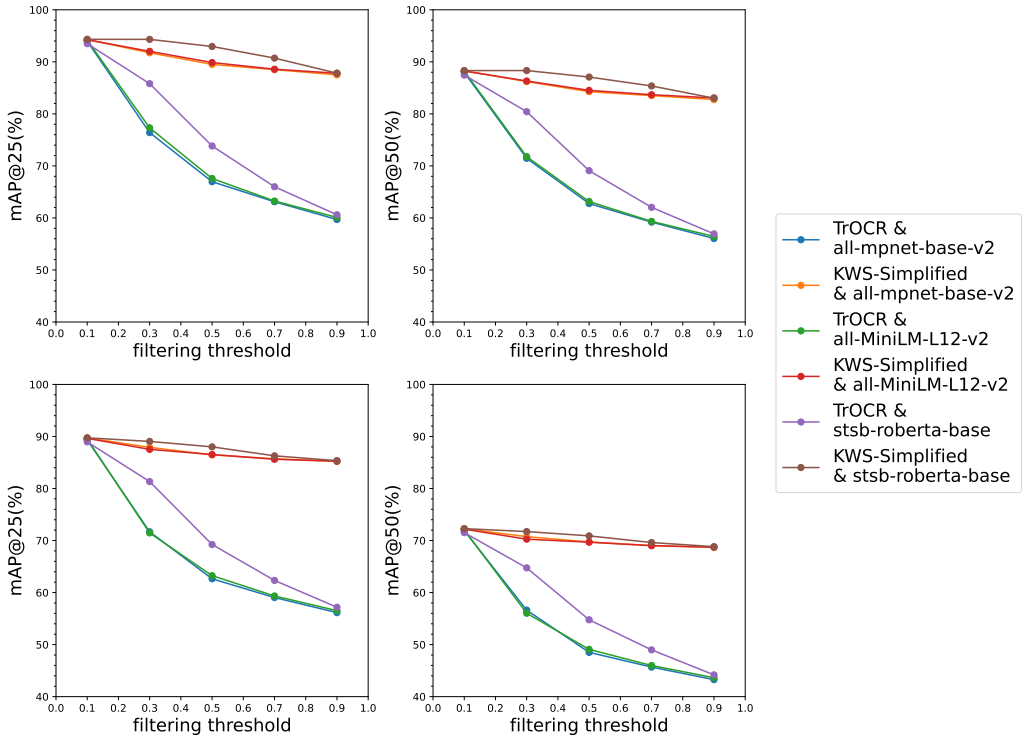


Figure 4.4: mAP@25 and mAP@50 curves for the *semantic pruning* strategy on GW reflecting the results from Tables 4.7 and 4.8 for the WordRetrievalNet (top) and KWS-Simplified (bottom) baselines.

4.6 Qualitative Analysis

Previous sections evaluated the quantitative performance of the proposed framework through mAP metrics. We now examine its qualitative benefits: the system successfully augments retrieval with semantic matches while preserving exact lexical matching capabilities. Visual examples demonstrate these enhancements, which transcend

standard evaluation indices of verbatim KWS.

Figure 4.5 shows three top-10 ranked lists for the query “forgot”. The first list is ranked by the conventional (verbatim) KWS similarity, the second by semantic similarity, and the third by the combined similarity used in our proposed system. For each instance, the top-left corner displays the global rank assigned by each method, using different colors: purple for verbatim, yellow for semantic, and blue for combined similarity. The bottom-right corner shows the similarity scores with respect to the query, following the same color scheme. The ground-truth bounding box is highlighted in green.

Due to a suboptimal bounding box prediction, the relevant instance corresponding to the query “forgot” is initially ranked tenth, following several instances of “fort”. Both the semantic ranking and the combined re-ranking successfully identify its true relevance, increasing its score while also reducing the scores of the visually similar false positive instances corresponding to “fort”. Yet, the visually similar results still dominate the combined list, due to several factors, including the limited size and semantic variability of the GW dataset, as well as a significant class imbalance. That

Figure 4.5: Top-10 ranked list for the query “forgot” ordered by verbatim (purple), semantic (yellow), and combined (blue) ranking. The rightmost part of each snippet represents corresponding similarities, while the ground truth is highlighted in green.

is, many words have either very few or no semantically related instances in the database. Although the re-ranking process successfully promotes the correct instance to the top rank, its bounding box remains unchanged. As a result, under stricter overlap thresholds, this correction is effectively disregarded, which explains the greater performance gain in mAP@25 compared to mAP@50.

