MARLOS ALEX DE OLIVEIRA MARQUES

RECONSTRUÇÃO DIGITAL DE DOCUMENTOS MUTILADOS COM FORMAS REGULARES

Dissertação apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática.

CURITIBA

2009

MARLOS ALEX DE OLIVEIRA MARQUES

RECONSTRUÇÃO DIGITAL DE DOCUMENTOS MUTILADOS COM FORMAS REGULARES

Dissertação apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática.

Área de Concentração: *Computação Forenses e Biometria*

Orientador: Profª. Dra. Cinthia Obladen de Almendra Freitas

CURITIBA

2009

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central

M357r
2009

Marques, Marlos Alex de Oliveira
Reconstrução digital de documentos mutilados com formas regulares / Marlos Alex de Oliveira Marques ; orientadora, Cinthia Obladen de Almendra Freitas. – 2009.
ix, 55 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2009
Bibliografia: f. 53-55

1. Processamento de imagens – Técnicas digitais. 2. Sistemas imageadores. 3. Informática. I. Freitas, Cinthia Obladen de Almendra. II. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. III. Título.

CDD 20. ed. – 004

Pontifícia Universidade Católica do Paraná
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Informática

ATA DE DEFESA DE DISSERTAÇÃO DE MESTRADO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

DEFESA DE DISSERTAÇÃO Nº 04/2009

Aos 20 dias do mês de fevereiro de 2009 realizou-se a sessão pública de Defesa da Dissertação **"Reconstrução Digital de Documentos Mutilados com Formas Regulares"**, apresentada pelo aluno **Marlos Alex de Oliveira Marques** como requisito parcial para a obtenção do título de Mestre em Informática, perante uma Banca Examinadora composta pelos seguintes membros:

Profª. Drª. Cinthia Obladen de Almendra Freitas PUCPR (Orientadora)	(*assinatura*)	APROVADO (*aprov/reprov.*)
Prof. Dr. Alceu de Souza Britto Junior PUCPR		APROVADO
Prof. Dr. João Marques de Carvalho UFCG		Aprovado

Conforme as normas regimentais do PPGIa e da PUCPR, o trabalho apresentado foi considerado APROVADO (*aprovado/reprovado*), segundo avaliação da maioria dos membros desta Banca Examinadora. Este resultado está condicionado ao cumprimento integral das solicitações da Banca Examinadora registradas no Livro de Defesas do programa.

Prof. Dr. Mauro Sérgio Pereira Fonseca
Diretor do Programa de Pós-Graduação em Informática

Rua Imaculada Conceição, 1155 Prado Velho CEP 80215 901 Curitiba Paraná Brasil
Tel.: (41) 3271 1669 Fax: (41) 3271 2121 www.ppgia.pucpr.br

Dedico esse trabalho à minha esposa Luciane, ao meu filho Gabriel, à minha mãe Creusa e meu irmão Marcelo pelo companheirismo, inspiração e carinho.

Agradecimentos

Agradeço, primeiramente, a Deus por ter me abençoado, me guiado e me ajudado em todos os momentos desta jornada.

À minha família que sempre incentivou a luta pelos meus objetivos.

À Professora Dra. Cinthia Obladen de Almendra Freitas pela grande paciência, pela ajuda na construção do trabalho, com sua presença constante, sempre motivando, transmitindo seus conhecimentos com segurança, incentivando a pesquisa.

Ao Professor Dr. Flávio Bortolozzi pelo apoio e incentivo.

À Pontifícia Universidade Católica do Paraná, através do Programa de Pós-Graduação em Informática Aplicada (PPGIA), pelo apoio estrutural que permitiu a minha participação no Mestrado e a realização desse trabalho.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro ao presente trabalho de pesquisa.

Aos meus colegas de estudos, professores e demais funcionários do PPGIA.

Finalmente a todos aqueles que de alguma maneira me ajudaram na concretização desse trabalho.

Sumário

Agradecimentos

Sumário i

Lista de Figuras iii

Lista de Tabelas v

Lista de Símbolos vi

Lista de Abreviaturas vii

Resumo viii

Abstract ix

Capítulo 1

Introdução **1**

1.1. Desafio 2

1.2. Motivação 3

1.3. Objetivo 4

1.4. Contribuição 5

1.5. Organização 5

Capítulo 2

Fundamentação Teórica **6**

2.1. Modelos de Cor 6

2.1.1. Modelo RGB 7

2.2.2. Modelo HSV 9

2.3.3. Conversão de RGB para HSV 10

2.2. Revisão de Sistemas de Reconstrução de Documentos 11

2.3. Comentários Finais 20

Capítulo 3

Método para Reconstrução de Documentos Mutilados em Formato "Spaghetti" 21

3.1. Aquisição de Imagens..... 21

3.2. Base de Documentos..... 22

3.3. Extração das Características..... 26

3.4. Reconstrução Digital..... 27

3.5. Comentários Finais..... 32

Capítulo 4

Resultados Experimentais 33

4.1. Resultados usando o modelo de cor RGB através da soma dos canais–(RGB SOMA)... 33

4.2. Resultados usando o modelo de cor RGB através da combinação dos canais – (RGB)... 37

4.3. Resultados usando o canal R do modelo RGB..... 37

4.4. Resultados usando o canal G do modelo RGB..... 39

4.5. Resultados usando o canal B do modelo RGB..... 41

4.6. Resultados usando o modelo de cor HSV..... 44

4.7. Comparação entre os modelos de cor..... 45

4.8. Problemas na reconstrução do documento..... 48

4.9. Comparação entre o Sistema Proposto e o Sistema de Skeok..... 49

Capítulo 5

Conclusão 51

Referências Bibliográficas 53

Lista de Figuras

Figura 1	Departamento de documentoscopia do FBI remontando documento mutilado [FBI, 2007] [SOLANA, 2005].	2
Figura 2	Exemplo de fragmentos do tipo “spaghetti” de um documento.	4
Figura 3	Modelo de cor subtrativa CMY [CALIXTO, 2005].	7
Figura 4	Modelo de cor aditiva RGB [CALIXTO, 2005]..	8
Figura 5	Representação tridimensional do modelo RGB [GONZALEZ e WOODS, 2000].	8
Figura 6	Representação das cores no modelo RGB [GONZALEZ e WOODS, 2000].	9
Figura 7	Representação das cores no modelo HSV [CALIXTO, 2005].	9
Figura 8	Passos do processo de reconstrução desenvolvido pelo Solana [SOLANA, 2005].	11
Figura 9	Exemplo da base de imagens da PUCPR [SOLANA, 2005].	12
Figura 10	Documento fragmentado através do método “spaghetti” [SOLANA, 2005].	13
Figura 11	Simulação do processo da ChurchStreet Technology, Inc. (a) documento em tiras; (b) tiras digitalizadas; (c) documento reconstruído [CHURCHSTREET, 2007] [SOLANA, 2005].	14
Figura 12	Simulação do processo da ChurchStreet Technology, Inc. de tiras recortadas na vertical e aleatoriamente na horizontal [CHURCHSTREET, 2007] [SOLANA, 2005]..	14
Figura 13	Documento reconstruído utilizando o método de Skeoch [SKEOCH, 2006].	16
Figura 14	Documento reconstruído utilizando o método de Skeoch [SKEOCH, 2006].	16
Figura 15	Imagem dividida no computador em 25 partes iguais [SKEOCH, 2006].	17
Figura 16	Documento reconstruído utilizando o método de Skeoch [SKEOCH, 2006].	18
Figura 17	Documento reconstruído utilizando o método de Skeoch [SKEOCH, 2006].	19
Figura 18	Documento reconstruído utilizando o método de Skeoch [SKEOCH, 2006].	19
Figura 19	Esquema geral da metodologia.	21

Figura 20	Exemplo de aquisição de fragmentos de um documento utilizando um scanner.	22
Figura 21	Documento original não mutilado.	24
Figura 22	Documento original não mutilado.	24
Figura 23	Documento mutilado e digitalização dos fragmentos na ordem correta. [MARQUES E FREITAS, 2009]	25
Figura 24	Vetores de características: borda esquerda (VCE) e borda direita (VCD).	26
Figura 25	Exemplo de encaixe.	28
Figura 26	Problemas com "falsa" borda.	29
Figura 27	Diagrama de blocos dos algoritmos.	30
Figura 28	Matriz de Distância Euclidianas.	31
Figura 29	Exemplo de matriz com distâncias entre fragmentos.	31
Figura 30	Imagem de um documento da Classe 02 - Tipo texto.	34
Figura 31	Reconstrução do documento da Figura 30 através do modelo RGB.	35
Figura 32	Imagem do tipo Folders, Flyers, Anúncios, etc.	36
Figura 33	Imagem do tipo Folders, Flyers, Anúncios, etc., reconstruída usando o modelo RGB.	36
Figura 34	Imagem com os 3 canais. Imagem composta somente pelo canal R.	38
Figura 35	Imagem reconstruída usando o canal R.	39
Figura 36	Imagem com os 3 canais. Imagem composta somente pelo canal G.	40
Figura 37	Imagem reconstruída usando o canal G.	41
Figura 38	Imagem com os 3 canais. Imagem composta somente pelo canal B.	42
Figura 39	Imagem reconstruída usando o canal B.	43
Figura 40	Reconstrução usando o modelo HSV.	45
Figura 41	Desempenho do sistema proposto para todos os modelos de cor.	46
Figura 42	Exemplo de reconstrução na qual as bordas opostas se encontram.	47
Figura 43	Exemplo de reconstrução na qual as bordas totalmente brancas são reunidas ao final do documento.	48
Figura 44	Exemplo de anomalias das tiras.	49
Figura 45	Comparação de Resultados.	50

Lista de Tabelas

Tabela 1	Resultados dos testes de Skeoch [SKEOCH, 2006].	18
Tabela 2	Categorização dos documentos da base de dados.	26
Tabela 3	Média do sistema usando o modelo RGB através da soma dos canais – Classe 01.	33
Tabela 4	Média do sistema usando o modelo RGB através da soma dos canais – Classe 02.	34
Tabela 5	Média do sistema usando o modelo RGB através da combinação dos canais – Classe 01.	37
Tabela 6	Média do sistema usando o modelo RGB através da combinação dos canais – Classe 02.	37
Tabela 7	Média do sistema usando o canal R – Classe 01.	38
Tabela 8	Média do sistema usando o canal R – Classe 02.	38
Tabela 9	Média do sistema usando o canal G – Classe 01.	40
Tabela 10	Média do sistema usando o canal G – Classe 02.	40
Tabela 11	Média do sistema usando o canal B – Classe 01.	42
Tabela 12	Média do sistema usando o canal B – Classe 02.	42
Tabela 13	Resumo dos Resultados com os Canais R, G e B.	43
Tabela 14	Média do sistema usando o modelo HSV – Classe 01.	44
Tabela 15	Média do sistema usando modelo HSV – Classe 02.	44
Tabela 16	Melhores taxas do sistema proposto.	45
Tabela 17	Tabela de acertos o sistema proposto.	46
Tabela 18	Comparação de Resultados Sistema Proposto e Sistema Skeoch.	50

Lista de Símbolos

Σ	Somatória
max	Valor máximo de um conjunto de valores
min	Valor mínimo de um conjunto de valores
DE	Distância Euclidiana
x_i	Uma coordenada espacial
y_i	Uma coordenada espacial
n	Número de pixels (altura do fragmento)
r	Cor vermelha
g	Cor verde
b	Cor azul
h	Matiz
s	Saturação
v	Valor

Lista de Abreviaturas

CMY	Ciano, Magenta, Yellow
FBI	Federal Bureau of Investigation
HSI	Hue, Saturation, Intensive
HSL	Hue, Saturation, Lightness
HSV	Hue Saturation Value
JPEG	Joint Photographic Experts Group
MPEG	Moving Picture Experts Group
PUCPR	Pontifícia Universidade Católica do Paraná
RGB	Red Green Blue
VC	Vetores de Características
VCE	Vetores de Características da Borda Esquerda
VCD	Vetores de Características da Borda Direita

Resumo

Este trabalho apresenta um processo de reconstrução de documentos mutilados através de máquinas fragmentadoras (formato “spaghetti”), o qual caracteriza-se por ser um problema na área de ciências forenses, relativo à análise de documentos questionados. O método proposto extrai características baseadas na cor das bordas, e em seguida calcula a Distância Euclidiana e determina através do algoritmo Vizinho-Mais-Próximo, os pares de fragmentos para realizar a reconstrução do documento. A cor é extraída aplicando-se dois modelos diferentes: RGB (Red-Green-Blue) e HSV (Hue-Saturation-Value). Desta forma, a complexidade global do problema pode ser drasticamente reduzida, visto que métodos simples são utilizados para realizar a reconstrução. Os resultados preliminares relatados no presente documento (documentos na classe 01: 97,42% - HSV, documentos na classe 02: 98,53% - HSV), levando em conta os 200 documentos da base de dados, demonstraram que o método baseado na característica da cor produz resultados interessantes para o problema da reconstrução de documento podendo ser interessante para os examinadores de documentos forenses, bem como, oferecer algumas soluções eficazes para a aplicação prática que venha auxiliar a área da computação forense.

Palavras-Chave: Análise de documentos questionados, Reconstrução de documentos mutilados, Modelos de cor e Distância Euclidiana.

Abstract

This work presents a procedure for reconstructing destroyed documents that have been strip-shredded, which is a frequent problem in forensic sciences. The proposed method first extracts features based on color of the boundaries and then computes the nearest neighbor algorithm to carry out the local reconstruction. The color has been extracted applying two different models: RGB (red-green-blue) and HSV (hue-saturation-value). In this way the overall complexity can be dramatically reduced because few features are used to perform the matching. The preliminary results reported in this paper (document in class 01: 97.42% - HSV and document in class 02: 98.53% - HSV), which take into account a two hundred documents database, demonstrate that the color-matching based method produces interesting results for the problem of document reconstruction and can be of interest to the forensic document examiners, as well as to provide some effective solutions for law enforcement practitioners.

Keywords: Analysis of questioned document, strip-shredded document, Color-based feature extraction, Euclidean Distance.

