CULTURE-ADAPTIVE DOCUMENT DESIGN ASSESSMENT

by

Alexis M. Rodriguez-Diaz

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE
in
COMPUTER ENGINEERING

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS
2010

Approved by:

_____	_____
Nayda G. Santiago-Santiago, PhD Member, Graduate Committee	Date
_____	_____
Manuel Rodriguez-Martinez, PhD Member, Graduate Committee	Date
_____	_____
Jose Fernando Vega-Riveros, PhD President, Graduate Committee	Date
_____	_____
Agustín Rullán-Toro, PhD Representative of Graduate Studies	Date
_____	_____
Hamed Parsiani, PhD Chairperson of the Department	Date

Abstract

An approach to analyze the aesthetics of a document's design is researched. The premise behind the developed methodology is that the perception of beauty and pleasantness of a document's design is dependent on the viewing audience, which can vary between cultures. The methodology was developed with people who are not document design experts in mind, including students and clerical office workers. The system applies concepts from Case-Based Reasoning and document matching methods to provide assessment data from previously evaluated documents that are similar to the document in question, but before selecting such a methodology a Layout Esthetics Measurement methodology was explored. The system's effectiveness is assessed by cross-comparing matching results between documents in a sample set. A variant version of a Confusion Matrix, which we name Distance Matrix, is employed in the visualization of these matching results.

Resumen

Un enfoque para el análisis de la estética del diseño de un documento es investigado. La premisa detrás de la metodología desarrollada es que la percepción de la belleza y agradabilidad del diseño de un documento depende de la audiencia que lo observa, la cual puede variar de cultura a cultura. La metodología fue desarrollada con personas que no son expertos en diseño de documentos en mente, incluyendo estudiantes y empleados de oficina. El sistema aplica los conceptos de Razonamiento Basado en Casos y métodos de pareo de documentos para proporcionar datos de evaluación de documentos previamente evaluados, similares al documento en duda, pero antes de seleccionar dicha metodología investigamos el uso de una metodología de Medición de Estética del Diseño. La eficacia del sistema es evaluada comparando los resultados del pareo entre documentos de un conjunto de muestras. Una variante de la Matriz de Confusión, la cual nombramos Matriz de Distancia, se emplea en la visualización de los resultados del pareo.

To life, and those who are a part of mine.

Acknowledgements

I cannot start without stressing the importance of having a continuous support from my family and friends during these past few years. That unconditional support was vital from beginning to end.

I must continue by not only thanking, but also providing my most sincere admiration and respect to my advisor, Dr. Fernando Vega-Riveros, who gave me the opportunity to work with him. His guidance and unconditional availability was key in the development of this research. Our conversations regarding subjects outside the research became an important part of my personal growth.

Special thanks to Dr. Jan P. Allebach, from Purdue University, who went out of his way to loan us a Mac computer to the Digital Publishing team in UPRM. This equipment became a quintessential tool in the design of several samples used for this research. This appreciation is extended to HP Labs who provided research funds to support the Digital Publishing team at UPRM.

Finally, but no less important, a recognition to the Electric and Computer Engineering department at UPRM for their financial support that allowed me to represent our school at the 2007 Chameleon Federation Conference, in Bologna, Italy. My most sincere thanks goes to my academic counselor, Mrs. Sandra Montalvo. Her willingness to help students and invaluable advice made some of the stresses of academia more bearable.

Glossary of Terms

Digital Publishing (DP) – the act of compiling written, visual or audio materials and releasing them in a cohesive form by using computer software and/or hardware.

Variable Data Printing (VDP) – a form of printing in which elements such as text, graphics and images may be changed from one printed piece to the next using information from a database or external file.

Variable Data Job (VDJ) – an efficient Variable Data Printing implementation where a master template is used to define each instance, which can contain a different set of elements from one another.

Case-Based Reasoning (CBR) – a reasoning paradigm where previous situations are represented as cases and are used to solve a new problem.

Component Block Projection (CBP) – concatenated directional projection vectors of the components of a document structure.

Autonomous Document Assessment Expert (ADAE) – a joint project between three universities to develop software tools that facilitate the Digital Publishing process to individuals considered non-experts in the publishing field.

Extensible Markup Language (XML) – a set of rules developed by the World Wide Web Consortium for structuring, transporting and storing data. It is used as a basis to develop other data representation standards.

eXtensible Stylesheet Language - Formatting Objects (XSL-FO) – an XML-based language that can format a given set of data elements for its visual presentation.

Personalized Print Markup Language (PPML) – an XML-based standard designed to represent Variable Data Jobs used by high-volume printers.

Case-Based Design Reasoning Environment (CADRE) – a Case-Based Reasoning adaptation specially designed for the domain of building structure design.

Case-Based Design Tool (CADET) – is a system that aids conceptual design of electro-mechanical devices and is based on the paradigm of Case-based Reasoning.

Knowledge-Based Artifact Recognition (KBAR) – a framework that combines several reasoning paradigms, including Rule-Based and Case-Based Reasoning, to detect errors in a Variable Data Job template's instance.

TABLE OF CONTENTS

TABLE LIST

Tables **Page**

FIGURE LIST

Figures **Page**

1 INTRODUCTION

Marketers, publishers, consumer product companies and consumers are increasingly using digital printing technology to produce everything from marketing collateral and direct mail to photo merchandise, books and manuals. InfoTrends reported that marketing collateral and direct mail collectively account for 44 percent of high-volume digital color press pages[1]. And while short-runs and on-demand printing are already popular applications, considerable growth was projected for personalization in Digital Publishing (DP), from 49 billion pages in 2004 to 138 billion pages in 2009[2].

The purpose of DP is to empower individuals to control all, or at least most of the publishing processes [1]. It is an end-to-end workflow where customers can obtain the desired service on demand, from document [3] design to printing. To satisfy this last requirement the technology should be as flexible and automated as possible. In short, DP is the integration of the necessary technologies to publish anything, at anytime, anywhere, and by anyone. Digital Publishing, however, is still in its infancy despite recent investigations [5] [6] [7] [8] [9].

In line with this vision, HP Research provided funds to support an inter-university effort led by Purdue University, under the Chameleon Federation (www.dp-chameleon.org), known as the "Autonomous Document Assessment Expert" (ADAE). The project was a joint

---

[1] "HP Breaks Record for Annual Page Growth in Digital Printing", http://www.hp.com/hpinfo/newsroom/press/2007/070223a.html, February 23, 2007, Retrieved March 8, 2010.
[2] "Xerox Completes Acquisition of XMPie", http://news.xerox.com/pr/xerox/NR_2006Nov10_XMPie.aspx, November 10, 2006. Retrieved March 8, 2010.
[3] The term "document" is a general term used in this thesis to refer to many types of publications like articles, brochures, posters, etc. When a specific type of document applies, it will be referred to with its name.

effort between Purdue, University of Puerto Rico at Mayaguez, and Pontifícia Universidade do Rio Grande do Sul (PUCRS) in Brazil.

The ADAE project's main objective and outcome was to develop tools that facilitate the end-to-end process of DP to individual users. However, the ADAE project went beyond creating software tools. To simply provide access to DP hardware and software is not enough to enable their users, i.e. students and educators, in incorporating DP as part of their daily learning and teaching activities. The document design knowledge applied by such a tool may be obvious to experts in the field, more specifically graphic artists and marketers. This, however, does not apply to non-experts; the complexity of this applied knowledge must be hidden to the non-expert. As a result, expert-level knowledge in design and publishing is available to a wider population, not only a select group. Achieving this would allow non-experts to prospectively improve the appearance of a design and create high quality content for homework assignments, course projects, lecture materials, and reference resources, in the form of reports, presentations, and posters[4]. In essence, the software tool would provide design "assessment" and apply the same kind of critical judgment that a good designer would bring, like identifying problems and offering suggestions to the document creator.

---

[4] "A proposal to create an Autonomous Document Expert Assessment (ADAE)", http://linus.ecn.purdue.edu/chameleon/documents/ADAE_proposal.pdf, 10 6-06, Retrieved September 19, 2007.

1.1 The Problem

> *Digital documents […] are created by human designers and humans make mistakes. Therefore many documents contain defects, such as missing context, wrong use of metrics, aesthetically unpleasant context arrangement, infringement of the page constraints, image resolution below requirements and so on.* [1]

Design defects, which we will refer to as artifacts from now on, take on many forms and shapes. The most difficult task is identifying and correcting them. Such a task is fairly simple to a Graphic Designer, Art Director or some other expert in Media design. Any of these experts might even be able to explain why a design is not "Aesthetically[5] Pleasant" with terminology well known in the media design industry. There are a couple of issues with this, however. First, this terminology is not necessarily obvious to amateur designers, much less to those who know nothing about design. Second, but no less important, there is ambiguity on the terms used, meaning that there is no quantitative standard or formula that can be used to measure artifacts. Two research works come to mind whose approaches developed methods to detect design defects in documents [1] [2], however their methods used an expert-approved template as reference to compare to other documents with a similar, if not equal structure. Thus, how can a non-expert determine if a design is "Aesthetically Pleasing"? Another consideration not taken on previous approaches is that of ever-changing trends in design. None of these methods are flexible enough to easily evolve or adapt to these trends. Also trends may vary between countries and cultures, which means that the solution should not only adapt easily to changes through time, but to differences between cultures too.

It is the main goal of this research to develop a method that determines if a particular design may or not be "aesthetically pleasing". To achieve this, design principles and their relation with document design composition were researched to define the logic behind the document analysis process, while considering variations of this process between cultures.

The next step was to analyze previous works with similar approaches, allowing us to determine the most feasible and flexible implementation for the analysis process. The result would be a design assessment method that can be useful for design experts and non-experts alike.

1.2 The Framework

One important feature considered in the analysis is its applicability to different document types tailored to different audiences and cultures. Thus, the method would allow a certain level of flexibility in handling a variety of cases. We have determined that a Case-based reasoning[6] system was a feasible choice due to its trainability, or ease of modification.

The framework then consists of a case-base (CB) of previously reviewed documents, judged as aesthetic (good) or non-aesthetic (bad). These would be compared with a document selected for analysis. A match may turn one of the following possible outcomes:

1. The analyzed document is similar to an aesthetic example, thus itself being aesthetically pleasing

---

[5] The terms *Aesthetic* and *Esthetic* are both used throughout this thesis. By definition on the Oxford Dictionary, both words are used interchangeably. However, we will use the word *Aesthetic* as defined by the Oxford Dictionary and *Esthetic* when referring to Esthetic Measurements.

[6] See "Case-based reasoning", by Janet L. Kolodner (1993).

2. The analyzed document is similar to a non-aesthetic example, thus itself being aesthetically unpleasing. An aesthetically pleasant example of this same document is provided, along with an explanation of the recommended changes.

The "cases" used to train the CB are Personalized Print Markup Language[7] (PPML) files, whose structure is XML-like. To obtain PPML files, documents were designed on QuarkXpress 6.5 and converted to PPML using a plug-in application from HP Anvil designed for QuarkXpress.

The analysis process is, of course, the main feature of the framework. The CB is what helps the analysis do its intended work. Because the amount of cases in a CB can be considerably high, the analysis process should be quick but effective. It should also be able to equally analyze any type of document, no matter how complex or simple it is. **Projection Vectors** [3] are an efficient structure that represents the layout of a document while at the same time providing an easily calculable structure for mathematic and logic operations. These vectors can easily be created with the data available in the CB and compared with other vectors (files).

The only assumption about the cases in the CB is that each non-aesthetic document in the database should have a reference to its corrected (aesthetic) version and also contain a small description of the recommended changes. This is not required for aesthetic cases.

---

[7] More information of PPML is found on www.podi.org

1.3 Main Contributions

The intention of this thesis is the discovery of a means to analyze any document to determine if it is aesthetically pleasing, without restricting the analysis to a particular audience or culture. Previous approaches assumed having the approved version of a document to be analyzed, a template version of the analyzed documents that followed certain design principles [1] [2]. Hall's [4] research on how culture, from a sociological perspective, influences perception supports our logic behind the analysis process. Furthermore, a modified implementation of Projection Vectors [3] was considered due to its effectiveness, efficiency, and practicality.

2 Literature Survey

Before the invention of digital printers, many individuals could not accessibly produce low-cost, complex, and graphically rich pages. Printing plates were the only medium able to produce high volumes of pages, but once digital (plate-less) printing devices became available there was an urge to exploit the ability to manipulate graphic elements – from half-tones to character outlines [5] – in very low quantities. These substantial changes in the document design tools and the plate-less press machinery have created the revolutionary Digital Publishing (DP) field. A few years ago, a group of researchers paved the way for the previously undeveloped Digital Publishing arena - undeveloped, since this was both a relatively new commercial and research field. Thanks to this group of researchers a considerable amount of literature related to DP is available today. In the pages that follow we will provide insight on topics that were relevant to this investigation along with others that are beyond the scope of this research but still provided some insight to our work.

2.1 Document Analysis

The recent focus of researchers has been the detection of unacceptable changes in template-dependent instances [1] [2] [6], which can be considered as defects in a document. These errors are implicit within the document and require different and unique approaches.

Our research begins following the footsteps of a previous work by Santos-Villalobos [1], demonstrating the effectiveness of using Case-based reasoning (CBR) to autonomously proof, without human intervention, instances of a Variable Data Job, or VDJ, a type of file that contains a document's template and its instances. To achieve this, Santos-Villalobos'

focus is to recognize style-arrangements by having the system becoming aware of design patterns inside the document page which provide information about a document's style and intent. Despite the overwhelming achievement of such approach, it is limited to marking defects without providing any suggestions to fix them. Ideally, the method should flag the mistakes and, if possible, provide solutions to fix them. The approach created by Santos-Villalobos is a framework tool known as the Knowledge-Based Artifacts Recognition, or KBAR for short. There are many reasons why CBR approaches are better in such cases, but among the most important is that it can handle a wide variety of situations without explicitly hard-coding them. This was a good start, yet our focus would not be the instances of a VDJ, but the document's template.

Two bold attempts to fix defects are based on automation approaches, the first known as Object Adaptation [7], where an arbitrarily shaped image is automatically fitted into an arbitrarily shaped image space in a template while maximizing use of available space; and the second, Content Fitting [5] where both text and images are resized to maximize the distribution of content in an instance. Improvements of these approaches and tight integration into a workflow are expected in the near future.

We can observe that in the development of DP technologies it is common to see how the implementation of this end-to-end process varies. Personalized Direct Marketing aims to create one-to-one marketing by leveraging customer information that a company may already have [8], while others have a more robust vision of Direct Marketing, focusing all efforts on running a targeted marketing campaign [9]. Marketing campaigns are complex, as they require three different organizations to communicate constantly:

- The *Enterprise* that funds the campaign
- The *Agency* that creates the campaign based on the Enterprise's goals
- The *Print Service Provider* which, of course, is responsible for producing the materials based on the specifications of the agency

The solution to this is five workflows running in a sequence, where each workflow handles the tasks of one or many of the three organizations previously mentioned. These tasks involve:

- Recipient Selection and Interest Topic Matching
- [Individualized] Material Creation
- Content Upload
- [Print] Production and Fulfillment and
- Response Analysis/Database Update

Yet another example is the Anvil toolset [10], where every part of the publishing process is tightly integrated to streamline Variable Data Printing[8] (VDP) production (see figure 2.1). The way Anvil is designed allows a wider array of VDP jobs to be created, while only concentrating on the vital components of a VDP workflow which are the Art Design, the Post-Design where the variable fields are created, and the Pre-Press where the VDP instances are created and sent to press for printing. Therefore, the workflow has the flexibility of outsourcing some tasks, as the Design phase can be assigned to a freelance graphic artist or the Pre-Press process can be done by a Print Service Provider; but they can all be done by the same company, from design to printing.

---

[8] Essentially, Variable Data Printing refers to personalizing a document from an existing template by inserting a different set of text and illustrations to each document created.

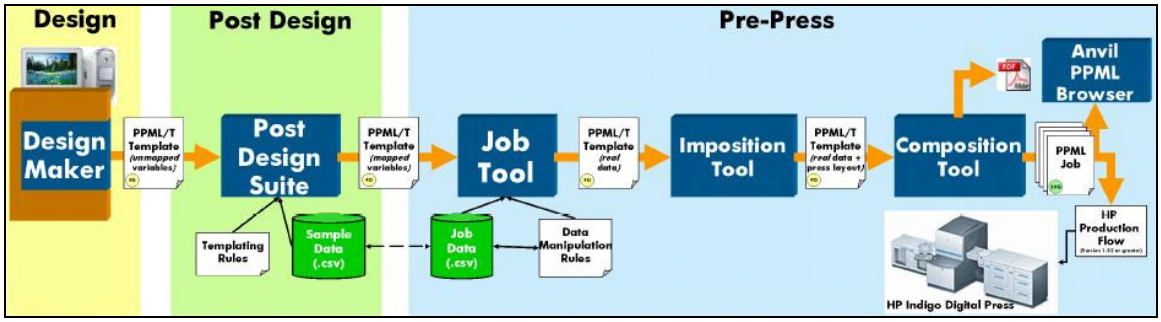

Figure 2.1 – The Phases and Components of the Anvil Segmented Workflow as seen in [10]

As pictured in figure 2.1, such a workflow allows individuals and companies to control all, or at least most of the publishing process. Depending on the needs of the user, some software tools or equipment may be required. Anvil, for example, requires that a template - or document layout - be created using an art design tool – specifically Quark XPress or Adobe inDesign -- and that the final output be printed in a PPML-compatible press, like Hewlett-Packard's Indigo [10]. Depending on the individual or the company's resources or needs, the workflow can be tailored to fit the user's advertising strategic model. But because this is a relatively new field, the technologies available are not mature enough. So even after a workflow is adapted to fit an advertising model, there are issues that will be encountered. One of these issues is encountered when a specific part of the workflow requires the intervention of an expert due to an error that can be somehow detected. For example, errors in a Variable Data Job[9] (VDJ) require that a human intervenes to find out the problem and its solution. Issues like this are costly in terms of human or time resources.

[9] A group of instances created from a single document template, where each instance is distinct due to varying content, like pictures and text.

To verify that a VDJ has little or no issues in it, two preventive steps are taken: **preflight** and **proofing**. Preflight was not given much attention by researchers, compared to proofing, mainly because most tasks were already automated –*i.e.* font type verification, margins, images. The layout design's creation and approval was assumed to be done by experts. If the document is not created or at least verified by an expert, the proofing process may be a total failure since the document instances will inherit the template's flaws. Even if the template is to be used for the creation of one instance, it is possible that the proofing mechanism will not find artifacts. Flawed or not, the instances will be compared with the template, thus creating "false positive" results[10]. Even if the template is to be used for the creation of one instance, the result will possibly be an aesthetically unpleasant document which provokes little or no interest to its intended audience.

In essence, and based on previous work, there is a sense of direction that can be followed. The Autonomous Document Assessment Expert (ADAE) group followed this direction by focusing efforts on document layout design assessment, an integral part that follows the design process and precedes the preflight phase.

2.2 Design Principles

Not just a few, but many principles exist. A few books were consulted to understand how these "principles" find their way through the many visual publications we see in our daily lives.

---

[10] A false positive is a result that is erroneously positive when actually it is the complete opposite.

To provide an idea of the amount of design principles that exist and in which ways they are used, we make reference of Lidwell et al. [11] book, Universal Principles of Design. To be precise, one-hundred (100) principles are presented in this book with visual examples. Of course, not all of these principles are useful for publishing purposes, but they do provide insight into the convolution of ideas and concepts that are promoted by design experts. At times, the principles presented in this book would appear confusingly similar. For example, it is common to see that objects in a layout that are somehow grouped together allow it to look more structured. Two principles mentioned in this book can be associated to this idea. *Proximity* relates objects by having them positioned next to each other while *Closure* achieves the same with other visual cues, like having a group of objects arranged in a recognizable pattern. In other instances, they might overlap. Take for example *Readability* and *Legibility*, the former being the degree of complexity in the text (i. e. literature complexity) while the latter refers to how easy or difficult is it to see the text. Among the many design principles used by Santos-Villalobos [1], *alignment* seemed predominantly the most relevant. At first thought, we might think alignment refers mostly to text - left, center, right, justify - but its application goes far beyond this. Alignment between elements runs at the edge or center of an invisible line, also referred to as a Strong Line [12]. Figure 2.2 shows two examples of Strong Lines.

