ANDRE MIRANDA PIMENTA

RECONSTRUÇÃO DIGITAL DE DOCUMENTOS MUTILADOS USANDO PROGRAMAÇÃO DINÂMICA

CURITIBA
2009

ANDRE MIRANDA PIMENTA

RECONSTRUÇÃO DIGITAL DE DOCUMENTOS MUTILADOS USANDO PROGRAMAÇÃO DINÂMICA

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

Área de Concentração: *Computação Forense e Biometria, Documentoscopia.*

Orientador: Prof. Dr. Edson José Rodrigues Justino.
Co-orientador: Prof. Dr. Luiz Eduardo Soares de Oliveira.

CURITIBA
2009

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 03/2009

Aos 20 dias do mês de fevereiro de 2009 realizou-se a sessão pública de Defesa da Dissertação **"Reconstrução Digital de Documentos Mutilados usando Programação Dinâmica"**, apresentada pelo aluno **André Miranda Pimenta** como requisito parcial para a obtenção do título de Mestre em Informática, perante uma Banca Examinadora composta pelos seguintes membros:

Prof. Dr. Edson José Rodrigues Justino PUCPR (Orientador)	(assinatura)	Aprovado (aprov/reprov.)
Prof. Dr. Luiz Eduardo Soares de Oliveira PUCPR		APROV
Prof. Dr. Jacques Facon PUCPR		Aprovado
Prof. Dr. Helio Pedrini UNICAMP	Hélio Pedrini	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

Programa de Pós-Graduação em Informática
PPGIa
PUCPR

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 este trabalho aos meus pais, Jovelino Caitano Pimenta e Maria de Fátima Miranda Pimenta, por terem me guiado e me educado da forma mais correta possível: através de seus próprios exemplos.

AGRADECIMENTOS

A Deus por permitir e por me tornar capaz de viver as oportunidades que surgiram em minha vida, cada uma no seu devido tempo, incluindo a oportunidade e capacidade de realização deste trabalho.

A minha família e aos meus amigos pela compreensão nas minhas ausências, e por nunca terem deixado de me apoiar e incentivar nas minhas caminhadas.

Ao Professor Dr. Edson José Rodrigues Justino pelo incentivo e apoio no desenvolvimento deste trabalho, e principalmente pela confiança depositada em mim que, mesmo nos momentos mais críticos, nunca foi abalada.

Ao Professor Dr. Luiz Eduardo Soares de Oliveira, pelo auxílio e apontamentos importantes na realização deste trabalho.

Finalmente agradeço aos meus colegas de estudo, colegas de trabalho, e a todas as pessoas que de alguma forma me auxiliaram e incentivaram na continuidade e conclusão deste trabalho.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

A_1, A_2	Valor da área.
A_i	Soma dos comprimentos das arestas do fragmento A.
B_j	Soma dos comprimentos das arestas do fragmento B.
C	Quantidade de verificações entre pares de fragmentos.
d	Distância entre dois pontos.
da	Valor atual da distância.
D_{a1}	Distância euclidiana entre o vértice posterior e seu vizinho anterior do fragmento A.
D_{a2}	Distância euclidiana entre o vértice anterior e seu vizinho posterior do fragmento A.
dab	Distância do ponto a até o ponto b.
D_{b1}	Distância euclidiana entre o vértice posterior e seu vizinho anterior do fragmento B.
D_{b2}	Distância euclidiana entre o vértice anterior e seu vizinho posterior do fragmento B.
dbc	Distância do ponto b até o ponto c.
dpr	Distância entre ponto e reta.
E_m	Erro de alinhamento.
Es	Número de arestas do grafo.
F	Número de fragmentos de um documento.
L	Comprimento da lateral de um ponto.
Np	Número de pontos.

Nv	Número de elementos para cálculo da variância.
P	Ponto inicial de um segmento.
Q	Ponto final de um segmento.
V	Medida de variância.
Vns	Número de vértices do grafo.
W_{angulo}	Resultado da combinação de ângulos.
$W_{matching}$	Melhor resultado de combinação.
X_f, Y_f	Coordenadas finais após rotação.
X_{Fai}, Y_{Fai}	Coordenadas do ponto do fragmento A.
X_{Fbf}, Y_{Fbf}	Coordenadas do ponto final de translação do fragmento B.
X_{Fbi}, Y_{Fbi}	Coordenadas do ponto inicial de translação do fragmento B.
α	Ângulo.

LISTA DE ABREVIATURAS E SIGLAS

BMP	Microsoft Windows Bitmap.
DNA	Ácido Desoxirribonucléico.
DPI	Dots Per Inch (pontos por polegada).
FBI	Federal Bureau Of Investigation
ICP-Brasil	Infra-estrutura de Chaves Públicas Brasileira.
LCS	Last Common Subsequence (maior subseqüência comum).
LOT	Left most-on-top (seqüência que inicia pelo canto superior esquerdo).
MPEG-7	Moving Picture Experts Group Number 7 (padrão ISSO-IEC).
PPM	Previsão por Combinação Parcial.
PUCPR	Pontifícia Universidade Católica do Paraná.
QDU	Questioned Documents Unit (unidade para documentos questionados).
RNA	Ácido Ribonucléico.

RESUMO

A reconstrução de documentos é uma tarefa importante no processo de perícia em documentos questionados, sendo um processo manual, demorado e de difícil execução que necessita de equipamentos e profissionais treinados.

Este trabalho propõe uma metodologia para a reconstrução de documentos mutilados baseado em programação dinâmica e numa versão modificada do algoritmo de Prim. O intuito é possibilitar a recomposição de documentos questionados, com base nos fragmentos digitalizados, voltados para o auxílio e agilidade na perícia em questões judiciais, utilizando métodos digitais e não destrutivos.

O método proposto utiliza poucas características as quais são extraídas dos contornos dos fragmentos, e a partir dessas características inicia-se o processo de análise e combinação para encontrar candidatos a parceiros.

Os resultados alcançados mostraram-se promissores, com taxa média de 75% de reconstrução dos documentos constantes na base de imagens da PUCPR, atingindo um ganho em 24% comparado com o método proposto por Solana [SOLANA, 2005], incluindo a redução da taxa de erro em 4% e de falsos candidatos em 20%.

Palavras-chave: Documentos questionados, reconstrução de documentos, documentos mutilados, computação forense, programação dinâmica.

ABSTRACT

The document reconstruction is an important task on the document questioned process. In general, it can be a very time consuming, very hard to execute and needs equipment and well trained experts.

This work proposes a methodology for document reconstruction using dynamic programming and a modified version of the Prim´s algorithm in order to improve the documents questioned process.

The proposed method uses a few features that are extracted from the boundaries of the fragments and them use this features to find the best match among all pieces.

The results reported are promising, with an average rate of 75% of reconstruction of the documents contained in the PUCPR database. It reaches a gain of 24% compared to previous method proposed by Solana (2005), including reducing the error rate in 4% and false candidates in 20%.

Key-words: Questioned documents, documents reconstructed, mutilated documents, forensic computing, dynamic program.

Capítulo 1

INTRODUÇÃO

1.1 CONTEXTO

Quando, a partir de fatos, busca-se fundamentar uma pretensão deduzida em juízo, considera-se o conceito de prova. Sabendo que as afirmações ou negativas dos fatos declaradas pelo autor podem ou não ser verdades, e que estas irão se contrapor às alegações do réu, cabe ao juiz, através das provas apresentadas pelas partes, solucionar a questão.

Prova é um instrumento do qual o juiz utiliza para formar seu livre convencimento acerca dos fatos alegados no processo.

Prova é a produção dos atos ou dos meios com os quais as partes e o juiz entendem afirmar a verdade dos fatos alegados. A prova pode ser testemunhal, onde se considera a afirmação pessoal do fato; documental, onde se exprime a afirmação do fato de forma escrita ou gravada; material, consiste em qualquer materialidade da prova do fato, como os exames periciais [CAMPELLO, 2005].

Existem várias formas de reunir provas ou indícios que possam suportar a convicção dos fatos pelo juiz, auxiliando na associação ou desassociação de provas de um fato alegado. Uma dessas formas é através do uso de ciência. Na ciência, o ramo do conhecimento que propõe o uso de conhecimentos

técnico-científicos para este fim em específico é conhecido como ciência forense.

As ciências forenses possuem uma grande área de abrangência, dentre elas, pode-se citar a psiquiatria, medicina legal e jurisprudência, patologias, toxicologia, padrões de manchas de sangue, DNA (ácido desoxirribonucléico) e sorologia, odontologia, antropologia, arqueologia entre outras. Nas ciências forenses, o ramo que está diretamente relacionado com análises de documentos é conhecido como documentoscopia.

Documentos podem estar expostos em diversos meios de suporte diferentes, como fitas de áudio, fotografias impressas ou digitais, fitas de vídeos, pinturas ou simplesmente documentos escritos em papel conhecidos como *documentos questionados*.

Documentos em papel comumente são usados como prova para compor os altos do processo. Como exemplos, são citados os bilhetes de extorsões, bilhetes de seqüestros, registros de negócios, registros de propriedades, documentos falsificados, cartas anônimas que possam destruir relacionamentos familiares ou relacionamentos em negócios, notas fiscais, bilhetes de loterias, formulários de seguros, registros médicos entre outros [ECKERT, 1992].

O FBI (*Federal Bureau Of Investigation*), assim como os demais departamentos de investigação espalhados por diversos estados americanos, possuem unidades chamadas de QDU (unidade de documentos questionados) com a finalidade de prover suporte para análise forense de documentos, tidos como evidências, coletados durante as investigações. O departamento federal

americano provê ainda treinamento especializado e suporte para implantação de novos departamentos regionalizados em todos os Estados americanos, evidenciando a importância das análises forenses nas investigações [FBI, 2005] [SCHMITKNECHT, 2004].

Porém, para se obter êxito em perícias de documentos, os peritos dependem diretamente da composição e do estado de conservação dos documentos submetidos à análise. Devido a vários fatores, é fato que documentos podem apresentar diversos problemas quanto a sua conservação, principalmente devido às intempéries sofridas durante os anos em seus locais de armazenamentos. Há também problemas de mutilação dos documentos de causa proposital ou criminal, com a intenção de ocultar ou representar informações que não condizem com a realidade, ou seja, resultando em falsificações.

Sendo assim, a reconstrução de documentos mutilados é necessária e, na maioria das vezes, é executada de forma manual através de um processo de difícil execução. Conforme mostra a figura 1.

Figura 1 - Reconstrução de documentos mutilados [FBI, 2004].

Processos de reconstrução manuais podem provocar diversas alterações nos documentos originais ocasionando perdas de informações. Processos manuais implicam em processos destrutivos, uma vez que os materiais utilizados para realizar a composição do documento, como cola e fitas adesivas ou até mesmo o manuseio negligente, não são adequados [SOLANA, 2005].

1.2 DESAFIO

Sabendo que a atividade de reconstrução de documentos mutilados em papel é realizada, em sua maioria, utilizando processo manual, onde os métodos são considerados destrutivos e não próprios, e de que o tempo necessário para a realização desta atividade é extremamente dispendioso, o desafio é desenvolver um método computacional, não destrutivo, para a execução total ou parcial desta atividade.

Serão analisados documentos que apresentam mutilações voluntárias, criadas através de equipamentos cortantes como tesoura, estilete, régua ou mutilações realizadas manualmente através de rasgos no documento. Na figura 2 é demonstrado um documento mutilado constante na base de dados da PUCPR.

Figura 2 - Um documento mutilado [SOLANA, 2005].

1.3 MOTIVAÇÃO

Em diversos casos, os fatos alegados pelas partes aos quais o juiz precisa conhecer, não podem ser provados por simples declarações das partes e/ou de testemunhas. Mesmo os documentos que sejam apresentados pelas partes, podem necessitar de uma avaliação técnica para averiguar a sua veracidade.

Caso o juiz não possua conhecimento técnico suficiente para uma análise adequada que seja capaz de avaliar a veracidade do documento, faz-se a necessidade do trabalho de perícia [CAMPELLO, 2005]. Sendo assim, é de suma importância que existam métodos automatizados ou semi-automatizados que auxiliem o processo de perícia, tanto para melhorar a qualidade da perícia quanto para diminuir o tempo dispensado nesta atividade.

Para que uma perícia seja considerada correta e imparcial, ela deve ser baseada em conhecimentos técnico-científicos confiáveis e relevantes perante a comunidade científica.

Analisando por outro aspecto, a reconstrução digital de documentos mutilados pode auxiliar não somente em questões judiciais, mas também nas mais diversas áreas do conhecimento, como em recuperações de peças de cerâmica em escavações arqueológicas, conforme figura 3, recuperação de material histórico como livros e documentos antigos, pinturas em murais, painéis, azulejos e quadros conforme figura 4.

(a) (b)

Figura 3 - Fragmentos arqueológicos: (a) Fragmentos com possíveis combinações; (b) Fragmentos combinados manualmente [KAMPEL & SABLATNIG, 2004].

(a) (b) (c) (d)

Figura 4 - Área de aplicação para reconstrução de documentos: (a) Análise de documentos questionados; (b) Recuperação de livros; (c) Remontagem de afrescos, painéis, murais, azulejos, etc; (d) Documentos históricos [SOLANA, 2005].

Observando as diversas áreas de aplicação para a reconstrução de documentos mutilados junto à carência de métodos científicos que auxiliem este processo e o aumento constante de processos envolvendo documentos que necessitam de análise pericial, percebe-se a necessidade e a importância

de realizar pesquisas na área forense. Principalmente valendo-se do auxílio e da evolução das técnicas dos sistemas computacionais.

1.4 PROPOSTA

O presente trabalho tem como finalidade apresentar um método de reconstrução digital de documentos em papel mutilados intencionalmente, por objetos cortantes, utilizando técnica de programação dinâmica aplicada à análise de borda dos fragmentos em documentos que possuem formas irregulares de mutilação.

