**Università degli Studi di Salerno**

.DIEM

**Dipartimento di Ingegneria dell'Informazione ed Elettrica e Matematica Applicata**

Dottorato di Ricerca in Ingegneria dell'Informazione
Ciclo 35

TESI DI DOTTORATO / PH.D. THESIS

ABSTRACT

# N-gram Retrieval for Word Spotting in Historical Handwritten Collections

**GIUSEPPE DE GREGORIO**

SUPERVISOR: **PROF. ANGELO MARCELLI**

PHD PROGRAM DIRECTOR: **PROF. PASQUALE CHIACCHIO**

Anno 2023

## Abstract

# N-gram Retrieval for Word Spotting in Historical Handwritten Collections

Collections of handwritten documents of historical interest are often small, ranging from a few dozen to a few hundred pages, but they may have features typical of the collection itself which make them interesting for scholarly groups. Word retrieval can be complicated, as handwriting recognition techniques applied to images of such text documents can produce unsatisfactory results. To circumvent this, KeyWord Spotting (KWS) techniques promise to retrieve words without having to perform the recognition explicitly. KWS systems often require the construction of a reference dictionary consisting of words that the system can search for in the document. To do this, a dictionary must be created, and usually, a small portion of the collection is transcribed by hand. This limits the search to words from the dictionary (InVoc) and introduces the problem of OOV (Out Of Vocabulary) words, which cannot be searched for. Intuitively, increasing the cardinality of the reference dictionaries by manually transcribing new examples of words may seem an immediate way of limiting the OOV problem.

As we can imagine, manually labelling pages is an expensive process that can take a lot of time. For small collections, the time required to transcribe and label even a few pages can prove to be a non-negligible obstacle, calling into question the usefulness of automated word retrieval systems. In this thesis, we will focus on a KWS system that can adapt to the lack of data. First, we show analytically, through the definition of a mathematical model, how the different components of the system affect the time of use of the KWS system by estimating the time gain that the system brings to the transcription of a small collection of handwritten historical documents.

After highlighting the importance of speeding up the manual annotation process, we then propose a semiautomatic method for image annotation. In particular, we present a learning-free end-to-end approach that includes a line segmentation algorithm and an algorithm for aligning transcripts to images with handwritten text. The former can extract lines of text with a curved baseline, while the latter allows us to easily get their transcript.

Finally, we propose a KWS system for word spotting that bases the search on recognizing sequences of characters (N-grams) rather than directly trying to find whole words. Studies on motor behaviour have shown that writing is the result of very fast and precise motor actions that can be automated. In the learning phase, an individual tends to develop motor programs associated with simple actions that are characterized by a high frequency of execution. It is plausible to assume that motor programs for writing develop in relation to sequences of a few characters. This would

mean that each time a subject writes an N-gram to which a motor program is associated, he or she produces an ink trace that is always compatible with and similar to all others. The repeated similarity in the execution of the same movements for each N-gram could make the N-grams recognizable, making them ideal candidates for handwritten cursive recognition.

The results of the experiments show how a KWS system can effectively reduce the time to the document collection transcription process. Moreover, it is shown that the process of labelling data and creating reference dictionaries is indeed an extremely costly operation and that methods that enable the acceleration of such processes are crucial. The experiments with the proposed KWS system have shown that focusing the search on the N-gram space enables the retrieval of InVoc and OOV words equally well, showing similar retrieval rates for both groups of words.

## Abstract - Italiano

# N-gram Retrieval per il Word Spotting in Raccolte di Manoscritti Storici

Le raccolte di documenti manoscritti di interesse storico sono spesso di piccole dimensioni, da poche decine a poche centinaia di pagine, ma possono presentare caratteristiche tipiche della raccolta stessa che le rendono interessanti per gruppi di studiosi. Il recupero delle parole può essere complicato, poiché le tecniche di riconoscimento della grafia applicate alle immagini di tali documenti possono produrre risultati insoddisfacenti. Per aggirare questo problema, le tecniche di KeyWord Spotting (KWS) promettono di recuperare le parole senza dover eseguire esplicitamente il riconoscimento. I sistemi KWS richiedono spesso la costruzione di un dizionario di riferimento composto da parole che il sistema può ricercare nel documento. Per fare ciò è necessario creare un dizionario e, di solito, una piccola parte della raccolta viene trascritta a mano. Questo limita la ricerca alle parole del dizionario (InVoc) e introduce il problema delle parole OOV (Out Of Vocabulary), che non possono essere recuperate. Intuitivamente, aumentare la cardinalità dei dizionari di riferimento trascrivendo manualmente nuovi esempi di parole può sembrare un modo immediato per limitare il problema dell'OOV.
Come possiamo immaginare, etichettare manualmente le pagine è un processo costoso che può richiedere molto tempo. Per le piccole raccolte, il tempo necessario per trascrivere ed etichettare anche poche pagine può rivelarsi un ostacolo non trascurabile, mettendo in discussione l'utilità dei sistemi automatizzati di recupero delle parole. In questa tesi, ci concentreremo su un sistema KWS in grado di adattarsi alla mancanza di dati. In primo luogo, mostriamo analiticamente, attraverso la definizione di un modello matematico, come le diverse componenti del sistema influiscono sul tempo di utilizzo del sistema KWS stimando il guadagno di tempo che il sistema apporta alla trascrizione di una piccola raccolta di documenti storici manoscritti.
Dopo aver evidenziato l'importanza di velocizzare il processo di annotazione manuale, proponiamo un metodo semiautomatico per l'annotazione delle immagini. In particolare, presentiamo un approccio end-to-end senza apprendimento che include un algoritmo di segmentazione della linea e un algoritmo per allineare le trascrizioni alle immagini con il testo scritto a mano. Il primo può estrarre righe di testo con una linea di base curva, mentre il secondo ci consente di ottenere facilmente la loro trascrizione.

Infine, proponiamo un sistema KWS per il word spotting che basa la ricerca sul riconoscimento di sequenze di caratteri (N-grammi) piuttosto che sul tentativo diretto di trovare intere parole. Gli studi sul comportamento motorio hanno dimostrato che la scrittura è il risultato di azioni motorie molto veloci e precise che possono essere automatizzate. Nella fase di apprendimento, un individuo tende a sviluppare programmi motori associati ad azioni semplici caratterizzate da un'elevata frequenza di esecuzione. È plausibile ipotizzare che i programmi motori per la scrittura si sviluppino in relazione a sequenze di pochi caratteri. Ciò significherebbe che ogni volta che un soggetto scrive un N-gramma a cui è associato un programma motorio, produce una traccia di inchiostro sempre compatibile e simile a tutti gli altri. La ripetuta somiglianza nell'esecuzione degli stessi movimenti per ogni N-gramma potrebbe rendere riconoscibili gli N-grammi, rendendoli candidati ideali per il riconoscimento del corsivo scritto a mano.

I risultati degli esperimenti mostrano come un sistema KWS possa ridurre efficacemente i tempi del processo di trascrizione della raccolta dei documenti. Inoltre, è dimostrato che il processo di etichettatura dei dati e di creazione dei dizionari di riferimento è effettivamente un'operazione estremamente costosa e che i metodi che consentono l'accelerazione di tali processi sono cruciali. Gli esperimenti con il sistema KWS proposto hanno dimostrato che concentrare la ricerca sullo spazio N-grammi consente il recupero di parole InVoc e OOV altrettanto bene, mostrando tassi di recupero simili per entrambi i gruppi di parole.