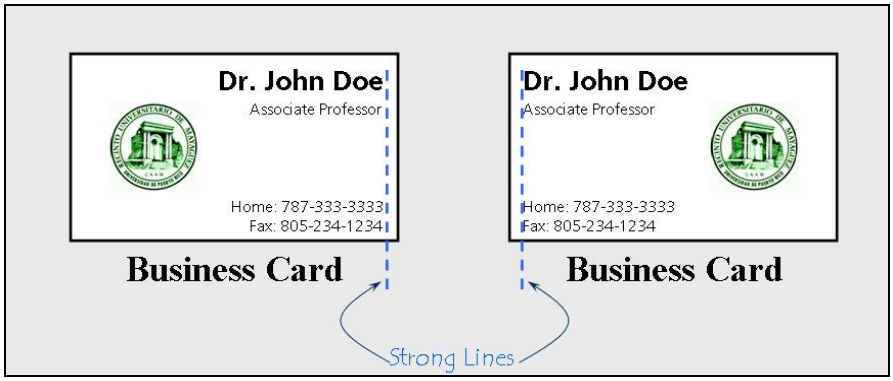

Figure 2.2 – Application of Strong Lines detection on a business card taken from [1]

Strong Lines is one of many techniques that help the designer to effectively lead the reader through the content of a document. Alternatively, white space "balances the context", giving the eyes a visual break [13]. White space is the area without content in a document relative to the areas where there is. Too much content and both, eyes and brain, are overloaded.

Out of that mix of principles, Lester [14] prefers a cleaner selection of principles that are more appealing to visual media: contrast, balance, rhythm and unity. More important than explaining how each of these are applied, Lester emphasizes that because good graphic design can follow or challenge these principles, he prefers calling them suggestions, not rules or principles.

This can be a problem as well as a challenge, because it is not feasible to understand them all and find a way to implement them in a systematic manner.

2.3 Perception and Culture

These two words, perception and culture, are considerably related when speaking about design in general. Lester [14] provides insight on visual perception as defined by

Aldous Huxley, who describes it as a 3-step process that repeats itself constantly, unconsciously:

1. Sense – regards the physical aspect of seeing, by letting light enter into the eyes to see the object being observed.
2. Select – is the process of isolating and looking for specifics. This step requires a combination of light-gathering and focus properties of the eye.
3. Perceive – requires a more extensive mental process of giving meaning to what is being observed, so that it becomes part of the individual's long-term memory.

Huxley curiously summarizes this process in one simple phrase: *"Sensing plus selecting plus perceiving equals seeing."* We can thus somehow argue that what we *perceive* will initially depend on what we *sense*, and what we *sense* can be different in another environment or culture. In fact, Lester argues that cultural influences have a tremendous impact on visual perception. To further support our argument, a more general perspective of *culture* is provided by anthropologist Edward T. Hall in his book *Beyond Culture* [4]. In it, he explains that there are rules which govern perception, within at least five categories of events that must be taken into account:

1. The subject or activity
2. The situation
3. One's status in a social system
4. Past experience
5. Culture

In combination, they bring upon *context*, whose function is to increase the brain's ability to supply missing information and understand the message being transmitted. Of the five categories, *Culture* is considerably one of the most important. One of its functions is to

provide a highly selective screen between a person and the outside world, therefore designating what we pay attention to and what we ignore. This “screen” provides structure for the individual's "world" and protects the nervous system from "Information Overload"[11]. Indeed, he argues that this “screening” function is brought upon that current *context*.

To provide some insight about these ideas, we came upon a somehow practical perspective on context, as explained by Toussaint [15]. He argues that the effect of context is that some object N that has certain properties when it is viewed in isolation can change when N is viewed in some context. Not only that, but an object N is seen as one thing in context A and another in context B. The effect can occur at many different levels including perceptual, cognitive, and “objective” mathematical levels [15]. One example is the Muller-Lyer illusion, as seen in figure 2.3.

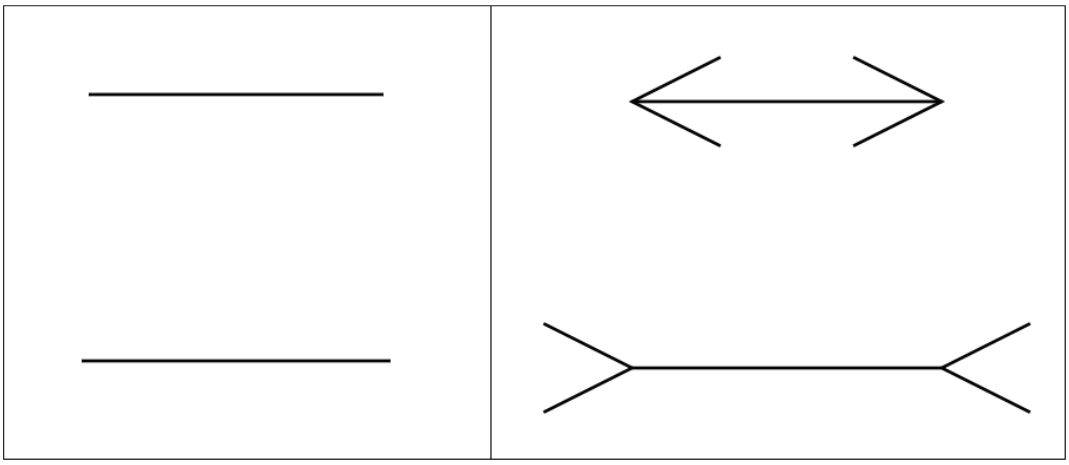

Figure 2.3 – Muller-Lyer illusion

[11] Hall refers to this term as the analogy of a computer system which can only handle a certain amount of data at once. A system that tries to process more information than it can handle is referred to as overload.

On the left side, the two horizontal lines' length is the same. Now add some context in the form of arrows at the end, having the figures on the right appear to be different in length. This is the result of context's effect on visual perception.

It is important to understand that culture strongly dictates what a person will perceive. Graphic designers keep this in mind, recurring to a mix of design principles and their own intuition and experience to compose a layout that, to their understanding, will captivate the audience's interest and attention.

2.4 Approaches to Layout Understanding and Aesthetic Analysis

Because there are many design principles, or suggestions as [14] defines them, and culture dictates their effectiveness, our approach cannot simply rely on the use of these principles. But before devising a different document layout analysis approach, we were able to find similar attempts during our research.

2.4.1 Layout Understanding

Esposito et al. [16] developed an efficient method of layout analysis for the purpose of classification and data gathering. The classification process is designed to handle many types of documents, each one having a distinct layout. *Context* is provided by means of knowledge about the document's hierarchical layout, which is the order in which certain object can be found in a document, from title to footnote. Santos-Villalobos' [1] approach was inspired by the method developed by Esposito et al; in fact he developed his own hierarchical layout knowledge tree which he called Document Ontology.

The success of this approach lies on the ability of knowing in advance the variety of documents layouts to be encountered during analysis. For our purposes, however, this is not desirable as the ability of managing any type of layout is reduced.

2.4.2 Previous attempts on the use of Case-Based Reasoning in design

At some point during our research we were attracted to the idea of using Case-based reasoning (CBR), mostly because of slightly successful attempts in the past [1]. Thus we decided to look into it from a specific perspective: CBR in design applications. Previous attempts to use CBR in analyzing a building's design parameters and providing assessment were found, which are comprehensively explained in [17].

While drafting a sketch for a new project, building architects often find themselves with structure design problems that have occurred in previous projects but are unable to recall them in detail. This leads them through the process of having to go through dozens, maybe hundreds of filed plans to find the previously applied solution. In CADRE (CAse-based Design Reasoning Environment), a case-based architectural design system, previous design knowledge can be applied to a new design within a similar context. Context is provided in the form of design parameters, including building geometry, structural parameters, constraints and environmental features, among others. CADRE falls on a category of CBR applications that are *episode-oriented*, meaning that the previous solutions to be selected will depend on the particular situation encountered in the new problem.

Another category of CBR applications is *model-oriented*, which use generic design models from causal knowledge. CADET – used for the conceptual design of electro-mechanical devices – is such an example whose methodology for design synthesis is viewed

as analogical reasoning used in applying known design models, rather than reusing specific design *episodes*. This is a very helpful method to aid the design by preventing problems from developing.

One aspect to note about these approaches is that, to provide the correct assessment, data about the *episode* or *model* is provided in advance. CADRE is able to search for similar issues encountered in the past with the building specifications provided from the very beginning. CADET can proactively let the designer know what is permissible or not because it already contains data about the parts being used in the design. This is somehow possible in document layout composition if the characteristics of the document are known beforehand – i.e. if it is a magazine article, then it should be aligned and structured. But this would vastly reduce the scope of the method being developed to a specific culture. The ideal approach should be able to handle a variety of document types, with a variety of arrangements, for a variety of cultures. This does not dismiss the use of CBR in the developed method; it must simply be approached from a different perspective.

2.4.3 Systematic design analysis approaches: the use of esthetic measurements and template matching for case retrieval

At some point it was more or less obvious that what we needed was a method to determine if a layout was aesthetic, based on similar layouts which have already been criticized as aesthetic or non-aesthetic, where the latter case should provide details on its defects and fixes. We considered Case-based reasoning an ideal paradigm for two reasons. First, these previously assessed layouts can easily be represented as cases, and second, it can easily be trained with more of these layouts.

The difficulty lied on the matching process, or finding a layout (case) that is similar to the one being analyzed. One common solution to do this in CBR is to categorize cases for a more effective selection. Ngo et al. [18] developed a method to evaluate the aesthetics of a computer window interface. The result of this evaluation was a numeric value between 0 (worst) and 1 (best), with a simple but long formula providing the resulting value. The variables of this formula came in the form of thirteen (13) esthetic measurements, also with values between 0 and 1, each one equally taking into consideration all the elements (e. g. text boxes, image boxes, lines) on the interface's layout. Even though we are not evaluating interfaces, the method is adaptable for measuring a document's layout. For proof of concept purposes, we selected only three (3) of these measurements which we considered relevant: Balance, Equilibrium and Symmetry. By measuring these values on a document layout we would expect to categorize it and find a similar case in the CBR. For example, if the document had a Symmetry value closer to 1, then it is easier to match with a case in the CBR by searching for cases with a similar Symmetry value. To zero in on the most similar case, one would also apply the same search process with the other two measurements, Balance and Equilibrium. We measure the similarity between two documents with respect to their three esthetic measurements by calculating their Euclidean distance. We named these distances Balance Distance, Equilibrium Distance, and Symmetry Distance. Finding acceptable threshold values for the resulting distances became a 3-dimensional optimization problem that, although not impossible to solve, would require extensive validation. We tried simplifying these distances by summing them up into one formula, that we named Esthetic Distance. But as we analyzed the resulting Esthetic Distance between documents we could

not find any patterns for the threshold selection from the results, possibly due to the dilution resulting from the sum of all three esthetic measurements into the esthetic distance. The distance formulas and some of the results are discussed in chapter 6.

We still needed a method that could measure the similarity of two documents without having to solve an n-dimensional problem of threshold validation. The solution comes in the form of a *Template Matching* method developed by Peng et al. [3], designed to "find the most similar template for any input document image in a pre-stored template document image data set". While complex to understand in detail, it is rather simple to implement. In short, a document layout is represented in numerical vectors, two to be precise: vertical and horizontal. These two vectors – where each element contains either a 0 or 1, representing the absence or presence of an element in a particular pixel-position of the document – are concatenated into a single vector A as illustrated in figure 2.4, which is then matched with vectors created from *n* cases in the CBR (B_n). The computation "MINIMUM (A - B_n)" would help obtain the closest matching layout in the CBR. The major advantages of this method, according to their authors, are two:

1. The spatial relationship between elements of a layout is better represented, thus providing a very high matching accuracy that can be obtained even for a large set of cases;
2. The representation of a document's layout as numerical vectors would allow for lower computational cost, because of simplicity in the computation

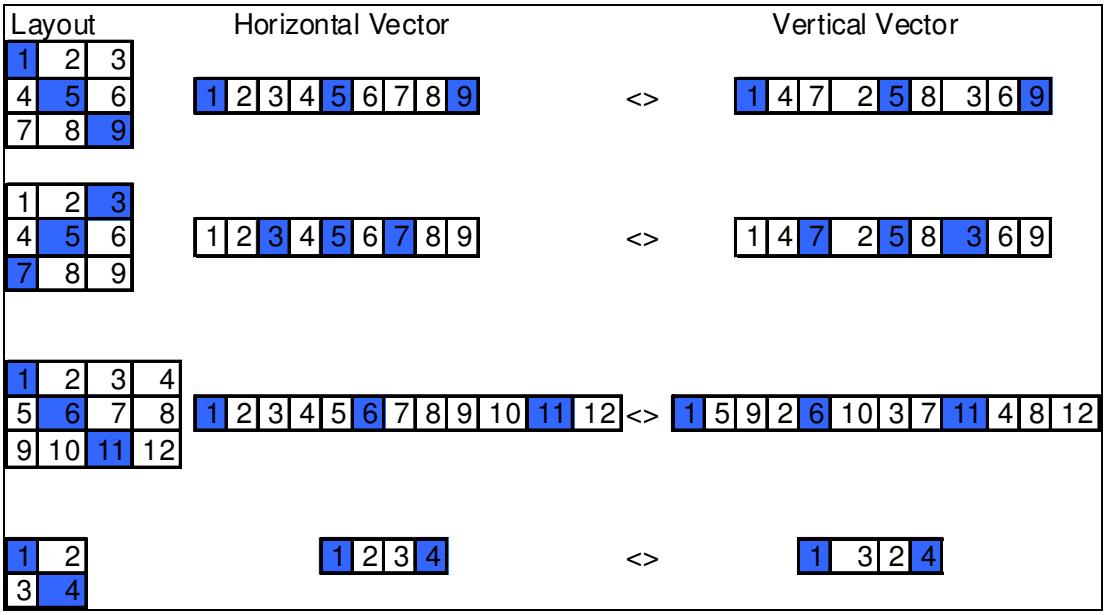

Figure 2.4 – Layout's vector representation
Layout example (extreme left) followed by its respective vectors representations to the right. Both vectors are joined to form a single vector A.

2.5 Chapter Review

It is important to note that when reviewing previous research on DP there is very little focus on the relevance of "culture" as a factor to consider in the perception of a document aesthetic appearance. Yet, it was still important to review what has previously been attempted to understand the circumstances and cases where a certain approach can or cannot work.

For example, research by [1] [2] and [6] focuses on the detection of unacceptable changes in instances of a VDJ, based on the document's template. This approach works well when the template is assumingly aesthetic. Their research also provide insight on how esthetic measurements allow to characterize a document by obtaining implicit-knowledge from it [1], by identifying the quality of an instance based on the number and severity of changes from the original template [2], or the quality of images in an instance which is

severely affected due to changes in their aspect ratio and others [6]. Other researchers take a more automated approach, which is the case of the Anvil toolset [10]. Font-type verification, margins and images in file are some of the tasks that require no human intervention. Anvil, however, assumes that the document template is designed by an expert, requiring only an analysis of the resulting instances that may contain artifacts.

Design Principles are an important part of the document design process. Graphic Designers used them frequently, but not all of them at a time. Lidwell et al. [11] suggest one-hundred (100) principles for a variety of purposes and cases, including printed media. Santos-Villalobos [1] also did some research on design principles, finding *Alignment*, *White Space* and *Strong Lines* very useful for the KBAR's approach. Although it is important to consider design principles for the analysis of a document's layout, we found even more interesting Lester's [14] argument who emphasizes that graphic designers can follow or challenge these principles, which is why he prefers calling them design suggestions, not rules or principles.

To support our logic on why the document analysis process should not depend solely on design principles, we did some research on perception from a sociological perspective. Lester [14] describes a 3-step process originally defined by Aldous Huxley: Sense, Select, Perceive. According to Huxley, the sum of the three equals "seeing". Another important contribution regarding perception comes from anthropologist E.T. Hall [4], who argues that *culture*, in its most general sense, is one of the most important factors that affect perception because it provides a highly selective screen between a person and the outside world. Toussaint [15] provides a more practical example of perception using the Muller-Lyer

illusion, adding that by understanding the *context* of what is being perceived can ultimately affect what is being perceived.

The research on the use of CBR in design [17] provided insight on practical approaches to our work. CADRE requires that *context* be provided in the form of design parameters of a building being designed, while CADET can proactively let the designer know what is permissible or not during the design process. The latter is very useful in preventing problems from developing during the design phase of electro-mechanical devices. Neither of these approaches can be used in a practical manner for our layout analysis process because this would imply requiring all the characteristics and specification of the document being designed beforehand, which is not common.

Finally, we provided information on a few of approaches that have influenced our layout analysis process. The use of esthetic measurements on a layout [18] is a practical method to characterize a document layout, which is one of many methods to search for cases in a CBR. Yet, these measurements would not be able to distinguish local variations in the layout because these measurements are more effective with Global variations. A more effective solution for our approach is that of a *Template Matching* method developed by Peng et al. [3], which is sensitive to local variations on a document layout due to better representation of spatial relationship with the elements of a layout. By representing a document layout as a numerical vector, the cost of computation is low due to the simplicity of such computations.

3 Towards a New Document-Analysis Approach

In design, ambiguity can be perfectly acceptable. "*Right and wrong does not exist in graphic design. There is only effective and non-effective communication*" says Peter Bilak as quoted on [19]. In science, however, ambiguity is not always welcome.

Although the subject of document-analysis is nothing new, there is much to be researched still as it involves many dimensions of complexity. While some approaches pretend to create a logical structure from a document image [16] [20] [21], others find errors in a document's layout design by comparing it to its approved template [1] [2]. However, when searching for a methodology to determine how pleasant, aesthetic or well designed a particular document is, very few alternatives can be found.

In [1], the Knowledge-Based Artifact Recognition (KBAR) model was designed to analyze changes in the layout of a document, known as the job instance, by using as a reference a template document, the approved instance. This approach is appropriate for Variable Data Printing. In contrast, the new document-analysis method is intended to analyze the design of a layout, a simple document. Therefore, the KBAR model cannot be used to develop our document-analysis approach. It becomes useful to identify the logical components of a document for applications like digital libraries. When these components are identified, documents can be analyzed, classified or understood [16] [22]; or a document's content and layout can be modified or reused [21]. The KBAR [1] applies the methodology of document understanding behind [16] to determine a document's type and thus the applicable document analysis. Although identifying logical components would allow making

inferences of a document's layout aesthetics, the question remains on what is the best way to make inferences about these components.

Books on Design Principles and how to apply them are abundant [11] [12] [13] [23] [24] [25]. Yet these principles are presented in a rather subjective manner, which is of little help when trying to develop an approach that can measure the presence of such principles in a particular document. A mathematical approach to measure design principles was needed.

3.1 Measuring Layout Aesthetics

In our research, we found a few sources and references on design composition and layout that from a theoretical standpoint could explain what a good design was. One of these sources used words like continuity, balance, unity, rhythm and sequence [26]. Another would consider a layout as "good" if it followed three basic criteria: It works, it organizes, and it attracts viewers [27]. This did provide some initial guidance, but it was still not practical – the reason being that none provided a quantitative method to apply these concepts in a computational manner. Finally we stumbled upon a method by Ngo et al [18] used originally to measure user interface layouts.