Para a reconstrução dos documentos será utilizada a análise de borda dos fragmentos digitalizados. As informações retiradas das bordas dos fragmentos compõem o vetor de características individuais, e a combinação das características será realizada submetendo os vetores de características à técnica de programação dinâmica.

A recomposição da imagem do documento será realizada através da rotulação da junção entre os fragmentos que compõem a imagem do documento original. Diferentemente dos métodos já propostos, o resultado do processamento não se resume em apenas realizar a rotulação dos fragmentos parceiros, mas sim recompor e apresentar digitalmente o resultado do processo de reconstrução.

A análise de borda será realizada de acordo com a irregularidade dos cortes nos fragmentos analisados, sendo excluído do objetivo desse trabalho os documentos mutilados que possuam fragmentos com bordas regulares ou documentos mutilados que possam apresentar características não analisadas

ou detectáveis pela borda, como mutilações por queimaduras, rasuras, raspagens, manchas e resíduos de qualquer natureza.

1.5 CONTRIBUIÇÕES

A contribuição deste trabalho está focada na reconstrução digital de documentos questionados voltados para o auxílio e agilidade na perícia em questões judiciais, utilizando métodos digitais e não destrutivos. Porém, a mesma técnica poderá ser aplicada e/ou adaptada para auxiliar na reconstrução digital de documentos ou fragmentos de diversas outras áreas, como recuperação de documentos históricos, painéis, quadros, cerâmicas e demais itens e peças arqueológicas limitando-se apenas em duas dimensões.

1.6 ORGANIZAÇÃO

O presente trabalho está organizado em 6 capítulos. O Capítulo 2 contém a fundamentação teórica dos principais métodos, procedimentos e algoritmos que serão utilizados para compor este trabalho. O Capítulo 3 contém o estado da arte em reconstrução de documentos, e os principais trabalhos já realizados nesta área. No Capítulo 4 serão demonstradas duas metodologias propostas para a reconstrução de documentos, uma metodologia sem resultados promissores e outra com resultados. No Capítulo 5 serão expostos os resultados encontrados nos experimentos. As conclusões na utilização dos métodos propostos, a contribuição efetiva deste trabalho de pesquisa e as prospecções para trabalhos futuros serão expostas no capítulo 6.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

2.1 INTRODUÇÃO

Neste capítulo é apresentada a revisão dos principais algoritmos que serão utilizados neste trabalho de pesquisa, divididas em 4 partes. Na primeira parte será demonstrada a análise dos contornos de documentos mutilados através da cadeia de códigos de Freeman como representação do contorno de imagens. Na segunda parte será apresentada a técnica de programação dinâmica e a sua importância para este trabalho no processo de reconstrução de documentos mutilados. Na terceira parte será apresentado o método de aproximação poligonal que será utilizado para representar em ângulos e arestas os contornos dos fragmentos. Na quarta parte será apresentado o algoritmo de Prim [PRIM, 1957] para a manipulação de grafos.

2.2 ANÁLISE DE CONTORNO

2.2.1 CADEIA DE CÓDIGOS DE FREEMAN

Existem diversas formas de se representar uma imagem através de seu contorno. Uma das formas mais simples e conhecida é utilizando uma lista de pixels representada pelos pontos cartesianos (x, y). Neste trabalho, a análise

do contorno das imagens será realizada utilizando o código de cadeia de Freeman [FREEMAN, 1974].

O código de cadeia de Freeman possui uma representação fiel do contorno de imagens utilizando apenas um caractere por ponto, caractere de direção, como identificação do próximo pixel do contorno, conforme figura 5.

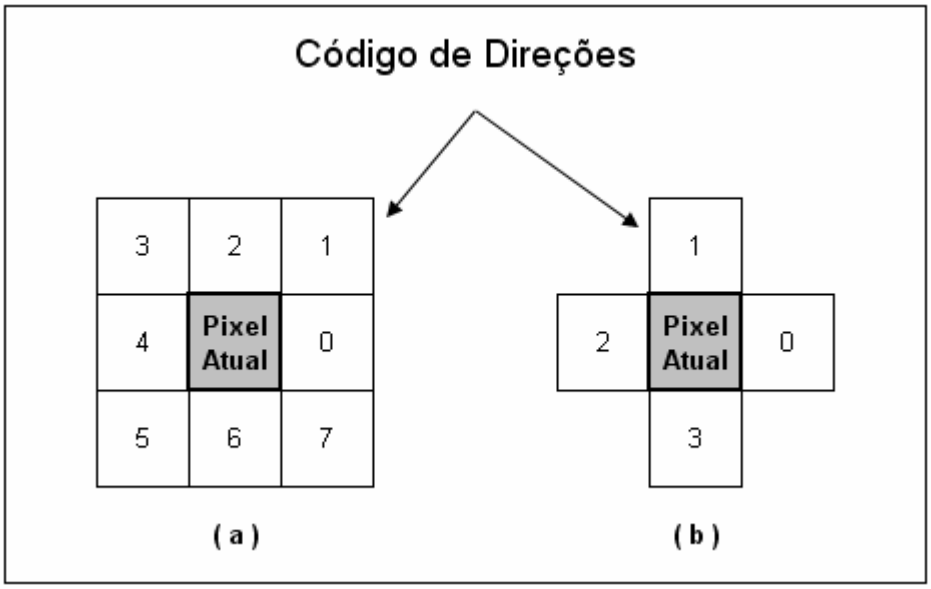


Figura 5 - Código de cadeia de Freeman: (a) Cadeia com 8 direções; (b) Cadeia com quatro direções.

Neste trabalho utilizar-se-á a cadeia de Freeman para realizar a análise da borda dos fragmentos de cada imagem constante na base de dados da PUCPR. Será utilizado o código de cadeia de 8 direções, representado todas as direções em que um pixel vizinho possa ser encontrado. A figura 6 demonstra como é feita a representação de um fragmento de imagem para uma cadeia de códigos de Freeman.

Na figura 6, no fragmento destacado (a), percebe-se que o primeiro pixel encontrado na imagem, realizando a varredura da esquerda para a direita e de cima para baixo, recebe o valor inicial zero. Na seqüência, cinco pixels no sentido horizontal caminhando para a direita, e de acordo com a definição da

figura 5, forma-se uma cadeia com seis pixels com valores zero. Ainda na seqüência, dois pixels que mudaram de direção em 90º, identificados pela direção 2. Formando assim a seqüência *00000022.* Dessa forma deve-se seguir o contorno formado pelos pixels até definir por completo a cadeia de códigos de Freeman para o contorno dos fragmentos.

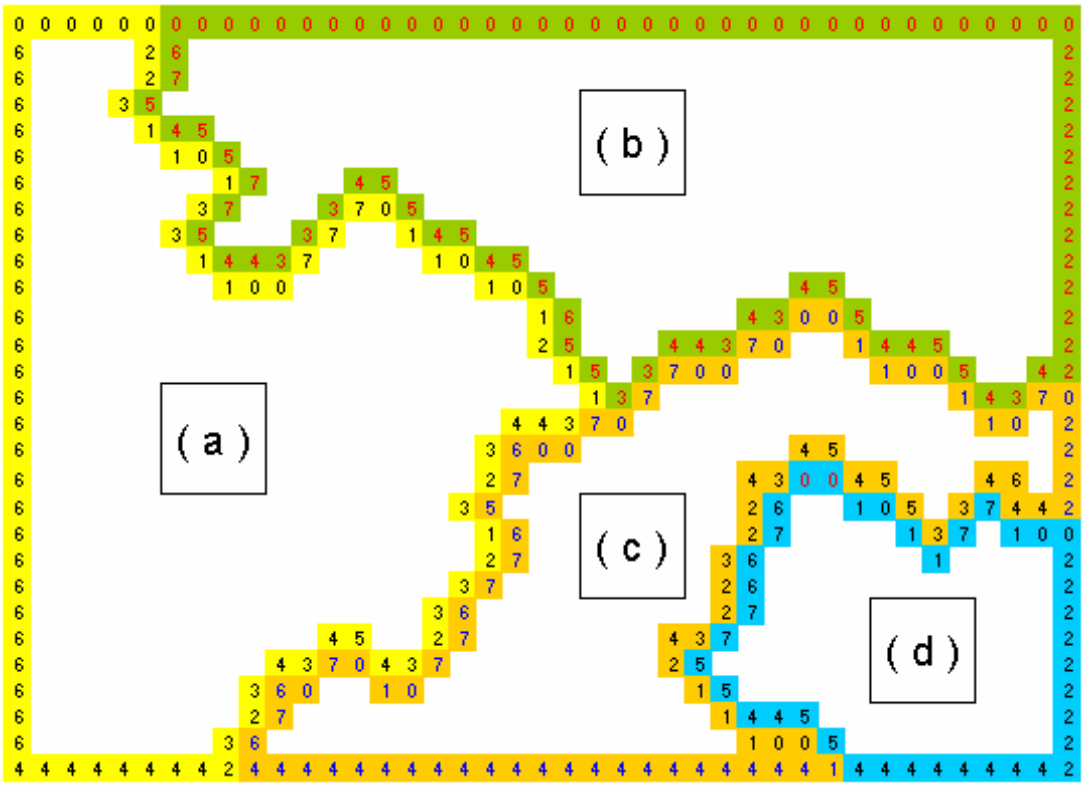


Figura 6 - Imagem em 4 partes com bordas representadas pela cadeia de código de Freeman com 8 direções.

A representação da cadeia de código de Freeman neste trabalho será sempre realizada a partir do primeiro pixel encontrado na imagem realizando a busca de cima para baixo e da esquerda para a direita. A figura 7 demonstra o resultado da cadeia de código de Freeman para os quatro fragmentos na figura 6.

(a)	H	000000223110133110077701101012113443231233234543432324444444466666666666666666666666666
	A	022222222222222222222222222200000000676707010767765767007556545455433344557754557664444
(b)	H	00000000000000000000000000000000000022222222222224345544554343443355655454554333445775545 76
	A	022310113310077701101011217700707011001107066666666666664444444444444444444444444444444
(c)	H	00110011070222244643355454342232234211100144444444444444444444446760701077677657600707700 70
	A	034344334344231233233454342320000000000000000000005445556076676607010107702006666434554455
(d)	H	00101177100222222222444444445544557766 76
	A	03232233110011000000006666666666445335545

Figura 7 - Resultado código de cadeia de Freeman: ‘H’ sentido horário; ‘A’ sentido anti-horário.

A representação da cadeia de códigos de Freeman em ambas as direções, horária e anti-horária, será importante para realizar a combinação (*matching*) de porções de seqüências de códigos dos fragmentos que forem adjacentes, ou candidatos a parceiros para reconstrução.

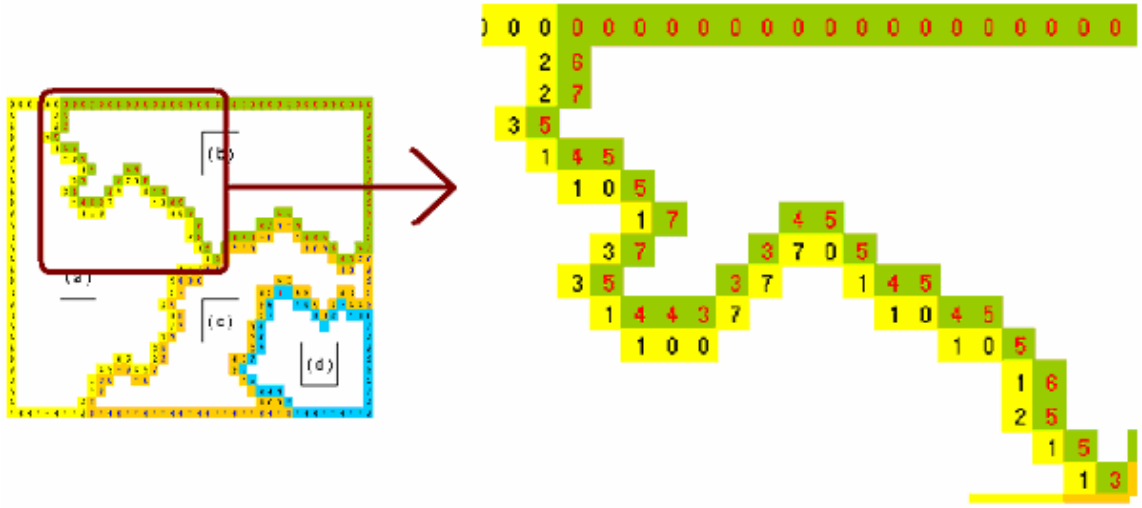


Figura 8 - Região de combinação (*matching*) entre o fragmento (a) e fragmento (b).

A figura 8 mostra a região de combinação entre o fragmento (a) e o fragmento (b) e a semelhança da seqüência de pixels que os compõe. Porém, observando a cadeia de código de Freeman formada pelos dois fragmentos na região de combinação, não há semelhança direta nas cadeias. Para resolver este problema e encontrar subcadeias de códigos em que se reconheça a semelhança e poder julgar os fragmentos corretamente como candidatos a

parceiros no processo de reconstrução, é preciso analisar necessariamente uma das cadeias no sentido horário e a outra no sentido anti-horário. Dessa forma as cadeias de combinação apresentam o mesmo sentido de leitura dos pixel pela cadeia de código de Freeman.

Na figura 9, em (1), é demonstrada a semelhança entre as cadeias de códigos de Freeman analisando a cadeia do fragmento (a) no sentido horário e o fragmento (b) no sentido anti-horário. Em (2), é demonstrada a semelhança entre as cadeias analisando o inverso, fragmento (a) sentido anti-horário e fragmento (b) sentido horário. As duas análises demonstram que os fragmentos são fortes candidatos a parceiros para a reconstrução. No próximo tópico será demonstrada uma técnica para encontrar computacionalmente as possíveis cadeias de combinação.

