How to best chain humans and machines together: From image identification to crowdsourcing to the social graph of European integration

Lars Wieneke and Marten Düring

With the increasing digitalization of contemporary historical sources, such as images, videos or sound recordings, new opportunities for research emerge. Nevertheless, automatic processes alone are often not reliable enough to extract high-level information, such as the identity of persons that are depicted in these resources. A fully manual validation of faces and identities on the other hand creates highly reliable information but leads at the same time to a bottleneck for the indexation of larger archives. In this paper we want to discuss how the histoGraph application overcomes these limitations through an adapted combination of human and machine computation.

Up until now the full potential of such combinations remains largely untapped due to the complexity and diverse demands of such projects: The implementation and integration of processing and recognition algorithms requires specialized know-how and users from the humanities are challenged with expressing requirements for unprecedented tasks and methods which haven't emerged yet while the final application should be useable for users who don't have a technical background. To overcome these issues histoGraph follows a decidedly user-centered approach that tries to fuse research in computer science, the design of human-computation tasks, data visualization, social engineering and the humanities in a coherent application.

The HistoGraph app overcomes the inherent complexity of hybrid human and machine computation in part through its integration in the CUbRIK framework. CUbRIK is an FP7-ICT funded 36-month long project that started in October 2011 and which focuses on building a flexible platform for multimedia search that combines human input and machine computation. In the proposed paper we discuss how CUbRIK succeeded in creating a workflow to which both humans and machines contribute in different stages, dwelling on the output of the other and preparing it for the next step. histoGraph's demo version is currently used to display a network of individuals pictured in photographs associated to the history of European integration. Based on the graph, users get an overview of who co-occurrs with whom and through which photographs two actors are connected. In the near future, we will expand the selection of primary sources, making the graph a much richer source for browsing primary sources related to the history of European integration.

To build the graph, a collection of 3.000 digitized photos had to be processed in a number of steps: 1) Identify faces in the photographs (machines, human clickworkers), 2) associate identities to these faces (machines, human experts), 3) create a network, 4) automatically enrich the network through other digitized sources, 5) visualize links and contextual information in a user-friendly environment. In the proposed paper we will discuss this workflow and point out the challenges from a technical perspective as well as in relation to the specific requirements of the humanities regarding source and information provenance and verification. The presentation will also try to critically reflect the approach and highlight several lessons learned for the future exploitation of large-scale multimedia archives.