Degraded document image enhancement

Gady Agam, Gideon Frieder, and Ophir Frieder

A system for processing hard-to-read but important materials has been applied to historical Holocaust documents.

Degraded documents are archived and preserved in large quantities worldwide. Electronic scanning is a common approach in handling such materials to facilitate public access to them. But the resulting images are often hard to read, have low contrast, and are corrupted by various artifacts. For example, the original document may be faded, washed out, damaged, crumpled, or otherwise difficult to discern (see Figure 1). Moreover, it may contain mixed handwriting in one or several languages, typed or printed text, or even pictures, tables, or diagrams. Machine-printed text may have been produced using different technologies with variable quality. Accordingly, readability and comprehensibility need to be increased.

Image enhancement normally focuses on minor deterioration in modern documents to improve optical character recognition. It often ignores difficult cases, such as those typical of historical and other highly degraded documents. Separating foreground and background can make an image more readable. A commercial-grade system that employs such separation for compression is the DjVu scanning technology, which is available through a Web interface.\(^1\)\(^-\)\(^3\) A simpler and more efficient approach for document image compression based on background-foreground separation is described by Simard and colleagues.\(^4\) Gatos’ team have carried out binarization of historical documents based on adaptive threshold segmentation and various pre- and postprocessing steps.\(^5\) (Binarization converts a color image to a bilevel image containing only black and white pixels.)

The enhancement system we are developing\(^6\) produces a document image that can be viewed in different ways using two interactive parameters that are simple and intuitive to interpret. The first parameter controls the decision threshold used in foreground segmentation, whereas the second controls the blending weight of the two channels. The decision threshold allows the user to increase or decrease the sensitivity of the foreground segmentation process. The blending factor provides control of the level of enhancement: from the original document image without any enhancements to the enhanced foreground displayed by itself. The application of these two adjustable thresholds is immediate once the document image has been processed. Moreover, parameter adjustment is not a mandatory step. It simply provides different views.

We employ parametric probabilistic models to perform foreground and background separation. The parameters of the models are set to maximize the likelihood of the observations. The well-known expectation maximization (EM) algorithm\(^7\) makes it possible to simplify the maximization of the complete data log likelihood by assuming a hidden feature that describes the unknown component in the mixture from which each observation was drawn.

Our system includes both color and edge features for two- and three-class segmentation. The EM algorithm is optimal as far as it goes. Yet the performance of this algorithm can be improved by considering local neighborhoods separately. This is due to the fact that degradation can vary across different areas of a document. The selection of a suitable window size can affect the segmentation results. If the window is too small, it might not have a sufficient number of pixels belonging to both the foreground and background classes. Conversely, if the window is too large, it might contain several different degradations and not perform ideally. To address this issue, we employ adaptive window selection (see Figure 2).

Continued on next page
We evaluated the performance of our system and compared it with those of known techniques using a subset of historical documents obtained from the Yad Vashem Holocaust memorial museum. The test collection contains 867 pages, written in several different languages, that were produced mainly by a typewriter. The documents contain numerous handwritten comments as well as pictures, logos, and signatures. To quantify the performance of the system, the degraded images were segmented using the different approaches, and the segmented results were compared with a known ground-truth image. Precision-recall graphs were then generated for each method by varying the decision thresholds individually.

In summary, we have developed a novel approach to enhancing deteriorated document images with multiple possible degradations. Our current efforts are focused on incorporating additional character features, and developing quantitative perceptual quality measures as well as techniques for selecting particular enhancement models in individual cases. We are also exploring techniques for exploiting internal coherence—i.e., using good-quality character images to correct poor-quality character images in the same document—to improve results. This project is related to a larger project of Complex Document Information Processing (CDIP) that is currently being developed at the Information Retrieval Lab at the Illinois Institute of Technology.9

The authors thank Otthein Herzog and the Image Processing Group at the University of Bremen for initial discussions and work on this problem.

Author Information

Gady Agam and Ophir Frieder
Department of Computer Science
Illinois Institute of Technology (IIT)
Chicago, IL

Gady Agam is an assistant professor in the Computer Science Department at the Illinois Institute of Technology and is the director of the Visual Computing Lab. He is a member of SPIE, the Institute of Electrical and Electronic Engineering, and ACM.

Ophir Frieder is the IIT Research Institute Professor of Computer Science and director of the Information Retrieval Laboratory at the Illinois Institute of Technology. He is a fellow of the AAAS, ACM, and IEEE.

References

8. The Diaries of Rabbi Dr. Avraham Abba Frieder. http://ir.iit.edu/collections/.

The authors thank Otthein Herzog and the Image Processing Group at the University of Bremen for initial discussions and work on this problem.