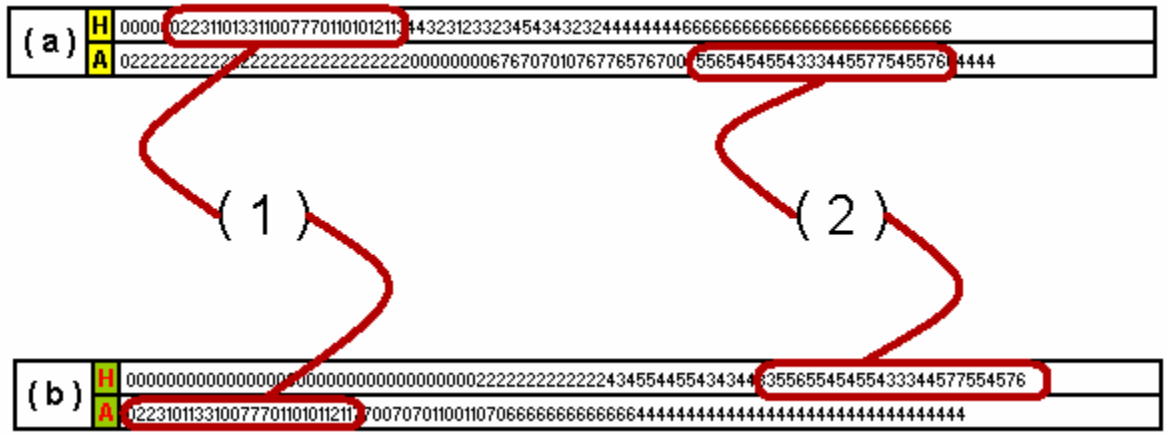


Figura 9 - Região de combinação entre os códigos de cadeia de Freeman entre as partes: (1), parte (a) sentido horário e (b) anti-horário; (2), parte (a) sentido anti-horário e (b) horário.

2.3 PROGRAMAÇÃO DINÂMICA

A programação dinâmica é uma técnica utilizada para resolver problemas de otimização onde a resolução total pode ser separada em problemas menores.

Cada subproblema deve poder ser processado independentemente dos demais subproblemas, e os resultados de cada um dos subproblemas precisam ser utilizados em diversos momentos até a resolução total do problema. Semelhante a um processo recursivo, porém cada resultado de um subproblema é armazenado em uma tabela, não sendo necessário realizar o reprocessamento a cada necessidade, o que ocorre em processos recursivos.

O termo programação dinâmica e o princípio de otimalidade surgiram em [BELLMAN, 1957] com o intuito de otimizar processos estocásticos. Porém, a técnica se tornou muito comum em diversas áreas do conhecimento, principalmente em pesquisas que envolvem análise e sequenciamento de cadeias de proteínas, DNA, RNA e na área computacional na construção de uma variedade de algoritmos.

Na programação dinâmica, diferentemente da programação linear, não há um padrão de necessidade que se possa aplicar a técnica. Apenas é necessário conseguir decompor a resolução de um determinado problema em problemas mais simples, porém isolados dos demais quanto a sua resolução, e ainda ligados entre si por uma recursividade para compor a solução total.

Para esta pesquisa, foi utilizado um algoritmo baseado em programação dinâmica conhecido como LCS (maior subseqüência comum). Este algoritmo tem o objetivo de apontar a maior subseqüência comum encontrada comparando a seqüência de duas cadeias [GREENBERG, 2003]. Esta técnica foi utilizada para encontrar as cadeias de seqüências comuns dentre as cadeias formadas pelos códigos de Freeman em fragmentos mutilados, com o

intuito de se avaliar as condições para considerar os fragmentos como candidatos a parceiros no processo de reconstrução.

O primeiro passo para executar o algoritmo está em criar uma matriz de tamanho (*M + 1*) x (*N + 1*), onde *M* e *N* são os respectivos comprimentos das duas cadeias a serem analisadas, sendo a primeira coluna e primeira linha preenchidas com zeros, conforme figura 10.

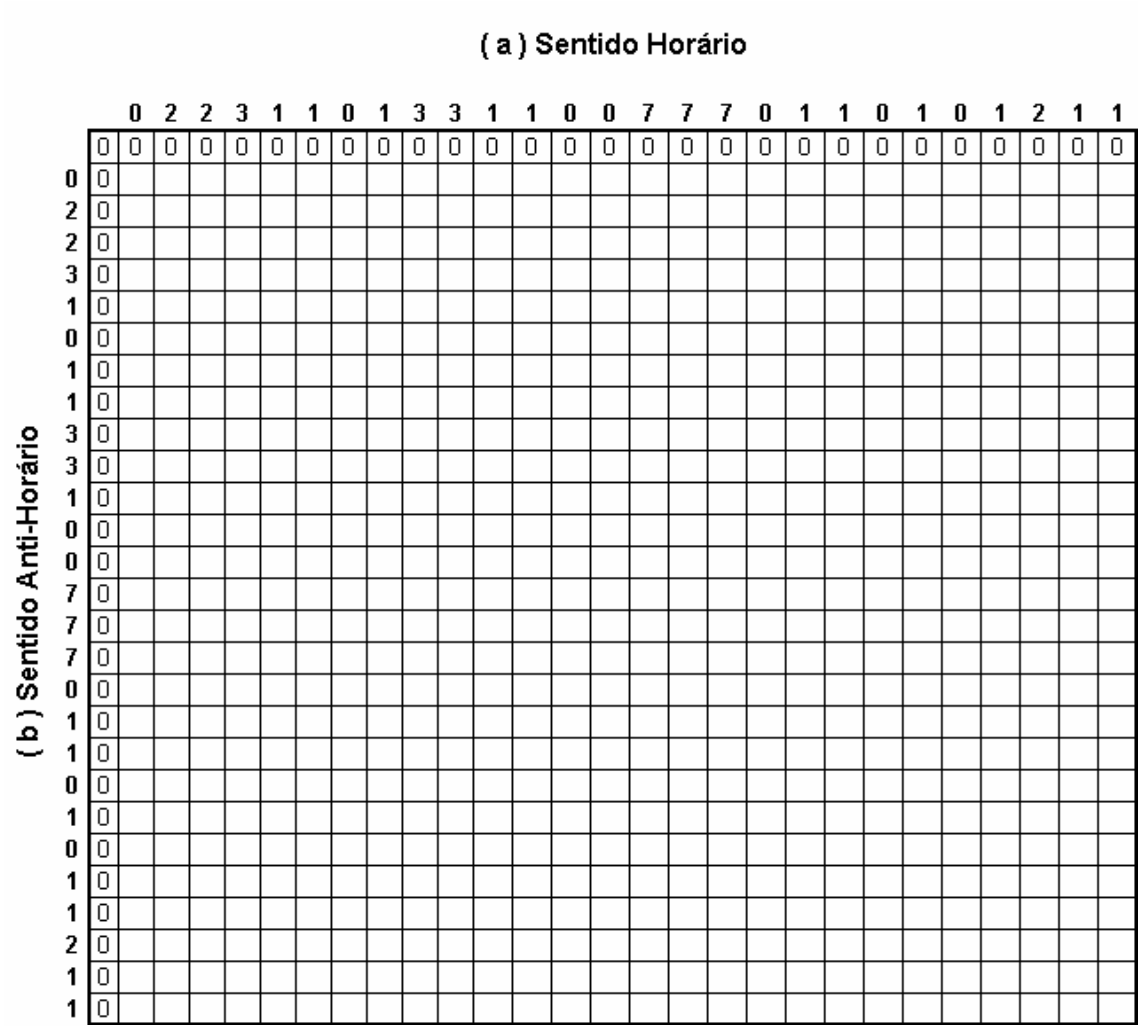


Figura 10 - Matriz inicial para o algoritmo de LCS: (a) Parte da seqüência no sentido horário; (b) Parte da seqüência no sentido anti-horário.

Após a criação da matriz inicial, o restante dos valores são completados segundo a regra de preenchimento da matriz de LCS, conforme equação 1.

$$E_{i,j} = Max\begin{cases} E_{i-1,j-1} + S_{i.j}; \\ E_{i,j-1} + P; \\ E_{i-1,j} + P; \end{cases} \quad (1)$$

Sendo:

- $E_{i,j}$ o elemento encontrado na linha i e na coluna j da matriz.
- $S_{i,j}$ = 1, se o valor de i na seqüência (a) for igual ao valor de j na seqüência (b). Ou seja, combinação (match).
- $S_{i,j}$ = 0, se o valor de i na seqüência (a) for diferente ao valor de j na seqüência (b). Ou seja, sem combinação (mismatch).
- P = 0, valor de penalidade.

Os valores de penalidade, combinação e não combinação podem ser alterados de acordo com o resultado que se espera alcançar.

O valor de penalidade tem a função de realizar um quebra na seqüência durante o processo de combinação. Utilizam-se valores baixos caso tenhamos uma tolerância maior a valores que não efetuaram combinação. Utilizam-se valores altos quando uma não combinação deve realizar uma quebra na seqüência de combinação e alinhamento.

Realizando o preenchimento dos valores na matriz como definido, a matriz do algoritmo fica conforme figura 11.

(a) Sentido Horário

(b) Sentido Anti-Horário		0	2	2	3	1	1	0	1	3	3	1	1	0	0	7	7	7	0	1	1	0	1	0	1	2	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	0
0		0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2		0	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2		0	1	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
3		0	1	2	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
1		0	1	2	3	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
0		0	1	2	3	4	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
1		0	1	2	3	4	5	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
1		0	1	2	3	4	5	6	6	7	7	7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
3		0	1	2	3	4	5	6	6	7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
3		0	1	2	3	4	5	6	6	7	8	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
1		0	1	2	3	4	5	6	6	7	8	9	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
0		0	1	2	3	4	5	6	7	7	8	9	10	10	11	11	11	11	11	11	11	11	11	11	11	11	11	11
0		0	1	2	3	4	5	6	7	7	8	9	10	10	11	12	12	12	12	12	12	12	12	12	12	12	12	12
7		0	1	2	3	4	5	6	7	7	8	9	10	10	11	12	13	13	13	13	13	13	13	13	13	13	13	13
7		0	1	2	3	4	5	6	7	7	8	9	10	10	11	12	13	14	14	14	14	14	14	14	14	14	14	14
7		0	1	2	3	4	5	6	7	7	8	9	10	10	11	12	13	14	15	15	15	15	15	15	15	15	15	15
0		0	1	2	3	4	5	6	7	7	8	9	10	10	11	12	13	14	15	16	16	16	16	16	16	16	16	16
1		0	1	2	3	4	5	6	7	8	8	9	10	11	11	12	13	14	15	16	17	17	17	17	17	17	17	17
1		0	1	2	3	4	5	6	7	8	8	9	10	11	11	12	13	14	15	16	17	18	18	18	18	18	18	18
0		0	1	2	3	4	5	6	7	8	8	9	10	11	12	12	13	14	15	16	17	18	19	19	19	19	19	19
1		0	1	2	3	4	5	6	7	8	8	9	10	11	12	12	13	14	15	16	17	18	19	20	20	20	20	20
0		0	1	2	3	4	5	6	7	8	8	9	10	11	12	13	13	14	15	16	17	18	19	20	21	21	21	21
1		0	1	2	3	4	5	6	7	8	8	9	10	11	12	13	13	14	15	16	17	18	19	20	21	22	22	22
1		0	1	2	3	4	5	6	7	8	8	9	10	11	12	13	13	14	15	16	17	18	19	20	21	22	22	23
2		0	1	2	3	4	5	6	7	8	8	9	10	11	12	13	13	14	15	16	17	18	19	20	21	22	23	23
1		0	1	2	3	4	5	6	7	8	8	9	10	11	12	13	13	14	15	16	17	18	19	20	21	22	23	24
1		0	1	2	3	4	5	6	7	8	8	9	10	11	12	13	13	14	15	16	17	18	19	20	21	22	23	24

Figura 11 - Matriz LCS calculada para as subseqüências de cadeias de códigos de Freeman.

O último elemento da matriz, destacado com o valor 25, indica a pontuação recebida no processo de combinação entre as seqüências. Esse valor é diretamente proporcional aos valores arbitrários estipulados para combinação, não combinação e penalidade. Em uma análise de combinação com cadeias é possível se conseguir várias combinações, sendo escolhida como melhor combinação àquela que apresentar a maior pontuação.

Para conseguir o resultado do alinhamento entre as cadeias, é necessário realizar um procedimento conhecido como *backtracking*. O processo de *backtracking* sugere percorrer uma trilha em sentido oposto dos valores encontrados, analisando o seu predecessor, ou seja, partindo do final

para o início. Para se conseguir este efeito, é criada uma nova matriz com as mesmas dimensões da matriz criada para a execução do algoritmo LCS conforme figura 11, porém o seu preenchimento será diferente, conforme a seguinte definição:

- $B_{i,j} = Diagonal,$ se o elemento da seqüência (a) no índice *i* for igual ao elemento da seqüência (b) no índice *j*. Ou seja, combinação.
- $B_{i,j} = Cima,$ se o valor de pontuação na matriz LCS no índice $E_{i-1, j}$ for igual ou maior o valor de pontuação do índice $E_{i,j.}$
- $B_{i,j} = Esquerda,$ se o valor de pontuação na matriz LCS no índice $E_{i, j-1}$ for igual ou maior o valor de pontuação do índice $E_{i,j.}$

Sendo:

- $B_{i,j}$ o elemento encontrado na linha i e na coluna j da matriz de backtracking.
- *Diagonal* é igual a direção *diagonal* de precedência na matriz de backtracking.
- *Cima* é igual a direção *cima* de precedência na matriz de backtracking.
- *Esquerda* é igual a direção *esquerda* de precedência na matriz de backtracking.