A method developed by Tullis in the 1980's was used to measure how well designed a screen was. At the time, good design was measured by predicting user performance using static alphanumeric displays, very common at that time. The research expands this concept with today's Graphical User Interfaces (GUIs), certainly more complex than yesteryear static alphanumeric displays, by focusing on the perception of structure created by such concepts as

spacing and borders. The application of aesthetic concepts, they argue [18], can aid an interface's:

- Acceptability – there's a high correlation between user's perceptions of interface esthetics and usability
- Learnability – esthetically pleasing layouts have a definite effect on user's motivation to learn how to use a system

Document aesthetics is not too distinct from GUIs, as document perception is equally important. In total, 13 characteristics or measurements are suggested. Among them we selected three for testing purposes: Balance, Equilibrium and Symmetry. These three were specifically selected because they have been suggested by others [23] [26] as important characteristics that must be present in a layout.

Preliminary results were as expected and in line with the method's description. But there were a few issues at hand. First, a fourteenth measure of *Order and Complexity* is written as an aggregate of the previously mentioned thirteen measurements. The equation for *Order and Complexity* is written as:

$$\frac{1}{13}\sum_{i}^{13}\alpha_i M_i \in [0,1], \qquad 0 \le \alpha \le 1 \qquad \text{Equation 3.1}$$

where each measure M_i has its own weighting component α_i, which is assumed to be a constant. The issue here is one of optimizing the value of every α_i for each type of document to be analyzed. Determining weights is one of the multidimensional optimization problems that are application specific, possibly solved using objective-based evolutionary programming [18] which is out of the scope of this research.

The second issue is created by the first one. If weights were to be optimized for specific types of document designs, it would limit the system's flexibility to analyze any type of document layout without preference to one style over any other. This issue also implies that it may constrain its application to a specific culture's aesthetics preconceptions, which is explained later in this chapter.

Up to a certain point esthetic layout measurements was an ideal method. In fact, it was one of very few who have presented aesthetics in a quantitative manner. In search of new clues for a different document-analysis approach, we found useful to know that the interpretation of a graphic design's intent can be considered as *quite subjective* [28]. To understand what influences subjectivity when interpreting a design it became necessary to analyze the subject from a very distinct point of view.

3.2 Subjectivity in the analysis

> *One of the functions of culture is to provide a highly selective screen between man and the outside world.[…] culture therefore designates what we pay attention to and what we ignore.* [4]

In chapter 2, the issue of how cultural influences have a tremendous impact on perception was presented. A document, in its most generalized term, is an arrangement of elements that in unison transmit a message to its intended audience, be it an individual or a specific group of people. The language used is a visual one and, similarly to spoken and written language, is subject to a variety of interpretations. In this sense, Hall [4] argues that interpretation is dependent on the current context provided and on the culture. To know how

efficient context is being communicated in a document's design, we would have to know design principles commonly used by expert designers under a specific culture.

To complement Hall's definition of what a culture is we cite anthropologist Franc Boas as quoted on [14] who says that culture is *the community of emotional life that rises from our everyday habits*. Today's habits may influence future ones, but they could hardly define them. In fact, Lester [14] argues that attempting to identify "good" graphic design is dangerous because, like beauty, it is a subjective determination. He continues with the following important assertion: "*What is considered good design changes over time and varies among cultures"*.

On the importance of context, Toussaint [15] notes how the magnitude of Muller-Lyer's illusion, presented in Chapter 2, varies between cultures. Toussaint also argues on the importance of using context for pattern recognition problems. For example, when an object Z is observed in isolation (see left side of Fig. 3.1) it may need the help of some context A (Fig. 3.1, right side) to see Z as what it is.

It is then clear from an anthropological perspective that culture defines an individual's perceptions and preconceptions, which may vary between cultures and with time. Thus, Esthetic Measurements as defined in [18] cannot be used for our purposes as they are both culture- and time-dependent. To develop a method that could be considered culture-adaptive, and thus style-adaptive, a different approach to document-analysis was necessary.

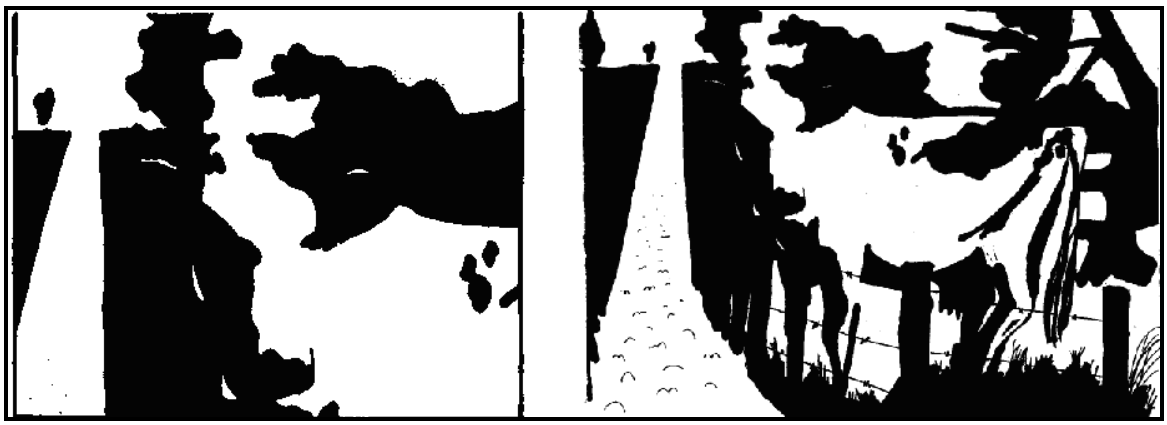

Figure 3.1 – The need of context in visual perception taken from [15]. Used with permission by Jim Adams.

3.3 Document template matching

> *A logical next step in exploring the theories of layout is the use of simple objects as pictorial elements on a page.* [23]

Questions on why was such a method preferred over Esthetic Layout Measurements will eventually be explained. The argument explains that the process of analyzing the aesthetics of a document can be subjective because of the viewer's perception, a concept influenced by culture. As such, esthetic measurements as defined in [18] are not ideal if we are to make a document-analysis method culture-adaptive. To this we add the fact that good design also changes over time, which would possibly require frequent recalibration of the weighting component α_i for those measurements. Then, for a method to be culture, style and time-adaptive it must infer that a document has a good design when its layout is similar to other documents previously considered as well designed within the same framework of time and culture. The same is true when inferring that a document has a bad design when a similar badly designed document is found. A database with a variety of documents, with both aesthetic and non-aesthetic layout designs, would represent the current trend of document

styles for a particular culture, at a particular time. Documents are represented as cases from a Case-based reasoning database. In [1], artifacts that can appear on a document are represented as cases in a Case-based reasoning database.

At first, the values obtained by the three esthetic measurements selected from [18] were used in the case search criteria. The argument was that similar documents would have similar esthetic measurements. However, there was no easy way to prove or disprove this assertion. Also, calculating the distance between documents – that is measuring the difference between the symmetry, balance and equilibrium of two documents – and determining threshold values for each was no simple task. So, instead of calculating esthetic differences between documents a better approach would be to calculate geometrically the difference between documents.

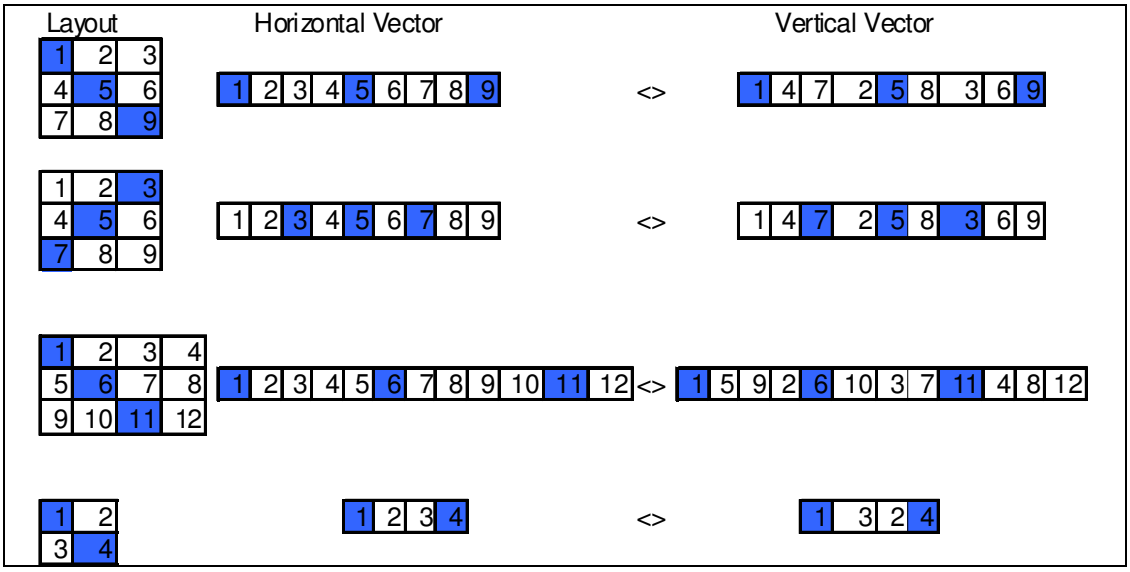

Figure 3.2 – Four CBR samples (Layout) with their respective CBP's

Not only would we need an effective method to calculate geometrical distance between documents, the algorithm should be simple to implement and efficient. The selected method is based on the global matching of Component Block Projections (CBP) shown in Figure 3.2, which are the concatenated directional projection vectors of the component blocks of a document layout [3]. Directional projection vectors are obtained from the Component Block Representation (CBR) in Figure 3.3, a binary image of rectangular boxes where foreground pixels – i.e. box edges – take value 1 and background pixels take value 0.

0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	1	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	0	0	1	0	0	0
0	0	1	0	0	0	0	0	0	0	1	0	0
0	1	0	0	0	0	0	0	0	0	0	1	0
0	1	1	1	1	1	1	1	1	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3.3 – A Component Block Representation (CBR) of a triangle shape

To calculate the geometrical distance between the document analyzed and those in the Case-base a simple computation of distance is calculated using their respective CBP's. Details on the calculation, test results, and comments on the implementation of this method are explained in subsequent chapters. Finding the case *T* in the database that best matches the document analyzed *Q* is equivalent to finding:

$$\Delta g(T) = g(Q) - g(T) \quad \rightarrow \text{Equation 3.2}$$

where $g(Q)$ and $g(T)$ are the concatenated vectors of the input CBR of Q, the document analyzed, and the corresponding CBR of *T*, one of the many cases in the database [3].

3.4 Chapter Review

We began by researching methods where design principles are quantitatively measured. A method in [18] proposes thirteen design principles that can quantitatively be calculated and expressed. From those thirteen measurements a fourteenth is created from the resulting values of the other measurements and each one has a weighting component, a constant value between 0 and 1, inclusively, that determines the level of importance of each measure for a specific document type. As explained, this is a multidimensional optimization problem that, first, is out of the scope of our research and, second, because of frequent changes in culture, the optimized values may lose their effectiveness due to changes in design trends.

A different perspective was needed, and as a result, we searched for a definition of perception and how it is affected by culture from an anthropological point of view. Hall [4] and Lester [14] provide relevant insight on the function of culture in perception and why deciding that a design is “good” can be a subjective determination, while Toussaint [15] argues in favor of the role of context in Pattern Recognition applications. The next step was to find a method that calculated how geometrically similar two documents are. In [3] we see both an easy to implement and a quick-execution algorithm to do such an operation.

4 Geometrical Document Matching Method

In the previous chapter it was explained how perception of aesthetics may vary between cultures. Although a method to measure layout esthetics [18] is a good alternative to quantify pleasantness of a design, the approach is culture dependent. We then argued and justified that a culture-adaptive method must quantify the document geometrically instead of measuring layout esthetics, as geometrical measurements are less biased. As a result, such a method was implemented based on an algorithm developed by Peng et al [3], which is described in this chapter.

4.1 Geometry of a Document

We can recall our definition of a document as having a certain amount and combination of content (i.e. text, images), arranged in a specific order known as layout. Take for example a marketing flyer and a newsletter article as seen in Figure 4.1.

The text-to-image ratio for both documents is noticeably opposite, the flyer having the most illustrations whereas the newsletter has more text than anything else. Also, the layout is considerably different, the newsletter being the most symmetric and structured. Aesthetics aside, we tend to look at the document's layout as an arrangement of blocks of text and/or illustrations. In fact, we don't necessarily judge how pleasant a document is based on its content, but on its layout. A similar premise was used to test the effectiveness of esthetic measurements in layouts by Ngo et al [18]. The experiment consisted of having a

group of people rate the “beauty” of a layout on a low/medium/high scale. At no moment they were instructed to consider the content.

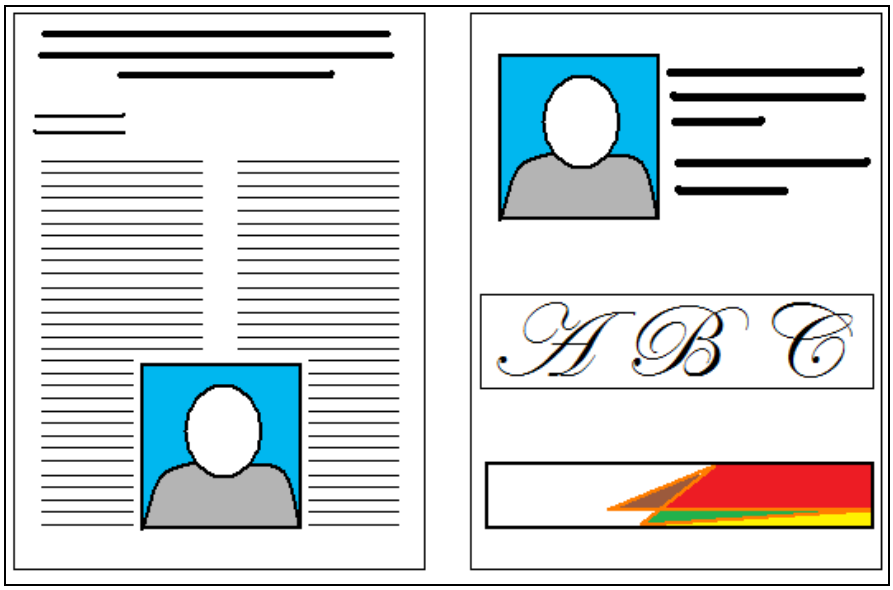

Figure 4.1 – Side-by-side document layouts
Newsletter (left), Marketing Flyer (right)

This arrangement of blocks in a document is what we call the **Geometry of a Document**, which is the basis of the implemented methodology of our work. By looking at a document as an arrangement of blocks we can then compare the similarity of one specific document instance with a second instance that has already been categorized as pleasant or not.

It is then clear that the method to determine if an analyzed document is pleasant or unpleasant is dependent on a matching algorithm that can find a highly similar instance as pictured in Figure 4.2. The best matched instance can tell us if the analyzed document is considered pleasant or not. A matching algorithm developed by Peng et al [3] was originally designed for document image *registration* and *retrieval* in document image databases. It becomes necessary to modify the original concept by [3] to be adjusted for our implementation. Let us begin with the *registration* phase.

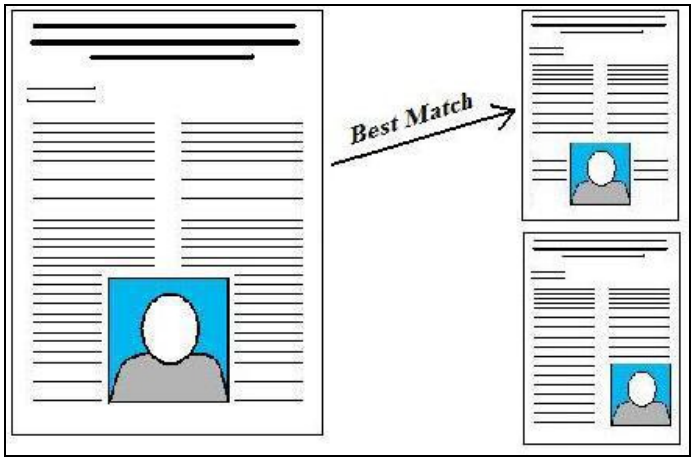

Figure 4.2 – Example of finding best match of a newsletter instance (left) and approved samples (right)

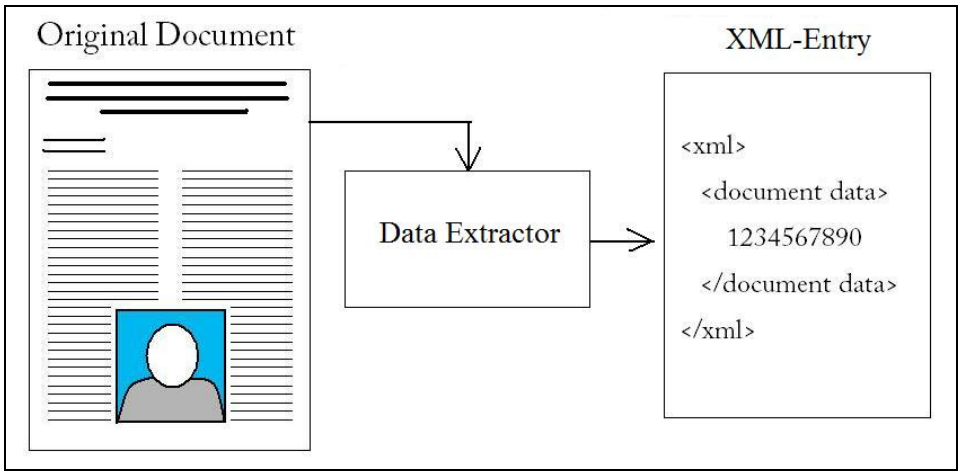

Figure 4.3 – Document registration process.

Because documents on our matching algorithm implementation are assumed to be in an XML-based file format – known as PPML – that contains data about the size, location and content of a document's composing blocks, document *registration*, depicted in figure 4.3, is reduced to simply obtaining the relevant data from the original document (PPML). This includes the size and location of the composing blocks. The content is not registered as it is not relevant during the matching process. A more thorough explanation of the case data is found in the next chapter, which describes the case-base structure in more detail.

4.2 Document Representation as Vectors

The authors in [3] suggest representing the layout's composing blocks of a document in a so called Component Block Projections, or CBP. Simply stated, a CBP is a one-dimensional data structure created from a two-dimensional pixel-by-pixel representation of a document. The concept was explained in Chapter 3 (see Fig. 3.3). These vectors are created from the XML-entries in the case-base.

4.3 Document Matching Algorithm

The *retrieval* part is mostly associated with the matching process, where a document instance represented as a CBP is compared to instances of a database of documents, which we name Template Component Block Projections or TCBP. A pseudo-code of this algorithm is illustrated in figure 4.4.

Distance between the CBP and a TCBP is calculated as denoted in equation 4.1:

$$\frac{\sum_{i=1}^{k} \left| CBP[i] - TCBP[i] \right|}{k}$$ Equation 4.1

where the dividend is the Sum of Absolute Differences (SAD) of each element in the CBP and TCBP vectors. The SAD is averaged with the divisor *k*, the length of the vectors. This should not be confused with *Mean Difference*, a measure of statistical dispersion equal to the average absolute difference of two independent values drawn from a probability distribution. Note that this applies when both CBP and TCBP have the same length *k*. If any of the two vectors is smaller than *k*, then we concatenate *k – length(shortest vector)* zeroes to the shortest vector.

Document Matching Algorithm

```
Create vector from document to be analyzed (CBP)

For each case in the database (all TCBP's)
   -Create vector from case (create TCBP)
   -Measure distance between CBP and current TCBP instance
   -If measured difference is the smallest so far, keep a copy of
   current TCBP instance

Next case

If smallest TCBP is within a specified threshold, return it with its
assessment data
```

Figure 4.4 – Document Matching algorithm

Having to measure only the distance between documents eliminates complexities found in the KBAR [1], which proposes the use of weights to determine a tolerance level for artifact detection, known as the Analysis Criterion (AC), for each of the 21-features of an artifact. For test purposes, the AC was manually selected based on trial and error observations.

Santos-Villalobos [1] suggests that this criterion be determined based on the target audience, the document's viewer. This, however, is an optimization problem. We already argued about the non-trivial problem of optimizing several weight values at once, taking as example the Order and Complexity measurement of a document layout as presented in [18] and detailed in equation 3.1. Under our proposed approach, only a simple distance measure between documents is needed as part of the document matching process.

