Social networks of the past: information extraction from historical demographic documents

Search centered at people is very important in historical research, including historical demography, people trajectories reconstruction and genealogical research. Queries about a person and his/her connections to other people allow to get a picture of a historical context: a person's life, an event, a location at some period of time. For this purpose, scholars use documents like birth, marriage, or census records.

We are conducting rearch on Document Image Recognition addressed to the extraction of information from demographic documents. In particular, we are working with the Barcelona Marriages Database (a source of handwritten marriage license records covering five centuries), and census documents from different places. From a technical point of view, the core methodology in our work in word spotting. Word spotting is the process of retrieving all instances of a queried keyword from a digital library of document images. We have proposed different word spotting approaches for historical manuscript retrieval.

We have also made contributions in context-aware word spotting. Usually word spotting is built based solely on the statistics of local terms. The use of correlative semantic labels between codewords adds more discriminability in the process. Three levels of context can be defined in a word spotting scenario. First, the joint occurrence of words in a given image segment. Second, the geometric context involving a language model regarding to the relative 1D or 2D position of objects. Third, the semantic context defined by the topic of the document. A number of document collections convey an underlying structure. We take advantage of the structure to boost the search of words, with a joint search of the query word and its context.

Josep Lladós

Computer Vision Center

Universitat Autònoma de Barcelona