Na figura 12, é demonstrada a cadeia de *backtracking* calculada para a cadeia de códigos de Freeman. A região destacada mostra a seqüência de combinação entre as duas cadeias. Os trechos onde a cadeia segue as posições na diagonal representam a combinação, e os trechos onde a cadeia segue no sentido esquerdo ou para cima representam as falhas, ou não combinação. Para realizar o *backtracking*, os valores de direção foram fixados em direita = 3, cima = 2 e esquerda = 1.

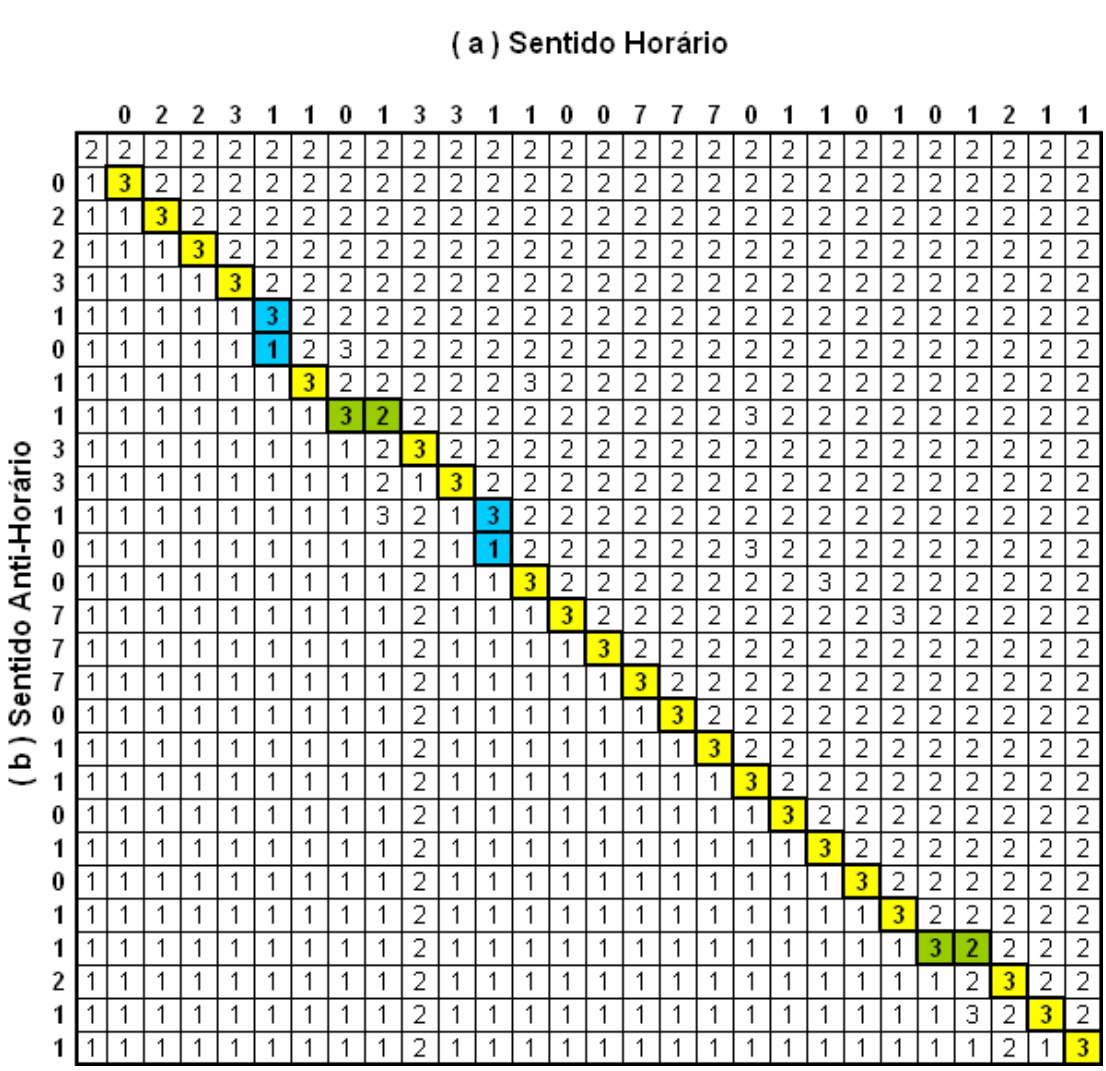


Figura 12 - Matriz de *backtracking* calculada para as subseqüências de cadeias de códigos de Freeman.

Após a matriz de *backtracking* montada, retira-se a cadeia de combinação resultante entre as duas cadeias questionadas. Neste exemplo, a cadeia de combinação resultante do *backtracking* pode ser retirada observando a cadeia no sentido horizontal ou vertical. Em ambas, quando a cadeia de *backtracking* não caminhar no sentido da cadeia analisada, ocorre a penalidade, onde a cadeia não forma combinação. Na figura 12, existe a ocorrência de quatro penalidades, duas no sentido horizontal, e duas no

sentido vertical. Na figura 13 estão os resultados das cadeias de *backtracking*. O símbolo "_" representa a penalidade.

Backtracking	022310133100777011010121 1
Backtracking Horizontal com penalidade.	0223_10133_1007770110101211
Backtracking vertical com penalidade.	0223101_331007770110101_211

Figura 13 - Resultado da cadeia de backtracking.

Neste exemplo pode-se identificar:

- Cadeia na horizontal com 27 caracteres e com combinação com a seqüência da vertical de 25 caracteres, resultando em um total de 92,6% de combinação. Número de penalidades igual a 2 ou 7,4%.
- Cadeia na vertical com 27 caracteres e com combinação com a seqüência da horizontal de 25 caracteres, resultando em um total de 92,6% de combinação. Número de penalidades igual a 2 ou 7,4%.

De posse desses dados pode-se julgar os fragmentos como bons ou ruins candidatos à combinação.

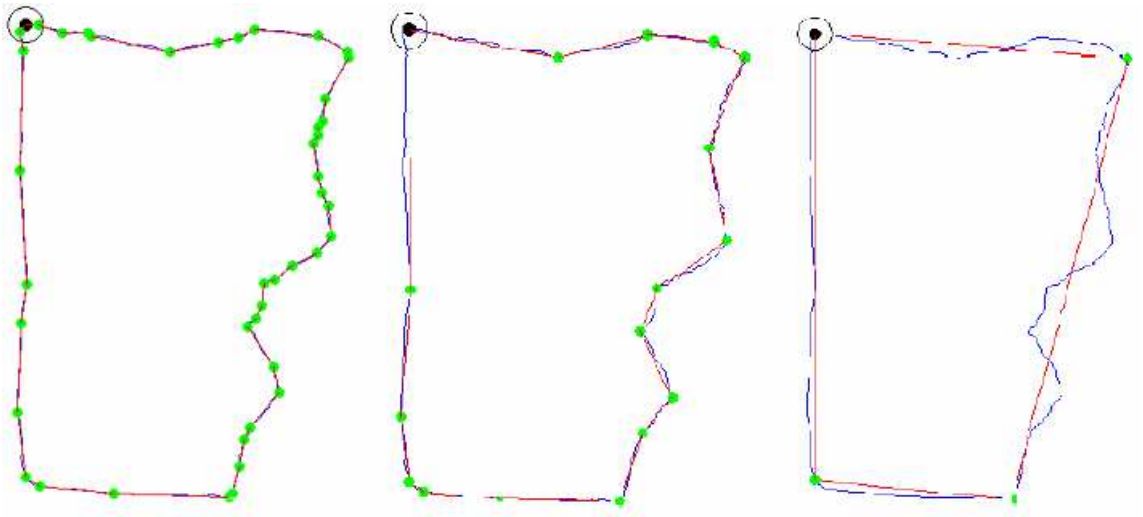
Para a análise de bordas de documentos mutilados, a quantidade de análises que deverão ser realizadas e o tamanho das cadeias poderiam impactar negativamente no tempo de processamento. Com o uso da técnica de programação dinâmica, o índice de complexidade para o processamento e a extração da cadeia de *backtracking* é linear dado por (*M x N*) processamentos, onde *M* é o comprimento da primeira cadeia analisada e *N* é comprimento da segunda cadeia. Em processamentos baseados em árvores de decisão recursivas, a complexidade para o processamento das cadeias é exponencial dado por ($M \times N^2$). Dessa forma a utilização da técnica de programação

dinâmica mostra-se como uma solução de extrema importância para o desempenho do processo de reconstrução.

2.4 APROXIMAÇÃO POLIGONAL

A aproximação poligonal é uma técnica utilizada para representar contornos de figuras complexas através de semi-retas, criando os vértices nos pontos onde ocorre as mudanças significativas de direção no contorno.

A técnica de programação dinâmica auxilia no trabalho de reconstrução de documentos pelo fato de criar uma representação do contorno contendo uma quantidade menor de pontos, não excluídas as características intrínsecas da forma do contorno original, mas apenas removendo os ruídos provenientes do processo natural de aquisição das imagens.



(a) (b) (c)

Figura 14 - Contorno aplicado ao algoritmo de aproximação poligonal [SOLANA, 2005].

A figura 14 demonstra um fragmento de documento submetido ao processo de aproximação poligonal. O algoritmo de aproximação poligonal utiliza uma medida de erro responsável por definir o quão os pontos do contorno irão ser fiéis aos detalhes do contorno do fragmento. Na figura 14 (a)

foi utilizada uma taxa baixa de erro, gerando diversos pontos representativos do contorno, em (b) foi utilizada uma taxa média e em (c) uma taxa alta, gerando poucos pontos para representar o contorno dos fragmentos.

Solana realizou um estudo aprofundado a respeito dos algoritmos de aproximação poligonal existentes, concluindo que para a representatividade de fragmentos de documentos, o algoritmo proposto por Douglas e Peucker [DOUGLAS & PEUCKER, 1973] é o mais indicado por melhor representar as características do contorno dos fragmentos. O algoritmo de Douglas e Peucker é o mais utilizado em sistemas comerciais e em sistemas de geoprocessamento [SOLANA, 2005], e será o algoritmo de aproximação dinâmica utilizado neste trabalho de pesquisa.

2.5 ALGORITMO DE PRIM

O algoritmo de Prim [PRIM, 1957] relacionado à teoria de grafos, propõe a criação de árvores geradoras mínimas a partir de grafos que possuem pesos em suas arestas. O objetivo do algoritmo é criar uma árvore passando por todos os vértices sem produzir repetições, onde o custo total seja o mínimo possível. Para a criação da árvore geradora mínima, o algoritmo segue a seguinte seqüência:

- Criar uma árvore resultante *Rs,* a próxima aresta a ser adicionada é sempre a aresta de menor peso conectando a árvore a um vértice que não esteja na árvore.
- Um vértice qualquer é selecionado para ser o vértice inicial da árvore resultante *Rs*.
- A cada ciclo do algoritmo, uma aresta é adicionada a árvore *Rs,* conectando *Rs* a um vértice de *Gs = (Vs; Rs),* Evitando-se ciclos.
- Quando não há mais vértices para adicionar a árvore resultante, o algoritmo termina e a árvore geradora mínima está criada.

Dependendo da forma de implementação, o algoritmo de Prim pode assumir diferentes índices de complexidade, confira na tabela 8.

Tabela 1 - Índice de complexidade na implementação do Algoritmo de Prim. *Es* é o número de arestas e *Vns* o número de vértices do grafo.

Método	Complexidade
Busca em matriz de adjacência	O (Vns^2)
Busca em árvore binária	O (Es * log (Vns))
Busca em árvores enárias	O (Es + V * log (Vns))

O algoritmo de Prim, iniciado a partir de um vértice aleatoriamente escolhido, aumenta a altura da árvore resultante até que todos os nós do grafo sejam visitados.

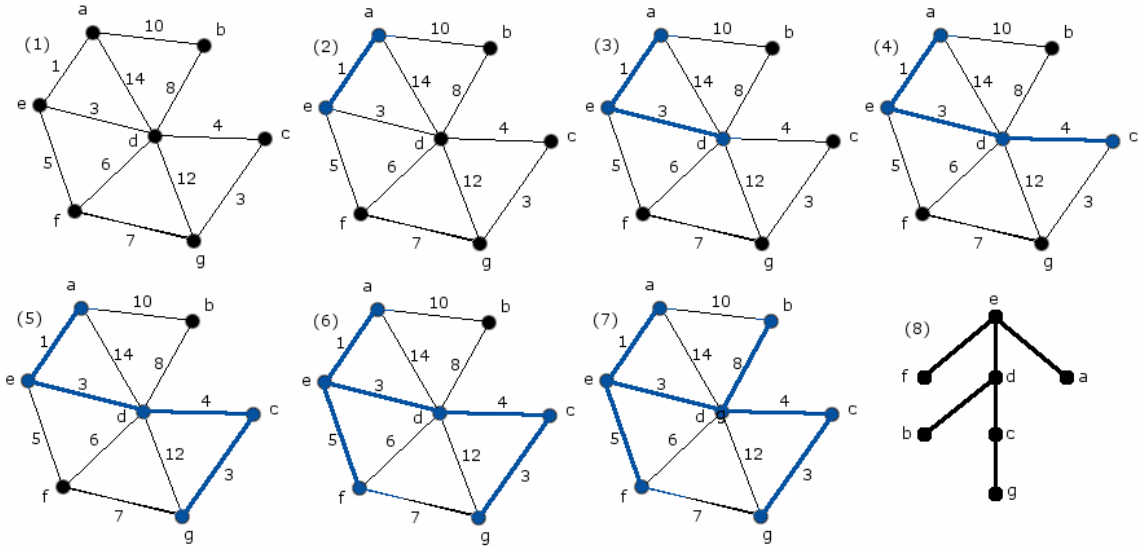


Figura 15 - Seqüência de formação da árvore geradora mínima a partir de um grafo utilizando o algoritmo de Prim.