Capítulo 1

Introdução

Na ciência forense a documentoscospia é a disciplina que trata do estudo ou análise de documentos e que possui grandes aplicações na criminalística, engenharia, artes, informática, arqueologia, biblioteconomia, entre outras. Além de representar uma área importante de pesquisa e que possui ainda um vasto inexplorado campo [MENDES, 2003].

Normalmente a documentoscopia é utilizada para determinar a autenticidade, a contemporaneidade, a associação ou dissociação da autoria do documento, em aplicações forenses [SOLANA, 2005].

Na documentoscopia um documento pode ser qualquer objeto ou fato que serve como prova, confirmação ou testemunho [LAROUSSE, 1998]. Entre outras situações a classificação do objeto ou fato pode estar associada ao material ou base onde o mesmo foi aposto.

O Departamento de Justiça dos Estados Unidos junto com o FBI (Federal Bureau of Investigation) [FBI, 2007], por exemplo, mantém um departamento para a análise de documentos questionados. As análises envolvem exames de escrita manuscrita, assinaturas, textos datilografados, impressos, rasuras, alterações e obliterações.

Os processos de reconstrução de documentos mutilados possuem um grau elevado de complexidade, sem considerar que muitos documentos são intencionalmente destruídos com a finalidade de ocultar informações que podem ser usadas como prova, dificultando a identificação ou interpretação de seu conteúdo, ou ainda escondendo ou disfarçando a verdadeira identidade do seu autor.

De acordo com Solana [SOLANA, 2005] a reconstrução de um documento mutilado é executada de forma manual, ou através de processos complexos e de difícil execução. E

dependendo da complexidade da mutilação e do tipo do documento isso pode levar dias de trabalho. Além do mais, os atuais processos de reconstrução alteram as propriedades do documento original, no qual provocam uma interferência nas propriedades químicas e nas eventuais impressões digitais no documento.

1.1. Desafio

Os documentos podem sofrer diversas mutilações durante sua vida útil podendo ser naturais ou involuntárias e intencionais ou voluntárias [UNB, 2007]. As mutilações naturais estão relacionadas com o aspecto da conservação do documento. Isto ocorre quando o documento sofre influência de fatores com a umidade, temperatura, poeira, poluição, fungos, insetos, microorganismos, catástrofes (enchentes, incêndios), etc. Já as mutilações intencionais ou voluntárias são efetuadas por pessoas com a ajuda de objetos como tesoura, régua, estilete, máquinas fragmentadoras, ou utilizando apenas as mãos para quebrar ou rasgar o documento, Tais mutilações são realizadas com o objetivo de destruir ou inutilizar documentos que poderiam ser utilizados como provas.

A identificação e catalogação dos fragmentos parceiros é o maior desafio para a remontagem de um documento mutilado [SOLANA, 2005], conforme mostrado na Figura 1.

Figura 1 – Departamento de documentoscopia do FBI remontando documento mutilado [FBI, 2007] [SOLANA, 2005].

Na reconstrução de um documento mutilado um aspecto a ser considerado é a conservação física dos fragmentos. Em alguns casos, o mau estado de conservação conduz a uma análise pericial não conclusiva.

A reconstrução digital de documentos fornece recursos que permitem a remontagem estrutural de um documento para auxiliar e viabilizar a análise pericial do mesmo, sem provocar modificações no documento original. Tal procedimento propõe um método não destrutivo para a reconstrução digital dos documentos mutilados, e ainda, visando à redução do tempo consumido por essa atividade. Alguns resultados neste sentido, para documentos mutilados de formas irregulares, podem ser obtidos em [SOLANA, 2005].

O presente trabalho considera documentos tendo como elemento físico de base o papel fragmentado de maneira regular intencional através de máquinas fragmentadoras de papel, constituindo fragmentos do tipo "spaghetti".

1.2. Motivação

O trabalho de reconstrução de documentos mutilados traz à discussão um campo de pesquisa abrangente, em função da existência de inúmeros interesses nesta recuperação. Neste contexto, se destaca: a aplicação em Ciências Forenses, na sub-área de documentoscopia, recuperação de documentos.

A reconstrução de documentos é desenvolvida, em geral, de forma artesanal e requer equipamento e pessoal especializado. Embora os documentos reconstruídos fiquem com o formato igual ao original, as partes danificadas ficam com textura e cores diferentes, além de que, os textos e figuras existentes nestas partes danificadas ficam incompletos.

As dificuldades existentes na reconstrução de documentos mutilados e a evolução constante da computação mostram o quanto é necessário desenvolver atividades de pesquisa para automatizar esses procedimentos. Isso ocorre principalmente na área forense, onde se encontra um grande volume de documentos em papel que podem ser utilizados como meio de prova.

Outro ponto relevante é que o presente projeto de pesquisa integrou o projeto "Desenvolvimento de Metodologias e Técnicas para Ciências Forenses" financiado pelo CNPq Proc. No. 476637/2006-6 (vigência 2006-2008).

1.3. Objetivo

O objetivo geral deste trabalho é descrever um método de reconstrução de documentos mutilados com formas regulares, no formato “spaghetti”, picotados por máquinas fragmentadoras, com o tamanho máximo de um papel A4, e coloridos como mostra a Figura 2. O método aplicado tem por base a cor dos fragmentos, considerando-se dois modelos de cores: RGB (Red-Green-Blue) e HSV (Hue-Saturation-Value).

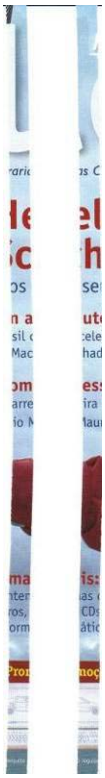


Figura 2 – Exemplo de fragmentos do tipo “spaghetti” de um documento.

A utilização dos modelos de cores RGB e HSV permite a identificação dos pixels das bordas dos fragmentos, sendo tais modelos utilizados como extratores de primitivas para que a semelhança entre as diferentes bordas possa ser analisada. Neste sentido aplica-se a Distância Euclidiana como método de medida de similaridade entre as bordas dos diversos fragmentos, permitindo verificar se estes fragmentos são ou não consecutivos (adjacentes), reconstruindo o documento original.

1.4. Contribuição

O presente trabalho apresenta um método para auxiliar na reconstrução digital de documentos com forma regular (formato "spaghetti"), visando possibilitar uma solução computacional sem a necessidade do manuseio direto dos documentos. A principal aplicação do método desenvolvido está relacionado com a área de perícias forenses de documentos questionados.

Outra contribuição importante é a criação da base de dados de documentos mutilados formada por 200 documentos diferentes entre si, a ser descrita posteriormente (Capítulo 3). Além disto, a implementação de algoritmos de extração de características baseados nos modelos de cores (RGB e HSV), e ainda, o procedimento para comparação dos fragmentos para reconstrução de documentos.

1.5. Organização

Este documento é composto por cinco capítulos. O Capítulo 2 é dedicado à revisão bibliográfica e apresentação dos sistemas existentes na área de digitalização e reconstrução de documentos mutilados. O Capítulo 3 mostra o método aplicado visando à reconstrução digital do documento e, ainda, são mostradas a criação e aquisição da base de imagens. O Capítulo 4 mostra os resultados experimentais e o Capítulo 5 conclui o documento e descreve possíveis trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste Capítulo é apresentada a revisão bibliográfica que foi utilizada no desenvolvimento deste trabalho. O Capítulo se divide em duas partes. Na primeira parte apresentam-se os modelos de cor e na segunda parte descrevem-se os métodos já existentes na literatura para a reconstrução digital de documentos.

2.1. Modelos de Cores

À análise da cor é de grande importância para extrair e identificar características, pois a cor está presente em tudo que se observa. A percepção da cor pelo homem é dada pela interação da luz com o sistema de visão, desta forma a interpretação das cores ocorre de maneira particular, sendo que seu significado depende das condições psicofísicas do observador. Para padronizar as cores foram criados modelos de cores [RAMOS, 2004].

Um modelo de cor facilita a especificação de cores respeitando um padrão de representação. O modelo, além de representar a cor propriamente dita, representa também os relacionamentos destas entre si. Mais especificamente, um modelo de cor é uma especificação de um sistema de coordenadas tridimensionais e um subespaço dentro deste sistema, onde cada cor é representada por um único ponto.

Diferentes sistemas de processamento de imagem utilizam diferentes modelos de representação de cores. Dentre os modelos mais utilizados no mercado encontram-se: RGB (Red, Green, Blue), CMY (Ciano, Magenta, Yellow), HSV (Hue, Saturation, Value) ou HSI (Hue, Saturation, Intensive), HSL (Hue, Saturation, Lightness). A escolha de um modelo de cores para um determinado sistema depende de diversas variáveis, como por exemplo: área de atuação do sistema, tempo necessário para o processamento de imagens, informações

relevantes da imagem para tratamento, condições ideais para o algoritmo de tratamento de imagens, entre outras.

Os modelos mais usados para imagens coloridas são o RGB, que são usados em monitores coloridos e câmeras de vídeos e o modelo HSV usado para manipulação de imagens coloridas [GONZALEZ E WOODS, 2002].

2.1.1. Modelo RGB

É a abreviatura do sistema de cores aditivas formado pelo vermelho (Red), verde (Green) e azul (Blue). O sistema aditivo de cores utiliza a projeção da luz, usado em monitores de vídeo, diferentemente do sistema subtrativo, que é usado em impressoras, que usa o modelo CMY. Neste modelo as cores primárias absorvem alguns comprimentos de onda da luz branca e refletem os comprimentos restantes para a retina como mostra a Figura 3 [CALIXTO, 2005].

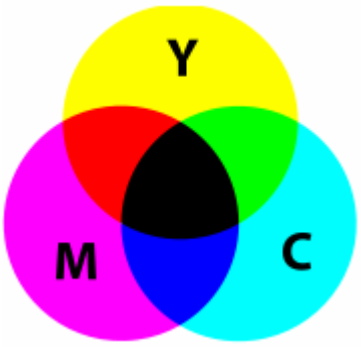


Figura 3 – Modelo de cor subtrativa CMY [CALIXTO, 2005].

O modelo de cores RGB é baseado na teoria de visão colorida tricromática, de Young-Helmholtz, e no triângulo de cores de Maxwell. Este modelo é o mais adequado e utilizado para a representação de cores em dispositivos de apresentação, como monitores de vídeo e alguns aplicativos gráficos computacionais. Entretanto, a leitura deste modelo não transmite uma idéia exata e natural de qual cor será percebida [CALIXTO, 2005].

No modelo RGB toda cor associada a um pixel é representada pela adição dos valores primários (vermelho, verde e azul) como mostra a Figura 4, mas estas três cores não devem ser confundidas com cores primárias usadas no mundo das artes. Os três valores são fortemente correlacionados entre si, de modo que uma variação no brilho, sem alterar a cor,

implica em uma variação não-linear em todos eles. Esta observação evidencia que não faz parte da linguagem natural descrever cores por esta terna.

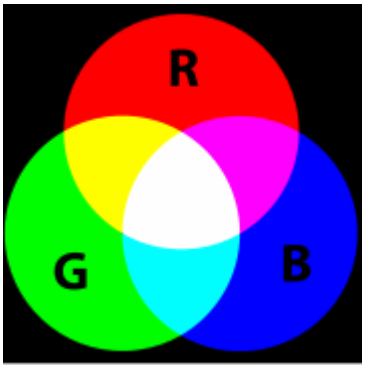


Figura 4 – Modelo de cor aditiva RGB [CALIXTO, 2005].

A representação do modelo RGB é feita através de um cubo em três dimensões, conforme ilustram as coordenadas a Figura 5 e as cores na Figura 6 [GONZALEZ E WOODS, 2000].

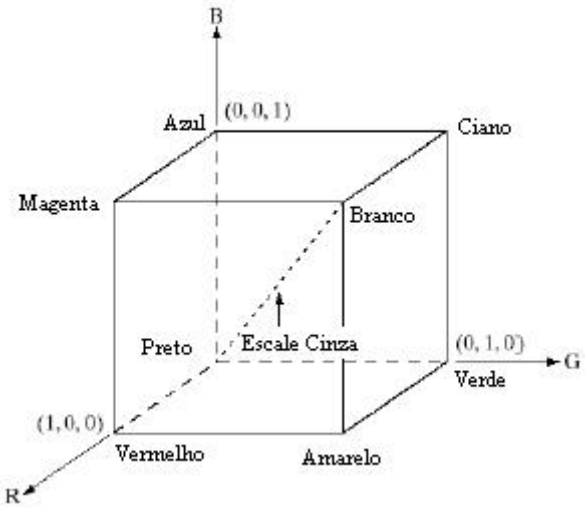


Figura 5 – Representação tridimensional do modelo RGB [GONZALEZ E WOODS, 2000].

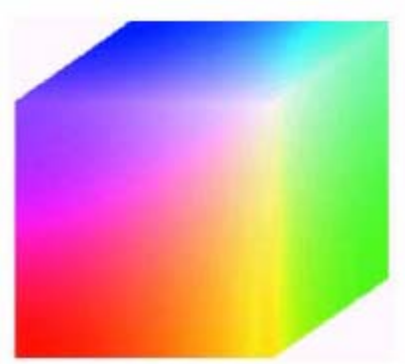


Figura 6 – Representação das cores no modelo RGB [GONZALEZ E WOODS, 2000].

Em geral define-se em três o número de cores primárias em um modelo, devido ao fato do olho humano possuir três tipos de fotoreceptores. Nem todos os modelos de cor possuem uma base: nos modelos de cores HSV e HSL não existe um grupo de cores primárias, pois este espaço não é obtido pela composição de cores [CALIXTO, 2005].

2.1.2. Modelo HSV

O Modelo HSV (Hue - matiz ou tonalidade, Saturation - saturação, Value - intensidade), foi desenvolvido em 1978 por Alvey Ray Smith, baseando-se nas misturas de cores. O matiz (H) é a cor pura da imagem, a saturação (S) indica o afastamento da cor e a intensidade (V) é a luz refletida pela superfície do objeto, ou o brilho [GONZALEZ E WOODS, 2000] [CALIXTO, 2005], tal qual representado na Figura 7.