Another noteworthy qualitative characteristic of the proposed pipeline, useful in recognition-free semantic retrieval, is illustrated in Figure 4.6. In a similar fashion to the previous example, we present three ranked lists for the query "soldiers". While the vocabulary of the GW dataset generally lacks semantic depth due to its limited size, this limitation is less pronounced in certain areas, such as military terminology. In such a scenario, the qualitative benefit is clear: when a user searches for military-related terms (e.g., "soldiers"), the system is more likely to retrieve other relevant military terms (e.g., "officers", "recruits", "military") rather than visually similar but semantically unrelated results. Hence, users receive results that better reflect the intended meaning of their query.

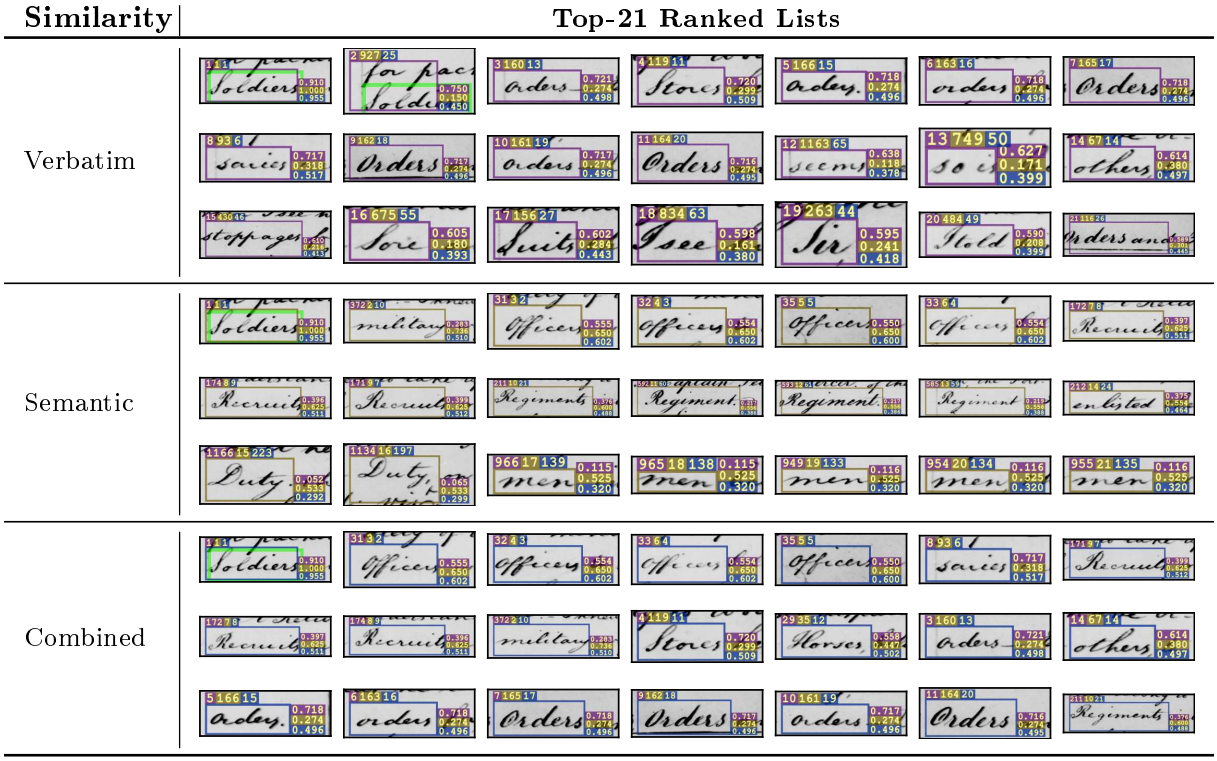


Figure 4.6: Top-21 ranked list for the query "soldiers" ordered by verbatim (purple), semantic (yellow), and combined (blue) ranking. The rightmost part of each snippet represents corresponding similarities, while the ground truth is highlighted in green.

Moreover, note that since there is only one instance of the word "soldiers" in the dataset, the second instance in the initial verbatim retrieval is an erroneous duplication of the first, possibly due to suboptimal NMS. The final re-ranking correctly discards this duplicate, a qualitative improvement not effectively measured by mAP.

4.7 Discussion

In a nutshell, our semantic re-ranking framework operates as a modular, plug-and-play extension to existing KWS systems, requiring neither retraining nor dataset-specific adaptation. In contrast with approaches that rely on fine-tuned embeddings or corpus-dependent training, our method is able to generalize effectively across datasets. This generalization capability is further supported by consistent performance gains observed across both employed datasets. To ensure robustness of the GW results despite the limited size of the dataset, we reported mAP scores averaged over four experimental trials. The consistently low standard deviation observed across these experiments indicate the reliability and stability of the proposed semantic augmentation pipeline. This robustness is further evidenced by the consistent effectiveness of the method across different semantic embedding models, highlighting its adaptability to varying NLP backends.