A figura 15 mostra a seqüência de formação da árvore geradora mínima a partir de um grafo cíclico com pesos nos vértices iniciando aleatoriamente pelo vértice “e”. Observando as opções de ligação partindo do vértice “e”, seleciona-se o que possuir menor peso. Nesse caso foi selecionado o vértice

que liga o ponto “e” ao ponto “a” por possui o menor peso, peso 1. Seguindo o mesmo processo, observa-se qual a próxima ligação que possui o menor peso analisando os pontos “e” e “a”. Esse processo se repete até que todos os nós tenham sido visitados.

A utilização do algoritmo de Prim é importante para compor a seqüência de reconstrução dos fragmentos dos documentos.

2.6 CONCLUSÃO

No presente capítulo foi apresentado na fundamentação teórica os procedimentos e os algoritmos mais importantes para este trabalho.

Apresentamos a técnica de análise do contorno através da cadeia de códigos de Freeman; o algoritmo de programação dinâmica que será utilizado para a análise de combinação de características; o algoritmo de aproximação poligonal com a finalidade de criar uma representação fiel do contorno dos fragmentos e removendo os ruídos existentes; e por último, o algoritmo de Prim que será utilizado para a verificação de oclusões e do sequenciamento dos fragmentos no momento da recomposição digital dos documentos.

No capítulo 3 será apresentado o estado da arte em reconstrução de documentos, mostrando os métodos propostos existentes. Serão demonstrados também alguns processos de reconstrução de peças arqueológicas por possuírem semelhança em reconstrução de documentos. Será abordada também uma proposta para a reconstrução de documentos originalmente digitais mostrando a preocupação eminente também em recuperar documentos que não estejam em papel.

Capítulo 3

ESTADO DA ARTE EM RECONSTRUÇÃO DE DOCUMENTOS

3.1 INTRODUÇÃO

Este capítulo aborda as técnicas e métodos já existentes para reconstrução de documentos mutilados, tanto para documentos com cortes irregulares como regulares.

Devido à natureza semelhante, destaca-se neste capítulo processo para a reconstrução de peças de cerâmica e procedimentos para a reconstrução de quebra-cabeças. Também é demonstrada uma técnica para a recuperação de arquivos, documentos digitais, armazenados em mídias eletrônicas.

3.2 RECONSTRUÇÃO DE DOCUMENTOS EM RETALHOS “SPAGHETTI”

Em 2002, a companhia Churchstreet Technology anunciou o desenvolvimento de um sistema inédito, semi-automatizado, para a reconstrução digital de documentos mutilados do tipo “Spaghetti” [SOLANA, 2005].

Mutilação do tipo “Spaghetti” são mutilações realizadas por máquinas picotadoras, largamente utilizadas pelas empresas gerando mutilações em formatos de tiras regulares.

Figura 16 - Exemplo de mutilação “Spaghetti”. Unshredder Systems.

A figura 16 mostra um exemplo de mutilação “Spaghetti” gerada por máquina picotadora. A proposta da Churchstreet Technology é a reconstrução de documentos retalhados regularmente através do desenvolvimento de tecnologia própria de hardware e de software. A metodologia empregada na reconstrução não foi divulgada, sendo a reconstrução do documento realizada pelo próprio laboratório da Churchstreet Technology após o envio dos fragmentos dos documentos aos quais se deseja reconstruir [SOLANA, 2005]. O processo realiza a reconstrução de documentos mutilados em tiras mesmo com mutilações em tamanhos e posições diferentes, conforme demonstra a figura 17.

Figura 17 - Tiras recortadas na vertical e aleatoriamente na horizontal [SOLANA, 2005].

Em 2007, a companhia Unshredder Systems lançou o “Sistema de Reconstrução de Documentos Retalhados”. Segundo a própria empresa, esta

seria a primeira ferramenta comercial de reconstrução de documentos retalhados.

Figura 18 - Exemplo de reconstrução realizado pela Unshredder Systems.

A ferramenta é comercializada através da compra de licenças de uso do software. A empresa atua em diversos segmentos, como agências governamentais, departamentos de polícia, agências de segurança, escritórios de advocacia, agências comerciais de informação e investigadores privados, etc.

A figura 19 apresenta a interface da ferramenta desenvolvida pela Unshredder Systems.

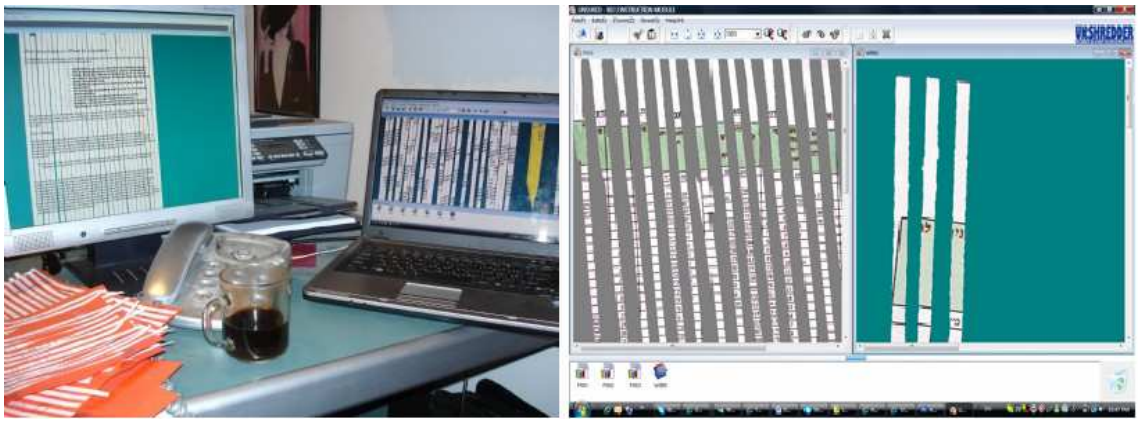


Figura 19 - Exemplo da interface da ferramenta de reconstrução desenvolvida pela Unshredder Systems.

Ukovich (2004) apresenta um método para a reconstrução de documentos retalhados em tiras utilizando descritores contidos no protocolo MPEG-7. Os testes foram realizados em apenas um documento retalhado, e os testes demonstraram que os descritores MPEG-7 podem ser utilizados para este trabalho de reconstrução, em particular na utilização dos descritores de cores, uma vez que os descritores de textura não apresentaram o resultado esperado. Ukovich conclui que existem diversas características que ainda podem ser exploradas em novos trabalhos.

Um dos trabalhos mais recentes para a reconstrução de documentos de texto retalhados em tiras é proposto por Prandtstetter, [PRANDTSTETTER, 2008], sendo uma solução híbrida entre o processo automatizado e a ação humana.

Máquinas com mutilações em tiras são as mais utilizadas comercialmente, porém não apresentam segurança para a informação que se deseja destruir. Existem diversas máquinas com modelos de mutilação diferenciados, além da mutilação em tiras, algumas delas são expostas abaixo:

- Corte cruzado: A máquina possui dois tambores para realizar os cortes com angulação diferenciada do corte, produzindo fragmentos retangulares, losangulares e filetes.
- Cortes em partículas: Cortes realizados em partículas pequenas em formatos geométricos, como círculos, quadrados, etc.
- Desintegração ou granulação: Corta repetidamente o papel aleatoriamente até que as partículas tornam-se suficientemente pequenas para passar através de uma malha.
- Furar e cortar: Lâminas rotativas furam o papel e depois o cortam aleatoriamente.

Estes métodos de mutilação são de difícil reconstrução, tornando-se indispensáveis na utilização em processos que requerem uma completa destruição da informação contida em documentos de papel.

3.3 RECONSTRUÇÃO DE DOCUMENTOS MUTILADOS IRREGULARMENTE

Diferente do processo de mutilação de documento regular, não existem produtos comerciais que realizam a tarefa de reconstruir documentos que tenham sido mutilados aleatoriamente, tanto por materiais cortantes quanto mutilações a mão.

Os trabalhos que mais se assemelham à reconstrução digital de documentos mutilados é a reconstrução de cerâmicas danificadas e a montagem automática de quebra-cabeças [LEITÃO, 2000], [KAMPEL & SABLATNIG, 2004], [YAO & SHAO, 2003], [TYBON, 2004], [KONG & KIMIA, 2001].

3.3.1 RECONSTRUÇÃO DE PEÇAS DE CERÂMICA FRAGMENTADAS

Leitão [LEITÃO, 2000] apresenta um método para reconstrução de cerâmicas quebradas ou partidas em formas irregulares. Leitão realizou dois testes. No primeiro teste, foi utilizado um documento em papel com 20 pedaços, o processo retornou 28 pares de fragmentos como candidatos a parceiros. Porém, desses 28 candidatos, 11 eram candidatos verdadeiros e 17 foram falsos positivos.

No segundo teste, foram utilizados 5 ladrilhos os quais foram quebrados em 112 fragmentos com cada fragmento com no mínimo 250 pixels de comprimento mínimo. Este processo retornou 22 pares de candidatos sendo 3 considerados falsos positivos.

A figura 20 apresenta os fragmentos de cerâmica para os testes de remontagem, a figura 21 (a) apresenta os possíveis encaixes e a figura 21 (b) apresenta os encaixes encontrados através do processo.

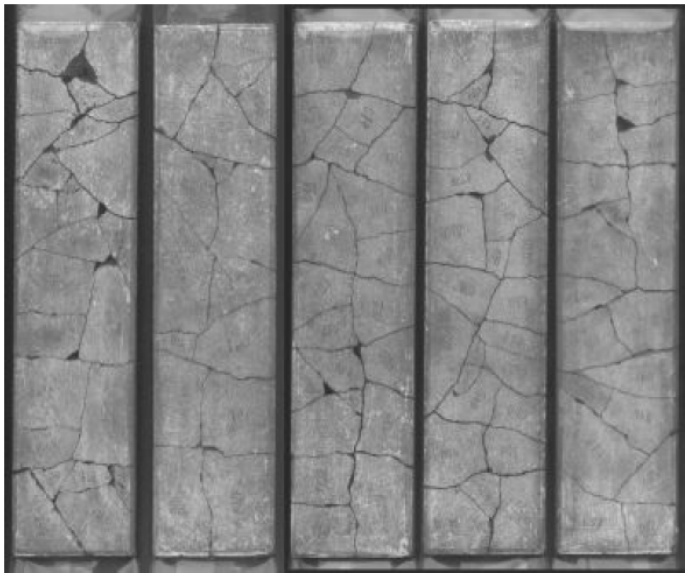
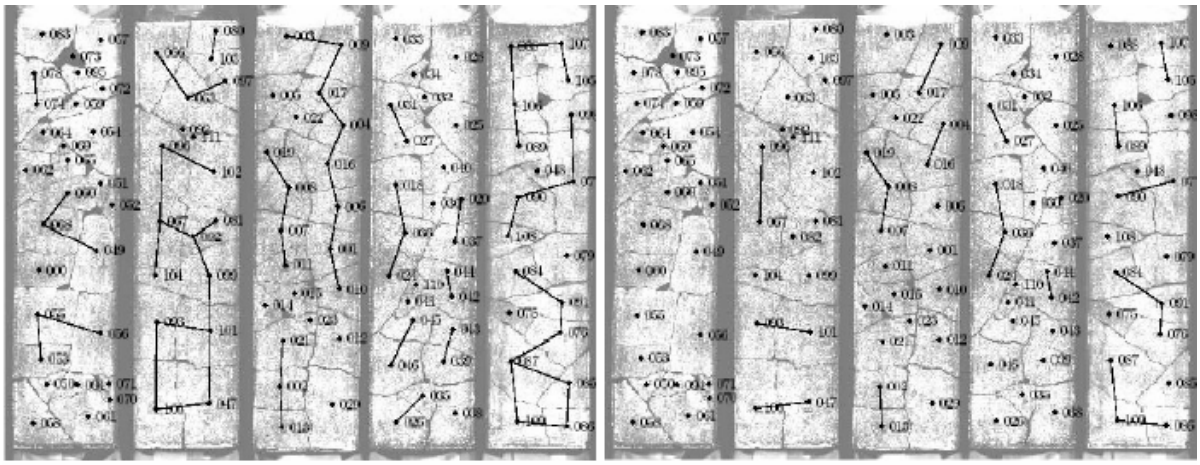


Figura 20 - Fragmentos de cerâmica para teste [LEITÃO, 2000].



(a) (b)

Figura 21 - (a) Fragmentos encaixáveis; (b) Resultado obtido [LEITÃO, 2000].

Leitão demonstrou ainda um segundo teste com modificações nos parâmetros que resultou em 277 pares de candidatos a parceiros. Dentre os 277 pares, apenas os 60 primeiros foram analisados, sendo 39 pares verdadeiros e 21 falsos verdadeiros.

A reconstrução dos fragmentos de cerâmica proposta por Leitão apenas realiza a rotulação dos candidatos a parceiros. A remontagem precisa ser realizada manualmente [LEITÃO, 2000].

Kampel e Sablatnig apresentam um método de reconstrução de potes de cerâmica fragmentados. Para a aquisição das imagens dos fragmentos, é utilizado o digitalizador 3D Minolta VIVID 900, equipamento de captura de imagens 3D [KAMPEL & SABLATNIG, 2004].

Após a digitalização em 3D dos fragmentos, o processo é iniciado realizando uma estimativa para analisar a correta posição em relação à inclinação axial dos fragmentos. Com a inclinação axial dos fragmentos alinhadas, é iniciado o processo de determinação dos candidatos a parceiros e os seus devidos encaixes.

