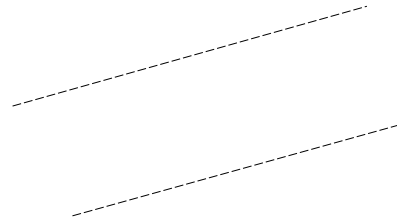


Efficient Conversion of Digital Documents to Multilayer Raster Formats

Léon Bottou, Patrick Haffner and Yann LeCun
AT&T Labs - Research
{leonb,haffner,yann}@research.att.com

Abstract

How can we turn the description of a digital (i.e. electronically produced) document into something efficient for multilayer raster formats [1, 6, 4]? It is first shown that a foreground/background segmentation without overlapping foreground components can be more efficient for viewing or printing. Then, a new algorithm that prevents overlaps between foreground components while optimizing both the document quality and compression ratio is derived from the Minimum Description Length (MDL) criterion. This algorithm makes the DjVu compression format significantly more efficient on electronically produced documents. Comparisons with other formats are provided.



1 Introduction

Most document description languages such as PostScript, PDF and MSWord are generally slow to render, may produce very large files and are often platform dependent. The distribution of electronic documents through the web is better achieved with efficient “rasterized” formats such as DjVu [1]. Such formats use different layers for text or graphic elements (i.e. foreground) and pictures or background. Obtaining a correct foreground/background segmentation has a critical impact on the performance of layered document raster formats such as DjVu or ITU T.44 [6]. The current DjVu segmenter [3] analyzes a raw high resolution color document image and produces a bitonal mask that relegates pixels to the background or the foreground with good accuracy.

In many practical cases, one would like to compress a document produced using computerized methods that do not rely on a pixel based representation of the document. A text processing software, for instance, represents a document using high level objects such as text, fonts, colors, embedded images, etc. This *structured page information* obviously provides considerable help for the segmenter and should yield very high quality compressed images.

Structured page information for electronic documents