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 1

Introduction

A good beginning makes for a good end.

—English proverb

1.1 Elements of style

Style is not an accidental phenomenon in document design, nor is its purpose purely aesthetic. Consider the document images of figure 1.1. The examples begin with a dense chunk of text, a bag of words completely undecipherable without extreme patience. In figure 1.1(b) some styling has been performed. Text that belongs together has been grouped, and spatial barriers introduced to isolate and associate them. The style indicates that it is some sort of addressed correspondence. It is still unclear what the document is, however, without analyzing the content; it could be an email message. The completely styled version in figure 1.1(c) is unmistakably a business letter. The content of each of the documents is identical in all of the examples. Style adds something; something that supplies immediate cues as to the purpose of document components and aids their interpretation.

![Figure 1.1: Incremental style. (a) a difficult to decipher bag of words. (b) an inkling of style. (c) a fully styled document](image)
The key observation is that style exists independently of content. In any of the examples of figure 1.1 the content could be endlessly changed while the type and purpose of each remained equivalently ambiguous or clear. We identify three measurable elements of style:

**Textual style** refers to the styling of individual characters, words, and textlines in document images. It a local phenomenon in documents. Document content is scanned or read with the aid of typographical signposts which take the form of keywords and emphasis. The large text of section headings is interpreted differently than the small print of the body. Emphasized text in the body is interpreted differently than the rest.

**Structural style** is organized textual style. It is most often conceptualized in the form of dotted-box-and-arrow annotations added to the first page of the prelude. Document pages are thought of as boxes of homogeneous content floating on a sea of paper. The structure of arrangement these boxes conform to rules of style. This page has a completely different organization of boxes floating on it than page 3, and is therefore in some way different.

**Visual style** refers to the overall visual impact of a document page. It is a grosser feature than structural style. When a page is first viewed, it imparts an immediate visual impression. Business letters and newspapers are distinguished at a glance, without having to analyze structure.

We will refer throughout this thesis to the agglomerated collection of stylistic elements as document *genre*. Genre is determined by association and similarity. A document genre is a category of documents characterized by similarity of expression, style, form, or content. Genre can be thought of as style consistency. Familiarity with a genre creates expectations that document instances conform to pre-conceived visual, structural and textual forms. Such conventions allow authors to effectively encode information according to styles that allow readers to decode it. These same conventions help machines to understand documents.

### 1.2 Context and scope

We begin with a description of the major components of modern document understanding systems, how they fit into the lifecycle of documents, and how they relate to the scope of this work.

#### 1.2.1 Context

Document understanding research is a very broad discipline. It brings together many different areas of research, such as image processing, computer vision, machine learning, and artificial intelligence An excellent review of the modern developments in the field of document understanding is the paper of Nagy [77].

Illustrated in figure 1.2 is a conceptualized version of a closed document lifecycle. The life of a document begins with the intent to express something in the form of written communication: a conceptual document. A genre for the document is selected, under
the constraints of the conceptual document. It would be absurd for this dissertation to be published in the form of a business letter or newspaper article. This is the highest level of stylistic discretion, determining the purpose and form of expression of the document. Once a genre has been selected, the logical entities possible in the document are constrained: business letters have a "To-Address," dissertations do not. Once this logical structure has been fixed, content generation begins, in which content is associated with all of the logical entities of the document genre. The document has expression and logical content, but no physical form. Half of the stylistic elements of the genre are determined. A line is drawn here separating document authoring from typesetting.

Layout mapping is the process by which logical entities are mapped to physical layout elements – floating boxes on paper. The structure of the text is fixed. Each type of layout component has internal rules according to which content is mapped and styled. Printing, in combination with document structuring, determine the final form and visual appearance of the document. Document typesetting ends at this point, and the document leaves the cycle for a time.
Upon re-entering, the document undergoes an analysis process that mirrors typesetting, and an understanding phase mirroring authoring. The major components of document analysis and understanding systems are:

1. **Document Acquisition**
   A document returns to the cycle and is acquired by imaging it with a scanner. Ideally, the document image after acquisition is identical to the one imaged during typesetting. Extensive preprocessing of the image is performed at this stage to remove real-world degradations. The image is binarized, if necessary, converting it from grayscale or color format [83, 80, 110]. Skew and other mis-registration errors are detected and removed [11, 51, 53]. There has been a recent trend toward document degradation modeling [52, 13]. The general goal of document degradation modeling is the development of accurate models of document image deformations such as mis-registration, low contrast, sensor noise, and physical paper distortions so that synthetically degraded document images can be generated for training, or that document analysis algorithms can be shown analytically to be robust with respect to the model. An excellent pictorial analysis and survey of how low-level image degradations can affect the local structure of printed characters is the book of Rice, Nagy, and Nartker [92].

