Feature recognition for herbarium specimens (Herbar-Digital)

Karl-Heinz Steinke, Robert Dzido, Martin Gehrke, Klaus Prätel

Abstract

Our research project, “Rationalizing the virtualization of botanical document material and their usage by process optimization and automation (Herbar-Digital)” started on July 1, 2007 and will last until 2012. Its long-term aim is the digitization of the more than 3,5 million specimens in the Berlin Herbarium. The University of Applied Sciences and Arts in Hannover collaborates with the department of Biodiversity Informatics at the BGBM (Botanic Garden and Botanical Museum Berlin-Dahlem) headed by Walter Berendsohn. The part of Herbar-Digital here presented deals with the analysis of the generated high resolution images (10,400 lines x 7,500 pixel).

The image of a herbarium specimen may contain different kinds of objects, such as: mounted plant material; envelopes with additional parts of the plant; printed or handwritten labels with printed headlines (or without); annotation labels; a metric chart; a colour chart; stamps indicating ownership, specimen donation, accession numbers, or other events in the history of the specimen (e.g. digitization); specific markers (such as the red "Typus" label in the Berlin herbarium); barcode; and handwritten annotations directly on the sheet. Known, recurring stable objects can be found automatically by template matching methods (strong classifiers), which compute the similarity between the image and the template at each position in the whole image. The similarity is transformed to a brightness value. By choosing the maximum brightness, the location of the template in the image is resolved, and the object can be cut out of the image. More variable objects require more sophisticated methods, based on classifiers, i.e. entities that serve to classify an object into categories. One approach uses boosting-algorithms from face recognition applications. Boosting-algorithms utilize an iterative approach that combines the results of several weak classifiers to arrive at a decision to a problem instead of matching a single strong classifier. The idea suggests that it is easier to find some rules of thumb than relying on a general rule to resolve the problem.

More than 30% of the images contain handwriting. Because commercial optical character recognition (OCR) software is not able to interpret handwriting in old documents, we developed an algorithm to distinguish handwriting in complex surroundings such as those on a herbarium sheet, where it may be mixed with leaves and roots. With the help of an interactive method, single characters of handwriting may be extracted from the text. The characters can be described mathematically e.g. by Legendre polynomials and the coefficients are then stored for comparison in a database. The aim is to identify the writer, for example "Alexander von Humboldt".

The computer demonstration will show several approaches for image analysis. First, stable objects like stamps are identified and eliminated. Afterwards, areas containing writing are localized and analysed by self-developed methods or by Optical Character Recognition. An adaptive boost classifier can find unstable objects. Identification of coloured objects is possible by means of a colour space transformation. Eventually, the separation of the plant from the background is done. All the approaches have to be combined in order to extract the largest amount of information from the image.

This project is financed by the state of Lower Saxonia and the Volkswagen Foundation.

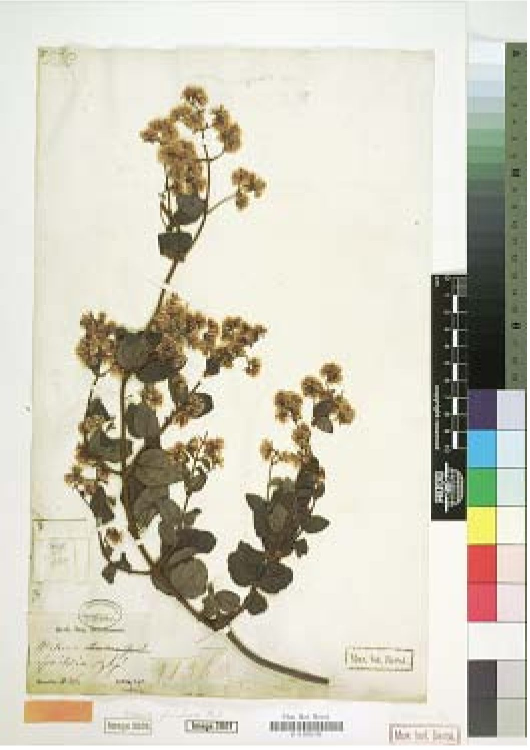

Fig. 1: Original image

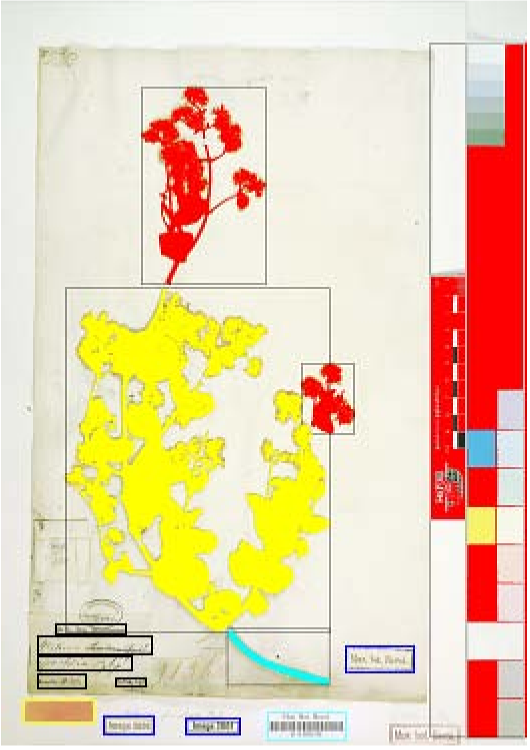

Fig. 2: Segmented image