J4 Automated Reconstruction of Strip Shredded Documents

Patrick De Smet, PhD, Ghent University, Department TELIN/TW07, Sint-Pietersnieuwstraat 41, Ghent, B-9000, Belgium*

The goal of this presentation is to present an overview on recent research regarding several semi-automatic reconstruction methods for recomposing strip-shredded documents.

This presentation will impact the forensic community and/or humanity by providing more detailed and objective information about automated reconstruction of strip-shredded documents, creating an incentive for further discussion and research

Up until recently, the forensic or investigative reconstruction of shredded documents has always been dismissed as an important but unsolvable problem. Manual reassembly of the physical shreds could always be considered, but for large amounts of shreds this problem would quickly become an intangible task, requiring vast amounts of time and/or personnel. Automated reconstruction methods have recently been studied, but little objective information about these techniques and the results that can be obtained with them, has been made available.

In this research several image processing and computer vision techniques that can be used to reconstruct strip-shredded documents contained within a digital image database are reviewed, proposed, and discussed.

As already indicated above, an important problem for reconstructing a digital database of document shreds is given by the computational burden of the shred matching process that has a complexity order of O(N^2) (with N the number of shreds). Hence, in order to reduce this complexity researchers have investigated the use of several feature based classification and matching steps.

The first processing step was designed to detect and eliminate document shreds that do not contain any valuable feature information; such shreds typically are the result of shredding (partially) blank pages or page margin areas. Additionally, each shred side is examined separately for determining if any reliable cross-shred features can be found; if not, that side of the shred does not need to be considered for further matching.

Next, a single binary feature vector for each shred is determined. These feature vectors contain a binary flag for each horizontal line of pixels on the surface of the shred image, indicating if any foreground colors can be found on that line of pixels. Thus, the feature vector encodes the detection of printed or written character lines and white space areas found within the shredded documents. The information contained within these feature vectors can be stored efficiently using a run length encoding method, and also allows efficient computation of the shred matching costs. This also allows classification of shreds e.g., based on their upper and lower page margins (header and footer detection).

So, by using these and other characteristics the shreds can pre-sorted after which they can be matched more quickly and more accurately using the pre-sorted classification data of the shreds. To determine these last sets of accurate matching costs, color based feature vectors are used that are computed for each pixel along each side of the full-length contours of the shreds.

The last step that discussed is the actual reassembly of the different shred images on top of a common image canvas. This requires an accurate determination of the translational and rotational repositioning parameters of all the shreds.

Obviously, the proposed algorithms are discussed with several matching and reconstruction results obtained for a real shred database containing various types of shredded document pages.

Finally, the impact of these findings on secure document management strategies and the possibilities for applying these techniques within the context of forensic and investigative applications are briefly discussed.

Reconstruction, Shredded Documents, Digital Image Processing

 ** Presenting Author*