2. **Layout Segmentation/Analysis**
   A document image is segmented, re-capturing the floating boxes representation. Homogeneously styled content is grouped together into rectangular regions. Regions are further subdivided into consistently styled textlines, words, and characters. There is a voluminous literature on document structure segmentation algorithms. The report by Cattoni et al. is an good survey [19]. The XY tree algorithm for document structure segmentation is a representative top-down approach [78, 79]. A document is cut recursively by alternating vertical and horizontal divisions. Cuts occur at horizontal and vertical lines of whitespace in the document image. Fitness of a potential cut position is determined by analysis of the horizontal and vertical projection profiles of the region being cut. The document spectrum method of O'Gorman is a bottom-up approach that clusters nearest neighbor connected components on the basis of distance and angular distribution [82]. With the increase in style variability in documents, attention has focused more recently on the segmentation of complex to very complex layouts. Ishitani proposes an emergent computational approach to layout segmentation and analysis [48]. Kise et al. suggest analysis of segmentation through Voronoi diagrams [55], and a novel approach using information about the background of document images is proposed by Antonacopoulos [5].

3. **Logical Analysis**
   As there is a distinction between document authoring and document typesetting, there is a distinction drawn between document image analysis and document image understanding. To use common computer science terminology, if document layout provides the syntax of a document structure, the logical structure of a
document attaches semantics to elements of document layout. Central to logical analysis of document are models linking physical to logical structure. The WISDOM++ system models the logical structure of classes of document as a set of inductively learned logical rules associating geometric concepts with logical ones [34]. Liang, et al. propose a graph matching approach where components are labeled by matching layout structure to a learned model graph for different classes [69]. The advent of the XML family of structured logical document standards has caused an increase in the use of syntactic models of logical structure to drive analysis [4, 116, 68].

4. Genre Classification
In the final stage of document understanding a complete stylistic summary of the document is made. The form, style, content and expression of the document have been recaptured, and the stylistic attributes of the document is integrated into its digital representation. In addition to the physical and logical structure, all information about its style has been re-captured.

1.2.2 Scope
Genre characterization is placed on the line separating document analysis from document understanding in figure 1.2. It is the natural analogue of the look-and-feel phenomenon associated with the proliferation and variability of house styles. There is a trend in document image understanding research toward model-based understanding of classes of documents [12, 57]. The adoption of a specific document model allows a document understanding system to focus on the recognition tasks unique to a specific class. Researchers have constructed models to solve difficult problems in table understanding [12], business letter analysis [26], character recognition [63], office mail flow automation [115], and postal automation [106, 84].

Another reason for this shift away from the unrestricted problem of document understanding systems is that there is a huge gap between an array of pixels in a scanned document image and the meaning of logical document elements. The gap is only partially bridged be analyzing the pixels into groups of layout components. It is referred to as the semantic gap in the content based image retrieval community [98]. Identifying a document as belonging to a known class at the earliest possible stage of analysis helps to bridge this gap by constraining the possible interpretation of document elements. Directly in the context of document understanding systems, this aids in the selection of models for understanding.

After each stage in the document analysis pipeline, style measurements are made that incrementally reconstruct a picture of a document's genre. Components corresponding to the stylistic elements identified in section 1.1 are arranged and inserted into the document lifecycle as shown in figure 1.3.

**Visual Style Characterization** is the first stage of style measurement. Immediately after document acquisition, the overall visual appearance of the document is assessed. Associations are made at this point between the incoming document and known classes of visually similar ones. Interpretation is constrained at this point, influencing the behavior of subsequent components. The assumption made here is that
documents that look the same should be interpreted the same.