Para o alinhamento e determinação dos candidatos a parceiros é utilizado um algoritmo baseado na pontuação obtida através de um valor de erro no alinhamento calculado, tendo como base as distâncias euclidianas dos pontos que formam a região de combinação, conforme a equação 2.

$$E_m = \frac{1}{Np} \sum_{i=1}^{N} \sqrt{(x_i - x'_i)^2 + (y_i - y'_i)^2} \quad (2)$$

Sendo:

- E_m o erro médio de alinhamento.
- Np número de pontos do alinhamento.
- x_i e x'_i coordenadas iniciais dos pontos de correspondência.
- y_i e y'_i coordenadas finais dos pontos de correspondência.

A equação de erro proposta é composta pelo inverso do número de pontos multiplicado pela somatória das distâncias euclidianas, calculadas através do teorema de Pitágoras aplicadas ao plano cartesiano, encontradas entre os pontos seqüenciais entre as arestas de encaixe.

(a) (b) (c)

Figura 22 - Partes encaixadas: (a) Fragmento 1 e fragmento 3; (b) Fragmento 1 e fragmento 5; (c) Reconstrução fragmentos 1, 2, 3 e 5.

Kampel e Sablatnig consideram os resultados promissores e o próximo passo seria testar a metodologia em uma base de dados com mais de 100 fragmentos de potes de cerâmica. Este resultado ainda não foi divulgado.

Outros trabalhos de reconstrução de objetos fragmentados também produziram resultados interessantes, como [WILLIS & COOPER, 2008], [LEITÃO & STOLFI, 2002] e [PAPAODYSSEUS, 2002].

3.3.2 RECONSTRUÇÃO DE QUEBRA-CABEÇAS

Diversos trabalhos já foram produzidos para tentar resolver o problema de reconstrução automática de quebra-cabeças, conhecidos como *Jigsaw Puzzle*.

O nome *Jigsaw Puzzle* (*Jigsaw* é uma ferramenta para realizar cortes precisos em objetos de madeira) provém do ato de recortar figuras pintadas em blocos de madeira em pequenos pedaços que se entrelaçam. É a origem dos quebra-cabeças comerciais atualmente encontrados em lojas de brinquedos. A figura 23 mostra um exemplo de *Jigsaw Puzzle* em madeira.

Figura 23 - Quebra-cabeça cortado à mão, em madeira. Golfistas no campo de golfe Prestwick na Escócia, construído em 1914. (www.britannica.com).

Atualmente, os quebra-cabeças comerciais seguem um padrão de formação das peças tornando os fragmentos mais regulares com extremidades lisas e bem definidas. Este fato diminui a complexidade para a resolução da remontagem de quebra-cabeças [SOLANA, 2005], sendo estas características bem exploradas, conforme a figura 24.

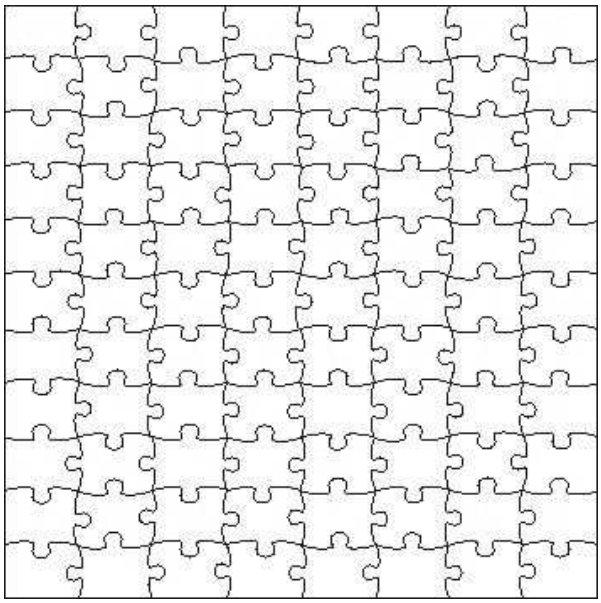


Figura 24 - Exemplo de formação das peças em quebra-cabeça.

Yao e Shao desenvolveram um método para a reconstrução automática de quebra-cabeças. O método é baseado em extração de características globais de cada peça do quebra-cabeça, incluindo características geométricas do contorno e características da imagem [YAO & SHAO, 2003]. Os candidatos a parceiros são encontrados através de algoritmos de reconhecimento de curvas (*curve matching*), e na seqüência os falsos candidatos podem ser analisados através de análises de contexto das peças. A figura 25 demonstra um exemplo de reconstrução.

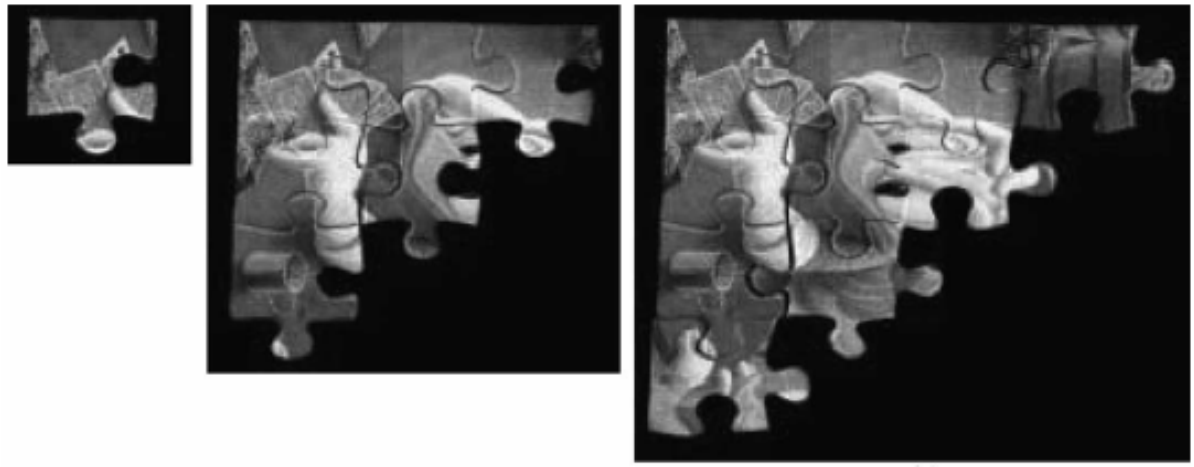


Figura 25 - Processo de reconstrução de quebra-cabeça. (@Disney) [YAO & SHAO, 2003].

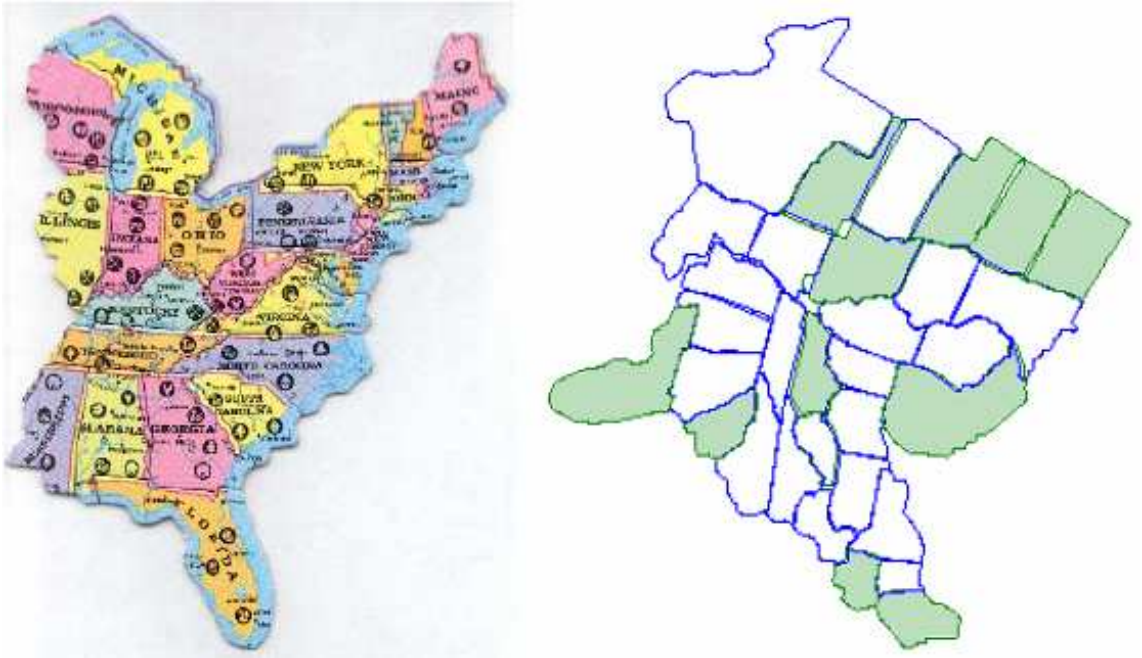
O método proposto por Yao e Shao realiza os encaixes em pares de peças, sendo um futuro trabalho, realizar os encaixes entre as 4 peças adjacentes na expectativa de melhora de resultados. Os resultados são promissores e considerados com sucesso, porém o método precisa ser testado com quebra-cabeças com maiores quantidades de peças [YAO e SHAO, 2003].

Kong e Kimia apresentam um método para a reconstrução de quebra-cabeça também baseado em reconhecimento de curvas, porém o trabalho foi realizado em quebra-cabeças com formato de peças diferenciados, diferente do proposto por YAO e SHAO [YAO e SHAO, 2003]. Kong e Kimia também utilizaram fragmentos de cerâmica para a realização dos testes da metodologia [KONG & KIMIA, 2001].

O método é baseado em duas etapas: a primeira etapa realiza um processo de aproximação poligonal no contorno dos fragmentos para diminuir a complexidade computacional da representação do contorno. Após o processo de aproximação poligonal, segmentos de curvas são analisados a partir dos pontos de vértices encontrados na aproximação poligonal. Estes segmentos

encontrados são então analisados com mais critérios e escalas mais sensíveis de aproximação e reconhecimento de curvas [KONG e KIMIA, 2001].

Kong e Kimia utilizam os valores de reconhecimento de curvas para realizar a reconstrução em grupo de 3 fragmentos por processo. Na figura 23 é apresentado o resultado de reconstrução em uma porção do mapa dos Estados Unidos com 32 fragmentos, sendo 11 erroneamente remontados destacados em 26 (b).



(a) (b)

Figura 26 - (a) Parte de quebra-cabeça; (b) Resultado obtido pelo método [KONG e KIMIA, 2001].

Nos demais testes efetuados foram remontadas peças de cerâmicas fragmentadas, porém não foram divulgados os valores estatísticos da reconstrução obtida pelo método proposto.

3.3.3 RECONSTRUÇÃO DE DOCUMENTOS EM PAPEL MUTILADOS

Solana descreve um método para a reconstrução de documentos mutilados utilizando aproximação poligonal aplicada na extração de características dos contornos dos fragmentos [SOLANA, 2005].

A metodologia proposta por Solana atua basicamente em dois passos: No primeiro passo, a aproximação poligonal é aplicada nos contornos dos fragmentos para reduzir a complexidade das bordas dos fragmentos. No segundo passo atua na extração das características encontradas nos vértices encontrados na aproximação poligonal. Algumas ambigüidades no processo são resolvidas buscando uma solução global para a reconstrução. O método proposto por Solana está organizado em 6 partes, conforme figura 27.

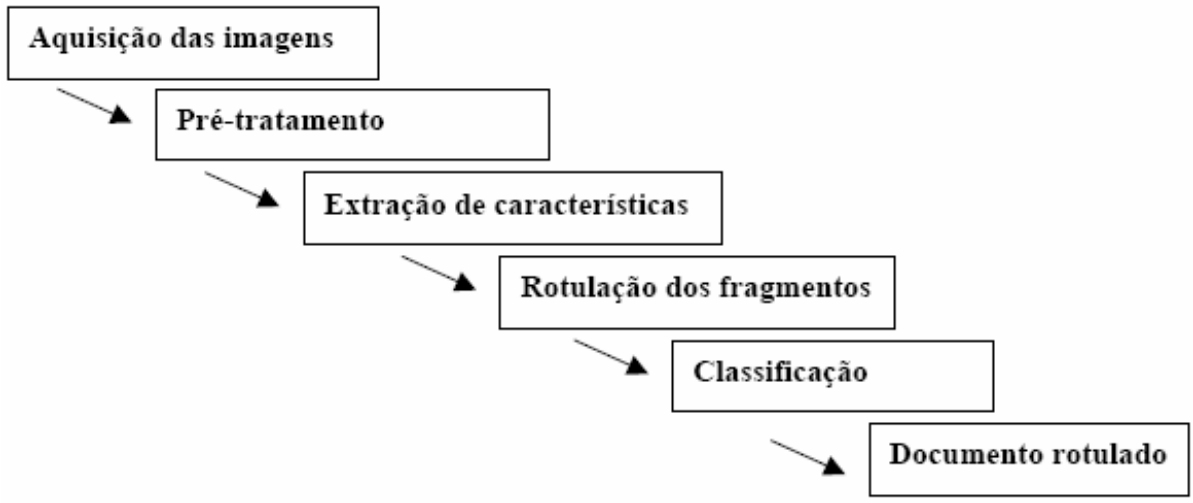


Figura 27 - Esquema geral da metodologia de reconstrução de documentos mutilados. [SOLANA, 2005].

Para a realização dos experimentos e testes, Solana criou a base de dados de documentos em papel mutilados da PUCPR, contendo 100 documentos fragmentados em 855 fragmentos. No desenvolvimento da base de dados, buscou-se a fidelidade com os documentos submetidos à perícia forense que geralmente são mutilados visando destruir ou inutilizar provas em

contestações judiciais. Cada tipo de mutilação, como rasgados à mão ou cortados com tesouras e estiletes, possui características próprias de mutilação. Os fragmentos de documentos da base de dados buscam representar estas características [SOLANA, 2005].

Nestas condições, a base foi formada contendo documentos manuscritos, documentos de textos, documentos tipografados e documentos contendo imagens. Os documentos sofreram mutilações através de tesoura, régua, estilete e também rasgados manualmente.

