Style characterization of machine printed texts
Bagdanov, A.D.

Citation for published version (APA):
Bagdanov, A. D. (2004). Style characterization of machine printed texts

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 7

Summary and concluding remarks

"The end."

—Anonymous

7.1 Summary

In this thesis we explored three measurable elements of style in machine printed documents. The visual style of a document imparts an immediate impression on the reader, allowing immediate discrimination without analyzing its deeper structural organization. Structural style is a measure of how the informational content of a document is organized into homogeneous regions, what their physical dimensions are, and their spatial relationships to each other. Textual style refers to the styling of the constituent elements of homogeneous regions, to the style of the textlines, words, and characters in them. The combination of these elements of style establish implicit rules which authors use to encode information in documents so that readers can decode it. Through characterization of the stylistic elements of machine printed texts, document understanding systems can exploit the implicit rules of style that humans take for granted.

We referred throughout this thesis to the agglomerated collection of stylistic elements as document genre. A document genre is a category of documents characterized by similarity of expression, style, form, or content. The textual, structural, and visual elements of style are the constituent elements of genre. Stylistic consistency defines a class of similar documents — a genre. Through characterization of these elements individually and by identifying consistency we developed several techniques for characterizing document genre in machine printed texts.

In chapter 2 we described an extension to discrete first order random graphs which uses continuous Gaussian distributions for modeling the densities of random elements in graphs. The First Order Gaussian Graph approach is particularly appealing for its simplicity in learning and representation. This simplicity is reflected in the ability to learn the distributions of random graph elements without having to worry about the discretization of the underlying feature space, and the ability to do so using few training samples. The use of an approximate matching strategy for comparing random graphs also enhances the efficiency of the technique while preserving discriminatory power. Experimental results establish the technique as an effective approach to the problem of document structure classification.
When comparing documents images based on visual similarity it is difficult to determine the correct scale and features for document representation. In chapter 3 we described a form of multivariate granulometries based on rectangles of varying size and aspect ratio. These rectangular granulometries are used to probe the visual structure of document images, and the rectangular size distributions derived from them are used as descriptors for document images. Experimental results indicate that rectangular size distributions are an effective way to characterize visual similarity of document images and provide insightful interpretation of classification and retrieval results in the original image space.

A morphological approach to textual style characterization was presented in chapter 4. We described several tools and techniques for measuring the style of characters, words, and textlines. A greedy algorithm for word spotting based on the linear properties of size distributions was presented that enables efficient searching of textline images for words. By approaching morphological size distributions in a slightly different way, and by introducing a generative model of size distributions of printed text, a perspective justifying the use principal component analysis on them for feature reduction was reached. These techniques expand upon the ideas developed in chapter 3, and also address some outstanding questions about the application of morphological size distributions to the problem of document style characterization.

Issues specific to the introduction of color to style characterization were addressed in chapter 5. At high scanning resolutions the shape of halftone dots is resolved, interfering with subsequent style measurements. We proposed a non-linear diffusion technique for recovering continuous tone color images from scanned color halftones. The technique uses a measure of local autocorrelation to drive and limit the diffusion. Experiments illustrated how the visual appearance of scanned halftones is improved when reproduced. Our diffusion technique mutes the high-frequency halftone signal, while preserving important visual details. By comparing the fidelity of images reconstructed from synthetically halftoned images, it has also been shown that our diffusion technique performs similarly on a variety of non-classical halftoning algorithms. We have also shown how diffused images can be much more effectively quantized to the number of perceptually salient colors in an observed document page. This results in images that are more visually appealing, reproduce better, and scale predictably because high frequency distortions induced by the halftone patterns are eliminated. It also allows us to simplify the representation of scanned color images, while preserving fidelity with the original.

Chapter 6 described a functional approach to experimental image processing and computer vision software design. The ideas described were directly motivated by our observations of how we and our colleagues conduct research on a daily basis. We illustrated how a modern, type-infering functional programming language can be used to construct image processing and computer vision software that provide a balance between meaningfulness and utility. The system described is able to provide functionality on demand, meaningful abstractions, and scales smoothly from prototype to production.
7.2 Concluding remarks

This thesis is about machine understanding of document images. Because we are characterizing phenomena in images, each stylistic component of document style is a visual phenomenon. Textual style is an expression of organized visual elements, and structural style is organized text. The important difference is the level of abstraction. Visual style characterization is often limited by the lack of prior assumptions that can be made. By assumption, nothing is known about the document when assessing its visual qualities. Because of this, many visual approaches fail to recognize the strong structural organization of information in their view. Structural characterization techniques are appealing because they most closely mirror their counterpart, layout mapping, on the other side of the document lifecycle. They are fragile, however, because they must make prior assumptions about admissible configurations of content regions. Textual style characterization is restricted by its limited purview within the vastness of the document image.

There is a notion of stylistic scale at work here. The methods proposed in the document understanding literature and every method described in this work operates at a single level of stylistic characterization, or over a limited range of scales. Integrated style characterization techniques that break out of the visual-structural-textual stratification are needed to advance the utility of style characterization as an aid to interpretation.

Style is used by authors and designers to facilitate the interpretation of content. Measuring style should constrain the interpretation of a document at the level the measurement is made. Determining that a document is visually similar to another constrains interpretation if the similar document is understood. Designating a document as belonging to a class of structurally similar documents narrows the range of sensible analysis. We have shown how style can be effectively characterized at the visual, structural and textual levels. In order for style measurements to be of use in document understanding systems, classes of style-conscious segmentation and layout analysis algorithms are needed that are able to articulate the style measurements most useful to their own operation.