A number of techniques have been proposed for characterizing visual similarity between document images. Soffe proposed texture measures to characterize visual similarity, extending the $N$-gram principle from linguistic analysis to two dimensional binary images [101]. Several distance measures are derived to measure the similarity between $N \times M$-grams (their term for the extension of $N$-grams to two dimensions). A notable feature of their approach is that it is not specific to document images, but can be applied effectively to arbitrary classes of binary images.

Hull describes visual similarity and equivalence measures computed directly from a compressed image representation [47]. The techniques use the pass-coding mode of the CCITT group 4 fax-compression standard. A document is represented by the locations where pass-coding occurs, taking advantage of the fact that the baseline of characters emit pass codes in predictable patterns. Similar documents are identified by computing the number of pass codes in every cell of a regular grid and computing the Euclidean distance between them. Document equivalence is determined by computing the modified Hausdorff distance between local patches of the two images. Patches with a low Hausdorff distance are considered to belong to different scanned images of the same physical document page. An advantage of these similarity measures is that they are robust to document image degradations such as photocopying.

Peng et al. developed an approach based on horizontal and vertical projections of bounding rectangles of connected content blocks [87]. They conjoin the horizontal and vertical projections into a single vector propose a maximum likelihood classifier as well as two distance-based classification rules. The computational simplicity of their technique is appealing and performed very well on a large sample of tax forms.

A method based on the partitioning of a document image into sequences of text and
non-text bins is proposed by Hu et al [46]. An edit distance defined on the intervals of text bins is given that they use to rank document images in a collection against a query image. Hidden Markov Models are also proposed for the task of assigning documents to learned classes of layout structure. Their techniques performed well at discriminating between structurally different classes of document layout styles.

Our approach to visual style characterization is motivated by the intrinsic multi-scale nature of documents. All of the techniques described above are sensitive to scale or characterize similarity at a single scale. Our focus has been on developing visual style characterizations capable of making discriminations at multiple scales of visual similarity.

**Structural Style Characterization** occurs once low-level document components have been grouped into homogeneous units. At this stage, important measurements are made on the properties of homogeneous regions and their relationships.

Shin et al. developed a system for classifying document page images using structural features [97]. It is a hybrid technique in that it uses texture features to measure visual properties locally, combining them with structural attributes of segmented document pages. The performance of their approach when ranking documents based on visual similarity correlates very high with human similarity judgments.

Generalized $N$-grams have been proposed as a statistical model for logical document structure [17]. The authors identify four hierarchical relationships in document structures and learn a probabilistic model of document structure from examples. This model is expressed in terms of the frequency of occurrence of each identified hierarchical relationship. A document of unknown type is compared against a learned model by testing its conformity to the model generalize $N$-gram distribution. An interesting aspect of this approach is that a global structural picture is built from observations of local structural relationships, allowing the models to capture sub-structural similarity.

A structural approach based on spatial relationships between segmented textblocks was proposed by Walischewski [114]. The technique uses a directed graphical model to represent document structure. Vertices are labeled with logical labels of textblocks and edges with the Allen interval relationship of the textblocks they connect [2]. A learning strategy is given for combining multiple instances of a document type. An advantage of the method is that the use of Allen interval relations allows the models to be formally precise about spatial relationships between document elements.

Another approach based on Allen interval representation is the reading order detection technique of Aiello et al. [1]. The technique uses a rule-base of reading order constraints defined over textblocks and their Allen relationship. Combined with linguistic constraints on consecutive textblocks in admissible reading orders, the technique performs well on complexly formatted document classes.

Several structural document type classification techniques are based on the XY tree segmentation algorithm, or variants thereof [78, 79]. The Geometric Tree (GTree) approach of Dengel and Dubiel arranges potential logical object arrangements hierarchically [28, 27]. Logical object arrangements are represented by XY cut patterns, and are arranged with document specific cuts close to the leaves, and class specific ones close to the root. Classification is done by traversing the GTree, evaluating the fitness of successor cuts with respect to the incoming document. Appiani, et al, pro-
pose a document structure classification technique based on a Document Decision Tree (DDT) [6]. Each node of the decision tree is a modified XY tree (MXY tree) which allows cuts to take place at non-white locations in the document depending on the local context of the potential cut. The DDT is learned from examples by hierarchically ordering the maximal common sub-trees between examples. An unknown structure is classified by comparing its MXY tree against the nodes in the DDTs of learned classes. Hidden Tree Markov Models (HTMMs) have been used to augment this approach [29]. The MXY tree representation was enriched with information about cut decisions, such as whether the cut was taken due to whitespace or ruling line, and geometric features of the cut region. A probabilistic architecture extending hidden Markov models to sets of trees is used to capture the dependence structure in these augmented MXY tree representations. HTMMs performed well in a comparison with DDTs on the task of classifying commercial invoices, particularly when the number of training samples was large.