The broader candidate pool and richer semantic space of IAM allow the semantic re-ranking to manifest more clearly, facilitating its ability to exploit higher-level meaning when more linguistic diversity is available. These results validate the generalization capacity of our method to adapt to heterogeneous data scenarios. Notably, WordRetrievalNet features improvements of 3% in mAP@25 (from 79.15% to 82.12%) and 2.6% in mAP@50 (from 72.85% to 75.43%), while KWS-Simplified achieves $+1.85\%$ (from 86.40% to 88.25%) and $+1.15\%$ (from 63.73% to 64.88%) respectively.

In our numerical results over GW, we observe consistent improvements for both baseline models when semantic relevance is incorporated. This behavior holds across both mAP@25 and mAP@50 metrics, regardless of the decoder architecture. WordRetrievalNet achieves the most substantial gain, with mAP@25 increasing by 2.3% (from 94.31% to 96.59%), while KWS-Simplified improves by 1.2% (from 89.74% to 90.94%).

Although the relative improvements in mAP@50 are more modest (typically less

than 1%), this does not fully capture the value of our method. The high baseline performance suggests that remaining false negatives are inherently difficult cases, often involving retrieved instances that fall below the IoU threshold due to imperfect bounding box predictions by the underlying word spotter. As a result, even when re-ranking elevates a semantically relevant instance, it may not be counted as correct under the strict overlap criterion. This evaluation limitation underscores the need for post-retrieval refinements that can adjust the predicted bounding boxes. Future extensions such as query expansion [12], late fusion [84], or joint re-ranking and refinement modules could alleviate this issue.

A natural direction for future exploration involves moving beyond late fusion toward fully trainable semantic re-ranking modules. Integrating fusion architectures inspired by recent vision-language models, such as CLIP-Rerankers [85] or BLIP [86], could enable richer cross-modal interactions between query and candidate embeddings. While such models introduce additional training complexity, they offer the potential to jointly optimize retrieval, decoding, and semantic alignment in a single unified pipeline.

Chapter 5

Conclusions and Recommendations for Future Work

This final chapter summarizes the key contributions and empirical findings of the thesis, and outlines potential directions and opportunities for future research.

5.1 Conclusions

This thesis investigates enhanced segmentation-free Keyword Spotting (KWS) through the integration of semantic understanding enabled by modern NLP breakthroughs, particularly large pre-trained language models. Motivated by the limitations of existing systems, especially their dependence on purely visual similarity during retrieval, we propose a lightweight, modular framework that projects retrieved word snippets into a latent semantic space, where a semantic similarity score is computed between each snippet and the query. The resulting semantic similarity scores are fused with the original visual similarities from the keyword spotter to enable joint re-ranking.

Experimental results on two standard KWS benchmarks, the George Washington (GW) and IAM datasets validate our initial hypothesis: semantic feedback can

systematically enhance retrieval precision while maintaining recall. Across two representative segmentation-free baselines (WordRetrievalNet and KWS-Simplified) and multiple decoding strategies, we observe consistent mAP improvements. Most notably, WordRetrievalNet achieves a $\sim 2.3\%$ absolute gain in mAP@25 on GW, as well as a $\sim 3\%$ improvement on IAM when paired with TrOCR. These findings demonstrate that integrating contextual word embeddings enhances Word Spotting retrieval quality even in recognition-free setups — a scenario where no prior work exists on semantic KWS to our knowledge.

The robustness of our framework is further supported by its low variance across cross-validation folds, and its insensitivity to the choice of semantic embedding model. The optimal fusion weights suggest that a balanced contribution between verbatim and semantic signals yields the best performance. Moreover, qualitative examples demonstrate that semantically relevant but visually dissimilar terms can be effectively elevated in the ranking, showcasing the method's ability to perform beyond mere pattern matching.

Overall, our findings encourage a shift in perspective wherein deep language models need not be confined to transcription, but can actively contribute to the semantic understanding and retrieval of handwritten documents. We hope this work paves the way toward more intelligent, generalizable, and semantically aware document analysis systems.

5.2 Future Directions

While this work demonstrates the viability of semantic-augmented segmentation-free KWS, several promising directions emerge for further research.