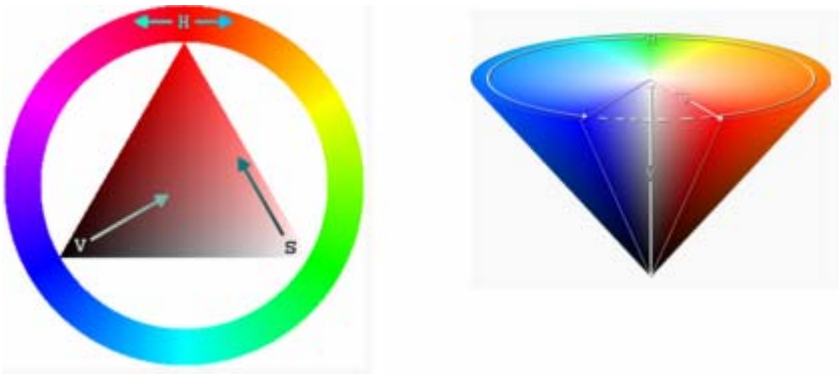


Figura 7 – Representação das cores no modelo HSV [CALIXTO, 2005].

A grande vantagem de se trabalhar com o modelo HSV é a possibilidade de separar a intensidade da informação tonalidade e saturação, e a relação que existe entre esses

componentes, que está muito próxima da forma pela qual o homem percebe a cor [RAMOS, 2004].

2.1.3. Conversão de RGB para HSV

As cores do modelo HSV são definidas matematicamente por transformações das coordenadas R, G e B do espaço RGB para as coordenadas H, S e V do espaço HSV, devidamente normalizadas, dadas por [CALIXTO, 2005]:

$$r = \frac{\mathrm{Re}\,d}{255} \tag{1}$$

$$g = \frac{Green}{255} \tag{2}$$

$$b = \frac{Blue}{255} \tag{3}$$

Após obter os canais R, G e B normalizados, a equação (4) é utilizada para o cálculo da matiz, saturação e intensidade.

Para encontrar a matiz H, do espaço HSV, deve-se encontrar o valor máximo e mínimo dos canais R, G, B e depois calcular [CALIXTO, 2005]:

$$h = \begin{cases} 0 & \text{if } \max = \min \\ 60^\circ \times \frac{g-b}{\max - \min} + 0^\circ, & \text{if } \max = r \text{ and } g \geq b \\ 60^\circ \times \frac{g-b}{\max - \min} + 360^\circ, & \text{if } \max = r \text{ and } g < b \\ 60^\circ \times \frac{b-r}{\max - \min} + 120^\circ, & \text{if } \max = g \\ 60^\circ \times \frac{r-g}{\max - \min} + 240^\circ, & \text{if } \max = b \end{cases} \tag{4}$$

Para encontra a saturação S e intensidade V, segue essa fórmula [CALIXTO, 2005].

$$v = \max \tag{5}$$

$$s = \begin{cases} 0, & \text{if } \max = 0 \\ \frac{\max - \min}{\max} = 1 - \frac{\min}{\max}, & \text{otherwise} \end{cases} \tag{6}$$

A escolha dos modelos RGB e HSV é devido ao fato que o RGB é o mais utilizado e adequado para a representação de cores em dispositivos de apresentação e o HSV possibilita a

separação das intensidades da tonalidade e saturação, deste modo a relação dos componentes se aproxima da forma pela qual o olho humano percebe as cores.

O uso dos dois modelos, RGB e HSV, tem a intenção de tornar complementares as informações obtidas através destes modelos de representação.

2.2. Revisão de sistemas de reconstrução de documentos

Já existem sistemas que realizam a reconstrução parcial ou semi-automática de documentos fragmentados com formas regulares ou irregulares.

O trabalho desenvolvido por Solana [SOLANA, 2005] consiste em um método para a reconstrução digital de documentos mutilados irregulares através da aproximação poligonal. O processo de reconstrução desenvolvido por Solana inicia com a conversão e tratamento das imagens para níveis de cinza buscando a eliminação do fundo, para depois fazer a extração do contorno do fragmento. A identificação dos fragmentos é realizada através da aproximação poligonal, buscando a vizinhança existente entre os fragmentos para posterior reconstrução do documento como mostra a Figura 8.

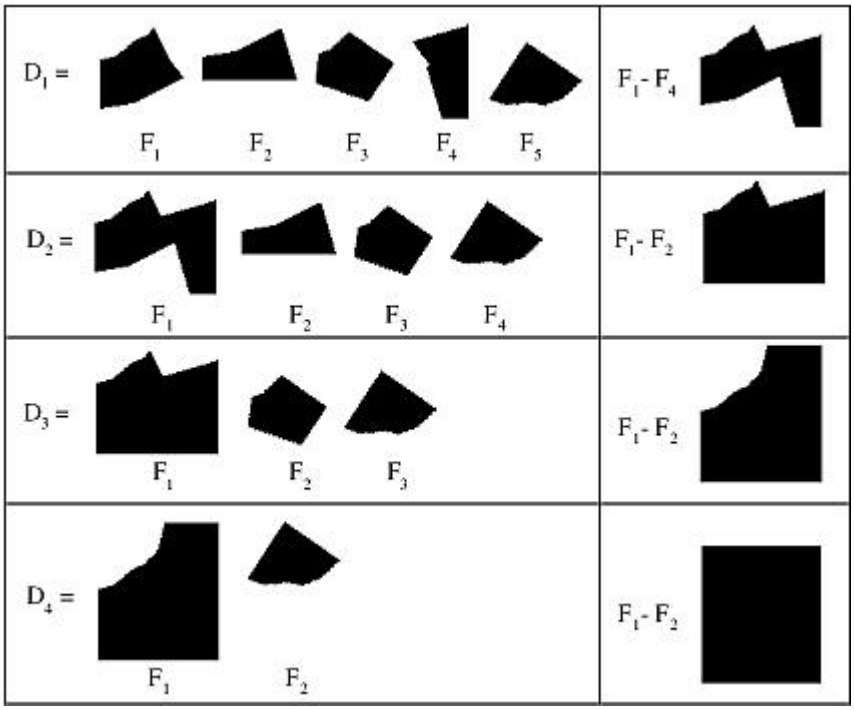


Figura 8 – Passos do processo de reconstrução desenvolvido pelo Solana [SOLANA, 2005].

O melhor resultado, obtido na base de imagens da PUCPR, Figura 9, foi de 5% de documentos sem nenhuma convergência, 50% de convergência parcial e 45% com convergência completa, sendo que, dos documentos que obtiveram convergência completa 86,67% desses documentos foram classificados corretamente e 13,33% apresentaram falsos candidatos [SOLANA, 2005]. Esse experimento foi realizado com um nível baixo de tolerância na aproximação poligonal.

Figura 9 – Exemplo da base de imagens da PUCPR [SOLANA, 2005].

Os documentos fragmentados ou triturados em formas regulares através de máquinas trituradoras, são também conhecidos como trituradores “spaghetti”, conforme Figura 10.

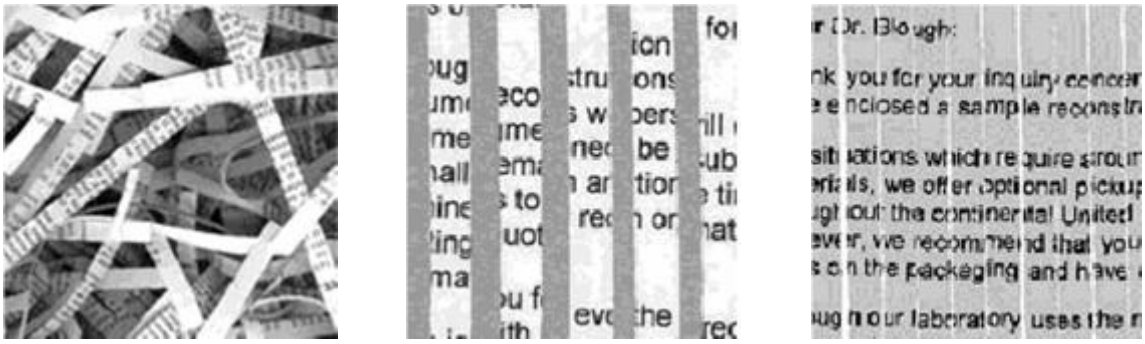
Existe um sistema semi-automático de reconstrução de documentos fragmentados ou triturados em formas regulares, desenvolvido pela empresa ChurchStreet Technology, Inc. [CHURCHSTREER, 2007]. Esse sistema funciona em três etapas:

- digitalização e conservação dos fragmentos em imagens digitais;
- catalogação e busca das características gráficas no contexto dos fragmentos;

- reconstrução e colocação dos fragmentos na ordem correta.

Figura 10 – Documento fragmentado através do método “spaghetti” [SOLANA, 2005].

O processo de reconstrução inicia lendo e atribuindo um identificador eletrônico único a cada fragmento; depois compilar os fragmentos que contêm traços similares. A tarefa seguinte é colocar os fragmentos na ordem apropriada para reconstrução da página [CHURCHSTREER, 2007], conforme mostra a simulação da Figura 11.



(a) (b) (c)

Figura 11 – Simulação do processo da ChurchStreet Technology, Inc. (a) documento em tiras; (b) tiras digitalizadas; (c) documento reconstruído [CHURCHSTREET, 2007] [SOLANA, 2005].

Devido ao sistema da ChurchStreet, os fabricantes de trituradores de papéis desenvolveram novos equipamentos. Os novos trituradores fazem os cortes das tiras na vertical e aleatoriamente na horizontal, resultando em pequenos retângulos de papel, de tamanho variado e formas regulares conforme a Figura 12.

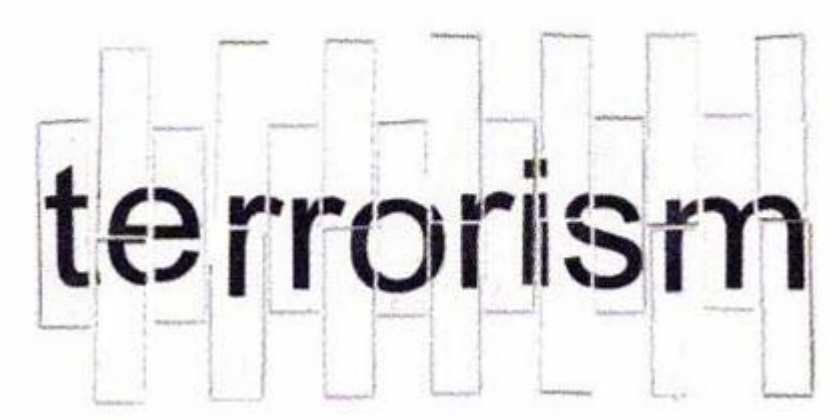


Figura 12 – Simulação do processo da ChurchStreet Technology, Inc. de tiras recortadas na vertical e aleatoriamente na horizontal [CHURCHSTREET, 2007] [SOLANA, 2005].

Decorrente dos novos trituradores de papel a ChurchStreet Technology, Inc. informou ter desenvolvido uma nova tecnologia, da qual não publicou a metodologia, passando a trabalhar com esse novo tipo de fragmento. Agora, para reconstrução dos fragmentos devem-se enviar as tiras dos documentos para seu laboratório, para depois devolverem os resultados da reconstrução através de arquivos magnéticos [CHURCHSTREET, 2007].

A vantagem do sistema da ChurchStreet Technology, Inc. é de fazer a reconstrução automática de documento em tiras, de formas regulares, usando as características das tiras.

As desvantagens são que, para reconstrução, devem existir todos os fragmentos do documento, as tiras precisam ser separadas e preparadas manualmente para digitalização e este método não preserva a integridade do documento devido ao manuseio do fragmento, podendo interferir na análise forense de outras características, como impressões digitais e elementos químicos. Outra desvantagem é que se houverem diversas partes de vários documentos misturados o sistema não reconstrói e nem faz a separação das tiras do documento [CHURCHSTREET, 2007].

Alguns estudos recentes abordam o problema da reconstrução automática de documentos mutilados com formas regulares. Os que se destacam são os trabalhos de Ukovich [UKOVICH ET AL., 2004] e Skeoch [SKEOCH, 2006].

Ukovich [UKOVICH ET AL., 2004] sugere a reconstrução de imagens baseada no seu conteúdo. A reconstrução de imagem baseada em conteúdo é uma forma de classificação que utiliza o conteúdo visual da imagem, como cor ou textura. Ukovick usa esta técnica para distinguir entre fragmentos de diferentes documentos, pois se deduz que imagens que contenham conteúdos semelhantes provavelmente irão pertencer ao mesmo documento. Em

especial usam o descritor de MPEG-7 como um método de distinguir fragmentos de uma imagem. O MPEG-7 é um tipo específico de descritor desenvolvido pelo Moving Picture Experts Group (MPEG), para descrever o conteúdo multimídia. Ukovich [UKOVICH ET AL., 2004] usa três descritores de cor, dois descritores de textura e dois descritores de forma.

O trabalho de Ukovich obteve alguns bons resultados usando os descritores de cor, mas não usando os descritores de textura e forma. A reconstrução de imagem baseada em conteúdo é uma técnica válida para resolver este problema. Experiências usando o descritor MPEG-7 padrão demonstraram que os recursos usados com a finalidade de reconstrução de imagem baseando-se em conteúdo podem ser usados para esta tarefa, em especial para os descritores de cor. Os descritores de textura de MPEG-7 usados não forneceram os resultados esperados, indicando a necessidade de encontrar outros descritores de textura, livre de problemas de dimensão de imagem. Os descritores de forma são úteis apenas no caso de uma forte curvatura nos restos de corte de documento e em algumas funcionalidades específicas, tal como OCR. Recursos que descrevam o conteúdo do fragmento na região perto das duas fronteiras horizontais, precisam também ser explorados [UKOVICH ET AL., 2004].

O trabalho desenvolvido por Skeoch [SKEOCH, 2006] apresenta um método para a reconstrução digital de documentos mutilados regulares através de um algoritmo de pesquisa heurística. O processo de reconstrução desenvolvido por Skeoch inicia com a aquisição dos fragmentos usando um *scanner* e tratamento das imagens, extração da característica da cor das bordas, combinação dos pares de pixels das bordas utilizando métricas de comparação e por fim a reconstrução dos fragmentos é realizada através de algoritmos genéticos.