Our approach to structural style characterization is similar to that of Walischewski [114] in that we use stochastic graphical models to represent classes of document style. One limitation of the approach of Walischewski is that logical labeling of all components is required for comparison. We focus on techniques to learn structural classes of document style in the absence of logical labeling.

A problem common to XY tree based structural characterizations is that relationships between segmented regions of the page are constrained by the recursively constructed hierarchy of rectangles. Regions are adjacent in the representation if they belong to a larger one that was cut during the segmentation process. Regions that are close on the physical page may be very distant in the XY tree. Our approach relates segmented regions according to spatial proximity on the page. It is based on rectangular segmentations, and is complementary to the XY tree approach in that it can be thought of as modeling spatial associations between the leaves of an XY segmentation tree.

Textual Style Characterization is where the lowest level measurements of style are made. At this point the deep structure of the homogeneous regions identified by layout segmentation is analyzed. Central to textual style characterization are characters, words, and textlines. The style of individual text elements influences their interpretation after recognition.

Typeface recognition is an important subject in textual style characterization [121, 14]. A significant sub-theme of typeface recognition is script determination. Spitz uses the locations of pass codes in CCITT group 4 compressed images and the distance of upward concavities from the baseline to discriminate many scripts and languages [105]. The identification of a document as belonging to a linguistic class greatly constrains the possibilities for logical and structural analysis.

Word spotting is another instance of textual style characterization. Optical character recognition is not possible for many document collections [66]. Spotting words in document images allow these images collections to be queried with textual queries. Discriminating between stop and non-stop words on the basis of textual style measurements also improves keyword spotting for information retrieval [45].
A common limitation of textual style characterization techniques is that they must be applied to entire words or textlines. We focus on approaches that allow local measurements to be combined into meaningful measurements of larger structures. In our approach, character style measurements combine into measurements of words.

This is the context of style characterization in the document lifecycle. Several classical applications in document analysis and understanding have similar purposes and goals as ours, and the positioning of document style characterization is not limited to the scenario in figure 1.3. Duplicate document detection, which is a very specific instance of document style characterization, is essential in large document collections to reduce needless storage and computational overhead [67]. Document image retrieval systems are of particular interest in some application areas [31]. Style is a crucial discriminating feature of document images in the absence of textual content.

1.3 Organization of this thesis

This dissertation examines elements of style in machine printed texts and proposes tools and techniques to characterize them. 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 refer 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. Characterizing these elements individually and identifying consistency characterizes a document genre.

In chapter 2 we describe a structural style characterization technique based on textblock segmentations of document page images. The rectangles that constitute structural style are not rigid entities, and we show how it is possible to construct stochastic models for document classes that capture the structural consistency and variability of document image segmentations.

A multi-scale technique for characterizing visual style is given in chapter 3. If accurate segmentation information is not available, or if for some reason we are unwilling to commit to a single segmentation of a document image, visual style consistency can be characterized through analysis of Morphological decompositions of the background of document images.

Textual style characterization is the subject of chapter 4. Characters are the brush strokes from which document style emerges. We consider how local style measurements
can be combined into meaningful measurements of higher-level style. In this chapter we extend the morphological techniques of chapter 3 to address this problem.

The techniques used to reproduce color in print make it difficult to acquire high resolution scans of color document pages that are perceptually accurate. This discrepancy between perceived and scanned document must be corrected before any style measurements can be made on color document images. Chapter 5 addresses the problem of obtaining perceptually salient versions of color document images from scanned halftones.

Document understanding is, in a general sense, a computer vision and image processing problem. In chapter 6 we reflect on the process of software design in support of computer vision and image processing research environments. Through observations of personal research practice we propose some new approaches to implementation of image processing software that allows for rapid prototyping and re-configuration of image processing functionality.