INSA Rennes recruits
a Research Engineer/Post-doctoral Researcher
in computer science

Analysis systems for serial sources in collections of historical image documents

Conditions

Establishment: INSA de Rennes, IRISA laboratory, France

Service: Intuidoc - research team of IRISA

Up to 36-months contract to take as soon as possible

IRISA - Intuidoc

IRISA is a joint research center for Informatics, including Robotics and Image and Signal Processing. 800 people, 40 teams, explore the world of digital sciences to find applications in healthcare, ecology-environment, cyber-security, transportation, multimedia, and industry... INSA Rennes is one of the 8 trustees of IRISA.

The Intuidoc team (https://www.irisa.fr/intuidoc) conducts researches on the topic of document image recognition. Since many years, the team proposes a system, called DMOS-PI method, for document structure analysis of documents. This DMOS-PI method is used for document recognition, or field extraction in archive documents, handwritten contents damaged documents (musical scores, archives, newspapers, letters, electronic schema, …).

EURHISFIRM project

EURHISFIRM European project aims at developing a research infrastructure to connect, collect, collate, align, and share reliable long-run company-level data for Europe to enable researchers, policymakers and other stakeholders to analyze, develop, and evaluate effective strategies to promote investment and economic growth. To achieve this goal, EURHISFIRM develops innovative tools to spark a “Big data” revolution in the historical social sciences and to open access to cultural heritage.

EURHISFIRM is a project funded by the European Commission within the Infrastructure Development Program of Horizon 2020. The first phase of the Infrastructure Development Program lasts for three years. It aims at developing an in-depth design study of the Research Infrastructure. After this phase, Development and Consolidation Phases follow if further applications will be successful. EURHISFIRM brings together eleven research institutions in economics, history, information technologies and data science from seven European countries.

Position to be filled

Position: Post-doctoral fellow / Research Engineer
Time commitment: Full-time
Duration of the contract: up to 36 months, starting as soon a possible
Supervisors: Bertrand Coüasnon and Aurélie Lemaitre
Indicative salary: Up to €36 000 gross annual salary (according to experience), with social security benefits
Location: IRISA – Rennes, France

INSA RENNES
20, Avenue des Buttes de Cöesmes
CS 70839 - 35 708 Rennes Cedex 7
Tél.+ 33 (0)2 23 23 82 00 - Fax + 33 (0)2 23 23 83 96
www.insa-rennes.fr

Missions

The post-doctoral fellow / research engineer will be working on two tasks of EURHISFIRM workflow: the architecture of an adaptable system for document recognition, and the implementation of a generic structure layout extraction module.
The scientific challenge will be to extract information from various printed serial sources. Due to the large variety of those documents, a flexible and easy-to-adapt document recognition system is designed. For that purpose, the system will be based on a modelling of knowledge not only at the page level but also at the collection level in interaction with experts of the historical sources. Thus, redundancies between pages will be used to make the system more reliable and reduce manual corrections while obtaining a high recognition quality.
The system will we based on the DMOS-PI method which gives a framework for the analysis of collections of documents. It enables to share information from the collection between the pages, thanks to an iterative mechanism of analysis. This mechanism also makes it possible to integrate an asynchronous interaction between automatic analysis and human operators in order to limit the time of interaction by avoiding mutual waiting.
This modelling of the global analysis must be able to adapt to very different kinds of documents: from very structured documents, like stock exchange lists with redundancy and strong consistency between sequences of data, up to less structured documents, like yearbooks even if, also for them, the sequence from one year to another is important for improving the recognition quality.
The implementation of a generic structure extraction module will be based on the DMOS-PI method. It uses a grammatical language, EPF (Enhanced Position Formalism), to describe a general page layout, with perceptive vision mechanisms, and an iterative analysis. The system will also combine structural method with Deep Learning. For new collections, an adapted description of the document layout will be developed. This has to be done on a large range of structure levels: from very structured pages like table structures from stock exchange lists, up to a paragraph-oriented structures from yearbooks.

Main Skills

PhD, Master degree or Engineering degree in computer science
Experience in document recognition, statistical analysis or deep learning.
Fluent English
Skills in grammars and languages and/or logical programming are nice-to-have.

For further information, please contact Bertrand Coüasnon (bertrand.couasnon@irisa.fr) and Aurélie Lemaitre (aurelie.lemaitre@irisa.fr). Applicants should send a curriculum-vitae with a list of publications and the names and email addresses of up to three references.

INSA RENNES
20, Avenue des Buttes de Cöesmes
CS 70839 - 35 708 Rennes Cedex 7
Tél.+ 33 (0)2 23 23 82 00 - Fax + 33 (0)2 23 23 83 96
www.insa-rennes.fr