

الجمهورية الشعبية الديمقراطية الجزائرية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
المدرسة العليا للإعلام الآلي ٨ ماي ٥٤٩١ سيدي بلعباس
Higher School of Computer Science
8 Mai 1945 - Sidi Bel Abbès



Master's Thesis

To obtain the diploma of Master's Degree

Field of Study: Computer Science

Specialization: Artificial Intelligence & Data Science (IASD)

Theme

**Historical Arabic Handwritten Text Recognition and
Document Restoration: A Survey of the State of the
Art and Open Challenges**

Presented by
Zellagui Mohamed Diao Eddine
Guechaoui Mohamed

Defended on: **July, 2026**
In front of the jury composed of

Abstract

Historical Arabic manuscripts constitute one of the largest and least computationally accessible repositories of human knowledge, spanning more than a millennium of continuous production across medicine, philosophy, law, and the sciences. Despite extensive digitization efforts, the automatic conversion of manuscript images into searchable text remains an open problem, held back by the scarcity of annotated training data, the cursive and contextual nature of the Arabic script, the diversity of calligraphic traditions, and the physical degradation of aged documents. This thesis surveys the state of the art that bears on this problem. We begin with the broad field of document image analysis and the general evolution of handwritten text recognition, then narrow to text line segmentation, to document image restoration, and to the datasets and benchmarks that drive progress, before focusing on the specific challenges of historical Arabic manuscripts. We review the metrics used to evaluate recognition, segmentation, and restoration, and we conclude by synthesising the literature into a set of open problems: the persistent data bottleneck, the absence of coverage for several Arabic calligraphic traditions, the lack of a standardised line-level restoration benchmark, and the gap between character-atomic restoration frameworks and connected scripts. These open problems motivate the engineering work presented in the companion report, which designs and builds a dataset and a restoration framework in response. As a work of synthesis, the contribution of this thesis is itself twofold: it consolidates a literature fragmented across the recognition, segmentation, and restoration communities into a single coherent account oriented towards the historical Arabic setting, and it distils that account into a precise, evidence-based statement of the open problems that a subsequent engineering effort must address.

Keywords: historical document analysis, Arabic handwritten text recognition, text-line segmentation, document image restoration, benchmark datasets, deep learning.

Résumé

Les manuscrits arabes historiques constituent l'un des réservoirs de connaissances humaines les plus vastes et les moins accessibles sur le plan informatique, couvrant plus d'un millénaire de production continue dans les domaines de la médecine, de la philosophie, du droit et des sciences. Malgré des efforts de numérisation considérables, la conversion automatique d'images de manuscrits en texte interrogeable demeure un problème ouvert. Elle est freinée par la rareté des données d'entraînement annotées, la nature cursive et contextuelle de l'écriture arabe, la diversité des traditions calligraphiques et la dégradation physique des documents anciens. Ce rapport passe en revue l'état de l'art lié à ce problème. Nous commençons par le vaste domaine de l'analyse d'images de documents et l'évolution générale de la reconnaissance de textes manuscrits, avant de nous recentrer sur la segmentation des lignes de texte, la restauration d'images de documents, ainsi que sur les ensembles de données et les benchmarks qui stimulent les progrès, pour enfin nous concentrer sur les défis spécifiques aux manuscrits arabes historiques. Nous examinons les métriques utilisées pour évaluer la reconnaissance, la segmentation et la restauration, puis nous concluons en synthétisant la littérature sous forme d'un ensemble de problèmes ouverts : le goulot d'étranglement persistant des données, l'absence de couverture pour plusieurs traditions calligraphiques arabes, le manque d'un benchmark de restauration standardisé au niveau de la ligne, et l'écart entre les cadres de restauration au niveau du caractère (approche atomique) et les écritures cursives. Ces problèmes ouverts motivent le travail d'ingénierie présenté dans le rapport d'accompagnement, qui conçoit et développe un ensemble de données et un cadre de restauration en réponse à ces défis.

Mots-clés : analyse de documents historiques, reconnaissance de textes manuscrits arabes, segmentation des lignes de texte, restauration d'images de documents, ensembles de données de référence, apprentissage profond.