A promising avenue for future research lies in transitioning from late fusion approaches to fully trainable end-to-end re-ranking systems. Recent advances in vision-language models — particularly architectures like CLIP-Rerankers [85] and BLIP [86] — demonstrate the potential for deeper, learnable interactions between visual and textual embeddings. While such models inevitably introduce greater computational complexity during training, they offer a significant advantage: the ability to jointly optimize document retrieval, text decoding, and semantic alignment within a unified neural framework.

Bibliography

[1] A. P. Giotis, G. Sfikas, B. Gatos and C. Nikou, "A survey of document image word spotting techniques", *Pattern Recognition*, vol. 68, pp. 310–332, 2017. DOI: 10.1016/j.patcog.2017.02.023 (cit. on pp. 3, 5, 10, 12, 13, 30).

[2] J. Almazán, A. Gordo, A. Fornes and E. Valveny, "Word Spotting and Recognition with Embedded Attributes", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 12, pp. 2552–2566, 2014. DOI: 10.1109/TPAMI.2014.2339814 (cit. on pp. 4, 10, 11, 16, 30, 31).

[3] A. Sharma and S. K. Pramod, "Adapting off-the-shelf CNNs for word spotting & recognition", in *Proceedings of the* 13^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 986–990. DOI: 10.1109/ICDAR.2015.7333909 (cit. on p. 4).

[4] P. Krishnan, K. Dutta and C. V. Jawahar, "Deep Feature Embedding for Accurate Recognition and Retrieval of Handwritten Text", in *Proceedings of the* 15^{th} *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2016, pp. 289–294. DOI: 10.1109/ICFHR.2016.0062 (cit. on p. 4).

[5] S. Sudholt and G. A. Fink, "Evaluating Word String Embeddings and Loss Functions for CNN-Based Word Spotting", in *Proceedings of the* 14^{th} *IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 493–498. DOI: 10.1109/ICDAR.2017.87 (cit. on pp. 4, 10, 16, 31).

[6] G. Retsinas, G. Louloudis, N. Stamatopoulos and B. Gatos, "Efficient Learning-Free Keyword Spotting", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1587–1600, 2019. DOI: 10.1109/TPAMI.2018.2845880 (cit. on p. 4).

[7] M. Mhiri, C. Desrosiers and M. Cheriet, "Word spotting and recognition via a joint deep embedding of image and text", *Pattern Recognition*, vol. 88, pp. 312–320, 2019. DOI: 10.1016/j.patcog.2018.11.017 (cit. on p. 4).

[8] F. Wolf, K. Brandenbusch and G. A. Fink, "Improving Handwritten Word Synthesis for Annotation-free Word Spotting", in *Proceedings of the* 17^{th} *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2020, pp. 61–66. DOI: 10.1109/ICFHR2020.2020.00022 (cit. on p. 4).

[9] F. Daraee, S. Mozaffari and S. M. Razavi, "Handwritten keyword spotting using deep neural networks and certainty prediction", *Computers & Electrical Engineering*, vol. 92, pp. 107–111, 2021. DOI: 10.1016/j.compeleceng.2021.107111 (cit. on p. 4).

[10] M. Rusiñol, D. Aldavert, R. Toledo and J. Llados, "Browsing Heterogeneous Document Collections by a Segmentation-Free Word Spotting Method", in *Proceedings of the* 11^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, 2011, pp. 63–67. DOI: 10.1109/ICDAR.2011.22 (cit. on pp. 4, 15).

[11] J. Almazán, A. Gordo, A. Fornes and E. Valveny, "Efficient Exemplar Word Spotting", in *Proceedings of the British Machine Vision Conference*, 2012, pp. 67.1–67.11. DOI: 10.5244/C.26.67 (cit. on p. 4).

[12] J. Almazán, A. Gordo, A. Fornes and E. Valveny, "Segmentation-free word spotting with exemplar SVMs", *Pattern Recognition*, vol. 47, no. 12, pp. 3967–3978, 2014. DOI: 10.1016/j.patcog.2014.06.005 (cit. on pp. 4, 15, 20, 49).

[13] A. Kovalchuk, L. Wolf and N. Dershowitz, "A Simple and Fast Word Spotting Method", in *Proceedings of the* 14^{th} *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2014, pp. 3–8. DOI: 10.1109/ICFHR.2014.9 (cit. on pp. 4, 15).

[14] L. Rothacker and G. A. Fink, "Segmentation-free Query-by-String Word Spotting with Bag-of-Features HMMs", in *Proceedings of the* 13^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 661–665. DOI: 10.1109/ICDAR.2015.7333844 (cit. on pp. 4, 30, 31).