Having a distance measure still poses an issue. Although it is not apparent from the algorithm, selecting the TCBP with the minimum distance does not necessarily mean that it is geometrically similar to the CBP in question. The possibility of having a CBP "match" with a 50% dissimilar TCBP exists when only a handful of TCBP's are registered in the case-

base. It becomes necessary to establish the α-threshold value that determines the minimum acceptable distance for two documents to be considered similar. To find this value some preliminary tests were done with the sample documents available in the database. Then, a confusion matrix with the distance between all sample documents was created, which can be seen in Appendix A. Based on observations of this matrix detailed in chapter 6 it is a safe assumption, for the sample set used, that documents with an α-threshold value lesser than *0.01* are considered geometrically similar.

Once an acceptable match is found, which is within the α-threshold value and has the minimum distance (closest match), then the matched TCBP's supplemental data is retrieved. This supplemental data has two important bits of information, one being an assessment if the document is well designed or not, while the second contains other details, including reasons about the problems the document that is not well designed may have. Figure 4.5 provides a simple flowchart of the process.

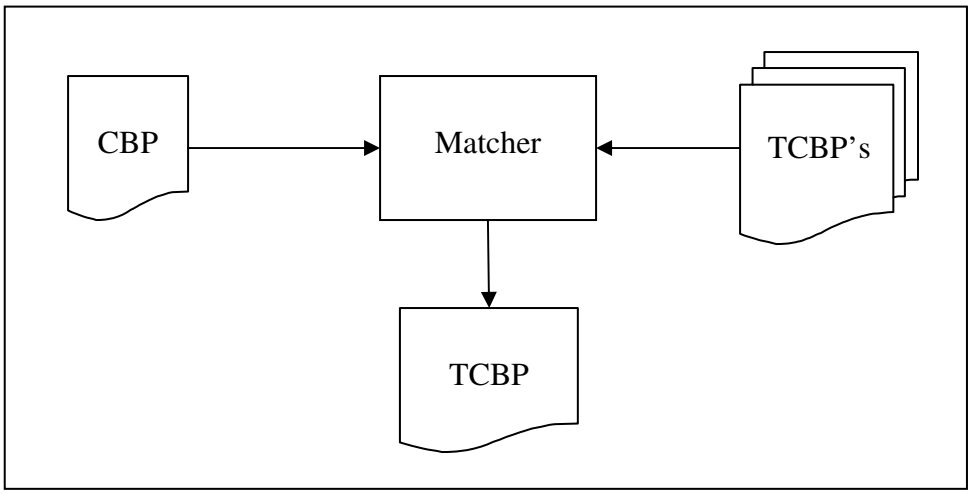

Figure 4.5 – CBL Matcher flowchart

4.3.1 Sensitivity to Geometrical changes

It is not possible to understand how sensitive the matching algorithm proposed by Peng [3] is without presenting a simple example. In figure 4.6 we present two sample documents which we will call document *A* (left) and *B* (right). Both documents have the same dimensions, and the same goes for their components. The only notable difference is that one of those components in *B* is on a different position when compared to *A*, specifically the center was moved one pixel both to the right and down.

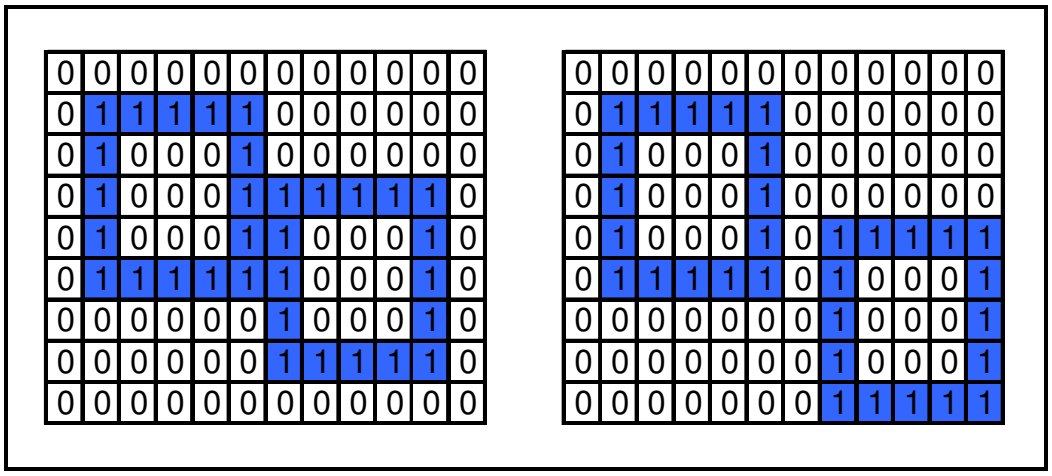

Figure 4.6 – Two small documents represented in pixels.
Pixels of components' borders are outlined in blue.
Boder pixels are denoted by 1's, and 0's for non-border pixels.

After creating the vectors from documents *A* and *B*, we get from equation 4.1 that the absolute difference sum (the dividend of the equation) is 56. The value k in the divisor is 216, and is determined by the length of the longest vector. The resulting difference is 0.259259259, a 26% difference between both documents. The reason behind such a big number is attributed to one important factor, the size of the components relative to the document's dimensions. Smaller components will have a smaller impact on the resulting difference. In fact, removing the translated component from document *B* will have a lesser

impact on the resulting difference, as its borders would be unaccounted for in the projection vector.

4.4 Chapter Review

In this chapter we provide more details regarding the Document Matching Method implemented. The matching algorithm is based on Peng's [3] document image template matching algorithm that considers the Geometry of a document while ignoring the content itself. This makes it a quick, efficient method to match document layouts using simple distance measurements.

An important observation regarding the method's sensitivity to geometrical changes was documented. Large components strongly influence the resulting distance difference with another document vs. missing components which have a much smaller influence in the resulting distance.

5 The Use of Case-Based Reasoning in Design

Case-based reasoning is the process of solving new problems based on the solutions of similar past problems. Among the numerous existing implementations of Case-based reasoning, its use in applications related to document design is not common. To put this idea into perspective, we compare some issues that exist on our implementation with those already present in *Archie*, a Case-Based Architectural Design Support system [29]. We also provide additional details on the conception of a case's structure and its limitations under our implementation as well. Possible scenarios are presented to explain some of these details.

5.1 Retrieve, Critique, Adapt: Archie's Cycle

Archie [29] is a support system used by architects during the design process of a building. It provides access to office building designs created by other architects (Retrieve) based on a set of design goals and constraints, and points to factors that must be considered in solving a given design problem (Critique).

For example, an architect concerned with the quality of lighting can select one of several qualitative domain models in Archie that represent the perspectives of different experts on that aspect of the design. The retrieved case contains comments as text annotations detailing what issues could arise (e.g. problems caused by artificial lighting) and an assessment (e.g. use diffusers over bulbs) [29]. After one aspect of the design is fixed, the architect can continue critiquing the design from other perspectives until satisfied with the design (Adapt).

There are more than 20 case features associated with lightning quality alone. There are also design cases with information of previous building design cases that have a broad list of features, including company information and type of furniture used. These features are divided in terms of Goals, Plan and Outcomes. The example of such a case present in Archie has almost 50 features. Despite that the case robustness found in Archie could serve as a foundation for our case-base structure, there are a handful of reasons why a similar implementation is not possible for our document design assessment method. Archie's supported user is an architect, an individual who has extensive knowledge on a particular subject (building design). Our implementation assumes no specific supported user; in fact it could be assumed that the user has little or no document design experience or knowledge about design principles, which is why the provided assessment should be simple to understand. Also, Archie's cases are used to support the conceptual design process – propose a design based on given constraints and other specifications – as opposed to the detailed design, such as drawing and drafting, numerical calculations, among others [29]. Our implementation assumes that a document is submitted for analysis after the design process is completed. Again, Archie is used during the design process while our implementation is used after the design process is completed.

Although supporting the document design process in the same way Archie supports building design would appear to be a good idea, it would become a daunting task that cannot be achieved due to the scope of our work. To put this into perspective, we refer to some of the difficulties found by Archie's designers [29]:

1) **Incomplete cases** – while there is an abundance of architectural designs that can be used as reference, it is not easy to find well-documented cases of office building designs.
2) **Large cases** – a well-documented Archie case might contain a huge amount of useful information about clients, design goals and constraints, designers, design history, the design plan, and so on. Much effort is required to gather all this information. This creates a case representation and indexing issue, as well.

To ease the process of documenting and adding cases to the database on our implementation, we made the cases smaller in terms of features and attributes. Smaller cases are easier to document, represent and index.

5.2 Conception and Structure of a Case

On designing the structure and features of cases, the initial objective – a culturally-adaptive assessment method – of our implementation was the most influential factor. Simplicity was also a considerable factor, as can be noticed from the relatively small amount of features in a case.

With this in mind, the structure of the cases and its features were considerably influenced by the two matching methods considered. The first one would use esthetic layout measurements [18] to categorize cases based on three esthetic measurements: balance, symmetry and equilibrium. The use of these features as case-search attributes was discarded for reasons already explained in chapter 3, even though these measurements are still included in the cases for any future use. The concept behind the second matching method was that of

geometrical comparison between the document being analyzed and a set of existing documents [3]. This is the reason why cases contain data about a document's dimensions and its composing object's position and size. To better understand this, observe the structure of a case in figure 5.1:

```
<assessment-case case-number="#">
	<case-id>CaseFileName.ppmlt</case-id>
	<assessment-data>
		<good-bad> </good-bad>
		<good-case-id/>
		<assessment/>
	</assessment-data>
	<esthetic-measures>
		<symmetry general="#.###" horizontal="#.###"
		radial="#.###" vertical="#.###"/>
		<balance general="#.###" horizontal="#.###"
		vertical="#.###"/>
		<equilibrium general="#.###" horizontal="#.###"
		vertical="-#.###"/>
	</esthetic-measures>
	<layout-objects page-height="###" page-width="###">
		<object height="##.#" id="1" type="image" width="##.#"
		x="##.#" y="##.#"/>
		<object height="##.#" id="2" type="text" width="##.#"
		x="##.#" y="##.#"/>
	</layout-objects>
</assessment-case>
```

Figure 5.1 – Structure of a case in XML-like format

The data inside the "<esthetic-measures>" tag refers to esthetic measurements calculated from [18] formula definitions of Balance, Equilibrium and Symmetry. A more thorough explanation of the formula definitions of these three measurements is found later in this chapter.

The "<layout-objects>" tag contains data about the composing blocks – position, dimensions and type, be it text or image – and the dimensions of the document. This data is used by the matching process implemented, based on [3] document matching algorithm.

The “<assessment-data>” is considered as an informational attribute, meaning that its data is not used to match a case but to instead provide information regarding the analysis previously made by an expert designer of that particular instance. In Archie, for example, the data composing a case depends on the availability of a design’s analysis, which may include justifications for an architect’s design decisions and outcomes of the design. The authors [29] argue about Archie that, despite the abundance of cases under the domain of architectural design, it is not easy to find well-documented cases of office building designs. It is not hard to believe then that the same situation occurs for document design. In any case, the assessment data of a case is provided, preferably, by an expert designer. This expert determines if the layout design of the document, to which a case refers to, is considered a good or bad design (see “<good-bad>” tag). If the layout, represented in the case, is considered as bad then an assessment is provided on why the analyzed document has design flaws (see “<assessment>” tag). It also provides, if available in the database, a similar case whose design is considered “good” that can be given as a reference (see “<good-case-id>” tag). The assessment data is not part of the original document, but is included as supplemental data in a separate file and inserted in the case-base during the case *Registration* process.

5.2.1 Extracting geometric data for Case registration

The process of registering a case in the database requires that geometric data from the PPML file be extracted. This extraction process is very similar, with a few adjustments to recognize PPML tags, to the analysis process of the document segmentation and

understanding module in the KBAR [1] that was originally designed to work with XSL-FO[12]. This module is detailed in figure 5.2.

In a top-down strategy, the Geometric Extractor is recursively applied to all children nodes of the PPML document. Relevant nodes are those that represent mainly text and image containers in the documents. For each of those nodes, the geometric features are extracted. For example, the node representing the document contains the width and height of the page while its children nodes (text/image containers) will have geometric information regarding their position and dimensions. After all geometric data is extracted from the PPML it is then registered in the case database with an XML-like format seen in figure 5.1.

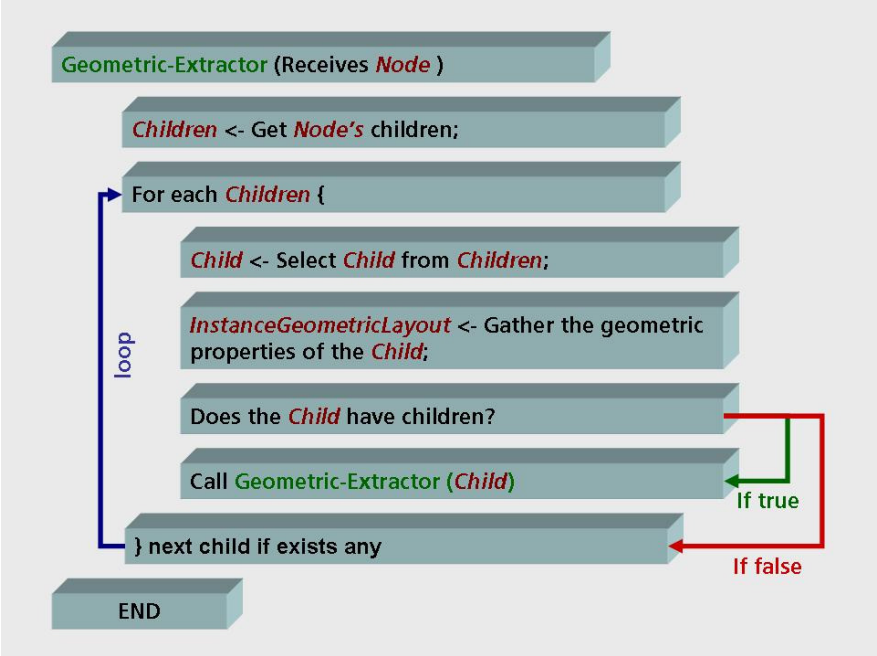

Figure 5.2 –Geometric properties extraction pseudo-code taken from[1]

[12] "eXtensible Stylesheet Language - Formatting Objects". More information can be found at the World Wide Web Consortium's website: http://www.w3.org/TR/xsl/, Retrieved March 10, 2010.

5.3 Esthetic Measurements explained: About Balance, Equilibrium and Symmetry

Research regarding aesthetics in design may very well be abundant, and a considerable amount focuses on its application to layout and graphic design [12] [13] [14] [16] [19] [23] [24] [25] [26] [27]. In spite of the overwhelming amount of literature available on the subject, most of the literature explains layout aesthetics conceptually. Yes, many agree that some design principles must be present in a layout's design like symmetry, closure, balance and unity, but none have actually dealt with or explain them in a quantitative manner.

As already explained in chapter 3, on our literature survey we found a method that can quantitatively measure several esthetic aspects of computer screen layouts [18], aspects that are easily transferable to document layouts. Out of 13 of these aspects, called esthetic measurements, we selected Balance, Equilibrium, and Symmetry, as they are prevalent in literature related to graphic design. The following explanations are obtained from [18]:

- **Balance** – defined as the distribution of optical weight, the perception that some objects appear heavier than others, in a picture. Balance is achieved by providing and equal weight of screen elements, left and right, top and bottom.
- **Equilibrium** – similar to balance in some ways, yet it goes beyond it because it considers page dimensions relative to all objects in the layout. Equilibrium is a midway center of suspension. It is accomplished through centering the layout itself.
- **Symmetry** – by dividing the layout into 4 quadrants – upper left, upper right, lower left, lower right – axial duplication, a unit on one side of the center line is

exactly replicated on the other side, can be determined. In this way, symmetry can measure a layout vertically, horizontally and diagonally (radial).

We then had a way to measure, more or less, aesthetics in a layout. The question remained on how to interpret the results of these measurements or how to use them in some other way. From the three esthetic measurement formulas we get 10 results. For Symmetry we get Vertical, Horizontal and Radial Symmetry, and of course the resulting Symmetry measure obtained by using the previous three measurements mentioned. In addition, Balance and Equilibrium yield 3 results, each. The range of values expected in seven of the results is a number within the range [-1,1], while three of them range within [0,1]. There is no specific procedure on how to interpret these values. Take Balance, for example. At times, it is a good idea to have large elements on the upper left quadrant of a layout as this is one of the first areas our eyes focuses on when looking at a layout. A desirable value for both Horizontal and Vertical Balance would be a number close to or lesser than 0, which implies that the optical weight lies in the upper half (horizontal) and left side (vertical) of the document. Therefore, interpreting a set of values that is acceptable for the esthetic measurements – based on a set of parameters and constraints that depend on the document type and document intent, among others – becomes an optimization problem. This problem may very well be outside the scope of our intended work as it would complicate the case-matching process – since the interpretation of such values would require a great deal of information, analysis and validation on a variety of document layout types. Moreover, it would limit the implementation to work for a specific set of layout types under a specific culture.

Human expertise will always be required to determine the current relevance of aesthetics features in a layout. In fact, Roger P. Schank, one of the leading researchers in the artificial intelligence field, wrote the following in regards to the *Generalization* of human expertise in understanding new problems by generalizing previous ones:

> *Ultimately human expertise is embodied not in rules but in cases. People can abstract rules about what they do of course, but the essence of their expertise, that part which is used in the most complex cases, is derived from particular and rather singular cases that stand out in their minds. The job of the expert is to find the most relevant case to reason from in any given instance.* [30]

The application of human expertise for analyzing a document's layout is best ascertained geometrically, while the representation of human expertise is best embodied as past cases. We use this argument to theoretically justify the use of a method that measures a layout geometrically.

It has already been explained in previous chapters that these esthetic measurements are not used during the matching process, but were left in case they are used in future work.

5.4 Limitations

As mentioned earlier, the structure of cases in the document design assessment implementation is influenced by the objective to be met, which is a culturally-adaptive assessment method. Deciding over a generalized case matching method for document design assessment was much more complex than its development, which was rather simple and straightforward. Yet, under all that simplicity of the implementation, limitations will certainly arise.

Despite not being able to support the design process – as Archie does with building design – due to a reduced amount of case robustness, our implementation is still able to provide some level of document design assessment without requiring an excessive amount of case data. In fact, the idea is to avoid as much as possible the need for case data to be provided when documenting a case a when searching for a case in the database. Still, having a reduced amount of documented data on a case does not make our implementation as useful in document design as Archie can be with architectural design.

Another aspect of the case structure that might create limitations on the case search process is the lack of case categorization. Although it is very common that cases are categorized to facilitate the querying and matching of cases in a database, we do not categorize cases in any way on our implementation for one simple reason. Aesthetics of a document's layout are not necessarily determined by its content. In fact, it is best determined by the arrangement of its composing blocks as noted on [18]. By not stamping cases with a category name, the matching process is not limited to any specific set of document types, but at the same time a more effective analysis becomes difficult as distance thresholds cannot be intuitively adjusted given certain parameters, which could include a document's type or category.

What happens when there are no matching cases? It is not mentioned explicitly in [29] what happens when there is no case that matches a particular design. Some insight can be provided from the perspective of our document design assessment implementation. The fact that no matching case could be found does not imply anything in particular about the aesthetics of the analyzed document. It doesn't mean that its design is either good or bad, but

simply that an assessment cannot be provided. The only action that can be taken is to have an expert analyze it, have him/her provide the proper assessment data and include this data when adding the case to the database.

5.5 Chapter Review

The purpose of this chapter was to provide additional insight on the structure of the cases in the database used on our implementation of a document design assessment method. The features and attributes of the cases were selected based on the matching method selected, which was already explained in detail in Chapter 4.