Embora o sistema de Skeoch [SKEOCH, 2006] tenha apresentado bons resultados em várias imagens, a conclusão final foi que o sistema não produz resultados corretos, isso devido a uma má adequação da função de estimação relacionada com o algoritmo genético e também por parte do processo de extração das características. O algoritmo genético se mostrou confiável para imagens recortadas por computador, pois geralmente produzem a solução correta. Apesar disto, alguns aspectos do sistema funcionam bem, podendo ser claramente visto que, quando a cor está presente, uma boa solução parcial pode ser alcançada, Figura 13. No entanto, o desempenho com base em imagens com texto é geralmente baixo, Figura 14.

O sistema de Skeoch pode ser muito lento na fase da extração das características dos fragmentos, especialmente quando se trata de imagens de tamanho A4 que tem um grande

número de pixels e normalmente contêm cerca de 26/27 fragmentos. Várias medidas foram contempladas para calcular a distância entre os valores cromáticos sendo elas a Distância Euclidiana, Manhattan, Chebychev, Minkowski, Mahalanobis, Canberra, Cosseno e a variação da Distância Euclidiana também foi considerada (NSR Euclideano).

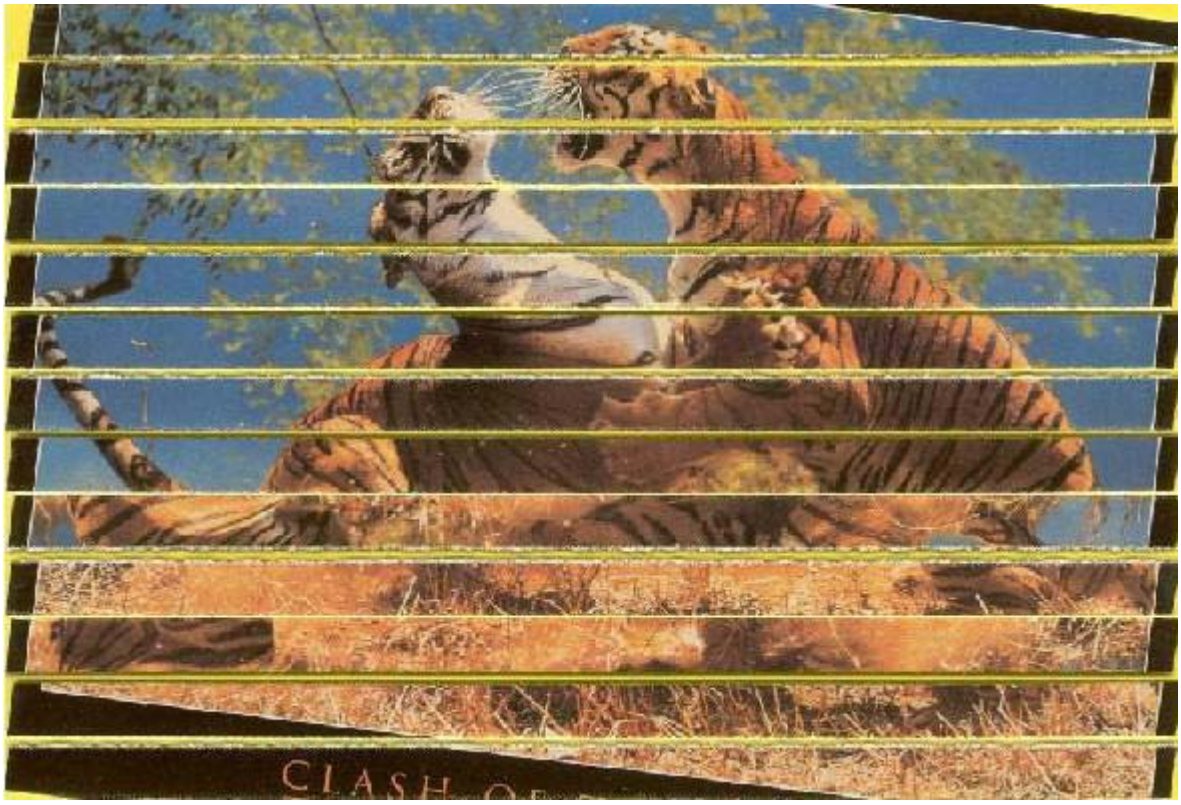


Figura 13 – Documento reconstruído utilizando o método de Skeoch [SKEOCH, 2006].

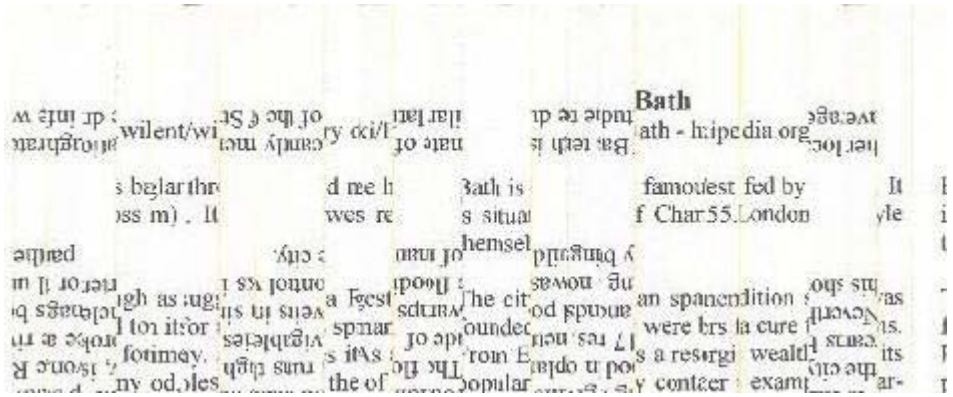


Figura 14 – Documento reconstruído utilizando o método de Skeoch [SKEOCH, 2006].

Skeoch [SKEOCH, 2006] comparou as diferentes medidas de distância, usando os espaços de cor HSV e RGB e pode se observar que, em geral, há pouca diferença na precisão entre os dois modelos. Usando o espaço de cor HSV a avaliação é mais precisa, mas não é tão maior do que o espaço cor RGB. Ao avaliar uma imagem com pouca variação de cor (Figura 15), os resultados são menos precisos, mas ainda dentro dos níveis aceitáveis.

Figura 15: Imagem dividida no computador em 25 partes iguais [SKEOCH, 2006].

Skeoch [SKEOCH, 2006] concluiu que o modelo HSV não oferece vantagens óbvias sobre o modelo RGB, além do excesso de computação necessário para converter imagens HSV para RGB. Do mesmo modo, nenhuma das medidas de distância superou significativamente as demais, mas a medida adotada pela autora foi Chebychev, devido à redução dos recursos computacionais. Na maioria das imagens, a avaliação da autora é que o sistema proposto funciona bem tanto nos espaços de cor RGB e HSV para todas as medidas. Os melhores resultados que Skeoch [SKEOCH, 2006] obteve com imagens fragmentadas computacionalmente encontram-se resumidos na Tabela 1.

Tabela 1. Resultados dos testes de Skeoch [SKEOCH, 2006].

Imagem (Fragmentos)	Reconstrução Correta
Waterfall(13)	0.9696
New York(25)	0.9462
Tower(20)	0.9646
Purple(15)	0.9758

Skeoch [SKEOCH, 2006] realizou também experimentos com mistura de documentos. As Figuras 16, 17 e 18 mostram os resultados para a mistura de 2 documentos. A autora realizou a mistura utilizando somente 2 documentos a cada experimento. Observa-se, então, as confusões realizadas entre os diferentes fragmentos e que o sistema não conseguiu ao menos separar os fragmentos pertencentes a cada um dos 2 documentos testados.

Figura 16 – Documento reconstruído utilizando o método de Skeoch [SKEOCH, 2006].

Figura 17 – Documento reconstruído utilizando o método de Skeoch [SKEOCH, 2006].

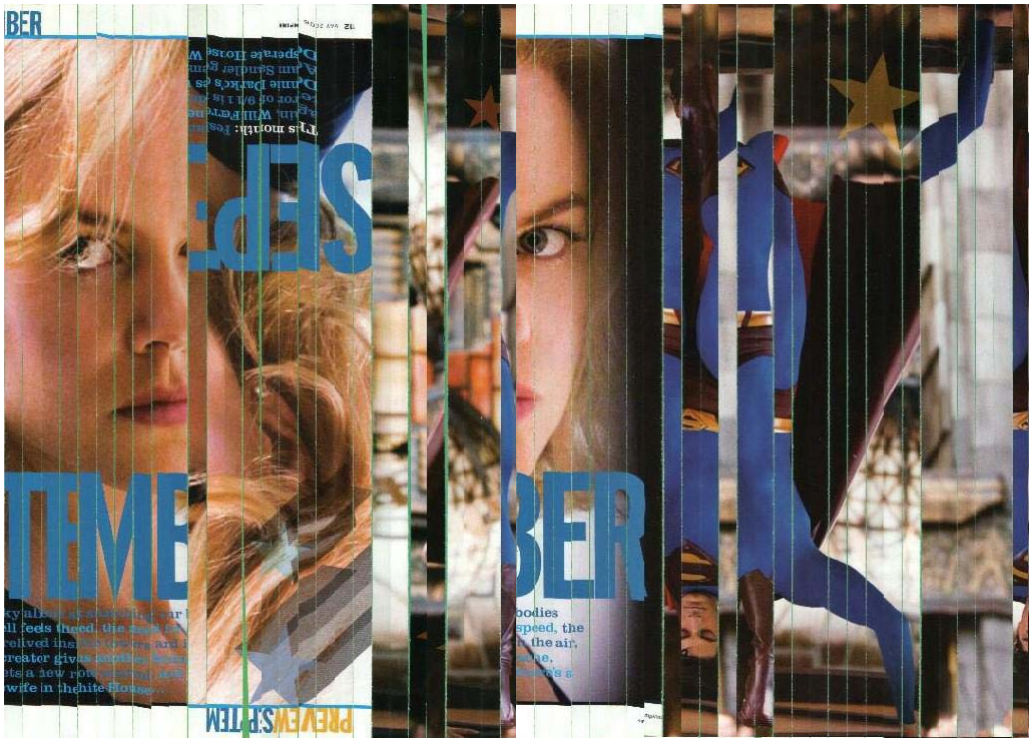


Figura 18 – Documento reconstruído utilizando o método de Skeoch [SKEOCH, 2006].

2.3. Comentários Finais

No presente Capítulo foi apresentado a revisão bibliográfica dos temas utilizados no desenvolvimento desse trabalho. Com a apresentação dos modelos de cores mais utilizados e que serão utilizados no desenvolvimento do sistema de reconstrução e também foi analisado o sistema desenvolvido pela ChurchStreet Technology, Inc. [CHURCHSTREET, 2007] destinado à reconstrução semi-automática de documentos triturados em formas regulares. Além disto apresentou-se o método apresentado de Solana [SOLANA, 2005] para reconstrução através da aproximação poligonal com formas irregulares. Ressalta-se o trabalho desenvolvido por Skeoch [SKEOCH, 2006] visto que o trabalho refere-se a documentos com forma regular tipo "spaghetti". No Capítulo 3 é abordado o método desenvolvido e aplicado para concretização do trabalho.

Capítulo 3

Método para Reconstrução de Documentos Mutilados em Formato "Spaghetti"

Este Capítulo apresenta o método para a reconstrução digital de documentos, que opera em três passos. O primeiro passo consiste na aquisição dos fragmentos e na localização das bordas das imagens. O segundo passo consiste na extração de características baseadas nas cores da borda do fragmento, utilizando os modelos de cor RGB e HSV. Finalmente o terceiro passo realiza a organização e junção dos prováveis fragmentos com base na Distância Euclidiana, como mostrado na Figura 19.

Figura 19 – Esquema geral da metodologia.

3.1. Aquisição de Imagens

A captura do documento pode ser realizada através do uso de câmera fotográfica, ou de *scanner*. Observa-se que na pesquisa para documentos com superfície lisa, como exemplo

a folha de papel, a melhor forma de aquisição é através de *scanner* de mesa [LEITÃO, 2000] [MELLO, 2002]. Assim, os fragmentos dos documentos foram digitalizados através de *scanner* de mesa conforme exemplo da Figura 20.

No início da criação da base de dados cada fragmento foi digitalizado separadamente. No decorrer do projeto verificou-se que seria mais ágil digitalizar todos os fragmentos do documento, para depois separá-los em arquivos. Para isso foram utilizadas as cores de fundo verde ou branco, em função da predominância de cores no documento original.

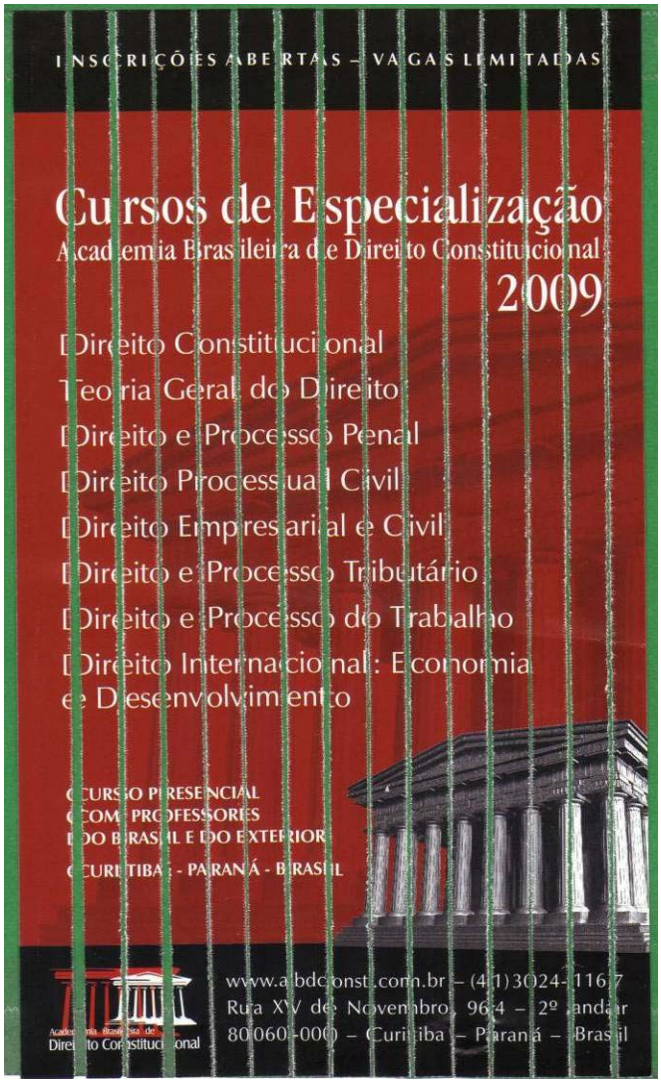


Figura 20: Exemplo de aquisição de fragmentos de um documento utilizando um *scanner*.