[15] T. Wilkinson, J. Lindström and A. Brun, "Neural Ctrl-F: Segmentation-Free Query-by-String Word Spotting in Handwritten Manuscript Collections", in

Proceedings of the 16^{th} *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4443–4452. DOI: 10.1109/ICCV.2017.475 (cit. on pp. 4, 12, 16, 30–32).

[16] T. Wilkinson, J. Lindström and A. Brun, *Neural Word Search in Historical Manuscript Collections*, 2020. DOI: 10.48550/arXiv.1812.02771 (cit. on pp. 4, 30, 32).

[17] P. Zhao, W. Xue, Q. Li and S. Cai, "Query by Strings and Return Ranking Word Regions with Only One Look", in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020, pp. 3–18. DOI: 10.1007/978-3-030-69544-6_1 (cit. on pp. 4, 6, 10, 16, 23, 30, 32).

[18] G. Retsinas, G. Sfikas and C. Nikou, "Keyword Spotting Simplified: A Segmentation-free Approach Using Character Counting and CTC Re-scoring", in *Document Analysis and Recognition - ICDAR 2023*, Springer Nature Switzerland, 2023, pp. 446–464. DOI: 10.1007/978-3-031-41676-7_26 (cit. on pp. 4, 6, 7, 10, 25, 26, 30–32).

[19] L. Rothacker, M. Rusiñol and G. A. Fink, "Bag-of-Features HMMs for Segmentation-Free Word Spotting in Handwritten Documents", in *Proceedings of the* 12^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, 2013, pp. 1305–1309. DOI: 10.1109/ICDAR.2013.264 (cit. on pp. 4, 15).

[20] P. Krishnan and C. V. Jawahar, "HWNet v2: an efficient word image representation for handwritten documents", *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 22, no. 4, pp. 387–405, 2019. DOI: 10.1007/s10032-019-00336-x (cit. on pp. 4, 10, 19).

[21] T. Wilkinson and A. Brun, "Semantic and Verbatim Word Spotting Using Deep Neural Networks", in *15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Shenzhen, China: IEEE, 2016, pp. 307–312. DOI: 10.1109/ICFHR.2016.0065 (cit. on pp. 4, 10, 11, 16, 17, 19).

[22] P. Krishnan and C. V. Jawahar, "Bringing Semantics in Word Image Retrieval", in *2013 12th International Conference on Document Analysis and Recognition*, 2013, pp. 733–737. DOI: 10.1109/ICDAR.2013.150 (cit. on pp. 4, 18).

[23] P. Krishnan and C. Jawahar, "Bringing semantics into word image representation", *Pattern Recognition*, vol. 108, p. 107 542, 2020. DOI: 10.1016/j.patcog.2020.107542 (cit. on pp. 4, 5, 18, 19).

[24] O. Tüselmann, F. Wolf and G. A. Fink, "Identifying and Tackling Key Challenges in Semantic Word Spotting", in *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, IEEE, 2020, pp. 55–60. DOI: 10.1109/ICFHR2020.2020.00021 (cit. on pp. 4, 5, 18, 19).

[25] T. Mikolov, K. Chen, G. Corrado and J. Dean, *Efficient Estimation of Word Representations in Vector Space*, 2013. DOI: 10.48550/arXiv.1301.3781 (cit. on p. 5).

[26] J. Pennington, R. Socher and C. Manning, "GloVe: Global Vectors for Word Representation", in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162 (cit. on p. 5).

[27] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, "Enriching Word Vectors with Subword Information", *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, Jun. 2017. DOI: 10.1162/tacl_a_00051 (cit. on p. 5).

[28] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186. DOI: 10.48550/arXiv.1810.04805 (cit. on p. 5).

[29] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach", 2019. DOI: 10.48550/arXiv.1907.11692 (cit. on p. 5).

[30] O. Tüselmann, F. Wolf and G. A. Fink, "Are End-to-End Systems Really Necessary for NER on Handwritten Document Images?", in *Document Analysis and Recognition – ICDAR 2021*, Springer International Publishing, 2021, pp. 808–822. DOI: 10.1007/978-3-030-86331-9_52 (cit. on pp. 5, 19).

[31] O. Tüselmann and G. A. Fink, "Named Entity Linking on Handwritten Document Images", in *Document Analysis Systems*, Springer International Publishing, 2022, pp. 199–213. DOI: 10.1007/978-3-031-06555-2_14 (cit. on pp. 5, 19).