A figura 28 mostra os vértices encontrados na aproximação poligonal para um fragmento de documento.

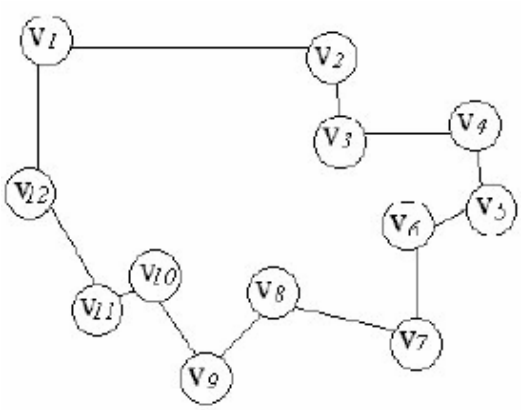


Figura 28 - Vértices de extração de características [SOLANA, 2005].

Para cada vértice exposto na figura 28, são retiradas as características do contorno do fragmento. As características são compostas pelo número do vértice, pelo ângulo externo ao fragmento formado pelo vértice, pela distância formada pelo vértice analisado e o vértice anterior e sua importância, pela distância formada pelo vértice analisado e o vértice posterior e sua importância, e pelas coordenadas cartesianas do vértice.

A importância das distâncias é dada pela relação do tamanho da distância e o tamanho total do contorno do fragmento.

O grau de similaridade entre características de fragmentos que possuem encaixe é dado pelas seguintes comparações:

- Ângulo**:** para haver similaridade entre a característica ângulo de dois vértices de fragmentos distintos, a soma entre os ângulos deve atingir 360º, porém com uma tolerância de 2º. Caso a semelhança ocorra, atribui-se o valor 1 para a variável W_{angulo}, caso contrário a variável W_{angulo} recebe o valor zero.
- Distância**:** para haver similaridade entre a característica distância de dois fragmentos, os comprimentos das distâncias são comparados conforme a figura 29. D_{b1} é a distância euclidiana entre o vértice A e seu vizinho anterior C, a qual é comparada com a distância D_{a1}. A distância D_{b2} é comparada com a distância D_{a2}. Existe uma tolerância no valor de 2 devido a arredondamentos.

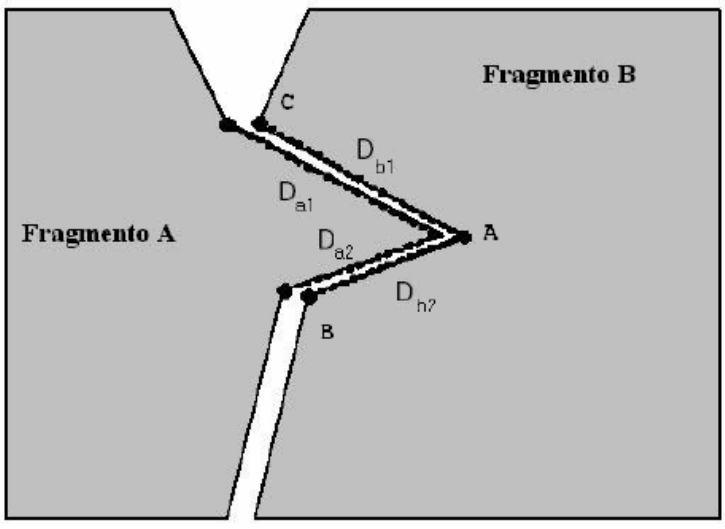


Figura 29 - Similaridade da característica distância [SOLANA, 2005].

Após o cálculo de verificação entre as características ângulo e distância, é calculada uma medida de semelhança $W_{matching}$ dada pela equação 3:

$$W_{matching} = \begin{cases} 1 \;\; se \;\; \left[(D_{b1} \cong D_{a1}) \;\; OU \;\; (D_{b2} \cong D_{a2}) \;\; e \;\; W_{ângulo} = 1 \right] \\ 5 \;\; se \;\; \left[(D_{b1} \cong D_{a1}) \;\; E \;\; (D_{b2} \cong D_{a2}) \;\; e \;\; W_{ângulo} = 1 \right] \end{cases} \quad (3)$$

- Importância da distância: Solana adicionou esta característica ao processo para analisar possíveis coincidências entre arestas e fragmentos [SOLANA, 2005]. Caso isso ocorra, é sabido que existe falso positivo. Sendo assim a escolha do candidato a parceiro é realizado, analisando a variável de $W_{matching}$ e a relação, em porcentagem, do tamanho das arestas em relação ao fragmento completo. Se o tamanho for superior ou igual a 20%, acrescentar 2 pontos ao $W_{matching}$, caso seja igual ou superior a 10%, acrescentar 1 ponto.

Os valores adicionados à medida de semelhança $W_{matching}$ são valores empíricos e através de experimentos mostraram-se adequados para o processo [SOLANA, 2005].

Para a rotulação da reconstrução dos documentos, Solana utilizou o algoritmo proposto por Leitão [LEITÃO, 2000], que busca a melhor combinação entre pares de fragmentos por vez através da utilização da medida de semelhança $W_{matching}$.

Tabela 2 - Resultados do experimento 1. Classificação com repetição de candidatos a parceiros [SOLANA, 2005].

Tolerância aproximação	Quantidade de	Erros durante o	Candidatos falsos	Candidatos corretos

poligonal	documentos	processo		
Baixa	81%	15%	34%	51%
Média	81%	17%	40%	43%

Tabela 3 - Resultados do experimento 2. Classificação sem repetição de candidatos a parceiros [SOLANA, 2005].

Tolerância aproximação poligonal	Quantidade de documentos	Erros durante o processo	Candidatos falsos	Candidatos corretos
Baixa	81%	19%	24%	57%
Média	81%	20%	31%	49%

Tabela 4 - Resultado do experimento 3. Classificação com convergência [SOLANA, 2005]. Apenas 45% dos documentos terminaram o processo.

Tolerância aproximação poligonal	Quantidade de documentos	Erros durante o processo	Candidatos falsos	Candidatos corretos
Baixa	45%	0%	13,33%	86,67%
Média	45%	0%	19,56%	80,44%

Os resultados obtidos por Solana demonstram um grande avanço na área de reconstrução de documentos mutilados. Principalmente pela construção e utilização da base de dados de documentos para a aferição do método proposto. Porém, o método não promove a reconstrução visual do documento, mas sim apenas a rotulação dos encaixes entre os fragmentos para que o documento possa ser reconstruído manualmente. Solana ainda conclui, sendo uma limitação, que o método se degrada à medida que o número de fragmentos dos documentos cresce [SOLANA, 2005].

Um dos trabalhos mais recentes para a reconstrução de documentos em papel mutilados é o proposto por Smet [SMET, 2007].

Smet demonstra uma metodologia para a reconstrução de documentos mutilados, quando a pilha de fragmentos rasgados pode ser recuperada

ordenadamente. Smet considera que, em campo, a recuperação dos fragmentos, o transporte e o armazenamento de documentos devem garantir a integridade da pilha de fragmentos rasgados.

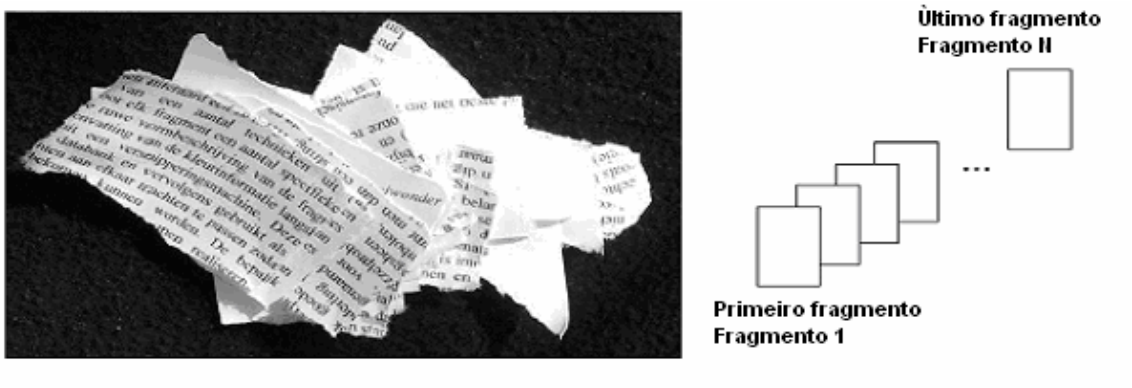


Figura 30 - Exemplo de pilha de fragmentos e a representação formal [SMET, 2007].

A figura 30 apresenta uma pilha de fragmentos de um documento mutilado e a sua representação formal para o método de Smet [SMET, 2007].

É pressuposto para este método que não há falso positivo no processo de reconstrução, ou seja, é pressuposto que o método não contém erros. A única consideração relevante realizada por este método é a verificação de casos em que os fragmentos possam ter sido rasgados em seqüências diferentes da esperada.

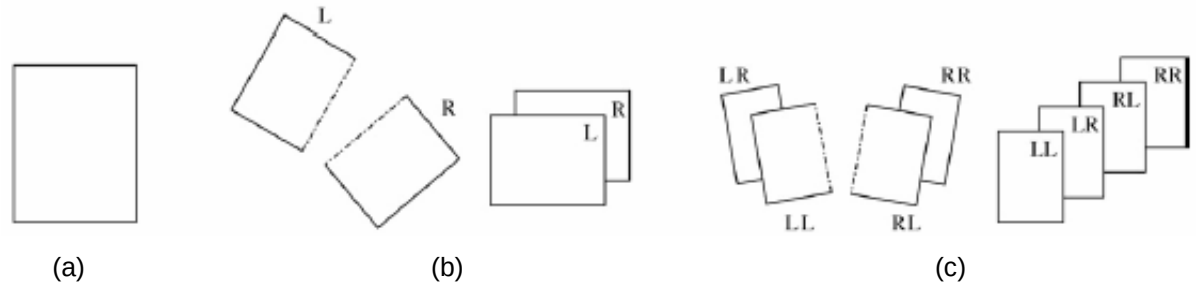


Figura 31 - Seqüência de rasgamento usando a seqüência LOT (leftmost-on-top) de posicionamento de fragmento: (a) Documento original; (b) Primeiro passo de rasgamento; (c) Segundo passo de rasgamento [SMET, 2007].

A figura 31 mostra a estratégia LOT (leftmost-on-top) de rasgamento. Porém, dependendo da seqüência e do rearranjo no momento de realizar os

rasgos no documento, a seqüência poderá ser: {*LL, LR, RL, RR*}, {*RL, RR, LL, LR*}, {*LR, LL, RR, RL*}, {*RR, RL, LR, LL*}, onde *L* é o fragmento da esquerda e *R* o fragmento da direita. Em casos de seqüências aleatórias, o algoritmo básico proposto não é efetivo.

Este método prevê também documentos rasgados que contenham mais de uma página. Em caso de papeis em tamanhos conhecidos, como A4 ou Letter, a detecção torna-se simples, pelo cálculo da área dos fragmentos.

Smet conclui o resultado de sua pesquisa como sendo eficiente para a reconstrução de documentos em papel em que os fragmentos tenham sido recuperados em pilha e em ordem. Porém Smet não expõe claramente qual a metodologia e quais as características analisadas nos fragmentos para encontrar os candidatos a parceiros assim como também não expõe se o resultado da metodologia apresenta a imagem digital remontada do documento ou apenas a sua rotulação.

3.3.4 RECONSTRUÇÃO DE DOCUMENTOS ELETRÔNICOS

Este trabalho tem a finalidade de propor um método de reconstrução de documentos originalmente criados em papel. Porém não podemos ignorar os esforços mundiais gastos para que se utilize, cada vez mais, documentos originalmente digitais, ou seja, documentos onde a origem é o meio eletrônico.

Normalmente, os documentos digitalizados são considerados meras cópias do documento original em papel, e em casos de comprovação de veracidade, o documento original em papel sempre deve ser apresentado. Porém, de acordo com o artigo 11 da lei 11.419/2006 que informatizou o

processo judicial, todos os documentos produzidos eletronicamente e juntados aos processos eletrônicos com garantia de origem e de seu signatário, na forma estabelecida na lei, serão considerados originais para todos os efeitos legais [FREITAS, 2008].

No Brasil, a partir da medida provisória nº 2200-2 de 24 de agosto de 2001, documentos digitais ou documentos eletrônicos já passam a ter veracidade jurídica, assim como aplicações de software habilitadas para operar com certificado digital emitido pelo ICP-Brasil (Infra-estrutura de Chaves Públicas Brasileiras) [GANDINI, 2002].

Dentro deste conceito, Kulesh e Memon (2003) propõem um método para a recuperação de documentos eletrônicos dispersos em mídias digitais. O propósito desse método é semelhante ao propósito de reconstrução de documentos em papel, porém visa conseguir reconstruir documentos que tenham sido apagados de suas mídias digitais, facilitando o trabalho do perito forense na busca por estes documentos.

O método de Kulesh e Memon demonstra que, devido à organização do sistema de arquivos utilizado, os arquivos são gravados em diversos blocos de dados, sendo esses blocos definidos pelo tipo de sistemas de arquivos. Alguns sistemas comerciais conseguem recuperar arquivos onde os blocos de dados estejam contíguos, sendo um ganho deste método, recuperar inclusive blocos de dados que estejam descontinuados.

Para a realização da determinação de candidatos a parceiros, simplesmente o método verifica se o trecho final do texto presente no fragmento A possui seqüência no fragmento B formando uma palavra existente

no dicionário. Se houver essa existência, o fragmento é dito candidato a parceiro.

Para a resolução de candidatos que sejam coincidentes, Kulesh e Memon propõe o uso de grafos, na tentativa de maximizar as probabilidades de reconstrução. Cada aresta do grafo representa uma ligação entre dois fragmentos, sendo o valor atribuído à nota, calculado através de uma verificação sintática e semântica da frase formada pelos encaixes. A figura 32 apresenta um grafo formado de 5 fragmentos.