3.2. Base de documentos

Para o desenvolvimento do trabalho foi criada uma base de documentos em papel, fragmentados intencionalmente através de uma máquina picotadora de papel tipo “spaghetti”.

A primeira atividade desenvolvida foi a criação desta base. Cada documento foi picotado entorno em 29 fragmentos medindo 0,7 cm na horizontal e 27 cm na vertical, conforme a especificação da máquina fragmentadora de papel Cadence, modelo FRG712. Os documentos contêm textos tipografados com imagens, tabelas e outras características. As páginas foram classificadas manualmente nas seguintes categorias:

a) página só com texto;

b) página com texto e figura ou tabela;

c) página só com figura ou tabela.

Esta classificação auxiliou no entendimento dos acertos e erros gerados pelo sistema, bem como, permitiu ponderar os erros e acertos.

A criação da base de documentos seguiu as seguintes especificações:

a) formato retangular ou quadrado;

b) tamanho máximo A4;

c) duas ou mais cores.

d) orientação do papel na vertical.

Os documentos que compõem a base foram digitalizados seguindo as seguintes especificações:

a) tipo de imagem: cor a 24 bits;

b) resolução: 300 dpi com 100% de qualidade da imagem.

Para se ter uma noção de tamanho das imagens a serem processadas, por exemplo, a Figura 21 é um folheto que mede 15cm (horizontal) por 21cm (vertical) e após digitalizado possui 501 kbytes em formato JPEG (1738x2430 pixels).

Figura 21: Documento original não mutilado.

Após a obtenção dos documentos os mesmos passaram pelos seguintes procedimentos:

a) digitalização do documento inteiro (numeração principal, exemplo, D01), Figura 22;

Figura 22: Documento original não mutilado.

b) fragmentação do documento utilizando uma máquina fragmentadora de papel Cadence, modelo FRG712;

c) numeração dos fragmentos (numeração secundária, exemplo, D0101, D0102... D0129, pois as máquinas geram até 29 fragmentos distintos);

d) digitalização dos fragmentos na ordem correta, da montagem do documento original, Figura 23;

e) armazenamento dos fragmentos de cada documento em um envelope de papel pardo;

f) realização do embaralhamento aleatório computacionalmente dos fragmentos de um mesmo documento (armazenar a rotulação da seqüência embaralhada, por exemplo, 06-01-03-05-02-04).

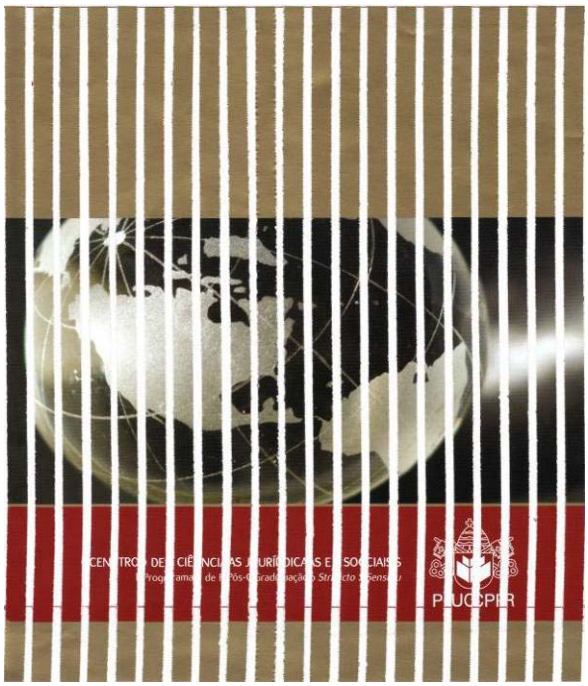


Figura 23: Documento mutilado e digitalização dos fragmentos na ordem correta. [MARQUES E FREITAS, 2009].

Assim, a base de dados contém atualmente 200 documentos categorizados de acordo com a Tabela 2. Esta categorização permite entender a complexidade da base de dados e, ainda, analisar os resultados obtidos. Uma descrição mais detalhada dos documentos que compõem a base é a seguinte:

- Documentos de textos tipografados - documentos contendo exclusivamente textos tipografados, por exemplo, ofícios, memorandos e outros;

- Documentos de textos com figuras - documentos que possuem textos, mas possuem também imagens (figuras, ilustrações);

- Documentos de textos com tabelas - documentos que possuem textos, mas possuem também tabelas;

- Documentos de textos com figuras e tabelas - documentos que possuem textos, mas possuem também imagens, tabelas e outras características que os diferenciam daqueles que possuem apenas textos.

Tabela 2. Categorização dos documentos da base de dados.

Classe 01	Quantidade	Classe 02	Quantidade
Somente texto	20	Revistas	110
Texto e figura	50	Ofícios, Memorandos, etc.	60
Texto e tabela	20	Folders, Flyers, Anúncios, etc	30
Texto, figura e tabela	110	----	----
Total	**200**	**Total**	**200**

3.3. Extração das Características

Para cada fragmento foram obtidos dois vetores de características (VC) correspondentes à dimensão vertical (eixo y) para cada uma das bordas do fragmento que serão usados para se calcular a distancia entre os dois VC. Cada vetor, por sua vez, recebe como informação a cor de cada pixel encontrado na borda esquerda (VCE) e na borda direita (VCD). A Figura 24 apresenta, de maneira ilustrativa, os vetores obtidos para o fragmento 01 do documento original mostrado na Figura 21.

Figura 24: Vetores de características: borda esquerda (VCE) e borda direita (VCD).

O VC no modelo RGB, extraído é a soma dos canais R, G e B, sendo que a cor é representada pela adição dos valores primários (vermelho, verde e azul). No modelo HSV o VC extraído será o canal H (hue - tonalidade), considerada como a cor pura. Através desta informação poder-se-á calcular a Distância Euclidiana e fazer a comparação dos resultados para reconstrução da imagem original.

Os VC's provenientes da extração de primitivas a partir dos dois diferentes modelos (RGB e HSV) foram comparados separadamente entre si, e posteriormente foi efetuada a combinação dos resultados.

Durante a extração das características das bordas foram desconsideradas 3 colunas de cada lado do fragmento para se evitar o problema da "falsa" borda que será explicada no item 3.4.

3.4. Reconstrução Digital

O método de classificação por Distância Euclidiana é um procedimento que utiliza esta distância para associar um "objeto" a uma determinada classe.

A Distância Euclidiana é uma das medidas de dissimilaridade entre dois pontos mais usados na prática [GAUCH, 1982]. De acordo com [BROWER E ZAR, 1977], quanto menor o valor da Distância Euclidiana entre dois pontos, mais próximos eles se apresentam em termos de parâmetros quantitativos por classe, logo, quanto menor a Distância Euclidiana, maior a eficiência do procedimento.

A busca pela seqüência correta de fragmentos tem por base o cálculo da Distância Euclidiana entre os VC extraídos para as bordas esquerda e direita de cada fragmento. A proposta é utilizar a Distância Euclidiana como medida da dissimilaridade entre os VC's das bordas. Assim, de acordo com [RAMOS, 2004], a Distância Euclidiana é definida como:

$$DE = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{7}$$

Sendo:

DE = Distância Euclidiana

x_i = valor do pixel *i* para o fragmento X;

y_i = valor do pixel *i* para o fragmento Y;

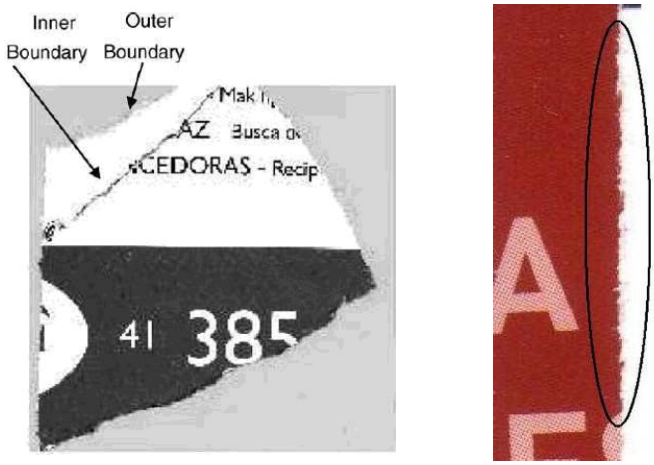
n = número de pixels (altura do fragmento).

A idéia deste tipo de medida é que quanto menor o valor obtido para DE maior será a similaridade entre as bordas dos fragmentos analisados, como exemplificado na Figura 25.

Figura 25: Exemplo de encaixe.

Uma das dificuldades encontradas ao se reconstruir digitalmente documentos mutilados é a "falsa" borda, tal qual descrito por Oliveira [OLIVEIRA et al, 2006]. Nestes casos o fragmento com forma irregular apresenta "falsa" borda devido a separação das fibras do papel durante a ação de rasgar o documento, como mostrado na Figura 26(a). Por outro lado, quando o fragmento possui forma regular a "falsa" borda é gerada pela ação da máquina fragmentadora que retira a película de tinta do papel, como exemplificado na Figura 26(b).

Após a reconstrução dos documentos aplicando-se os dois modelos de cor, os resultados foram comparados com o documento original para analisar as taxas de acerto e os tipos de erros cometidos pelo método proposto, visando avaliar qual entre os dois modelos de cor (RGB e HSV) é mais adequado ao problema em questão, a influência das "falsas" bordas, bem como peculiaridades relacionadas aos tipos de documentos que compõem a base de dados.



(a) (b)

Figura 26: Problemas com "falsa" borda: a) irregular [OLIVEIRA et al, 2006] e b) regular.

Para a realização dos testes com o sistema proposto, foram criados 3 algoritmos, a saber, como mostrado na Figura 27:

- extração das características da borda gerando os VCE e VCD de cada fragmento para os modelos de cor considerados (RGB e HSV) através da criação de arquivos com os respectivos vetores. Este arquivo contém também a informação da altura de cada fragmento, importante para que o sistema não compare e tente realizar o encaixe de fragmentos de alturas distintas;

- leitura dos arquivos para cálculo das distâncias entre o VCE de um fragmento F_1 e os VCDs dos demais fragmentos F_2 até F_n gerando uma matriz que é armazenada em outro arquivo no formato tabular, como ilustrado na Figura 28;

- comparação das distâncias para cada uma das tabelas geradas, visando estabelecer os pares de fragmentos para reconstrução dos documentos analisados.

Figura 27: Diagrama de blocos dos algoritmos.

O objetivo da organização e implementação de algoritmos específicos para estas tarefas advém da necessidade de se melhorar a performance durante a execução de tais procedimentos, visto que para encontrar os pares de fragmentos necessita-se buscar entre todas as opções as menores distâncias, para então determinar os fragmentos que se encaixam dois-a-dois.

A Figura 28 representa uma matriz quadrada de dimensão igual ao número de fragmentos. Para cada modelo de cor utiliza-se uma matriz específica. Cada posição da matriz contém a Distância Euclidiana entre os VCE de um fragmento com os VCD de outro fragmento. O sistema não realiza o cálculo das distâncias para VCs pertencentes a um mesmo fragmento, não gerando valores na diagonal principal, como representado na Figura 29. Além disto, ao gerar a matriz para documentos misturados, pois o sistema não sabe *a priori* se existe mais de um documento a ser tratado, o sistema identifica os fragmentos de alturas distintas e

armazena na matriz de distâncias um valor considerado elevado, mas na Figura 29 esse valor é representado pelo caracter **X**, não permitindo que fragmentos de tamanho distinto sejam encaixados.

	F1	F2	F3	F4	F5	F6	F7	F...	F29
F1	----								
F2		----							
F3			----						
F4				----					
F5					----				
F6						----			
F7							----		
F...								----	
F29									----

Figura 28: Matriz de Distância Euclidianas.

Assim, o sistema percorre a matriz utilizando lista duplamente encadeada para estabelecer qual a menor distância entre, por exemplo, o VCE de um fragmento F_1 e os VCDs dos demais fragmentos F_2 até F_n. Deste modo, considerando a matriz ilustrada na Figura 29 para o VCE do fragmento F_1 o encaixe será realizado com o VCD do fragmento F_{29}.

	F1	F2	F3	F4	F5	F6	F7	F...	F29
F1	----	14,177	15,458	9,178	34,158	12,000	44,715		4,715
F2	1,197	----	25,147	0,000	54,717	36,222	21,000		23,000
F3	36,222	4,175	----	17,117	0,500	13,175	21,150		15,120
F4	17,185	1,157	4,158	----	36,222	65,115	36,222		37,252
F5	1,137	6,159	54,115	14,590	----	X	13,144		32,144
F6	36,222	19,147	X	X	X	----	19,778		18,789
F7	17,767	31,127	15,654	19,592	7,777	8,587	----		36,258
F...								----	5,259
F29	1,767	13,127	35,435	15,215	8,712	3,123	43,212	13,132	----

Figura 29: Exemplo de matriz com distâncias entre fragmentos.

Na Figura 29 como exemplo, verifica-se que a menor distância Euclidiana encontrada para o fragmento F_1 é de magnitude 4,715, que corresponde ao fragmento F_{29}. A partir daí atualizam-se os índices apontando a borda esquerda do fragmento F_1 para a borda direita do fragmento F_{29} e passando a buscar os próximos fragmentos à esquerda do fragmento F_{29} e assim por diante, até que se tenha percorrido toda a matriz de distâncias. Ao término do procedimento de busca o algoritmo apresenta a ordem dos fragmentos para a reconstrução do documento.

