Image processing and semantic technologies in digital humanities: The NOTAE experience

Eleonora Bernasconi, Tiziana Catarci, Francesco Leotta, Antonella Ghignoli, Massimo Mecella, Silvestro Veneruso, Zahra Ziran
Sapienza Università di Roma, Italy

The ERC project NOTAE (for details see http://www.notae-project.eu/further-info/project), means also the opportunity to test the application of techniques from different areas of computer science, in order to address some potential issues that could be generated by investigating the documentary records of all possible kinds (legal contracts, petitions, official letters, private letters, lists, authentics from relics, etc.) preserved in original or in contemporary copies on papyrus, wooden tablet, slate, parchment, from 5th to 8th centuries AD, in particular as regards graphic symbols of all kinds traced out by individuals (professional scribes, literates, illiterates) within the written records in question, and whose origin, meanings and functions are object of historical investigation.
In particular, a meta-digital library of specific categories of documents has been proposed and is currently under development, allowing both curators/experts to deal with meta-data management and editing, and an advanced web-based visualization tool has been proposed which allows users to see the ancient documents that are geographically distributed in different locations.
The project outcome is not only a meta-data repository though. In particular, it integrates data with other authoritative information sources in the form of a knowledge graph. In addition, the project applies image processing techniques, made available to curators, for:
1) document preprocessing including removing duplicated images and removing image noises. Scraped Images, which usually come from various sources with different features (e.g., size, encoding/decoding compression algorithms), can be duplicated. In order to remove duplicated images, firstly a fixed-length hash for each image is calculated using the difference hash (dHash) algorithm, based on the visual features of images; secondly, a relative Hamming distance is calculated for each pair of image hash. If the calculated distance for a pair of images is less than a specific threshold, they are considered as duplicated, and the one with a higher resolution is kept. Using image binarization, given an image, we extract the foreground (e.g., handwritten, painting) from the image, and then, remove the noise in the background.
2) image enhancement. Using segmentation-based image enhancement algorithms (e.g., adaptive equalization, watershed), the quality of degrading images is improved when allowing the user to select a desired image area.
3) graphic symbol identification and classification. We have developed an interactive symbol identification system that investigates the presence and the position of possible graphic symbols in documents using two training approaches: a) feature discrimination, to train a model when it learns a set of introduced meta templates (e.g., stroke, simple curve), each is associated with a graphic symbol, in an unsupervised fashion and b) match, considering the teacher-student method, in which the model is trained, during an interactive training session, when a human expert teaches the model until its performance reaches the asymptotic limit of the model. In fact,

during a training session, the system asks the expert to provide her opinion by comparing the results of symbol detection in a document.

Eleonora Bernasconi is a PhD candidate in Engineering in Computer Science at Sapienza Università di Roma. Her research focuses on digital libraries and semantic technologies and advanced visual interfaces applied to digital humanities

Tiziana Catarci is full professor in Engineering in Computer Science at Sapienza Università di Roma. She was vice-rector for ICT in 2010 - 2014, currently she is the Director of the Dipartimento di Ingegneria informatica automatica e gestionale Antonio Ruberti and seat in the academic Senate. Her research interests spans from data management to advanced user interfaces, HCI, ethics in AI

Francesco Leotta is assistant professor in Engineering in Computer Science at Sapienza Università di Roma His research concerns algorithmic, methodological, experimental and practical aspects of different areas of information systems, including ubiquitous computing, human-computer (and robot) interaction and digital humanities. Such topics are challenged in the application domains of smart spaces, smart manufacturing and cultural heritage.

Antonella Ghignoli is full professor in Paleography at Sapienza Università di Roma, and has been and Senior Research Fellow at Sapienza School for Advanced Studies - SSAS (2018-2021). She is currently the Principal Investigator of the ERC funded project NOTAE (ERC Advanced Grant 2017 - GA nr. 786572), "NOT A writtEn word but graphic symbols. NOTAE: An evidence-based reconstruction of another written world in pragmatic literacy from Late Antiquity to early medieval Europe" (2018-2023). She is scientific co-coordinator of the research program "DiploMA. Diplomatics in Mediterranean Area" in partnership with the École Française de Rome (2022-2026).

Massimo Mecella is full professor in Engineering in Computer Science at Sapienza Università di Roma. His research interests are in software engineering, software architectures, information systems engineering, BPM - Business Process Management, smart environments. He has a vast experience in EU and National grants in all roles. He regularly serves as expert reviewers for EU bodies and national ones.

Silvestro Veneruso is a PhD student in Engineering in Computer Science at Sapienza Università di Roma. His research interests are on AR/VR and smart environments. In the context of the NOTAE project, has is also the leader of the Web digital archive.

Zahra Ziran earned a PhD in Engineering in Computer Science (2020) on Deep learning-based object detection models applied to document images at the University of Florence. She joined Sapienza as a research fellow in the context of NOTAE for investigating historical graphic symbols. Her research interests are document image analysis, image processing, AI, and machine learning.