New computer techniques to analyze historic Hebrew, Arabic documents under development

August 14 2009

Researchers at Ben-Gurion University of the Negev (BGU) will combine the scientific and scholarly expertise of their humanities and computer science experts in a new project to analyze degraded Hebrew documents.

The effort to develop new computer algorithms combines BGU's scientific expertise in computer vision, computer graphics, image processing and computational geometry with the scholarly expertise of historians and liturgy scholars to provide valuable answers regarding Jewish liturgical texts and Arabic historical texts that advance scholarship in these fields.

The technical goal of the research is to develop new state of the art algorithms for analyzing text and combine them into an easy to operate, open source system of tools to aid historical document research throughout the world.

Experiments are being conducted on degraded documents from sources such as the Cairo Geniza, copies of which are located at the national liturgy project at BGU, the El-Aqsa manuscript library in Jerusalem and the Al-Azar manuscript library in Cairo. Most fragments that have been discovered at the Geniza are now in libraries at Cambridge and Oxford universities, the Jewish Theological Seminary in New York, The British Library and in Israel and Paris.

Until now the documents have not been researched systematically. Prof. Uri Ehrlich of the Goldstein-Goren Department of Jewish Thought is the head of the Prayer Research Project at BGU. He explains that, "There was one book that was originally used as a Hebrew prayer book from the 12th century, but had been scratched off, and the parchment used to write an Arabic text (called a palimpsest). Our aim was to read the first book and not the second book. So we needed to find out how the Arab book could disappear and would leave only the Hebrew letters of the original book. This is why the computer sciences and humanities departments at BGU decided to collaborate."

"To solve the problem, we created an algorithm to cover the text in a dark grey color, which then highlights lighter colored pixels as background space and identifies the darker pixels as outlining the original Hebrew lettering," said Prof. Klara Kedem of the Department of Computer Sciences and one of the system's creators.

Many of the new methods will apply to other languages as well, including binarization of highly degraded documents (converting up to 256 grey colors to black and white to facilitate digitization), segmentation of skewed and curved lines and word spotting in both curved and highly degraded documents. Other algorithms will be more language specific, such as paleographic analysis of Hebrew and Arabic historical documents that will include automatic indexing of document collections, determining authorship, location and date of the documents.

The research is being funded by the Israel Science Foundation (ISF). Prof. Ehrlich and other BGU scholars in the humanities will be among those to evaluate the system to be built by Prof. Klara Kedem and Dr. Jihad El-Sana of the Department of Computer Sciences and Prof. Emeritus Tsiki Dinstein from Electrical Engineering.

The group is part of the emerging global effort to understand, manipulate and archive historical documents so that they are available to researchers in paleography, archaeology and historical research.

Source: American Associates, Ben-Gurion University of the Negev (news : web)

Citation: New computer techniques to analyze historic Hebrew, Arabic documents under development (2009, August 14) retrieved 30 April 2024 from https://phys.org/news/2009-08-techniques-historic-hebrew-arabic-documents.html