[32] O. Tüselmann, F. Müller, F. Wolf and G. A. Fink, "Recognition-Free Question Answering on Handwritten Document Collections", in *Frontiers in Handwriting Recognition*, Springer International Publishing, 2022, pp. 259–273. DOI: 10.1007/978-3-031-21648-0_18 (cit. on pp. 5, 19).

[33] O. Tüselmann and G. A. Fink, "Exploring Semantic Word Representations for Recognition-Free NLP on Handwritten Document Images", in *International Conference on Document Analysis and Recognition*, Springer, 2023, pp. 85–100. DOI: 10.1007/978-3-031-41685-9_6 (cit. on pp. 5, 16, 19).

[34] O. Tüselmann and G. A. Fink, "Neural models for semantic analysis of handwritten document images", *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 27, no. 3, pp. 245–263, 2024. DOI: 10.1007/s10032-024-00477-8 (cit. on p. 5).

[35] S. Dey, A. Nicolaou, J. Lladós and U. Pal, "Evaluation of word spotting under improper segmentation scenario", *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 22, pp. 361–374, 2019. DOI: 10.1007/s10032-019-00338-9 (cit. on pp. 5, 41).

[36] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", Nov. 2019. DOI: 10.18653/v1/D19-1410 (cit. on pp. 7, 27, 28).

[37] K. Song, X. Tan, T. Qin, J. Lu and T.-Y. Liu, "MPNet: Masked and Permuted Pre-training for Language Understanding", in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020, pp. 16 857–16 867. DOI: arXiv:2004.09297 (cit. on p. 7).

[38] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang and M. Zhou, "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers", in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020, pp. 5776–5788. DOI: arXiv:2002.10957 (cit. on p. 7).

[39] M. Li et al., "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models", in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, 2023, pp. 13 094–13 102. DOI: 10.1609/aaai.v37i11.26538 (cit. on pp. 7, 34).

[40] A. Graves, S. Fernández, F. Gomez and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks", in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376. DOI: 10.1145/1143844.1143891 (cit. on pp. 7, 26).

[41] V. Lavrenko, T. M. Rath and R. Manmatha, "Holistic word recognition for handwritten historical documents", in *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries*, 2004, pp. 278–287. DOI: 10.1109/DIAL.2004.1263256 (cit. on pp. 7, 30).

[42] U.-V. Marti and H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition", *International Journal on Document Analysis and Recognition*, vol. 5, no. 1, pp. 39–46, Nov. 2002. DOI: 10.1007/s100320200071 (cit. on pp. 7, 30).

[43] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms", *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979. DOI: 10.1109/TSMC.1979.4310076 (cit. on p. 9).

[44] S. Sudholt, L. Rothacker and G. A. Fink, "Query-by-Online Word Spotting Revisited: Using CNNs for Cross-Domain Retrieval", in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, vol. 1, 2017, pp. 481–486. DOI: 10.1109/ICDAR.2017.85 (cit. on p. 11).

[45] M. Mathew, D. Gomez Lluisand Karatzas and C. V. Jawahar, "Asking questions on handwritten document collections", *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 24, no. 3, pp. 235–249, Sep. 2021. DOI: 10.1007/s10032-021-00383-3 (cit. on pp. 13, 19).

[46] J. Rohlicek, W. Russell, S. Roukos and H. Gish, "Continuous hidden Markov modeling for speaker-independent word spotting", in *International Conference on Acoustics, Speech, and Signal Processing,*, vol. 1, 1989, pp. 627–630. DOI: 10.1109/ICASSP.1989.266505 (cit. on p. 14).

[47] F. Chen, L. Wilcox and D. Bloomberg, "Word spotting in scanned images using hidden Markov models", in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, 1993, pp. 1–4. DOI: 10.1109/ICASSP.1993.319732 (cit. on p. 14).

[48] R. Manmatha, C. Han and E. Riseman, "Word spotting: A new approach to indexing handwriting", in *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1996, pp. 631–637. DOI: 10.1109/CVPR.1996.517139 (cit. on p. 14).

[49] Y. Leydier, F. L. Bourgeois and H. Emptoz, "Text search for medieval manuscript images", *Pattern Recognition*, vol. 40, no. 12, pp. 3552–3567, 2007. DOI: 10.1016/j.patcog.2007.04.024 (cit. on p. 15).

[50] Y. Leydier, A. Ouji, F. LeBourgeois and H. Emptoz, "Towards an Omnilingual Word Retrieval System for Ancient Manuscripts", *Pattern Recognition*, vol. 42, no. 9, pp. 2089–2105, 2009. DOI: 10.1016/j.patcog.2009.01.026 (cit. on p. 15).