A brief conceptual discussion of the esthetic measurements that compose the cases in the database was also included. We argue, among other things, that using Esthetic Measurements "as is" does not consider currently accepted trends, thus requiring human subjective knowledge in providing layout assessment within a culture. Subjectivity of aesthetics made Case-based Reasoning a more effective approach.

In addition, we made parallels with Archie, an architectural design support system that also uses Case-based reasoning to achieve its objectives. Although our implementation is not as robust as Archie, we did consider some of the difficulties encountered when handling cases with an immense amount of features that require a great deal of documentation. The simplicity in our implementation looks to avoid these difficulties.

Finally, some possible limitations inherited from the simplicity of the implementation are detailed. Case robustness is reduced in favor of keeping simplicity of the matching and assessment process. Case categorization is not implemented so as to avoid constraining the

implementation to specific types of documents. Also, when no case match is obtained, no assessment can be provided.

6 Tests, Results, and Analysis

Analyzing results from tests become a challenge when there is no comparable system to match it up with. Our document design analysis implementation is no exception. Nevertheless, results must provide a palpable picture of what the system can and cannot do.

6.1 Test sample Distribution

Twenty-three (23) distinct documents were created, the majority within three types of commonly known layout designs: Newsletters, Flyers and Brochures. A small number of samples that do not fall in any of these categories were added to create some "noise" during the testing, to better analyze the effectiveness of our implemented method. Figure 6.1 provides a more visual representation of such distribution by means of a pie chart.

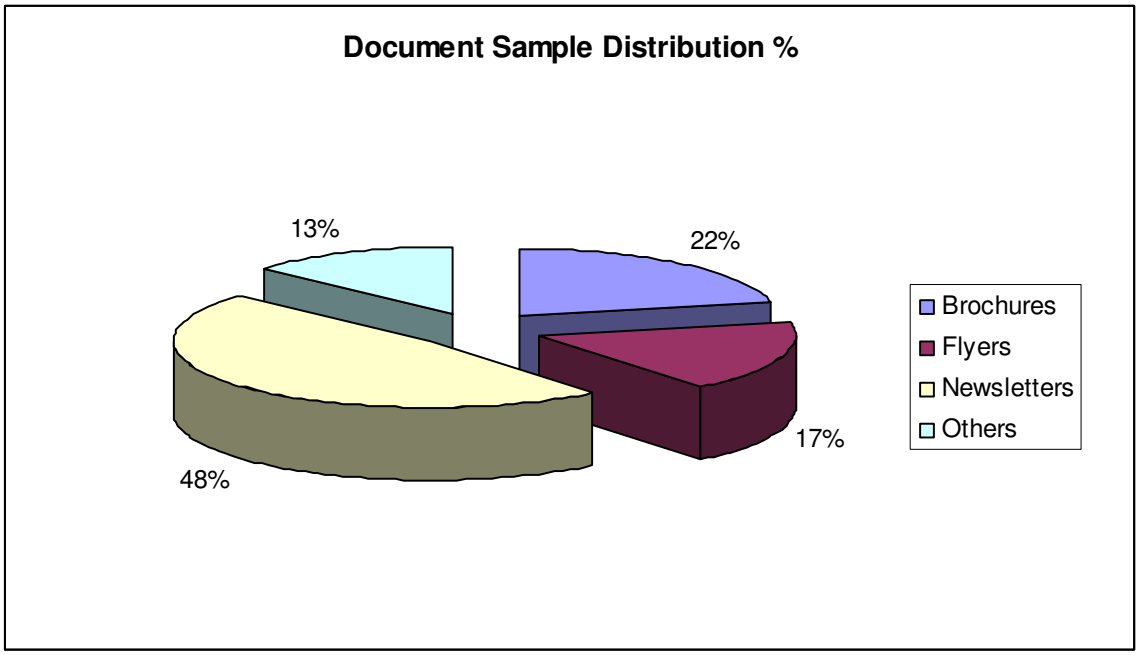

Figure 6.1 – Distribution of document samples in percentage terms

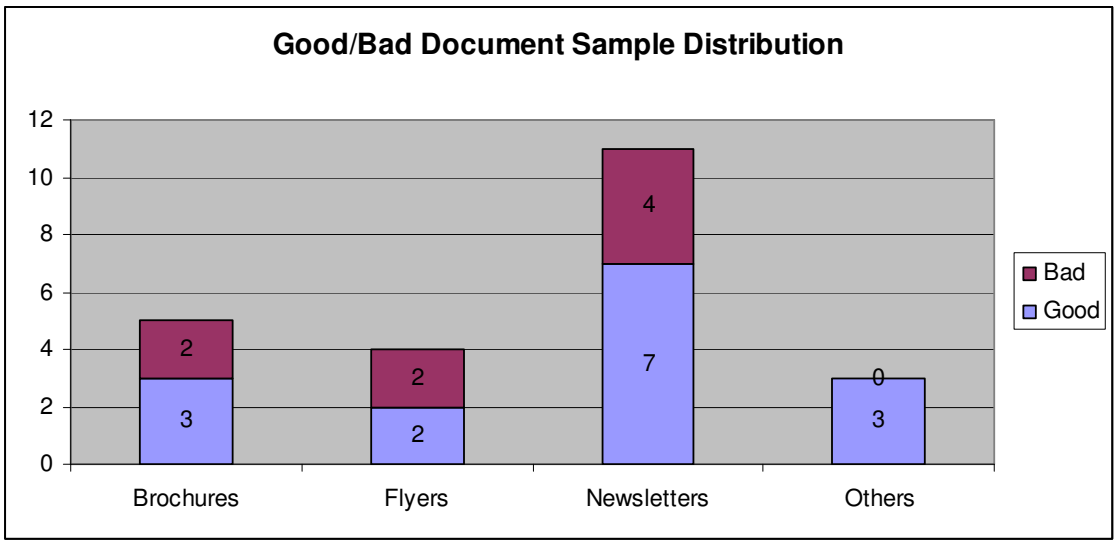

Figure 6.2 – Absolute distribution of good/bad documents for each document category within the sample set.

Even though the distribution does not appear to be equally divided, as almost fifty percent of the tests samples are Newsletters, this was done on purpose. Such a distribution represents a more realistic situation, where document cases may be added through time under no specific criteria. Also, we want to test if the results do not slant to favor Newsletters over other types of documents. This is analyzed in the different subsets of the Distance Matrices later explained in this chapter.

Figure 6.2 graphically shows the distribution, within the sample set, of good and bad documents for each category. A small variety of document samples were taken from various sources, including magazines and the internet. These samples are considered aesthetically well-designed, or referred to in the rest of the chapter as having a good design. The badly designed samples were obtained by making modifications to the well-designed samples in such a way that would categorize a sample as bad. For example, the space between the text columns in our newsletter example, called the gutter, should not be either too small or too big.

There is no formula to determine how small can "small" be, but visually we can agree that it can be categorized as small if there isn't enough space to distinguish where the text of one column ends and the other begins.

Each document in the sample set has a design assessment, which identifies the document as having a good or bad design. When a document is submitted for evaluation by the method implemented and a **match**[13] is found within the case database, the case's assessment data will help determine if such a document's design is acceptable or not. If according to the assessment data the design is not acceptable, then an explanation and possible suggestions can be obtained from the assessment data. An explanation about this assessment data was detailed in Chapter 5.

Documents are categorized in this chapter for explanation purposes. However, the implemented method does not distinguish between document categories. The document matching process is the same for any type of document. Within each category of document samples, variations between samples are minimal. Take for example the two newsletter samples pictured in figure 6.3.

The brochures that are part of the samples set have a 3-sided format. An example of such format is seen in figure 6.4. We assume that the brochure document is of one page. A separate sample could represent the other side of such document. There are many other format variations for brochures, including 2-sided and 4-sided versions, but for testing purposes we decided on 3-sided due to their common use.

---

[13] We refer to a match in the rest of this chapter as the document in the case-base that is most similar with the analyzed document.

Figure 6.3 – Similarly formatted one-page newsletters.
On the left, a newsletter with a runaround has a blue box representing a picture. The text is wrapped around the picture box.

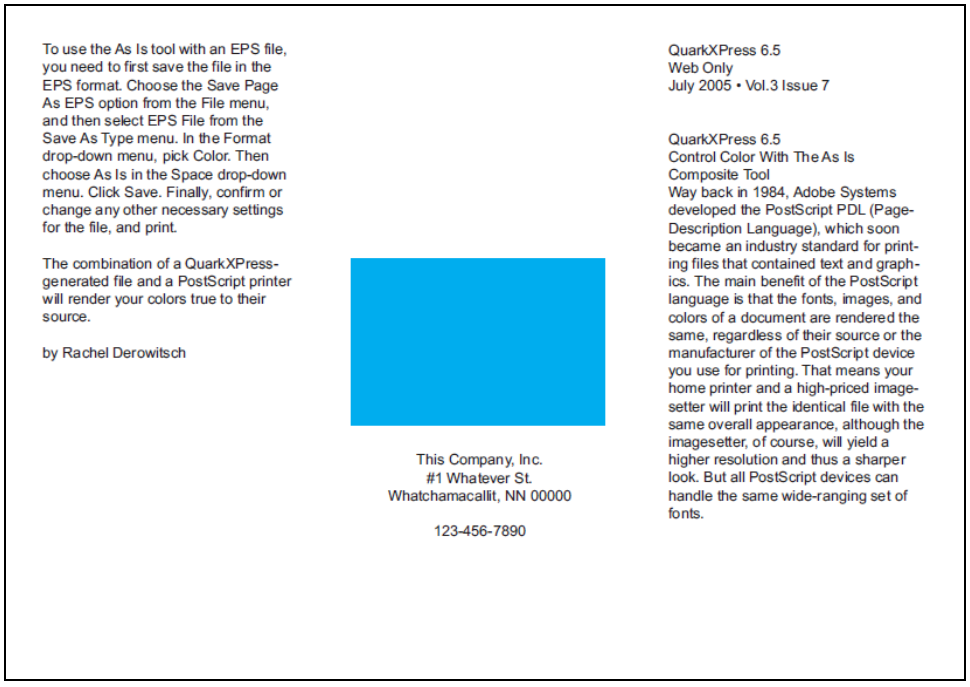

Figure 6.4 – Example of a 3-sided brochure.
The blue box in the middle is an image placeholder.

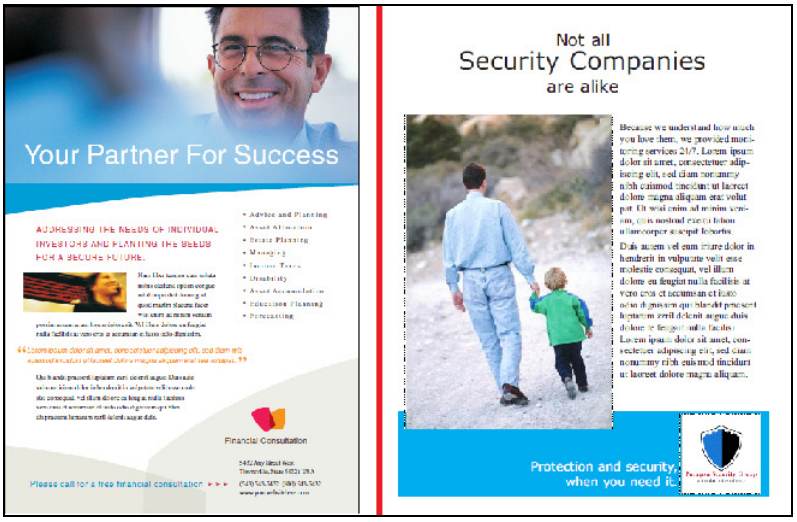

Figure 6.5 – Flyer examples.
Both have very different format structures, but text-to-image ratio is similar.

Flyers, which may also go by the name of ad or poster, have no specific structure as newsletters or brochures may have. They serve a different purpose. The newsletter format is ideal when presenting large amounts of text in a limited area like a letter sized page. Also, for a newsletter format it is assumed that the viewer has enough time to navigate the design. For a flyer, it is assumed that the viewer will only navigate the design for a few seconds, possibly tenths of a second. Due to this reason, the text-to-images ratio for newsletters vs. flyers is almost the opposite. Also, flyers are designed to stand out from everything else around it, which is why no specific design format is followed. Figure 6.5 provides two examples from the sample set, where each one has a very different structure, yet they are consistent having a higher amount of images compared to text.

Design principles are commonly applied during the design process of flyers, but we have argued already that design principles are culture-dependent and thus not easy to universally quantify.

6.2 Esthetic Measurements

We again retake the subject of esthetic measurements, but now from a results standpoint. We have already conceptually described in chapter 5 what Balance, Equilibrium and Symmetry represent in a layout as explained in [18]. In this section we describe them quantitatively, briefly of course, starting with Balance.

Balance is computed as the difference between total weighting of components on each side of the horizontal and vertical axis and is given by

$$Balance = 1 - \frac{|VerticalBalance| + |HorizontalBalance|}{2} \in [0,1]$$

VerticalBalance is the normalized difference between total weighting of objects on each side of the vertical axis, while the same applies to *HorizontalBalance* with respect to the horizontal axis. Negative values indicate in *VerticalBalance*, that the left side is heavier than the right side of the axis, and in the case of *HorizontalBalance* the top half of the frame is heavier than the bottom. Values closer to 0 are desirable if Balance of objects is convenient.

We calculate equilibrium as the difference between the center of mass of the displayed elements and the physical center of the screen and is given by

$$Equilibrium = 1 - \frac{|EquilibriumX| + |EquilibriumY|}{2} \in [0,1]$$

When not fully described, Equilibrium might appear to be computed similar to Balance. But Equilibrium goes beyond by considering both the width and height of the page to compute *EquilibriumX and EquilibriumY*, respectively. *EquilibriumX* is the normalized x-coordinate of the center of mass of the objects. Better values in *EquilibriumX* are related to

how closely the center coincides with that of the page. Positive values indicate that the center is situated on the right side of the frame, and negative values on the left side. *EquilibriumX* is 0 when the center lies somewhere along the y-axis. Conversely, the same applies to *EquilibriumY* but the center of mass is in respect to the x-axis.

Finally, Symmetry is the extent to which the screen is symmetrical in three directions: vertical, horizontal and diagonal (radial), and is given by

$$Symmetry = 1 - \frac{|VerticalSymmetry| + |HorizontalSymmetry| + |RadialSymmetry|}{3} \in [0,1]$$

VerticalSymmetry is the extent to which the layout is symmetrical with respect to the y-axis, while *HorizontalSymmetry* is with respect to the x-axis. *RadialSymmetry* consists of equivalent elements balanced about two or more axes that intersect at a central point, which is why it is also known as diagonal symmetry.

6.2.1 Tests and Analysis

In implementing a proof of concept to calculate the selected esthetic measurements, a few test cases were created to validate the results obtained from the implementation. In fact we created two cases, illustrated in figure 6.6, with the same layout properties defined in [18] as used to demonstrate the computed results of Symmetry. One case is considered a good version of the case while the other is the bad version.

To our surprise, the results we obtained were inconsistent with those in [18], even after numerous discussions about the interpretation of the formulas, and the debugging and reprogramming of the code.

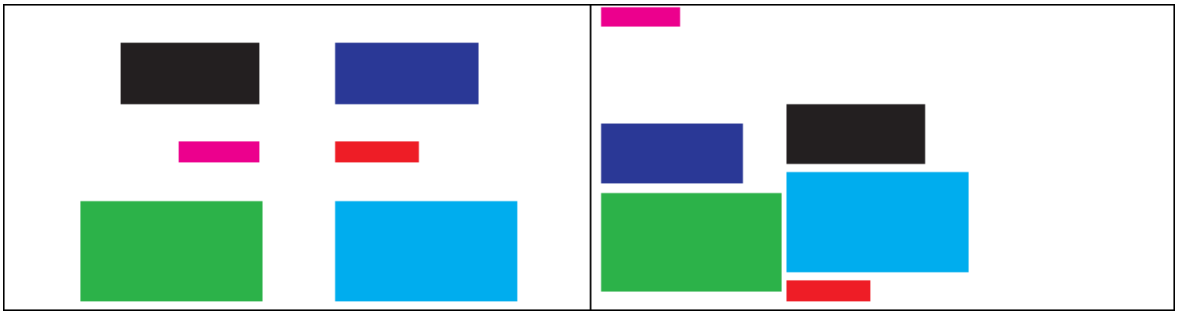

Figure 6.6 – Documents representing a good (left) and bad (right) version of Symmetry. Both versions are part of the "Others" documents sample category.

If we focus only on comparing the results of the "Good" versions in table 6.1, we might consider the possibility of data type constraints having an effect on the precision of results. The computed Symmetry from their implementation was 0.847474 while ours is 0.8499955, a considerable but still relatively small 0.0025255 absolute difference. But if we do the same comparison for the results of the "Bad" versions – same implementation with no adjustments at all – the absolute difference of 0.16600393 is quite alarming and hardly attributable to data type constraints. There is the possibility that their algorithm suffered some modifications in terms of the equations they described, or the test layout used for the "Bad" version computation was not the same one described in their paper, among others possibilities that could not be verified with the authors. Nevertheless, the rationale behind the equations was solid enough to consider them "as is", yet we already argued in chapter 5, section 3, against using these measurements "as is" and rely instead on a CBR approach.

TABLE 6.1 – Side-by-side comparison of results obtained vs. results published by Ngo

	Ngo et al. results [18]		Our results	
	Good version	**Bad version**	**Good version**	**Bad version**
Vertical Symmetry	0.01229	0.79328	0.015767774	0.69525814
Horizontal Symmetry	0.22264	0.60672	0.21712287	0.30474183
Radial Symmetry	0.22264	0.79328	0.21712287	0.69525814
Symmetry	0.84747	0.26891	0.8499955	0.43491393

TABLE 6.2 – Summary of layout properties

Layout Name	Object	X	Y	Height	Width
News-2-Runaround	1	325.5	492.5	236	250
	2	325.5	212.976	77	250
	3	209	28	107	368
	4	37	756	3	541
	5	36	143	12	540
	6	37	154	3	541
	7	37	141	3	541
	8	36	28	110	540
	9	36	180	27	540
	10	36	213	87	249
	11	183	292	192	246
	12	36	488.5	240	249
	13	36	299.5	188	133
	14	443	290.5	201	133
News-2	1	306	224	522	270
	2	36	224	522	270
	3	36	151	12	540
	4	37	162	3	541
	5	37	149	3	541
	6	36	36	110	540
	7	36	188	34	540

TABLE 6.3 – Summary of Esthetic Measurements Results

	News-2	News-2-Runaround
Vertical Symmetry	0.36104727	0.47469425
Horizontal Symmetry	0.63895273	0.36858985
Radial Symmetry	0.63895273	0.56018704
Symmetry	0.45368242	0.53217626
Vertical Balance	-2.55782860E-04	-0.15916744
Horizontal Balance	-0.045403257	0.39462638
Balance	0.9998721	0.9204163
Equilibrium in X	4.30785840E-05	0.027329115
Equilibrium in Y	0.007787341	-0.11113999
Equilibrium	0.9960848	0.93076545

A way to reduce the need to make inferences about the obtained measurements results is to use them to find cases with similar measurements, where these cases can provide assessment regarding the layout's design. Let us observe the results of esthetic measurements between the newsletter documents illustrated in Figure 6.3. An outline of the layout properties for each of those documents can be observed in Table 6.2. Esthetic measurements results are outlined in Table 6.3.