3.5. Comentários Finais

No presente Capítulo foi apresentado o método proposto e implementado para reconstrução de documentos mutilados em formato “spaghetti”, destacando-se a criação da base de dados contendo 200 documentos diferentes, totalizando cerca de 5400 fragmentos. Tais fragmentos foram utilizados nos experimentos realizados e descritos no Capítulo 4. A base de dados de documentos mutilados encontra-se sob os cuidados do Laboratório de Computação Forense e Biometria do PPGIA na PUCPR, podendo ser utilizada em outros experimentos ou projetos de pesquisa.

Capítulo 4

Resultados Experimentais

Este Capítulo apresenta os resultados experimentais obtidos através do sistema proposto. Os testes realizados encontram-se organizados pelos modelos de cores (RGB e HSV) e pelas classes de documentos que formam a base de dados (Tabela 02 – subseção 3.2).

4.1. Resultados usando o modelo de cor RGB através da soma dos canais – (RGB SOMA)

As Tabelas 3 e 4 e, ainda, as Figuras 33 e 34 apresentam os resultados obtidos com o modelo de cor RGB em função dos tipos de documentos que fazem parte da base de dados. Utilizando o modelo RGB os resultados foram bem próximos aos do HSV (ver item 4.3). Notou-se que as taxas de acerto na reconstrução dos documentos foi inferior à alcançada com o modelo de cor HSV devido à soma dos 3 canais de cor do modelo RGB conforme o método proposto. Sabe-se que a cor resulta da soma dos 3 canais RGB e no modelo HSV utilizou-se somente o canal H que é a cor propriamente dita. Assim, cores opostas obtiveram o mesmo valor numérico na representação total do modelo RGB. Os resultados apresentados levaram em conta a reconstrução individual de cada documento.

Tabela 3. Média do sistema usando o modelo RGB através da soma dos canais – Classe 01.

Tipos	**Média de Acerto (%)**
Ofícios, Memorandos, etc.	85,29
Revistas	92,10
Folders, Flyers, Anúncios, etc	97,68

Tabela 4. Média do sistema usando o modelo RGB através da soma dos canais – Classe 02.

Documentos	**Média de Acerto (%)**
Texto	79,66
Texto & Tabela	87,07
Texto & Figura	90,07
Texto & Figura & Tabela	93,95

Observa-se na Tabela 3 que o melhor resultado para o modelo RGB ocorre com os documentos do tipo Folders, Flyers, Anúncios, etc. Isto pode ser explicado devido ao fato destes documentos utilizarem muitas cores, além de textos, figuras e tabelas, com o intuito de despertar a atenção das pessoas. Observa-se, também, na Tabela 4, que os documentos contendo somente texto alcançaram a taxa mais baixa de acertos na reconstrução do tal tipo de documento, assim como já esperado, visto que estes documentos utilizam poucas cores como mostra a Figura 30. A Figura 31 mostra o resultado da reconstrução com base no modelo RGB sendo que ocorre confusão entre alguns fragmentos e, ainda, destaca-se que fragmentos do início do documento foram encaixados ao final, visto que tais fragmentos possuem cor branca e que o sistema não faz tratamento específico para este tipo de fragmento, como já mencionado anteriormente.

MARLOS ALEX DE OLIVEIRA MARQUES

RECONSTRUÇÃO DIGITAL DE DOCUMENTOS MUTILADOS COM FORMAS REGULARES

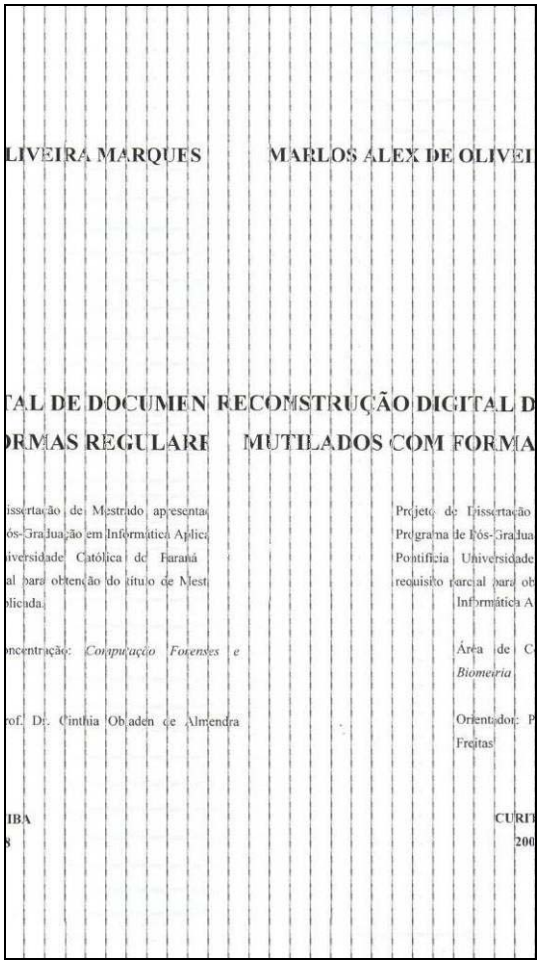
Projeto de Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática Aplicada.

Área de Concentração: *Computação Forenses e Biometria*

Orientador: Prof. Dr. Cinthia Obladen de Almendra Freitas

CURITIBA
2008

Figura 30: Imagem de um documento da Classe 02 - Tipo texto.


LIVEIRA MARQUES MARLOS ALEX DE OLIVEI

TAL DE DOCUMEN RECONSTRUÇÃO DIGITAL D
ORMAS REGULARE MUTILADOS COM FORMA

issertação de Mestrado apresenta
ós-Graduação em Informática Aplic
iversidade Católica do Paraná
al para obtenção do título de Mest
olicada.

Projeto de Dissertação
Programa de Pós-Graduaç
Pontifícia Universidade
requisito parcial para ob
Informática A

oncentração: *Computação Forenses e*

Área de C
Biometria

rof. Dr. Cinthia Obaden ce Almendra

Orientador: P
Freitas

IBA
8

CURI
200

Figura 31: Reconstrução do documento da Figura 30 através do modelo RGB.

As Figuras 32 e 33 apresentam, respectivamente, o documento original e o resultado da reconstrução de um documento do tipo Folders, Flyers, Anúncios, etc, através do modelo RGB. Pode-se notar que ocorreu um acerto de 24 em 26 fragmentos, perfazendo 92,3% de acerto, o que resulta em uma taxa de acerto maior ao comparar-se tal documento com o documento das Figuras 30 e 31, o qual resultou em 16 fragmentos corretos de um total de 26 fragmentos, perfazendo uma taxa de acerto igual a 61,5%.

Figura 32: Imagem do tipo Folders, Flyers, Anúncios, etc.

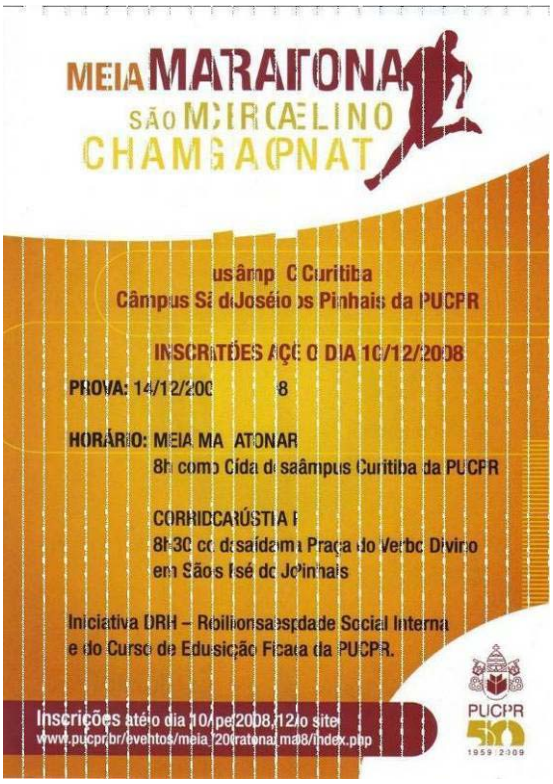


Figura 33: Imagem do tipo Folders, Flyers, Anúncios, etc., reconstruída usando o modelo RGB.

4.2. Resultados usando o modelo de cor RGB através da combinação dos canais - (RGB)

Neste experimento os canais R, G e B não foram somados como do item anterior, mas considerou-se que cada canal corresponde a um VC sendo estes vetores organizados da seguinte forma: primeiro o canal R, segundo o canal G e por último o B. Assim, o procedimento de determinação das distâncias permitiu a geração de três matrizes de distâncias, uma para cada canal de cor (R, G e B). Na seqüência, considerou-se que o par de fragmentos candidato ao encaixe era o que produzia a menor distância analisada entre estas três matrizes.

As Tabelas 5 e 6 apresentam os resultados obtidos com o modelo de cor RGB em função dos tipos de documentos que fazem parte da base de dados. Utilizando o modelo RGB, combinando os canais, os resultados foram bem próximos aos do HSV (ver item 4.3). Notou-se que as taxas de acerto na reconstrução dos documentos foi inferior a alcançada com o modelo de cor HSV. Os resultados apresentados levaram em conta a reconstrução individual de cada documento.

Tabela 5. Média do sistema usando o modelo RGB através da combinação dos canais – Classe 01.

Tipos	**Média de Acerto (%)**
Ofícios, Memorandos, etc.	86,50
Revistas	94,25
Folders, Flyers, Anúncios, etc	98,15

Tabela 6. Média do sistema usando o modelo RGB através da combinação dos canais – Classe 02.

Documentos	**Média de Acerto (%)**
Texto	82,50
Texto & Tabela	91,07
Texto & Figura	91,07
Texto & Figura & Tabela	95,00

4.3. Resultados usando o canal R do modelo RGB

Os resultados obtidos para os canais R, G e B separadamente alcançaram taxas mais elevadas de acerto do que a representação total do modelo RGB. Assim, as seções 4.3, 4.4,

4.5 mostram que com a decomposição dos canais do modelo RGB pode-se evitar que cores opostas com a vermelha, o verde e a azul obtenham o mesmo valor numérico na representação da soma dos 3 canais RGB. Na Figura 34 apresenta-se um exemplo de decomposição do canal R de uma imagem colorida.

(a) (b)

Figura 34: a) Imagem com os 3 canais. b) Imagem composta somente pelo canal R.

As Tabelas 7 e 8 apresentam os resultados obtidos usando somente o canal R do modelo de cor RGB em função dos tipos de documentos que fazem parte da base de dados. Os resultados apresentados levaram em conta a reconstrução individual de cada documento.

Tabela 7. Média do sistema usando o canal R – Classe 01.

Tipos	**Média de Acerto (%)**
Ofícios, Memorandos, etc.	86,29
Revistas	93,15
Folders, Flyers, Anúncios, etc	98,00

Tabela 8. Média do sistema usando o canal R – Classe 02.

Documentos	**Média de Acerto (%)**
Texto	82,36
Texto & Tabela	90,07
Texto & Figura	90,07
Texto & Figura & Tabela	94,95

Observa-se na Tabela 7 que o melhor resultado do canal R novamente ocorre com os documentos do tipo Folders, Flyers, Anúncios, etc. com 98 % de acerto, sendo a menor taxa

de acertos dos documentos do tipo texto, ou seja, 82,36% de acerto, como mostra na Tabela 8. A Figura 35 mostra o resultado da reconstrução com base no modelo RGB usando somente o canal R.

Figura 35: Imagem reconstruída usando o canal R

4.4. Resultados usando o canal G do modelo RGB

Na Figura 36 apresenta-se um exemplo de decomposição do canal G de uma imagem colorida.

(a) (b)

Figura 36: a) Imagem com os 3 canais. b) Imagem composta somente pelo canal G.

As Tabelas 9 e 10 apresentam os resultados obtidos usando somente o canal G do modelo de cor RGB em função dos tipos de documentos que fazem parte da base de dados. Os resultados apresentados levaram em conta a reconstrução individual de cada documento.

Tabela 9. Média do sistema usando o canal G – Classe 01.

Tipos	**Média de Acerto (%)**
Ofícios, Memorandos, etc.	85,99
Revistas	92,15
Folders, Flyers, Anúncios, etc	97,07

Tabela 10. Média do sistema usando o canal G – Classe 02.

Documentos	**Média de Acerto (%)**
Texto	81,36
Texto & Tabela	90,00
Texto & Figura	90,00
Texto & Figura & Tabela	94,00

Observa-se na Tabela 9 que o melhor resultado do canal G novamente ocorre com os documentos do tipo Folders, Flyers, Anúncios, etc. com 97,07 % de acerto, sendo também a menor taxa de acertos com os documentos do tipo texto que resultou em 81.36% como mostra

na Tabela 10. A Figura 37 mostra o resultado da reconstrução com base no modelo RGB usando somente o canal G.

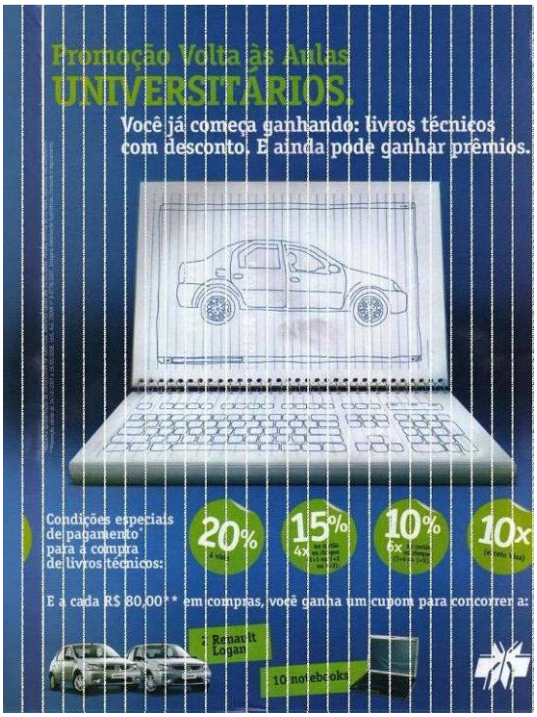


Figura 37: Imagem reconstruída usando o canal G.