[51] X. Zhang and C. Tan, "Segmentation-Free Keyword Spotting for Handwritten Documents Based on Heat Kernel Signature", in *Proceedings of the* 12^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, 2013, pp. 827–831. DOI: 10.1109/ICDAR.2013.169 (cit. on p. 15).

[52] X. Zhang and C. L. Tan, "Handwritten word image matching based on Heat Kernel Signature", *Pattern Recognition*, vol. 48, no. 11, pp. 3346–3356, 2015. DOI: 10.1016/j.patcog.2014.10.028 (cit. on p. 15).

[53] B. Gatos and I. Pratikakis, "Segmentation-free Word Spotting in Historical Printed Documents", in *Proceedings of the* 10^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, 2009, pp. 271–275. DOI: 10.1109/ICDAR.2009.236 (cit. on p. 15).

[54] P. Riba, J. Llados and A. Fornes, "Handwritten Word Spotting by Inexact Matching of Grapheme Graphs", in *Proceedings of the* 13^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 781–785. DOI: 10.1109/ICDAR.2015.7333868 (cit. on p. 15).

[55] K. Zagoris, I. Pratikakis and B. Gatos, "Unsupervised Word Spotting in Historical Handwritten Document Images Using Document-Oriented Local Features", *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 4032–4041, 2017. DOI: 10.1109/TIP.2017.2700721 (cit. on p. 15).

[56] S. K. Ghosh and E. Valveny, "R-PHOC: Segmentation-Free Word Spotting Using CNN", in *Proceedings of the* 14^{th} *IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 801–806. DOI: 10.1109/ICDAR.2017.136 (cit. on p. 16).

[57] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90 (cit. on pp. 16, 24, 26).

[58] L. Rothacker, S. Sudholt, E. Rusakov, M. Kasperidus and G. A. Fink, "Word Hypotheses for Segmentation-Free Word Spotting in Historic Document Images", in *Proceedings of the* 14^{th} *IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 1174–1179. DOI: 10.1109/ICDAR.2017.194 (cit. on p. 16).

[59] A. Gordo, J. Almazán, N. Murray and F. Perronin, "LEWIS: Latent Embeddings for Word Images and Their Semantics", in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1242–1250. DOI: 10.1109/ICCV.2015.147 (cit. on p. 18).

[60] G. A. Miller, "WordNet: a lexical database for English", *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995. DOI: 10.1145/219717.219748 (cit. on p. 18).

[61] M. Sahlgren, "The distributional hypothesis", *Italian Journal of linguistics*, vol. 20, pp. 33–53, 2008 (cit. on p. 18).

[62] C. Adak, B. B. Chaudhuri, C.-T. Lin and M. Blumenstein, "Detecting Named Entities in Unstructured Bengali Manuscript Images", in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, pp. 196–201. DOI: 10.1109/ICDAR.2019.00040 (cit. on p. 19).

[63] P. Krishnan and C. Jawahar, "Generating synthetic data for text recognition", 2016. DOI: 10.48550/arXiv.1608.04224 (cit. on p. 19).

[64] B. He, "Rocchio's Formula", in *Encyclopedia of Database Systems*. Springer US, 2009, pp. 2447–2447. DOI: 10.1007/978-0-387-39940-9_932 (cit. on p. 20).

[65] A. Bhardwaj, D. Jose and V. Govindaraju, "Script Independent Word Spotting in Multilingual Documents", in *Proceedings of the* 2^{nd} *Workshop on Cross Lingual Information Access (CLIA)*, 2008, pp. 48–54. [Online]. Available: https://aclanthology.org/I08-6007/ (cit. on p. 20).

[66] H. Cao, V. Govindaraju and A. Bhardwaj, "Unconstrained handwritten document retrieval", *International Journal on Document Analysis and Recognition*, vol. 14, no. 2, pp. 145–157, 2011. DOI: 10.1007/s10032-010-0139-z (cit. on p. 20).

[67] M. Rusiñol and J. Llados, "Boosting the handwritten word spotting experience by including the user in the loop", *Pattern Recognition*, vol. 47, no. 3, pp. 1063–1072, 2014. DOI: 10.1016/j.patcog.2013.07.008 (cit. on p. 20).

[68] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis and S. Perantonis, "Keyword-guided word spotting in historical printed documents using synthetic data and user feedback", *International Journal on Document Analysis and Recognition*, vol. 9, no. 2-4, pp. 167–177, 2007. DOI: 10.1007/s10032-007-0042-4 (cit. on p. 20).