At first glance, the values in table 6.3 for all three measurements do not seem to be too far apart, their absolute difference being 0.079456 for Balance, 0.065319 for Equilibrium and Symmetry with 0.07849384. To calculate the esthetic difference of two documents with greater precision, we created three formulas to measure the Euclidean distance between them. A fourth formula, which we call the *Esthetic Distance*, is the sum of the previous three formulas. The *Balance Distance* formula is given by

$$\sqrt{(VB1 - VB2)^2 + (HB1 - HB2)^2} \quad \text{Equation 6.1}$$

where VB is *Vertical Balance*, HB is *Horizontal Balance*, and the digits 1 and 2 identify the two documents compared. This numbering applies also to the following 3 equations. Similar to *Balance Distance*, the *Equilibrium Distance* formula is given by

$$\sqrt{(EX1 - EX2)^2 + (EY1 - EY2)^2} \quad \text{Equation 6.2}$$

where EX is *Equilibrium in X* and EY is *Equilibrium in Y*. The *Symmetry Distance* is

$$\sqrt{(VS1 - VS2)^2 + (HS1 - HS2)^2 + (RS1 - RS2)^2} \quad \text{Equation 6.3}$$

where VS is *Vertical Symmetry*, HS is *Horizontal Symmetry* and RS is *Radial Symmetry*. The *Esthetic Distance* is computed by

$$\sum (Eq6.1 + Eq6.2 + Eq6.3) \quad \text{Equation 6.4}$$

Going back to our newsletter documents, the values resulting from computing Equations 6.1, 6.2, 6.3 and 6.4 (see table 6.4) between these two documents are:

- Eq1 (Balance Distance): 0.46784504192135395
- Eq2 (Equilibrium Distance): 0.12201736567229514
- Eq3 (Symmetry Distance): 0.30367047915094375
- Eq4 (Esthetic Distance): 0.8935329

Note that the range of values for *all 4 equations* is [0,∞). Values closer to 0 are desirable.

TABLE 6.4 – Balance, Equilibrium, Symmetry and Esthetic Distance measurements between one instance and all other documents in the sample set

InstanceID: Newsletter-2				
Page Size Dimensions: (h, w)792.0 612.0				
Balance: 0.9998721 (Vertical: -2.5578286E-4, Horizontal: -0.045403257)				
Equilibrium: 0.9960848(Vertical: 4.3078584E-5, Horizontal: 0.007787341)				
Symmetry: 0.45368242 (Vertical:0.36104727, Horizontal:0.63895273, Radial:0.63895273)				
Case Name	Balance Distance	Equilibrium Distance	Symmetry Distance	Esthetic Distance
Asymmetry	1.299636159797	0.478328694910	0.475987535277	2.253952389985
BigAnvilTest	0.467330709202	0.067999411372	0.129931731124	0.665261851698
Brochure-test1-p1	1.045614995393	0.045647304920	0.193551415553	1.284813715866
Brochure-test1-p2	1.045426285214	0.009531781052	0.259912788244	1.314870854510
Brochure-test1b-p1	1.045567034520	0.010531215534	0.688303732652	1.744401982706
Brochure-test2-p1	0.543518937607	0.007095983319	0.462533713714	1.013148634640
Brochure-test2b-p1	0.927534871377	0.019363132782	0.538510849022	1.485408853181
Flyer-test	0.610339643919	0.165982903509	0.415528334617	1.191850882045
Flyer-testb	0.710221362003	0.178548968731	0.300316172076	1.189086502810
FN9981502-LAYOUT-MQ1	0.374151236400	0.042449377590	0.075336902720	0.491937516710
FN9981502-LAYOUT-MQ2	0.393119844204	0.039381678420	0.087473779668	0.519975302292
Newsletter-2-BigGutter	0.279124797838	0.053129025313	0.005775975660	0.338029798810
Newsletter-2-Runaround	0.467845041921	0.122017365672	0.303670479151	0.893532886745
Newsletter2SmallGutter	0.000000000000	0.000000000000	0.000000000000	0.000000000000
Newsletter2SmallHeader	0.325674220321	0.046775358304	0.002012653262	0.374462231887
Newsletter-Blocks	0.122342872475	0.023659777496	0.364468761398	0.510471411369
Newsletter-Sided	0.150242171020	0.016553754681	0.389683992018	0.556479917719
Newsletter-SmallSide	0.383545238669	0.083736338969	0.424525919269	0.891807496907
Newsletter-SmallSide-b	0.977076037514	0.137546761028	0.460219636358	1.574842434901
Newsletter2-SmallGutLine	0.256534554602	0.049926338615	0.088669428534	0.395130321751
NewsSided-SmallParagraph	1.032267237362	0.206931296174	0.439872246271	1.679070779807
SymmetryTest	0.485082312802	0.184049890829	0.689274123166	1.358406326797

On Table 6.4, we can see these distance measurements at work. All four formulas are calculated for the “Newsletter-2” document, illustrated on the right side of figure 6.3, and all other documents in the sample set. Their interpretation is rather simple. The matching document – being any one of the documents in the sample set compared with “Newsletter-2” – with the smallest calculated distance is the closest match. For the sake of argument, we will consider that the “Newsletter2SmallGutter” document is not part of the samples set due to its impressive aesthetic similarity with “Newsletter-2”.

Which of the four distances should be used? *Esthetic Distance* was our first choice, as it is composed of all other three distances. We argue over one of the many similar inconsistencies in the results shown in table 6.5 which contains a subset of the Esthetic Distance matrix, seen in Appendix B. The closest match of “Flyer-testb” is “Flyer-test” with a 0.654714942 esthetic distance, which is to be expected. However, for “Flyer-test” the best match is “Brochure-test2-p1” with a 0.387923032 esthetic distance, an inconsistency for two reasons: 1) The closest match is not within the “Flyers” category, and 2) the distance is much smaller (0.38) than the distance for the match found for “Flyer-testb” (0.65). This inconsistency is possibly due to the dilution of the distances summed in Equation 6.4 (Esthetic Distance). In the end, the idea was to select a good threshold range of values for Esthetic Distance, but it became apparent that no pattern existed in these results that could help identify a good match.

TABLE 6.5 – Esthetic Distance measurements between documents in the "Flyers" category and all other documents in the sample set

	Flyer-test	Flyer-testb	FN9981502-LAYOUT-MQ1	FN9981502-LAYOUT-MQ2
Asymmetry	1.889389634	1.312706232	1.957354546	2.605370045
BigAnvilTest	0.977282166	0.693151236	0.398860276	0.993719876
Brochure-test1-p1	2.436211586	2.414496183	1.427054405	1.377624512
Brochure-test1-p2	2.456690311	2.438342094	1.447740674	1.422839165
Brochure-test1b-p1	2.116531372	2.455330372	1.980403304	1.986098766
Brochure-test2-p1	0.387923032	0.861495733	1.133938551	1.173787832
Brochure-test2b-p1	0.616107643	1.073089361	1.523300648	1.565784931
Flyer-test		0.654714942	1.170478225	1.335234046
Flyer-testb	0.654714942		1.014969826	1.469762683
FN9981502-LAYOUT-MQ1	1.170478225	1.014969826		0.849344313
FN9981502-LAYOUT-MQ2	1.335234046	1.469762683	0.849344313	
Newsletter-2	1.191850901	1.189086556	0.491937518	0.519975305
Newsletter-2-BigGutter	1.516272783	1.475439191	0.679525018	0.716848731
Newsletter-2-Runaround	1.793059826	1.602509737	1.279674172	1.079804659
Newsletter2SmallGutter	1.191850901	1.189086556	0.491937518	0.519975305
Newsletter2SmallHeader	1.5526582	1.556509137	0.793327928	0.618748188
Newsletter-Blocks	1.309136987	1.177107453	0.79650563	0.970139086
Newsletter-Sided	1.513738036	1.405712843	1.00939858	0.848199248
Newsletter-SmallSide	1.774454832	1.626873732	1.271573782	1.140481114
Newsletter-SmallSide-b	2.251141787	2.240777016	2.015994787	1.529500127
Newsletter2-SmallGutLine	0.957735062	0.843268454	0.361732662	0.765985131
NewsSided-SmallParagraph	2.370357037	2.2003088	1.845459223	2.023303509
SymmetryTest	0.975722015	1.132171035	1.409192204	1.383730888

6.3 Geometric Distance

Geometric Distance is a term we use to refer to the document matching method proposed in [3] that represents a document's layout as one long vector to easily determine their geometrical variations. The methodology was explained with greater detail in Chapter 4. This section focuses on displaying and explaining the results obtained from our tests using a slightly modified version of this method.

To better visualize how the implemented geometrical analysis based on Peng's matching method [3] holds, we created a variation of the Confusion Matrix, which we called a Distance Matrix. We took the concept of Confusion Matrices – which are used in artificial

intelligence as a visualization tool – to aid in visualizing the distance between all samples in the set. In the Confusion Matrix, each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class. Let us explain the example below, taken from.

In the example in table 6.6 the confusion matrix says that of the 8 actual cats (eight being the sum of the Cat column), the system predicted that three were dogs, and of the six dogs, it predicted that one was a rabbit and two were cats. We can see from the matrix that the system in question has trouble distinguishing between cats and dogs, but can make the distinction between rabbits and other types of animals pretty well. One benefit of a confusion matrix is that it is easy to see if the system is confusing two classes, which is exactly the type of result that needs to be analyzed.

The results shown in our Distance Matrices represent the distance, Geometric or Esthetic, between the two documents outlined in the cell's row and column. Two Distance Matrices were created using the 23 document samples, one being the Geometric Distance matrix whose values are calculated using equation 4.1 while with the values in the Esthetic Distance matrix are obtained by equation 6.4. These matrices can be seen in the Appendix section. As mentioned earlier, each document in the sample set is compared with every other document in the sample, including itself.

TABLE 6.6 – Confusion Matrix Example

		Actual		
		Cat	Dog	Rabbit
Predicted	Cat	5	2	0
	Dog	3	3	2
	Rabbit	0	1	11

This visualization helps analyzing, for example, how different are newsletter documents within their own category and also compared with documents in other categories.

In regards to the observations of the distance matrices, we once again refer to the categories of the document types described earlier: Brochures, Flyers, Newsletters and Others. The documents are analyzed without regard to their type but to explain how we selected the threshold value previously mentioned (0.01) we first make comparisons within each category.

6.3.1 Flyers

Analyzing a small, controlled category of documents such as “Flyers” we can get a glimpse of how the matching is done with documents that are similar in terms of its structure. The “Flyers” subset is composed of four documents, each one having a sibling. For example, “flyer-test” is a sibling of “flyer-testb” and vice versa. In figure 6.7, note how the sibling documents are practically the same, in terms of its structure, with a few variations.

Figure 6.7 – Sibling flyers.
Variations between them are circled in the left example.

From the distance matrix subset shown in table 6.7 we see that the only match for each flyer is with its own sibling. This is to be expected from flyers and any other type of document that does not have a specific structure. Notice also how the difference with other documents is greater than 0.2, and that the matched value is a value smaller than 0.01.

6.3.2 Brochures

A slightly more diverse subset is the one found within the "Brochure" category. Sibling documents are also found within this subset, one good example being "Brochure-test1-p1", "Brochure-test1b-p1" and "Brochure-test2b-p1", as seen in figure 6.8.

TABLE 6.7 – Subset from distance matrix of documents in the "Flyers" category. The closest match within the threshold is marked in black.

	Flyer-test	Flyer-testb	FN9981502-MQ1	FN9981502-MQ2
SymmetryTest	0.02066622	0.02043309	0.03406491	0.03269252
News2SmallHeader	0.03043920	0.03020606	0.03683279	0.03592914
Brochure-test2b-p1	0.02386199	0.02361648	0.02778397	0.02677717
Flyer-test	0.00000000	0.00367853	0.03632526	0.03541749
News-2-Runaround	0.03703085	0.03679772	0.04407432	0.04313354
News-SmallSide	0.02706188	0.02682874	0.03264054	0.03159248
News2SmallGutter	0.02829356	0.02806042	0.03639953	0.03542781
News-2-BigGutter	0.03030716	0.03007402	0.03599723	0.03507501
Asymmetry	0.02071161	0.02048054	0.03424642	0.03285204
Flyer-testb	0.00367853	0.00000000	0.03613958	0.03520293
FN9981502-MQ1	0.03632526	0.03613958	0.00000000	0.00981573
News-SmallSide-b	0.02528347	0.02505859	0.03090133	0.02984915
NewsSided-SmallParagraph	0.02560944	0.02536806	0.03163580	0.03056711
Brochure-test1-p1	0.02363504	0.02335446	0.02760860	0.02660180
Brochure-test2-p1	0.02450980	0.02427667	0.02851018	0.02750132
FN9981502-MQ2	0.03541749	0.03520293	0.00981573	0.00000000
News-2	0.02829356	0.02806042	0.03639953	0.03542781
Brochure-test1b-p1	0.02199280	0.02176586	0.02596430	0.02494306
Brochure-test1-p2	0.02499051	0.02475119	0.02897232	0.02794283
News-Sided	0.02649658	0.02626345	0.03247962	0.03143155
News2-SmallGutLine	0.02924672	0.02901358	0.03736095	0.03636652
News-Blocks	0.03762090	0.03732175	0.04381643	0.04287359
BigAnvilTest	0.02878664	0.02854526	0.03300984	0.03193702

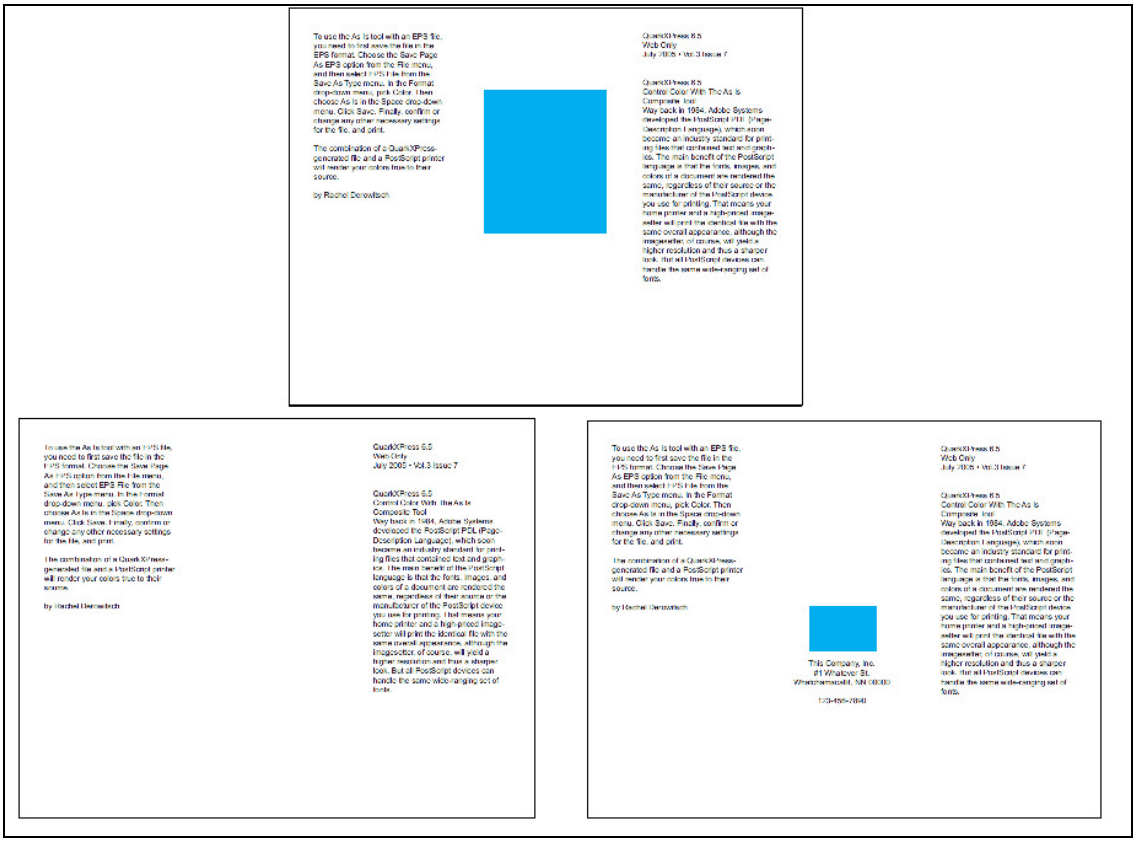

Figure 6.8 – Sibling brochures.
“Brochure-test1-p1” above. “Brochure-test1b-p1” lower left. “Brochure-test2b-p1” lower right.

Limiting our observations to these siblings we can better explain the behavior of the implemented method. Let us start by renaming the siblings in figure 6.8 as follows: “Brochure-test1-p1” (above) will be known as *A*, “Brochure-test1b-p1” (lower left) as *B*, and “Brochure-test2b-p1” (lower right) as *C*. From table 6.8 we get that the closest match for *A* is *B*, and for *C* is also *B*. Initially, one would assume that *A* and *C* should best match each other because both have objects in the middle whereas *B* does not. But because the implemented method considers only the borders of the objects, not the object’s area, then the difference between *A* and *C* is actually greater. Notice also that in table 6.3 the closest match for each document has a distance value smaller than 0.01.

6.3.3 Newsletters

The "Newsletter" subset has the greatest amount of sample documents. For this category we will contrast a pair of siblings that are not even close to a match, although they look similar in structure. Observe in figure 6.3 how both documents have a very similar format and two variations, one being in the header ("Hot Cuts!") and the other being the image box between the text columns. In table 6.9, these two documents are known as "News-2-Runaround", pictured on the left side of figure 6.3, and "News-2", pictured opposite.

TABLE 6.8 – Subset from distance matrix of documents in the "Brochure" category. The closest match within the threshold is marked in black.

	Brochure-test2b-p1	Brochure-test1-p1	Brochure-test2-p1	Brochure-test1b-p1	Brochure-test1-p2
SymmetryTest	0.01288828	0.01264070	0.01357117	0.01096546	0.01408488
News2SmallHeader	0.02449743	0.02422303	0.02512874	0.02261380	0.02565071
Brochure-test2b-p1		0.00445220	0.002116756	0.001931075	0.01261801
Flyer-test	0.02386199	0.02363504	0.02450980	0.02199280	0.02499051
News-2-Runaround	0.03165643	0.03141505	0.03233107	0.02979963	0.03266942
News-SmallSide	0.02151003	0.02122945	0.02216198	0.01958515	0.02252715
News2SmallGutter	0.02395689	0.02338128	0.02463565	0.02211246	0.02512874
News-2-BigGutter	0.02360822	0.02335033	0.02429111	0.02173079	0.02470374
Asymmetry	0.01290685	0.01260563	0.01358148	0.01098815	0.01413027
Flyer-testb	0.02361648	0.02335446	0.02427667	0.02176586	0.02475119
FN9981502-MQ1	0.02778397	0.02760860	0.02851018	0.02596430	0.02897232
News-SmallSide-b	0.01953770	0.01924061	0.02020408	0.01760868	0.02059608
NewsSided-SmallParagraph	0.02027010	0.01998539	0.02092824	0.01834728	0.02129753
Brochure-test1-p1	0.00445220		0.00436142	0.00252938	0.01321631
Brochure-test2-p1	0.00211676	0.00436142		0.00262016	0.01330709
FN9981502-MQ2	0.02677717	0.02660180	0.02750132	0.02494306	0.02794283
News-2	0.02395689	0.02338128	0.02463565	0.02211246	0.02512874
Brochure-test1b-p1	0.001931075	0.002529379	0.00262016		0.010686935
Brochure-test1-p2	0.01261801	0.01321631	0.01330709	0.01068693	
News-Sided	0.02115518	0.02084778	0.02180919	0.01922823	0.02217848
News2-SmallGutLine	0.02494512	0.02438189	0.02562389	0.02312133	0.02613348
News-Blocks	0.03144187	0.03118398	0.03213301	0.02959126	0.03245692
BigAnvilTest	0.02110566	0.02086222	0.02178649	0.01918284	0.02230433

Having that image box positioned between the text columns makes it necessary to have the text contoured around the image box. This, of course, disrupts a part of the document's structure enough to have these two siblings become distant. We can observe from table 6.9 that the geometric distance between these two documents is of 0.03590026, one of the highest distances among other matched documents within the same category.

Yet another example of distant siblings is shown in figure 6.9, each one with a similar structure but varying gutter size. The gutter is the space between the text columns. A document of this type should have enough gutter space for the reader to easily determine where the left column ends and the right one begins, and the small gutter example in the figure barely has any. In fact, a line is inserted in the middle to make it easier to visually separate the columns. The gutter should not be too big either as this makes the columns look as if they are not related when in fact they are. This last example also goes to show how sensitive the implemented method is to geometrical differences, evidence of this being the 0.02632534 distance between both of them. Notice that the distance of the closest matching documents in table 6.9 are all under 0.01.