4.5. Resultados usando o canal B do modelo RGB

Na Figura 38 apresenta um exemplo de decomposição do canal B de uma imagem colorida.

(a) (b)

Figura 38: a) Imagem com os 3 canais. b) Imagem composta somente pelo canal B.

As Tabelas 11 e 12 apresentam os resultados obtidos usando somente o canal B do modelo de cor RGB em função dos tipos de documentos que fazem parte da base de dados. Os resultados apresentados levaram em conta a reconstrução individual de cada documento.

Tabela 11. Média do sistema usando o canal B – Classe 01.

Tipos	**Média de Acerto (%)**
Ofícios, Memorandos, etc.	86,00
Revistas	92,15
Folders, Flyers, Anúncios, etc	97,05

Tabela 12. Média do sistema usando o canal B – Classe 02.

Documentos	**Média de Acerto (%)**
Texto	81,35
Texto & Tabela	90,00
Texto & Figura	90,00
Texto & Figura & Tabela	94,00

Observa-se na Tabela 11 que o melhor resultado do canal B novamente ocorre com os documentos do tipo Folders, Flyers, Anúncios, etc., alcançando 97,05% de taxa de acerto. Já os documentos do tipo texto obtiveram a menor taxa de acertos, igual a 81,35%, como mostra

na Tabela 12. A Figura 39 mostra o resultado da reconstrução com base no modelo RGB usando somente o canal B.

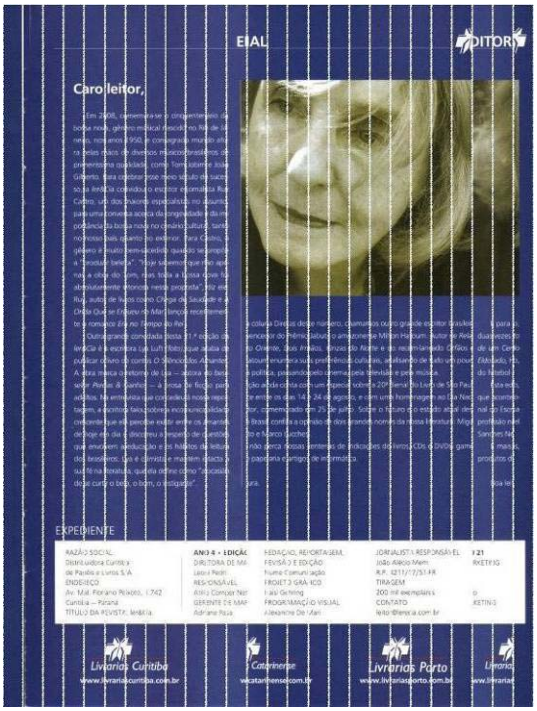


Figura 39: Imagem reconstruída usando o canal B.

A Tabela 13 apresenta um resumo dos resultados obtidos com os canais R, G e B separadamente. Observa-se que os resultados com os canais separados são praticamente iguais entre si o que demonstra que a base de dados não tem uma predominância de uma cor em detrimento de outras e, ainda, que o tratamento com os canais de cores separados é interessante para resolver o problema dos documentos tipo texto (classe 02), visto que a taxa de acerto está em torno de 82%, enquanto que para o modelo RGB, todos os canais reunidos, a taxa de acerto é de 79,66%.

Tabela 13: Resumo dos Resultados com os Canais R, G e B.

Tipos – Classe 01	**R**	**G**	**B**
Ofícios, Memorandos, etc.	86,00	85,99	86,00
Revistas	92,15	92,15	92,15
Folders, Flyers, Anúncios, etc	97,05	97,07	97,05
Tipos – Classe 02	**R**	**G**	**B**
Texto	82,36	81,35	81,35
Texto & Tabela	90,07	90,00	90,00
Texto & Figura	90,07	90,00	90,00
Texto & Figura & Tabela	94,95	94,00	94,00

4.6. Resultados usando o modelo de cor HSV

Os resultados obtidos utilizando o modelo HSV alcançaram taxas de acerto mais elevadas do que as alcançadas com o modelo RGB e de seus canais separados.

As Tabelas 14 e 15 apresentam os resultados obtidos com o modelo de cor HSV em função dos tipos de documentos que fazem parte da base de dados. Os resultados apresentados levaram em conta a reconstrução individual de cada documento.

Tabela 14. Média do sistema usando o modelo HSV – Classe 01.

Tipos	**Média de Acerto (%)**
Ofícios, Memorandos, etc.	89,14
Revistas	95,99
Folders, Flyers, Anúncios, etc	98,53

Tabela 15. Média do sistema usando modelo HSV – Classe 02.

Documentos	**Média de Acerto (%)**
Texto	82,76
Texto & Tabela	90,00
Texto & Figura	93,79
Texto & Figura & Tabela	97,42

Observa-se na Tabela 14 que o melhor resultado do modelo HSV novamente ocorre com os documentos do tipo Folders, Flyers, Anúncios, etc. com 98,53% de acerto, sendo a menor taxa de acertos a dos documentos do tipo texto, 82,76%, como mostra na Tabela 15. A Figura 40 mostra o resultado da reconstrução com base no modelo HSV.

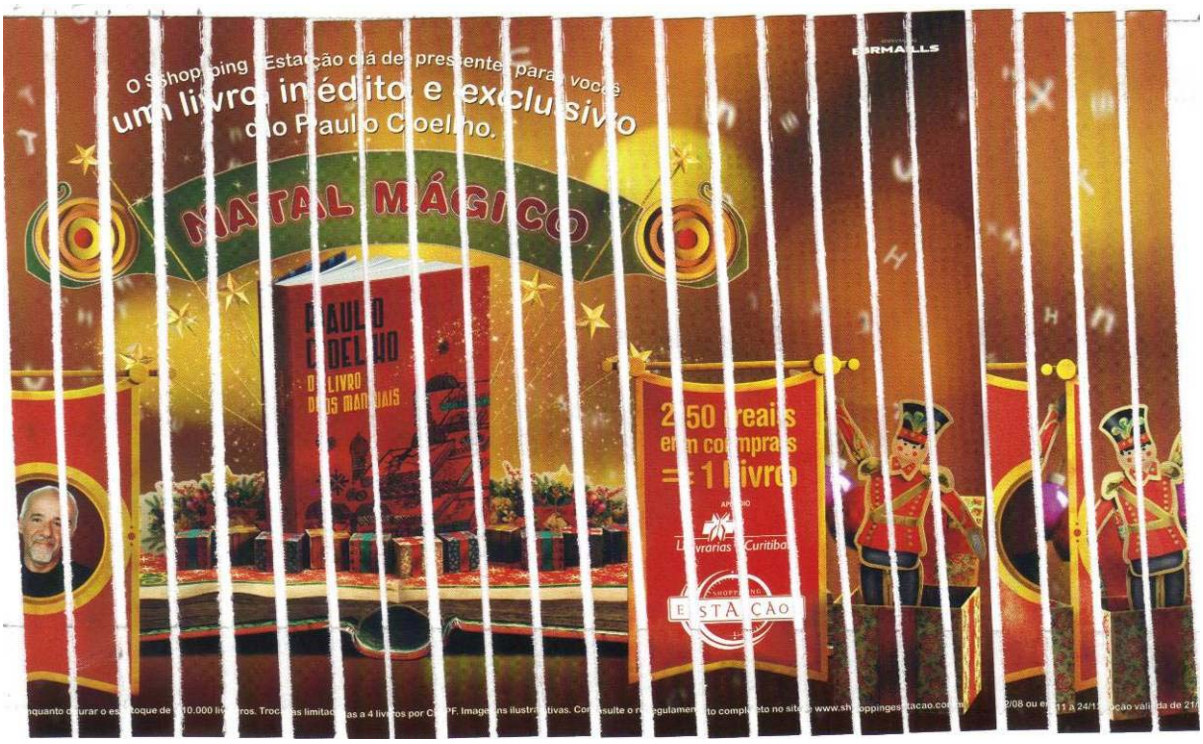


Figura 40: Reconstrução usando o modelo HSV.

4.7. Comparação entre os modelos de cor

A Tabela 16 e a Figura 41 apresentam as taxas de acerto e erro do sistema proposto considerando os dois modelos RGB e HSV e, ainda, os canais de cores isolados do modelo RGB. Observa-se que em sua totalidade o HSV obtém um maior número de ordenações corretas do que o modelo RGB e seus canais isoladamente.

Tabela 16. Melhores taxas do sistema proposto.

Modelo	Média de Acerto (%)
HSV	98,53
RGB	98,15
RGB-Soma	97,68
Canal R	98,00
Canal G	97,07
Canal B	97,07

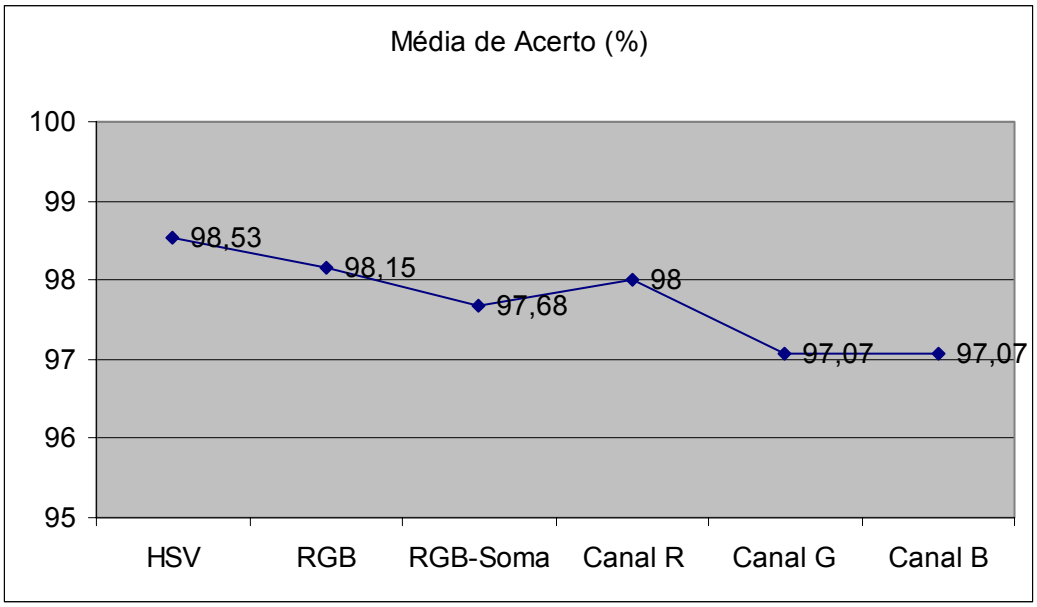


Figura 41: Desempenho do sistema proposto para todos os modelos de cor.

Outro experimento realizado foi o embaralhamento de todos os documentos (a base de dados contém 200 documentos) para avaliar o grau de confusão gerado pelo sistema. A Tabela 17 mostra os resultados obtidos para os modelos HSV, RGB e seus canais separados. Assim, o modelo HSV obteve os melhores resultados e o canal R apresentou melhor resultado que o modelo RGB em sua totalidade.

Tabela 17. Tabela de acertos do sistema proposto.

Modelo	**Média de Acerto (%)**
HSV	94,28
RGB	92,63
RGB-Soma	90,83
Canal R	92,53
Canal G	91,71
Canal B	91,71

Observou-se que o sistema não gerou a confusão entre fragmentos de documentos diversos em imagens ricas em informação de cor, ou seja, se o sistema não reconstruiu todo o documento corretamente conseguiu separar os fragmentos pertencentes a cada um dos documentos. Porém, em imagens contendo só texto e com a cor preta e branca predominantes

foram observadas as menores taxas de acertos e quando foram realizado os testes com toda base houve algumas junções de fragmentos de outros documentos.

Observou-se também que o menor número de erros, em quantidade de fragmentos corretamente encaixados, foi de 1 fragmento no modelo HSV. Porém, o maior número de erros, considerando o modelo RGB, foram 9 fragmentos posicionados erroneamente em um mesmo documento. Ambos os casos para os 200 documentos com aproximadamente 29 fragmentos [MARQUES E FREITAS, 2009].

Notou-se também que a maioria dos erros de ordenação dos fragmentos encontra-se no inicio ou no final do documento. Isso ocorre devido a que não se sabe qual é o fragmento inicial ou final do documento, notou-se também que as bordas inicial e final do documento tendem a serem reconstruídas como bordas consecutivas, como mostra a Figura 42. Outro problema detectado é que quando os fragmentos iniciais e finais são totalmente brancos, não se tem como saber se estes pertencem ao início ou ao final do documento, como mostra a Figura 43. O procedimento implementado não desconsidera estes fragmentos ou atribui tratamento especial a este tipo de problema. Caso isto seja realizado, o desempenho do sistema será ainda melhor.

Figura 42: Exemplo de reconstrução na qual as bordas opostas se encontram.

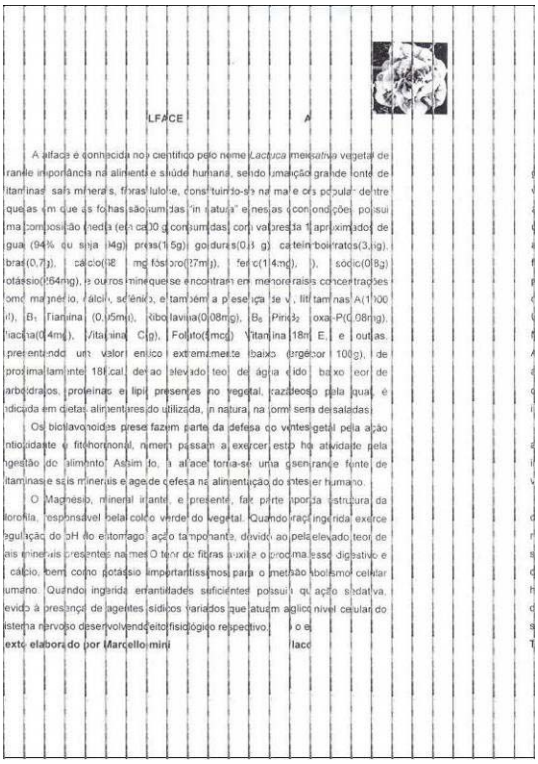


Figura 43: Exemplo de reconstrução na qual as bordas totalmente brancas são reunidas ao final do documento.