[69] A. L. Kesidis, E. Galiotou, B. Gatos and I. Pratikakis, "A word spotting framework for historical machine-printed documents", *International Journal on Document Analysis and Recognition*, vol. 14, no. 2, pp. 131–144, 2011. DOI: 10.1007/s10032-010-0134-4 (cit. on p. 20).

[70] F. Wolf, P. Oberdiek and G. Fink, "Exploring Confidence Measures for Word Spotting in Heterogeneous Datasets", pp. 583–588, 2019. DOI: 10.1109/ICDAR.2019.00099 (cit. on p. 20).

[71] R. Shekhar and C. Jawahar, "Word Image Retrieval Using Bag of Visual Words", in *Proceedings of the* 10^{th} *IAPR International Workshop on Document Analysis Systems (DAS)*, 2012, pp. 297–301. DOI: 10.1109/DAS.2012.96 (cit. on p. 20).

[72] S. Ghosh and E. Valveny, "A Sliding Window Framework for Word Spotting Based on Word Attributes", in *Proceedings of the* 7^{th} *Iberian Conference on Pattern Recognition and Image Analysis (PRAI)*, 2015, pp. 652–661. DOI: 10.1007/978-3-319-19390-8_73 (cit. on p. 20).

[73] E. Vats, A. Hast and A. Fornés, "Training-Free and Segmentation-Free Word Spotting using Feature Matching and Query Expansion", in *Proceedings of the* 15^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, 2019, pp. 1294–1299. DOI: 10.1109/ICDAR.2019.00209 (cit. on p. 20).

[74] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection", in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017, pp. 936–944. DOI: 10.1109/CVPR.2017.106 (cit. on p. 24).

[75] W. Wang et al., "Shape Robust Text Detection with Progressive Scale Expansion Network", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9336–9345. DOI: 10.1109/CVPR.2019.00956 (cit. on p. 24).

[76] F. Milletari, N. Navab and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation", in *2016 fourth international conference on 3D vision (3DV)*, Ieee, 2016, pp. 565–571. DOI: 10.1109/3DV.2016.79 (cit. on p. 24).

[77] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye and D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression", in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 12 993–13 000. DOI: 10.1609/aaai.v34i07.6999 (cit. on p. 24).

[78] J. Tian, B. Yan, J. Yu, C. Weng et al., "Bayes Risk CTC: Controllable CTC Alignment in Sequence-to-Sequence Tasks", in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. DOI: 10.48550/arXiv.2210.07499 (cit. on p. 26).

[79] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale", 2020. DOI: 10.48550/arXiv.2010.11929 (cit. on p. 28).

[80] Y. Wu et al., "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation", 2016. DOI: 10.48550/arXiv.1609.08144 (cit. on p. 28).

[81] G. Washington, *George Washington Papers, Series 2, Letterbooks 1754-1799: Letterbook 1, Aug. 11, 1754 - Dec. 25, 1755*, pp. 270–279, 300–309, Last accessed 30 April 2025. [Online]. Available: https://www.loc.gov/item/mgw2.001 (cit. on p. 30).

[82] I. Loshchilov and F. Hutter, *Decoupled Weight Decay Regularization*, 2019. DOI: 10.48550/arXiv.1711.05101 (cit. on p. 34).

[83] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio and L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation", in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Aug. 2017, pp. 1–14. DOI: 10.18653/v1/S17-2001 (cit. on p. 35).

[84] G. Louloudis, A. Kesidis and B. Gatos, "Efficient Word Retrieval Using a Multiple Ranking Combination Scheme", in *Proceedings of the* 10^{th} *IAPR International Workshop on Document Analysis Systems (DAS)*, 2012, pp. 379–383. DOI: 10.1109/DAS.2012.34 (cit. on p. 49).

[85] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision", in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, vol. 139, 2021, pp. 8748–8763. DOI: 10.48550/arXiv.2103.00020 (cit. on pp. 49, 51).

[86] J. Li, D. Li, S. Savarese and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models", in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, vol. 202, 2023, pp. 19 730–19 742. DOI: 10.48550/arXiv.2301.12597 (cit. on pp. 49, 51).

Short Biography

Born in Ioannina in 1995, Stergios Papazis received his B.Sc. and M.Sc. from the Department of Mathematics at the University of Ioannina in 2017 and 2020, respectively.

A self-taught programmer with a strong interest in computer science and software development, he completed an M.Sc. in Data and Computer Systems Engineering at the Department of Computer Science and Engineering in 2025, specializing in Data Science and Engineering.

His primary academic interests lie in Machine Learning, Computer Vision, and Mathematical Optimization.