TABLE 6.9 – Subset from distance matrix of documents in the "Newsletter" category.
The closest match within the threshold is marked in black.

	News2 SmallHeader	News-2-Runaround	News-SmallSide	News2 SmallGutter	News-2-BigGutter	News-SmallSide-b	NewsSided-SmallParagraph	News-2	News-Sided	News2-SmallGutLine	News-Blocks
SymmetryTest	0.02122533	0.02847511	0.01824619	0.02073843	0.02036707	0.01631718	0.01702689	0.02073843	0.01792022	0.02173285	0.02819246
News2SmallHeader		0.03274782	0.02745799	0.02939939	0.02270664	0.02524221	0.02598080	0.02939939	0.02524221	0.03078167	0.03032572
Brochure-test2b-p1	0.02449743	0.03165643	0.02151003	0.02395689	0.02360822	0.01953770	0.02027010	0.02395689	0.02115518	0.02494512	0.03144187
Flyer-test	0.03043920	0.03703085	0.02706188	0.02829356	0.03030716	0.02528347	0.02560944	0.02829356	0.02649658	0.02924672	0.03762090
News-2-Runaround	0.03274782		0.02944065	0.03590026	0.01872689	0.02851637	0.02847924	0.03590026	0.02852463	0.03737332	0.02507097
News-SmallSide	0.02745799	0.02944065		0.02348031	0.02531648	0.00684129	0.00669687	0.02348031	0.00759226	0.02498638	0.03459018
News2SmallGutter	0.02939939	0.03590026	0.02348031		0.02632534	0.02427255	0.02336890	0.00000000	0.02427255	0.00389929	0.03635002
News-2-BigGutter	0.02270664	0.01872689	0.02531648	0.02632534		0.02485022	0.02520507	0.02632534	0.02485022	0.02786855	0.01619339
Asymmetry	0.02125215	0.02854319	0.01831014	0.02077557	0.02039595	0.01641827	0.01708672	0.02077557	0.01799036	0.02176792	0.02826055
Flyer-testb	0.03020606	0.03679772	0.02682874	0.02806042	0.03007402	0.02505859	0.02536806	0.02806042	0.02626345	0.02901358	0.03732175
FN9981502-MQ1	0.03683279	0.04407432	0.03264054	0.03639953	0.03599723	0.03090133	0.03163580	0.03639953	0.03247962	0.03736095	0.04381643
News-SmallSide-b	0.02524221	0.02851637	0.00684129	0.02427255	0.02485022		0.00467089	0.02427255	0.00212501	0.02579512	0.03399188
NewsSided-SmallParagraph	0.02598080	0.02847924	0.00669687	0.02336890	0.02520507	0.00467089		0.02336890	0.00254588	0.02489148	0.03308824
Brochure-test1-p1	0.02422303	0.03141505	0.02122945	0.02338128	0.02335033	0.01924061	0.01998539	0.02338128	0.02084778	0.02438189	0.03118398
Brochure-test2-p1	0.02512874	0.03233107	0.02216198	0.02463565	0.02429111	0.02020408	0.02092824	0.02463565	0.02180919	0.02562389	0.03213301
FN9981502-MQ2	0.03592914	0.04313354	0.03159248	0.03542781	0.03507501	0.02984915	0.03056711	0.03542781	0.03143155	0.03636652	0.04287359
News-2	0.02939939	0.03590026	0.02348031	0.00000000	0.02632534	0.02427255	0.02336890		0.02427255	0.00389929	0.03635002
Brochure-test1b-p1	0.02261380	0.02979963	0.01958515	0.02211246	0.02173079	0.01760868	0.01834728	0.02211246	0.01922823	0.02312133	0.02959126
Brochure-test1-p2	0.02565071	0.03266942	0.02252715	0.02512874	0.02470374	0.02059608	0.02129753	0.02512874	0.02217848	0.02613348	0.03245692
News-Sided	0.02524221	0.02852463	0.00759226	0.02427255	0.02485022	0.00212501	0.00254588	0.02427255		0.02579512	0.03397537
News2-SmallGutLine	0.03078167	0.03737332	0.02498638	0.00389929	0.02786855	0.02579512	0.02489148	0.00389929	0.02579512		0.03787260
News-Blocks	0.03032572	0.02507097	0.03459018	0.03635002	0.01619339	0.03399188	0.03308824	0.03635002	0.03397537	0.03787260	
BigAnvilTest	0.02938082	0.03569601	0.02669671	0.02867936	0.02738785	0.02475325	0.02536806	0.02867936	0.02626758	0.02964902	0.03536591

Figure 6.9 – Newsletters siblings with varying gutter sizes. "News-2-BigGutter" to the left, "News2SmallGutter" opposite.

6.3.4 Others

Moving on to analyzing the effect of the documents in the "Others" category, we can argue that the "noise" these documents may cause is close to nil, assuming we select a good threshold. In fact, if we were to remove this subset of documents from the sample set, the difference values between the other documents will not change. As we have already explained, the implemented method compares the analyzed document with all other documents in the case database, with no specific regards to its category. Seeing the sample set as it is, not being divided by categories, no document in it can be considered as "noise", as long as the selected threshold is not too big. In fact, the rationale goes by saying that the greater the number of documents in the set, the greater the chances of providing a design judgment (good/bad) and an assessment, if any. Also, we can observe from the subset of

matrix illustrated in table 6.10 that for all three documents within this category, the difference values with all other documents are not within the threshold. This means that these "other" documents are not even considered as potential matches.

Based on our previous observations, a threshold value smaller than 0.01 is enough to consider two documents to be a match. The case with the smallest value calculated with equation 4.1 is considered as the matching case, as long as the value is within the selected threshold. If there is no matching case, then no assessment can be provided, meaning that the document being analyzed cannot be judged by our method as either having a good or bad design.

TABLE 6.10 – Subset from distance matrix of documents in the "Others" category. No match is determined as none of the values is within the threshold.

	SymmetryTest	Asymmetry	BigAnvilTest
SymmetryTest		0.02330208	0.01792228
News2SmallHeader	0.02122533	0.02125215	0.02938082
Brochure-test2b-p1	0.01288828	0.01290685	0.02110566
Flyer-test	0.02066622	0.02071161	0.02878664
News-2-Runaround	0.02847511	0.02854319	0.03569601
News-SmallSide	0.01824619	0.01831014	0.02669671
News2SmallGutter	0.02073843	0.02077557	0.02867936
News-2-BigGutter	0.02036707	0.02039595	0.02738785
Asymmetry	0.02330208		0.01796148
Flyer-testb	0.02043309	0.02048054	0.02854526
FN9981502-MQ1	0.03406491	0.03424642	0.03300984
News-SmallSide-b	0.01631718	0.01641827	0.02475325
NewsSided-SmallParagraph	0.01702689	0.01708672	0.02536806
Brochure-test1-p1	0.01264070	0.01260563	0.02086222
Brochure-test2-p1	0.01357117	0.01358148	0.02178649
FN9981502-MQ2	0.03269252	0.03285204	0.03193702
News-2	0.02073843	0.02077557	0.02867936
Brochure-test1b-p1	0.01096546	0.01098815	0.01918284
Brochure-test1-p2	0.01408488	0.01413027	0.02230433
News-Sided	0.01792022	0.01799036	0.02626758
News2-SmallGutLine	0.02173285	0.02176792	0.02964902
News-Blocks	0.02819246	0.02826055	0.03536591
BigAnvilTest	0.01792228	0.01796148	

6.4 Chapter Review

We have presented in this chapter details regarding the tests of the implemented method. Results were explained with the aid of a Distance Matrix which allowed us to visualize the calculated differences between all documents in the set using the Geometric Distance and the Esthetic Distance equations.

We argued in regards to the inconsistencies we found with the Esthetic Distance results seen in Appendix B, were no pattern could help us determine acceptable threshold values. This made us consider another method, which we called Geometric Distance, for case matching.

To better understand the results of the Geometric Distance Matrix, documents were separated into categories and the analysis was performed for each category. Two important observations mentioned are: 1) an increased amount of documents in the case database increases the chances of a successful matching; 2) the implemented method is highly sensitive to geometrical variations between documents.

A detailed description of the documents that composed the samples set was provided, including illustrations of a few of these documents and charts detailing the distribution between categories.

7 Conclusions

In the past few chapters we have presented in detail a methodology that can be used in analyzing a document's layout to consequently provide assessment of a design's aesthetics. Results were shown on how this method is more effective for a culture-adaptive layout analysis than using esthetic measurements [18] "as is". A few ideas are presented for future research and consideration.

7.1 Implementation Limitations

We start with the not-so-obvious but still relevant limitations of our methodology. Our implementation suggests new ideas yet it assimilates existing ones, mainly those from [1]. In it, for example, documents are of one page length. Some document types, i.e. magazine articles, extend into several pages and the analysis of its content may depend seeing all its pages as a whole. Yet, this type of analysis is more complex than it already is for a one page document, so future work should focus on expanding the functionality of our methodology before moving into such arena.

Second, documents are represented in PPML. We wanted to build upon an already set document representation framework which would facilitate our work. Although PPML was designed to represent a set of similarly structured documents known as Variable Data Job (VDJ), it can also be used to represent one particular document instead of a set. Each document has several containers, or boxes, where illustrations and texts could be included into. However, these containers can only be shaped in a rectangular form, limiting the possibility of including irregularly shaped objects in a document, a very common practice in

designing professional-level Flyers. Fortunately, the layout analysis in our methodology is done at the pixel-level, allowing the document representation component to be changed into something more flexible.

Finally, our methodology has no regard to content, including text and illustration. Analyzing the text's appropriateness is a complex subject that is outside the scope of our research. There are books on the subject of typography and text readability, one being [31] which dedicates a complete chapter on the subject. For example, a simple suggestion is that the use of *Sans Serif* font types should be limited to titles or small text boxes as they are very useful for illustrative purposes but are hard to read as part of a long paragraph, in which case *Serif* types are more useful as they ease the task of reading[14]. Other suggestions mentioned by on [31] are:

- Start any document with only two font types. Use more only when you are sure you need it.
- Use proven and efficient combinations, for example, a *Serif* type like *Bembo* for long paragraphs with a *Sans Serif* like *Franklin Gothic Heavy* or *Gill Sans Extra Bold* for headers.
- Keep body text font size between 9 points (for books) and 12 (for marketing material or informational material)
- Do not use uppercase nor underlining to highlight text. Bold or Italic can provide emphasis, but never use them together.

The use of colors and its combinations in a document's design is a subject on its own. Research on this is widely available. One good source on the application of Color Contrast,

---

[14] This was not included in Chapter 2 (Literature Survey) as it wasn't directly relevant to our methodology but could prove useful in future work.

also known by the Color Theory concept, in the design of web pages is provided by the World Wide Web Consortium's in terms of an algorithm[15]. Again, future work should include some type of content analysis. The correct selection of color and font in certain situations can make a flyer more attractive or newsletter more readable, respectively.

7.2 Contributions

There are only a handful of investigations on our subject, as explained in chapter 2, which despite the aforementioned limitations make our contributions considerably unique. Although recent approaches detected artifacts in a document [1] [2], they were designed with Variable Data Printing (VDP) in mind. As such, they assume that an approved instance, a document that is pleasantly designed, is provided upon submitting the instances for analysis. Our approach makes no such assumption. In fact, our assumption is that a document is independently submitted for analysis, which is why our approach differs from [1] [2]. At the time of this thesis' writing, there were no document analysis approaches that could analyze a document that was not from a Variable Data Job (VDJ).

We consider the influence of culture over aesthetics perception. Consequently, a Case-Based Reasoning approach was selected over the use of Esthetic Layout Measurements [18] "as is" since the process of analyzing the aesthetics of a document can be subjective due to a viewer's perception within a cultural setting.

As such, esthetic measurements as defined in [18] were not an ideal option when considering them for a document-analysis method that was culture-adaptive. Design trends

---

[15] Color Contrast Algorithm. http://www.w3.org/TR/AERT#color-contrast, Retrieved March 10, 2010.

tend to change over time, and adjusting the relevance of Esthetic Measurements to follow these trends may become cumbersome and unpractical. The analysis had to be dependent on previous assessments.

During our literature survey we found very few approaches that we could build upon. Consequently, it became necessary to look outside the boundaries of this research. In it we investigated the following subjects:

- Perception of aesthetics [14]
- Use of context in pattern recognition[15]
- Perception influenced by culture [4]

Not only did these resources provide insight into how aesthetics perception varies between audiences or cultures, it defined the analysis process in our methodology and justified its use. This is our second most important contribution to our research.

7.3 Final Thoughts

As we began our research, there was a clear idea of the need of finding issues with the design of a document's layout. The "how" was not clear at the time, however. There was a clear effort during the years previous to our research to define a methodology where defects, also known as artifacts [1] [2] [6], in the instances of a VDJ could be found. Later on we found the need of a methodology that could be used in documents submitted independently for aesthetics analysis. This concept had not been researched before, which required new ideas from other fields.

We began investigating more about design principles. This was the starting point for many a researcher [1] [2] [6], and we did accordingly by finding new ideas in [11] [12][13].

But ambiguity in the application of design principles, or detecting if such principles have been applied to a document's layout, was not helpful in determining defects in a document. It was necessary to investigate how perception of aesthetics is determined at various levels, the psychological [14], the contextual [15] and the cultural [4]. It became apparent that design principles, although useful in understanding well known design aesthetics, would limit our methodology to a specific culture. Also, as mentioned earlier, design trends tend to change over time and the relevance of some design principles change accordingly. Keeping track of these changes is not practical in any sense.

In the end it was clear what we wanted to achieve, a culture-adaptive document layout aesthetics analysis method. By geometrically analyzing a document to match it with a previously analyzed one, this could be achieved. To show how this could be achieved we implemented a matching algorithm defined by [3] to find a match within a set of documents in a Case-base. Our results confirm what our methodology is capable of, the detection of layout aesthetics given the availability of a design assessment for an unpleasant document analyzed.

When working with concepts that are predominantly subjective as design principles are, it becomes increasingly difficult to determine a quantitative method to measure layout esthetics. Moreover, when such quantitative methods are available [18] its application is limited to a specific set of situations, in this case a specific cultural setting. The importance of expanding our scope of literature references is evident in our work and that of others [1] [2] [6]. We certainly stress this consideration to be undertaken on related works in the future.

8 References

[1] H.J. Santos-Villalobos, "Style-Dependent Artifact Recognition For Digital Variable Data Printing", MS Thesis, University of Puerto Rico, Mayaguez, Puerto Rico, 2005.

[2] A.C. Faria and J.B.S. de Oliveira, "Measuring Aesthetic Distance Between Document Templates and Instances", Proceedings of the 2006 ACM symposium on Document Engineering, 2006, Vol. 1, pages 13 – 21.

[3] H. Peng, F. Long and Z. Chi, "Document Image Recognition Based on Template Matching of Component Block Projections", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, Vol. 25 No. 9, pages 1188 – 1192.

[4] E. T. Hall, "Beyond Culture", Anchor Press, 1981.

[5] D.F. Brailsford, "Laying Out the Future of Final-form Digital Documents", Digital Publishing Proceedings of SPIE-IS&T Electronic Imaging, January 2006, Vol. 6076, pages 60760I-1 – 60760I-13.

[6] X. Feng and J.P. Allebach, "Measurements of Ringing Artifacts in JPEG Images", Digital Publishing Proceedings of SPIE-IS&T Electronic Imaging, January 2006, Vol. 6076, pages 60760A-1 – 60760A-10.

[7] X. Lin, "Intelligent Content Fitting for Digital Publishing", Digital Publishing Proceedings of SPIE-IS&T Electronic Imaging, January 2006, Vol. 6076, pages 60760J-1 – 60760J-9.

[8] C.L. Kutty and J.K. Prabhakaran, "Personalized Direct Marketing using Digital Publishing", Digital Publishing Proceedings of SPIE-IS&T Electronic Imaging, January 2006, Vol. 6076, pages 607604-1 – 607604-8.

[9] G. Vondran, H. Chao, X. Lin, D. Beyer, P. Joshi, B. Atkins and P. Obrador. "Automated Campaign", Digital Publishing Proceedings of SPIE-IS&T Electronic Imaging, January 2006, Vol. 6076, pages 607605-1 – 607605-6.

[10] F. Giannetti and R. Sellman, "Anvil: VDP Segmented Workflow Toolset", HP White Paper, 2006.

[11] W. Lidwell, K. Holden and J. Butler, "Universal Principles of Design", Rockport, 2003.

[12] R. Williams, "The Non-Designer's Design Book", Peachpit Press, 2nd Edition, 2004.

[13] S.M. Topping, "Graphic design and color in today's office", KODAK Publication, Edition W-628, 1990.

[14] P.M. Lester, "Visual Communication: images with messages", Cengage Learning, 4th Edition, 2005.

[15] G.T. Toussaint, "The Use of Context in Pattern Recognition", Pattern Recognition, January 1978, Vol. 10, pages 189 – 204.

[16] F. Esposito, D. Malerba and G. Semerano, “A Knowledge-Based Approach to the Layout Analysis”, Document Analysis and Recognition, August 1995, Vol. 1, pages 466 – 471.

[17] M. L. Maher and A. Gomez de Silva Garza, “Case-Based Reasoning in Design”, IEEE Expert, AI in Design, March-April 1997, pages 34 – 41.

[18] D.C.L. Ngo, L.S. Teo and J.G. Byrne, “Evaluating Interface Esthetics”, Knowledge and Information Systems, 2002, Vol. 4, 46 – 79.

[19] E. Lupton, “D.I.Y, Design It Yourself”, Princeton Architectural Press, 2006.

[20] D. Malerba, F. Esposito, O. Altamura, M. Ceci and M. Berardi, “Correcting the document layout: a machine learning approach”, Document Analysis and Recognition, 2003, Vol. 1, pages 97 – 102.

[21] H. Chao and J. Fan, “Layout and Content Extraction for PDF Documents”, Document Analysis Systems VI, 2004, Vol. 3163, pages213 – 224.

[22] A. Ting, “Form recognition using linear structure”, Pattern Recognition, 1998, Vol. 32 No. 4, pages 645 – 656.

[23] R. A. Ballinger, “Layout and Graphic Design”, Studio Vista, 1970.

[24] J. V. White, “Editing By Design: For Designers, Art Directors, and Editors”, Allworth Press, 3rd Edition, 2003.

[25] T. Antin, “Great Print Advertising: Creative Approaches, Strategies, and Tactics”, John Wiley & Sons, 1993.

[26] E. Resnick, “Design for Communication: Conceptual Graphic Design Basics”, John Wiley and Sons, 2003.

[27] L. Siebert and L. Ballard, “Making a Good Layout”, North Light Books, 1992.

[28] X. Lin, H. Chao, G. Nelson and E. Durante, “Active Document Versioning: from layout understanding to adjustment”, Document Recognition and Retrieval, January 2006, Vol. 13, pages 15 – 19.

[29] M. Pearce, A.K. Goel, J.L. Kolodner, C. Zimring, L. Sentosa and R. Billington, “Case-Based Design Support: A Case Study in Architectural Design”, IEEE Expert, October 1992, 14 – 20.

[30] D. Partridge and Y. Wilks, “The Foundations of artificial intelligence: a sourcebook”, Cambridge University Press, 1990.

[31] A. Dabbs and A. Campbell, “The digital designer's bible”, Harper Design, 2005.