4.8. Problemas na reconstrução do documento

Existem problemas relacionados com a extração da informação da cor das bordas. Entre estes problemas destaca-se a existência de ruídos e o fato dos fragmentos terem sido mau digitalizados, resultando em procedimentos de comparação (pesquisa dos fragmentos parceiros) imprecisos.

Durante o processo de fragmentação do documento podem ocorrer algumas anomalias nos fragmentos, os quais podem dificultar o processo de reconstrução do documento, como exemplificado na Figura 44. Entre estas anomalias pode-se citar:

- os fragmentos não serem exatamente retangulares;
- os fragmentos não terem a mesma forma entre os lados de cada fragmento;
- os fragmentos apresentarem uma ligeira curvatura;
- os fragmentos são manipulados e podem estar incompletos ou dobrados;

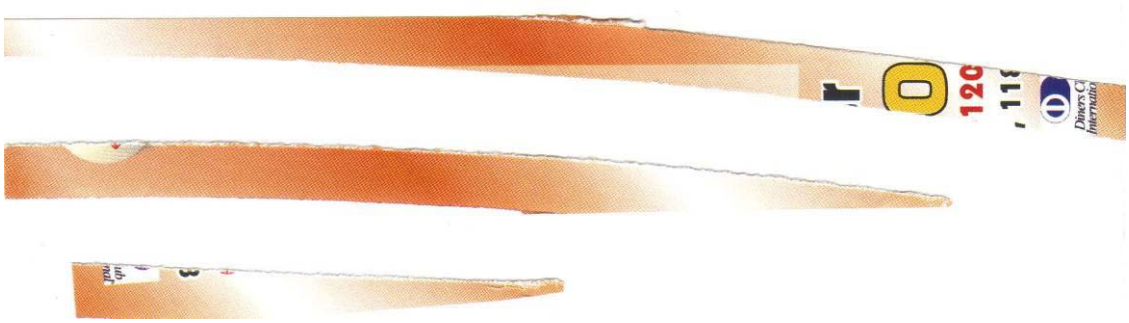


Figura 44: Exemplo de anomalias das tiras.

4.9. Comparação entre o Sistema Proposto e o Sistema de Skeok

Essa comparação é necessária visto que são trabalhos que abordam o mesmo problema e, portanto, pode-se avaliar os pontos fracos e fortes de cada sistema. A Tabela 18 e a Figura 45 resumem os resultados de ambos os sistemas: proposto e Skeok.

O trabalho desenvolvido pela Skeok [SKEOCH, 2006] obteve resultados no mesmo patamar de taxas de acerto que o sistema aqui proposto e testado. Porém, o trabalho de Skeok além de utilizar os dois modelos de cores RGB e HSV e medidas de dissimilaridade, também utilizou algoritmo genético para comparação dos pares de fragmentos, sendo este um ponto forte do seu trabalho. Apesar disto, o trabalho Skeok utiliza uma base de dados pequena em torno de 15 documentos somente, sendo que a maioria dos testes realizados nestas imagens considera a geração dos fragmentos computacionalmente, ou seja, não foi utilizada uma máquina fragmentadora. A utilização de uma máquina fragmentadora permite que os testes sejam realizados com situações reais, inclusive com a observação dos problemas já mencionados.

Tabela 18. Comparação de Resultados Sistema Proposto e Sistema Skeoch.

Documento	**Média de Acerto (%)**
Purple com 15 fragmentos [SKEOCH, 2006]	97,58
RGB SOMA - Folders, Flyers, Anúncios, etc	97,68
RGB SOMA – Texto & Figura & Tabela	93,95
RGB - Folders, Flyers, Anúncios, etc	98,15
RGB – Texto & Figura & Tabela	95,00
R - Folders, Flyers, Anúncios, etc	98,00
R - Texto & Figura & Tabela	94,95
G - Folders, Flyers, Anúncios, etc	97,07
G - Texto & Figura & Tabela	94,00
B - Folders, Flyers, Anúncios, etc	97,07
B - Texto & Figura & Tabela	94,00
HSV - Folders, Flyers, Anúncios, etc	98,53
HSV - Texto & Figura & Tabela	97,42

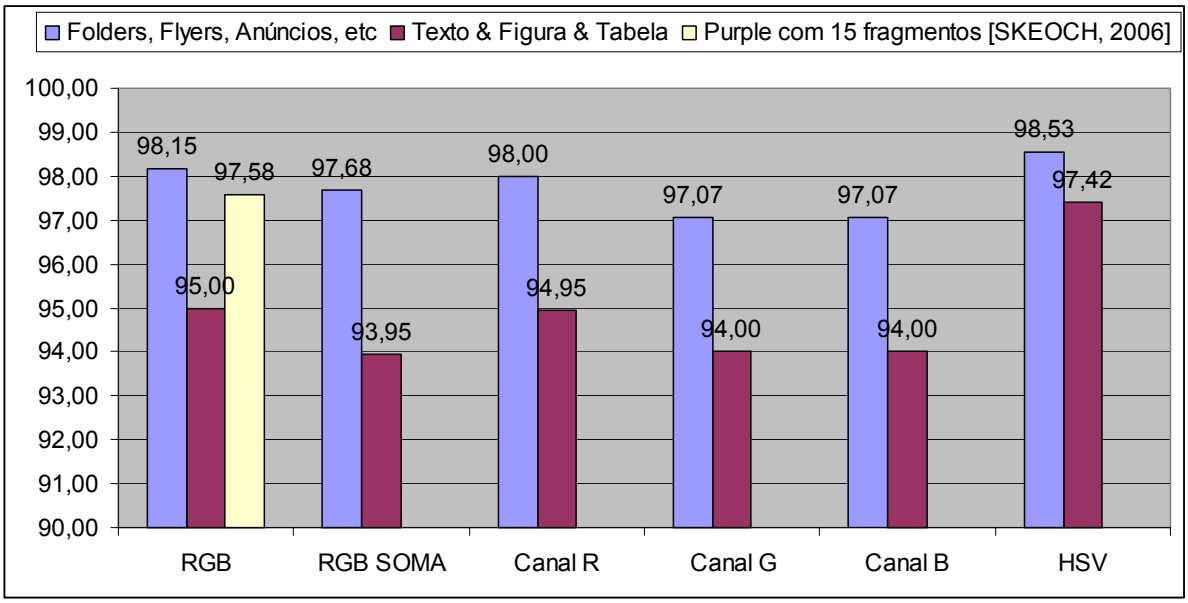


Figura 45: Comparação de Resultados.

O ponto forte do sistema aqui apresentado é a base de dados contendo 200 documentos de diferentes origens, cores, modelos e conteúdo, a qual possibilitará outros estudos e pesquisas.

Outro ponto forte do presente trabalho é o fato do sistema ser capaz de separar os fragmentos pertencentes a documentos distintos, quando diversos documentos encontram se misturados entre si. O sistema pode confundir fragmentos de um mesmo documento, mas não o faz entre documentos distintos.

Capítulo 5

Conclusão

Neste trabalho foi apresentado um método para reconstrução digital de documentos mutilados em formato "spaghetti" utilizando a características da cor das bordas dos fragmentos. Os resultados demonstram que as características utilizadas para identificar os fragmentos parceiros são promissores para dar continuidade nas pesquisas relacionadas com reconstrução de documentos mutilados e demonstram que existe um campo em aberto para trabalhos futuros [MARQUES E FREITAS, 2009].

Ressalta-se que o método desenvolvido apresenta resultados aceitáveis em termos de taxa de acerto de reconstrução de documentos mutilados, considerando 29 fragmentos para cada documento, sendo que para a Classe 01 de documentos da base de dados obteve-se 97,42% com o modelo HSV e para a Classe 02 de documentos obteve-se 98,53%, também para o modelo HSV.

Mesmo não atingindo 100% de acerto na reconstrução dos documentos mutilados, observou-se que o maior número de fragmentos posicionados erroneamente, utilizando-se o modelo RGB, corresponde e 9 fragmentos pertencentes a um mesmo documento. Além disto, foi observado também que os erros ocorrem com mais freqüência nos fragmentos iniciais dos documentos do que nos fragmentos finais. Isto devido ao fato de que os fragmentos iniciais ou finais são desconhecidos. O sistema não recebe nenhuma informação *a priori* sobre qual é o fragmento que deve iniciar a reconstrução. Outra observação importante é que quando os fragmentos são totalmente brancos (sem informação) o sistema não consegue estabelecer se este pertence ao início ou ao final do documento, gerando posicionamento incorreto. O sistema desenvolvido não descarta ou faz tratamento diferenciado para estes fragmentos em branco.

As maiores contribuições desse trabalho são:

- a criação da base de dados de documentos mutilados formada por 200 documentos diferentes entre si, descrita no Capítulo 3:

- a implementação de algoritmos de extração de características baseados nos modelos de cores (RGB e HSV):

- o procedimento para comparação dos fragmentos para reconstrução de documentos:

- o fato do sistema ser capaz de separar os fragmentos pertencentes a documentos distintos, quando diversos documento se encontram misturados entre si. O sistema pode confundir fragmentos de um mesmo documento, mas não o faz entre documentos distintos.

Como futuros trabalhos podem ser destacados a criação de características globais para identificar os fragmentos candidatos pertencentes a um mesmo documento, utilização de outras primitivas baseadas em contexto (*context-based*) a exemplo da textura, aplicação de outros métodos de classificação baseados em treinamento, estudar o comportamento do método em outros tipos de fragmentos com meio físico diferente da mídia papel, estudar o comportamento do método proposto em função do instrumento utilizado para fragmentação do documento visando estabelecer padrões para convergência por tipo de mutilação e, finalmente, ampliar a base de imagens de documentos mutilados.

Referências Bibliográficas

[BROWER E ZAR, 1977] BROWER, J.E.; ZAR, J.H. *Field & laboratory methods for general ecology***.** 2.ed. Dubuque: Wm. C. Brown Publishers, 1977. 226p.

[CALIXTO, 2005] CALIXTO, E., Orientador: Conci, A. (2005) Dissertação: *Granulometria morfológica em espaços de cores: estudo da ordenação espacial*, Universidade Federal Fluminense, Departamento da Computação, Niterói, Agosto 2005.

[CHURCHSTREET, 2007] CHURCHSTREET TECHNOLOGY, Inc. Disponível em: <http://www.churchstreet-technology.com>, acessado em 12/01/2007.

[FACON, 1996] FACON, J., *Morfologia Matemática: Teoria e Exemplos.* Editora Universitária Champagnat da Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, 1996.

[FBI, 2007] FEDERAL BUREAU OF INVESTIGATION. *Handbook of Forensic Services (revised 2003)*. FBI Laboratory Publication – http://www.fbi.gov/hq/lab/handbook/forensics.pdf, Quantico, Virginia, 2007.

[GAUCH, 1982] GAUCH, H.G. *Multivariate analysis in community ecology***.** Cambridge University Press, 1982. 298p.

[GONZALEZ E WOODS, 2000] GONZALEZ, R. e WOODS, R. E. *Processamento de Imagens Digitais*. Ed. Edgard Blücher Ltda., São Paulo, 2000.

[LAROUSSE, 1998] GRANDE ENCICLOPÉDIA LAROUSSE CULTURAL. Editora Nova Cultural Ltda. Pinheiros. São Paulo (SP), 1998.

[LEITÃO, 2000] LEITÃO, H. C. G., *Reconstrução automática de objetos fragmentados*, Tese de Doutorado de 21/10/1999, Instituto de Educação, Universidade Estadual de Campinas UNICAMP, Campinas, São Paulo, 2000.

[MARQUES E FREITAS, 2009] MARQUES, M A. O.; FREITAS, C. O. A. O., *Reconstructing Strip-shredded Documents Using Color as Feature Matching,* In: 24th Annual ACM Symposium on Applied Computing, 2009, Honolulu. Proc. of 24th Annual ACM Symposium on Applied Computing. New York : ACM, 2009. v. 2. p. 893-894.

[MELLO, 2002] MELLO, CARLOS A. B., *Filtragem, Compressão e Síntese de Imagens de Documentos Históricos,* Tese de Doutorado de 27 de Maio de 2002, Centro de Informática, Universidade Federal de Pernambuco UFPE, Recife, Pernambuco, 2002.

[MENDES, 2003] MENDES, L.B., *Documentoscopia. 2ª ed*., Editora Millenium, Campinas, SP, 2003.

[OLIVEIRA ET AL. 2006] OLIVEIRA, L. E. S.; JUSTINO, E. J. R.; FREITAS, C. O. A. . *Reconstructing Shredded Documents trough Feature Matching*. Forensic Science International, Ireland, v. 160, p. 140-147, 2006.

[RAMOS, 2004] RAMOS, F. R., Orientador: Borges, D. L. Dissertação: *Recuperação de informação baseada em conteúdo; Analisando imagens priorizando a característica cor*, Dissertação de Mestrado no Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná PUCPR, Curitiba, Paraná, 2004.

[SKEOCH, 2006] SKEOCH, A., *An Investigation into Automated Shredded Document Reconstruction using Heuristic Search Algorithms,* Dissertation is submitted to the University of Bath, in accordance with the requirements of the degree of Batchelor of Science in the Department of Computer Science, Bath, UK, 2006.

[SOLANA, 2005] SOLANA, C. D. O., R*econstrução Digital de Documentos por Aproximação Poligonal,* Dissertação de Mestrado de 01/08/2005 no Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná PUCPR, Curitiba, Paraná, 2005.

[UKOVICH, ET AL. 2004] UKOVICH, A., RAMPONI, G., DOULAVERAKIS, H., KOMPATSIARIS, Y.: *Shredded Document Reconstruction using MPEG-7 Standard Descriptors*. In: IEEE Int. Symp. on Signal Processing and Information Technology (ISSPIT-04), pp. 18--21. 2004.

[UNB, 2007] UNIVERSIDADE DE BRASÍLIA, *Conservação/Preservação de Documentos*, Centro de Documentação, disponível em: <http://www.unb.br/cedoc/convervacao.htm>, consultado em 01/03/2007.