APPENDIX A. GEOMETRIC DISTANCE MATRIX

Non-categorized
Newsletter
Brochure
Flyer

	SymmetryTest	News2SmallHeader	Brochure-test2b-p1	Flyer-test	News-2-Runaround	News-SmallSide	News2SmallGutter	News-2-BigGutter
SymmetryTest	0.00000000	0.02122533	0.01288828	0.02066622	0.02847511	0.01824619	0.02073843	0.02036707
News2SmallHeader	0.02122533	0.00000000	0.02449743	0.03043920	0.03274782	0.02745799	0.02939939	0.02270664
Brochure-test2b-p1	0.01288828	0.02449743	0.00000000	0.02386199	0.03165643	0.02151003	0.02395689	0.02360822
Flyer-test	0.02066622	0.03043920	0.02386199	0.00000000	0.03703085	0.02706188	0.02829356	0.03030716
News-2-Runaround	0.02847511	0.03274782	0.03165643	0.03703085	0.00000000	0.02944065	0.03590026	0.01872689
News-SmallSide	0.01824619	0.02745799	0.02151003	0.02706188	0.02944065	0.00000000	0.02348031	0.02531648
News2SmallGutter	0.02073843	0.02939939	0.02395689	0.02829356	0.03590026	0.02348031	0.00000000	0.02632534
News-2-BigGutter	0.02036707	0.02270664	0.02360822	0.03030716	0.01872689	0.02531648	0.02632534	0.00000000
Asymmetry	0.02330208	0.02125215	0.01290685	0.02071161	0.02854319	0.01831014	0.02077557	0.02039595
Flyer-testb	0.02043309	0.03020606	0.02361648	0.00367853	0.03679772	0.02682874	0.02806042	0.03007402
FN9981502-MQ1	0.03406491	0.03683279	0.02778397	0.03632526	0.04407432	0.03264054	0.03639953	0.03599723
News-SmallSide-b	0.01631718	0.02524221	0.01953770	0.02528347	0.02851637	0.00684129	0.02427255	0.02485022
NewsSided-SmallParagraph	0.01702689	0.02598080	0.02027010	0.02560944	0.02847924	0.00669687	0.02336890	0.02520507
Brochure-test1-p1	0.01264070	0.02422303	0.00445220	0.02363504	0.03141505	0.02122945	0.02338128	0.02335033
Brochure-test2-p1	0.01357117	0.02512874	0.00211676	0.02450980	0.03233107	0.02216198	0.02463565	0.02429111
FN9981502-MQ2	0.03269252	0.03592914	0.02677717	0.03541749	0.04313354	0.03159248	0.03542781	0.03507501
News-2	0.02073843	0.02939939	0.02395689	0.02829356	0.03590026	0.02348031	0.00000000	0.02632534
Brochure-test1b-p1	0.01096546	0.02261380	0.00193108	0.02199280	0.02979963	0.01958515	0.02211246	0.02173079
Brochure-test1-p2	0.01408488	0.02565071	0.01261801	0.02499051	0.03266942	0.02252715	0.02512874	0.02470374
News-Sided	0.01792022	0.02524221	0.02115518	0.02649658	0.02852463	0.00759226	0.02427255	0.02485022
News2-SmallGutLine	0.02173285	0.03078167	0.02494512	0.02924672	0.03737332	0.02498638	0.00389929	0.02786855
News-Blocks	0.02819246	0.03032572	0.03144187	0.03762090	0.02507097	0.03459018	0.03635002	0.01619339
BigAnvilTest	0.01792228	0.02938082	0.02110566	0.02878664	0.03569601	0.02669671	0.02867936	0.02738785

	Asymmetry	Flyer-testb	FN9981502-MQ1	News-SmallSide-b	NewsSided-SmallParagraph	Brochure-test1-p1	Brochure-test2-p1
SymmetryTest	0.02330208	0.02043309	0.03406491	0.01631718	0.01702689	0.01264070	0.01357117
News2SmallHeader	0.02125215	0.03020606	0.03683279	0.02524221	0.02598080	0.02422303	0.02512874
Brochure-test2b-p1	0.01290685	0.02361648	0.02778397	0.01953770	0.02027010	0.00445220	0.00211676
Flyer-test	0.02071161	0.00367853	0.03632526	0.02528347	0.02560944	0.02363504	0.02450980
News-2-Runaround	0.02854319	0.03679772	0.04407432	0.02851637	0.02847924	0.03141505	0.03233107
News-SmallSide	0.01831014	0.02682874	0.03264054	0.00684129	0.00669687	0.02122945	0.02216198
News2SmallGutter	0.02077557	0.02806042	0.03639953	0.02427255	0.02336890	0.02338128	0.02463565
News-2-BigGutter	0.02039595	0.03007402	0.03599723	0.02485022	0.02520507	0.02335033	0.02429111
Asymmetry	0.00000000	0.02048054	0.03424642	0.01641827	0.01708672	0.01260563	0.01358148
Flyer-testb	0.02048054	0.00000000	0.03613958	0.02505859	0.02536806	0.02335446	0.02427667
FN9981502-MQ1	0.03424642	0.03613958	0.00000000	0.03090133	0.03163580	0.02760860	0.02851018
News-SmallSide-b	0.01641827	0.02505859	0.03090133	0.00000000	0.00467089	0.01924061	0.02020408
NewsSided-SmallParagraph	0.01708672	0.02536806	0.03163580	0.00467089	0.00000000	0.01998539	0.02092824
Brochure-test1-p1	0.01260563	0.02335446	0.02760860	0.01924061	0.01998539	0.00000000	0.00436142
Brochure-test2-p1	0.01358148	0.02427667	0.02851018	0.02020408	0.02092824	0.00436142	0.00000000
FN9981502-MQ2	0.03285204	0.03520293	0.00981573	0.02984915	0.03056711	0.02660180	0.02750132
News-2	0.02077557	0.02806042	0.03639953	0.02427255	0.02336890	0.02338128	0.02463565
Brochure-test1b-p1	0.01098815	0.02176586	0.02596430	0.01760868	0.01834728	0.00252938	0.00262016
Brochure-test1-p2	0.01413027	0.02475119	0.02897232	0.02059608	0.02129753	0.01321631	0.01330709
News-Sided	0.01799036	0.02626345	0.03247962	0.00212501	0.00254588	0.02084778	0.02180919
News2-SmallGutLine	0.02176792	0.02901358	0.03736095	0.02579512	0.02489148	0.02438189	0.02562389
News-Blocks	0.02826055	0.03732175	0.04381643	0.03399188	0.03308824	0.03118398	0.03213301
BigAnvilTest	0.01796148	0.02854526	0.03300984	0.02475325	0.02536806	0.02086222	0.02178649

	FN9981502-MQ2	News-2	Brochure-test1b-p1	Brochure-test1-p2	News-Sided	News2-SmallGutLine	News-Blocks	BigAnvilTest
SymmetryTest	0.03269252	0.02073843	0.01096546	0.01408488	0.01792022	0.02173285	0.02819246	0.01792228
News2SmallHeader	0.03592914	0.02939939	0.02261380	0.02565071	0.02524221	0.03078167	0.03032572	0.02938082
Brochure-test2b-p1	0.02677717	0.02395689	0.00193108	0.01261801	0.02115518	0.02494512	0.03144187	0.02110566
Flyer-test	0.03541749	0.02829356	0.02199280	0.02499051	0.02649658	0.02924672	0.03762090	0.02878664
News-2-Runaround	0.04313354	0.03590026	0.02979963	0.03266942	0.02852463	0.03737332	0.02507097	0.03569601
News-SmallSide	0.03159248	0.02348031	0.01958515	0.02252715	0.00759226	0.02498638	0.03459018	0.02669671
News2SmallGutter	0.03542781	0.00000000	0.02211246	0.02512874	0.02427255	0.00389929	0.03635002	0.02867936
News-2-BigGutter	0.03507501	0.02632534	0.02173079	0.02470374	0.02485022	0.02786855	0.01619339	0.02738785
Asymmetry	0.03285204	0.02077557	0.01098815	0.01413027	0.01799036	0.02176792	0.02826055	0.01796148
Flyer-testb	0.03520293	0.02806042	0.02176586	0.02475119	0.02626345	0.02901358	0.03732175	0.02854526
FN9981502-MQ1	0.00981573	0.03639953	0.02596430	0.02897232	0.03247962	0.03736095	0.04381643	0.03300984
News-SmallSide-b	0.02984915	0.02427255	0.01760868	0.02059608	0.00212501	0.02579512	0.03399188	0.02475325
NewsSided-SmallParagraph	0.03056711	0.02336890	0.01834728	0.02129753	0.00254588	0.02489148	0.03308824	0.02536806
Brochure-test1-p1	0.02660180	0.02338128	0.00252938	0.01321631	0.02084778	0.02438189	0.03118398	0.02086222
Brochure-test2-p1	0.02750132	0.02463565	0.00262016	0.01330709	0.02180919	0.02562389	0.03213301	0.02178649
FN9981502-MQ2	0.00000000	0.03542781	0.02494306	0.02794283	0.03143155	0.03636652	0.04287359	0.03193702
News-2	0.03542781	0.00000000	0.02211246	0.02512874	0.02427255	0.00389929	0.03635002	0.02867936
Brochure-test1b-p1	0.02494306	0.02211246	0.00000000	0.01068693	0.01922823	0.02312133	0.02959126	0.01918284
Brochure-test1-p2	0.02794283	0.02512874	0.01068693	0.00000000	0.02217848	0.02613348	0.03245692	0.02230433
News-Sided	0.03143155	0.02427255	0.01922823	0.02217848	0.00000000	0.02579512	0.03397537	0.02626758
News2-SmallGutLine	0.03636652	0.00389929	0.02312133	0.02613348	0.02579512	0.00000000	0.03787260	0.02964902
News-Blocks	0.04287359	0.03635002	0.02959126	0.03245692	0.03397537	0.03787260	0.00000000	0.03536591
BigAnvilTest	0.03193702	0.02867936	0.01918284	0.02230433	0.02626758	0.02964902	0.03536591	0.00000000

APPENDIX B ESTHETHIC DISTANCE MATRIX

Non-categorized
Newsletter
Brochure
Flyer

	Asymmetry	BigAnvilTest	Brochure-test1-p1	Brochure-test1-p2	Brochure-test1b-p1	Brochure-test2-p1	Brochure-test2b-p1	Flyer-test
Asymmetry	0.00000000	1.719134331	3.280611038	3.296463251	3.375676155	2.0160532	2.102156639	1.889389634
BigAnvilTest	1.71913433	0.00000000	1.782753587	1.800032973	2.210331917	1.012600303	1.372133851	0.977282166
Brochure-test1-p1	3.28061104	1.782753587	0.00000000	0.130925506	0.862537265	2.268520117	2.724840403	2.436211586
Brochure-test1-p2	3.29646325	1.800032973	0.130925506	0.00000000	0.907704651	2.293478012	2.745695591	2.456690311
Brochure-test1b-p1	3.37567616	2.210331917	0.862537265	0.907704651	0.00000000	1.879821062	2.161696434	2.116531372
Brochure-test2-p1	2.0160532	1.012600303	2.268520117	2.293478012	1.879821062	0.00000000	0.520097613	0.387923032
Brochure-test2b-p1	2.10215664	1.372133851	2.724840403	2.745695591	2.161696434	0.520097613	0.00000000	0.616107643
Flyer-test	1.88938963	0.977282166	2.436211586	2.456690311	2.116531372	0.387923032	0.616107643	0.00000000
Flyer-testb	1.31270623	0.693151236	2.414496183	2.438342094	2.455330372	0.861495733	1.073089361	0.654714942
FN9981502-LAYOUT-MQ1	1.95735455	0.398860276	1.427054405	1.447740674	1.980403304	1.133938551	1.523300648	1.170478225
FN9981502-LAYOUT-MQ2	2.60537004	0.993719876	1.377624512	1.422839165	1.986098766	1.173787832	1.565784931	1.335234046
Newsletter-2	2.2539525	0.665261865	1.284813762	1.314870834	1.744401932	1.013148665	1.485408902	1.191850901
Newsletter-2-BigGutter	2.46418452	0.917713046	0.978908539	1.076158524	1.497438073	1.339901447	1.812968969	1.516272783
Newsletter-2-Runaround	2.56486917	1.32436502	1.180405259	1.287391901	1.432805061	1.550672054	2.053264618	1.793059826
Newsletter2SmallGutter	2.2539525	0.665261865	1.284813762	1.314870834	1.744401932	1.013148665	1.485408902	1.191850901
Newsletter2SmallHeader	2.59019017	1.010356784	0.964346409	1.05525136	1.474534392	1.365380287	1.834214926	1.5526582
Newsletter-Blocks	2.2638154	0.832005501	1.520310402	1.561365724	1.673060894	1.14018321	1.609269977	1.309136987
Newsletter-Sided	2.11591029	1.084154487	1.463589072	1.49469018	1.738852262	1.239338517	1.751518726	1.513738036
Newsletter-SmallSide	2.22037435	1.368271947	1.292405248	1.389767289	1.498703361	1.486467957	2.004739523	1.774454832
Newsletter-SmallSide-b	2.99328661	2.082666636	1.422925711	1.52663064	1.452097416	1.953344226	2.434293509	2.251141787
Newsletter2-SmallGutLine	1.93423939	0.316232771	1.607186556	1.633349895	2.009338379	0.892910659	1.331586003	0.957735062
NewsSided-SmallParagraph	2.74157238	2.009652615	1.077015281	1.143519759	1.064472318	2.088635921	2.602909088	2.370357037
SymmetryTest	2.33420038	1.140051484	2.54771924	2.554892778	2.739751101	1.085118413	1.498669267	0.975722015

	Flyer-testb	FN9981502-LAYOUT-MQ1	FN9981502-LAYOUT-MQ2	Newsletter-2	Newsletter-2-BigGutter	Newsletter-2-Runaround	Newsletter2SmallGutter	Newsletter2SmallHeader
Asymmetry	1.3127062	1.957354546	2.605370045	2.2539525	2.464184523	2.564869165	2.2539525	2.59019017
BigAnvilTest	0.6931512	0.398860276	0.993719876	0.66526186	0.917713046	1.32436502	0.66526186	1.01035678
Brochure-test1-p1	2.4144962	1.427054405	1.377624512	1.28481376	0.978908539	1.180405259	1.28481376	0.96434641
Brochure-test1-p2	2.4383421	1.447740674	1.422839165	1.31487083	1.076158524	1.287391901	1.31487083	1.05525136
Brochure-test1b-p1	2.4553304	1.980403304	1.986098766	1.74440193	1.497438073	1.432805061	1.74440193	1.47453439
Brochure-test2-p1	0.8614957	1.133938551	1.173787832	1.01314867	1.339901447	1.550672054	1.01314867	1.36538029
Brochure-test2b-p1	1.0730894	1.523300648	1.565784931	1.4854089	1.812968969	2.053264618	1.4854089	1.83421493
Flyer-test	0.6547149	1.170478225	1.335234046	1.1918509	1.516272783	1.793059826	1.1918509	1.5526582
Flyer-testb	0.00000000	1.014969826	1.469762683	1.18908656	1.475439191	1.602509737	1.18908656	1.55650914
FN9981502-LAYOUT-MQ1	1.0149698	0.00000000	0.849344313	0.49193752	0.679525018	1.279674172	0.49193752	0.79332793
FN9981502-LAYOUT-MQ2	1.4697627	0.849344313	0.00000000	0.5199753	0.716848731	1.079804659	0.5199753	0.61874819
Newsletter-2	1.1890866	0.491937518	0.519975305	0.00000000	0.338029802	0.893532872	0.00000000	0.37446222
Newsletter-2-BigGutter	1.4754392	0.679525018	0.716848731	0.3380298	0.00000000	0.610077024	0.3380298	0.17573303
Newsletter-2-Runaround	1.6025097	1.279674172	1.079804659	0.89353287	0.610077024	0.00000000	0.89353287	0.52415872
Newsletter2SmallGutter	1.1890866	0.491937518	0.519975305	0.00000000	0.338029802	0.893532872	0.00000000	0.37446222
Newsletter2SmallHeader	1.5565091	0.793327928	0.618748188	0.37446222	0.17573303	0.524158716	0.37446222	0.00000000
Newsletter-Blocks	1.1771075	0.79650563	0.970139086	0.5104714	0.591531813	0.752713203	0.5104714	0.69738591
Newsletter-Sided	1.4057128	1.00939858	0.848199248	0.55647993	0.611858904	0.723900795	0.55647993	0.59919208
Newsletter-SmallSide	1.6268737	1.271573782	1.140481114	0.8918075	0.595817566	0.374707162	0.8918075	0.54590172
Newsletter-SmallSide-b	2.240777	2.015994787	1.529500127	1.57484245	1.378498316	0.942098677	1.57484245	1.23355877
Newsletter2-SmallGutLine	0.8432685	0.361732662	0.765985131	0.39513034	0.678398609	1.125607252	0.39513034	0.75981796
NewsSided-SmallParagraph	2.2003088	1.845459223	2.023303509	1.67907083	1.351086259	1.092588544	1.67907083	1.41361523
SymmetryTest	1.132171	1.409192204	1.383730888	1.35840631	1.688794374	1.826288223	1.35840631	1.70280647

	Newsletter-Blocks	Newsletter-Sided	Newsletter-SmallSide	Newsletter-SmallSide-b	Newsletter2-SmallGutLine	NewsSided-SmallParagraph	SymmetryTest
Asymmetry	2.263815403	2.115910292	2.220374346	2.99328661	1.934239388	2.74157238	2.334200382
BigAnvilTest	0.832005501	1.084154487	1.368271947	2.082666636	0.316232771	2.009652615	1.140051484
Brochure-test1-p1	1.520310402	1.463589072	1.292405248	1.422925711	1.607186556	1.077015281	2.54771924
Brochure-test1-p2	1.561365724	1.49469018	1.389767289	1.52663064	1.633349895	1.143519759	2.554892778
Brochure-test1b-p1	1.673060894	1.738852262	1.498703361	1.452097416	2.009338379	1.064472318	2.739751101
Brochure-test2-p1	1.14018321	1.239338517	1.486467957	1.953344226	0.892910659	2.088635921	1.085118413
Brochure-test2b-p1	1.609269977	1.751518726	2.004739523	2.434293509	1.331586003	2.602909088	1.498669267
Flyer-test	1.309136987	1.513738036	1.774454832	2.251141787	0.957735062	2.370357037	0.975722015
Flyer-testb	1.177107453	1.405712843	1.626873732	2.240777016	0.843268454	2.2003088	1.132171035
FN9981502-LAYOUT-MQ1	0.79650563	1.00939858	1.271573782	2.015994787	0.361732662	1.845459223	1.409192204
FN9981502-LAYOUT-MQ2	0.970139086	0.848199248	1.140481114	1.529500127	0.765985131	2.023303509	1.383730888
Newsletter-2	0.510471404	0.556479931	0.891807497	1.574842453	0.395130336	1.67907083	1.358406305
Newsletter-2-BigGutter	0.591531813	0.611858904	0.595817566	1.378498316	0.678398609	1.351086259	1.688794374
Newsletter-2-Runaround	0.752713203	0.723900795	0.374707162	0.942098677	1.125607252	1.092588544	1.826288223
Newsletter2SmallGutter	0.510471404	0.556479931	0.891807497	1.574842453	0.395130336	1.67907083	1.358406305
Newsletter2SmallHeader	0.697385907	0.599192083	0.545901716	1.233558774	0.759817958	1.413615227	1.702806473
Newsletter-Blocks	0.00000000	0.703057408	0.853983343	1.572804451	0.640918791	1.553529382	1.193684101
Newsletter-Sided	0.703057408	0.00000000	0.449054778	1.188802481	0.866192937	1.424999952	1.667140365
Newsletter-SmallSide	0.853983343	0.449054778	0.00000000	0.897431016	1.160457134	1.040137649	1.943719745
Newsletter-SmallSide-b	1.572804451	1.188802481	0.897431016	0.00000000	1.864457011	1.23030746	2.48065567
Newsletter2-SmallGutLine	0.640918791	0.866192937	1.160457134	1.864457011	0.00000000	1.861505747	1.112047911
NewsSided-SmallParagraph	1.553529382	1.424999952	1.040137649	1.23030746	1.861505747	0.00000000	2.73588419
SymmetryTest	1.193684101	1.667140365	1.943719745	2.48065567	1.112047911	2.73588419	0.00000000