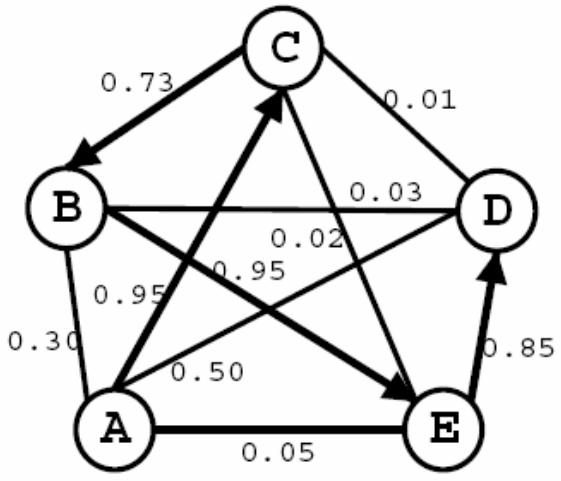


Figura 32 - Grafo completo com 5 fragmentos e o caminho hamiltoniano (ACBED) que maximiza os pesos dos vértices {0,95 + 0,73 + 0,95 + 0,85 = 3,48} [KULESH e MEMON, 2003].

Porém, Kulesh e Memon perceberam que utilizar palavras do dicionário para analisar candidatos a parceiros não seria tão efetivo, pois dependeria da língua dos documentos, e ainda documentos que não são baseados em textos, como códigos binários, etc, não seriam atingidos pelo método. Assim, alteraram o método e adicionaram o algoritmo para analisar o contexto através de PPM (previsão por combinação parcial). PPM é uma técnica estatística de análise de dados baseada na modelagem do contexto e previsão já conhecida e

largamente utilizada. Basicamente utiliza um conjunto de símbolos anteriores conhecidos e prediz estatisticamente qual o próximo símbolo que deve ocorrer.

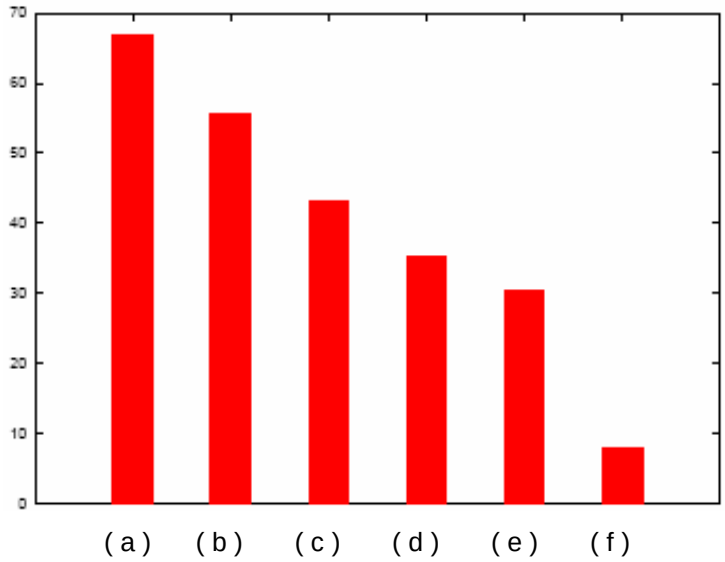


Figura 33 - Média de reconstrução dos fragmentos: (a) Arquivos de rastreio; (b) Arquivos de código fonte; (c) Arquivos de códigos binários; (d) Documentos de código binário; (e) Arquivos textos puros; (f) Arquivos criptografados ou comprimidos.

Kulesh e Memon concluem que os resultados são promissores, conforme resultados na figura 33, porém ainda é necessário estabelecer processos heurísticos efetivos para diversos outros tipos de arquivos na ânsia de construir um método independente de tipo de arquivo introduzindo meta informações ao processo.

3.4 CONCLUSÃO

Neste capítulo foram apresentadas algumas técnicas importantes para o desenvolvimento deste trabalho. Também foi apresentada a revisão bibliográfica dos temas, incluindo diversos trabalhos e resultados já alcançados na área de reconstrução de fragmentos, tanto de documentos quanto de cerâmicas, esta última está sendo considerada por apresentar semelhanças relevantes e aplicáveis também a documentos em papel.

Neste capitulo também abordamos casos de reconstrução de documentos originalmente eletrônicos. Tais documentos são reconstruídos em perícias forenses após a busca e apreensão de mídias digitais.

No capítulo 4, serão abordadas as metodologias para a concretização deste trabalho. Primeiramente será abordada a metodologia de reconstrução baseada na cadeia de códigos de Freeman [FREEMAN, 1974]. Esta metodologia não apresentou resultados satisfatórios no processo de reconstrução de documentos, porém o processo será apresentado. Na seqüência será abordada a metodologia baseada na técnica de aproximação poligonal [DOUGLAS & PEUCKER, 1973], programação dinâmica [BELLMAN, 1957] e o algoritmo de Prim [PRIM, 1957] para a resolução de árvores geradoras mínimas. Esta metodologia apresentou resultados significativos na reconstrução de documentos, incluindo a reconstrução visual do documento.

Capítulo 4

METODOLOGIA PROPOSTA

4.1 INTRODUÇÃO

Para a realização do processo de reconstrução dos documentos mutilados, duas abordagens distintas foram realizadas utilizando apenas as informações e características disponíveis na extração do contorno dos fragmentos.

A primeira abordagem para a reconstrução de documentos mutilados utiliza as cadeias de códigos de Freeman geradas pelos contornos dos fragmentos aplicados ao algoritmo de programação dinâmica. Esta metodologia não trouxe resultados promissores para a reconstrução de documentos e, por tanto, foi descartada para compor o resultado final deste trabalho. A abordagem e os problemas encontrados nesta metodologia estão expostos no item 4.3.

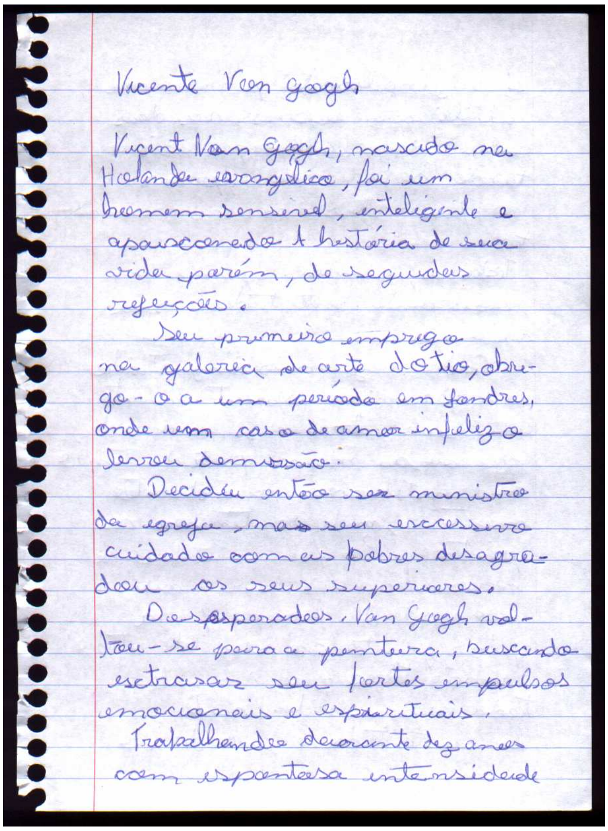
A segunda abordagem para a reconstrução utiliza as características geométricas extraídas das arestas e dos ângulos do contorno dos fragmentos conforme a técnica proposta por [SOLANA, 2005], sendo esta considerada como uma extensão de seu método. As características são aplicadas também à técnica de programação dinâmica. Esta técnica produziu resultados

promissores para a reconstrução de documentos mutilados. A abordagem dessa metodologia está exposta no item 4.4.

Este capítulo também apresenta a base de dados de imagens da PUCPR que foi utilizada para realizar os experimentos e a validação dos métodos propostos.

4.2 BASE DE DADOS DE IMAGENS PUCPR

A base de dados de imagens da PUCPR possui atualmente 855 imagens de fragmentos representativos de 100 documentos mutilados. As mutilações dos documentos constantes na base foram realizadas utilizando materiais cortantes como tesouras, réguas, estiletes ou simplesmente rasgados manualmente [SOLANA, 2005]. A base de dados possui documentos separados por tipos. Do total de documentos, 25% são manuscritos originados de trabalhos escolares, provas, exercícios, folhas de caderno e folhas avulsas, conforme figura 34. 25% são documentos textos tipografados como páginas de livros, páginas de documentos digitais, páginas impressas. 25% são documentos de texto contendo imagens conforme a figura 35. Os demais 25% são documentos variados.



Vicente Van gogh

Vicent Van Gogh, nascido na Holanda evangélico, foi um homem sensível, inteligente e apaixonado pela história de sua vida porém, de seguidas rejeições.

Seu primeiro emprego na galeria de arte do tio, obrigo-o a um período em Londres, onde um caso de amor infeliz o levou demissão.

Decidiu então ser ministro da igreja, mas seu excessivo cuidado com os pobres desagradou os seus superiores.

Desesperados, Van Gogh voltou-se para a pintura, buscando extravasar seu fortes impulsos emocionais e espirituais.

Trabalhando durante dez anos com espantosa intensidade

Figura 34 - Base de imagens PUCPR. Documento manuscrito.

SOPA DE LETRAS

série literatura brasileira

Carta de Pero Vaz de Caminha

Rir é o melhor remédio

"Dizem que ele é maravilhoso com crianças."

Figura 35 - Base de imagens PUCPR. Documentos textos com figuras.

4.2.1 AQUISIÇÃO E PRÉ-TRATAMENTO DE IMAGENS

As imagens constantes na base de dados foram digitalizadas utilizando *scanner* de mesa com o fundo preto. As imagens dos fragmentos foram geradas coloridas com 24 bits por pixel, resolução de 150 dpi no formato BMP, conforme figura 36 (a).

(a) (b)

Figura 36 - (a) Fragmento de documento conforme digitalização original da base de imagens; (b) Imagem convertida em níveis de cinza.

Para reduzir a complexidade, as imagens foram convertidas para escala de cinza com 256 tons, conforme figura 36 (b). Após a imagem convertida em escalas de cinza, a extração do fundo da imagem é realizada. Para a extração do fundo preto das imagens foram utilizadas duas etapas; A primeira etapa tem a função de eliminar todas as aglutinações de pixels pretos menores que nove pixels agrupados três a três. A segunda etapa tem a função de eliminar os demais ruídos ainda existentes nas bordas das imagens deixando apenas os pixels que possuem dois vizinhos caracterizando o contorno, conforme figura 37.

Após a remoção do fundo da imagem por completo, realiza-se a operação de extração do contorno na imagem, utilizando o algoritmo de Freeman [SOLANA, 2005].

Figura 37 - Fragmento de documento com o fundo eliminado.

Com a utilização do algoritmo de Freeman, é extraído o contorno externo da imagem contendo apenas um pixel na linha de borda, conforme a figura 38. A imagem contendo apenas a borda é salva e faz parte da base de dados da PUCPR.

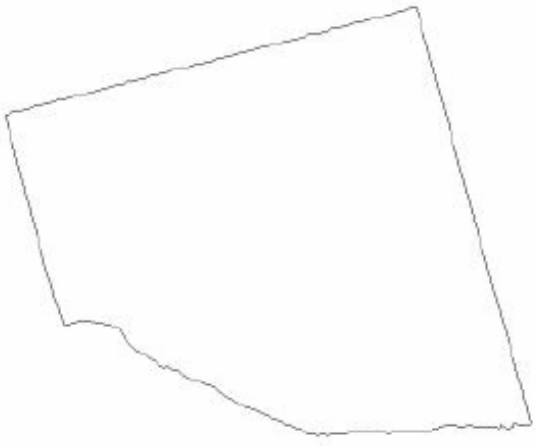


Figura 38 - Contorno do fragmento contendo apenas um pixel na borda.

4.3 METODOLOGIA BASEADA NA CADEIA DE CÓDIGOS DE FREEMAN

Devido ao tamanho, forma de aquisição e irregularidade dos fragmentos de documentos mutilados, existem diversos pontos a serem considerados para realizar a análise do contorno e a reconstrução.

Cadeias muito longas, poucos símbolos para realizar a representação – apenas oito símbolos na cadeia de códigos de Freeman –, ruídos nas bordas e diferenças na inclinação axial dos fragmentos, leva-se a não conseguir realizar a combinação dos fragmentos diretamente através das informações contidas em cada pixel que forma o contorno.

Para tanto, é necessário que se realize pré-processamentos nas cadeias antes de buscar os candidatos a parceiros dos fragmentos.

A metodologia utilizada para analisar e minimizar os problemas encontrados nesta abordagem, será dividida em quatro partes: a primeira parte apresenta os problemas encontrados para realizar a análise de borda dos fragmentos; a segunda parte demonstra os problemas da aquisição em diferentes inclinações axiais; a terceira etapa é demonstra a técnica para remover as cadeias longas com tendências retilíneas; a última parte apresenta a seqüência dos passos de pré-processamento do contorno dos fragmentos para então seguir o processamento de definição dos candidatos a parceiros.

4.3.1 IRREGULARIDADES NA BORDA

Na figura 39, está a representação do contorno de um fragmento demonstrando as irregularidades que ocorrem nas bordas. Estas irregularidades, ou ruídos de borda, comprometem a busca de candidatos a

parceiros. Apesar da representatividade da borda demonstrar uma tendência na formação do fragmento, a cadeia de pixels formada pela região possui elementos que não a representam. Quando esta cadeia é submetida ao algoritmo de LCS, causa diversas penalidades dificultando a análise e definição de candidatos a parceiros.

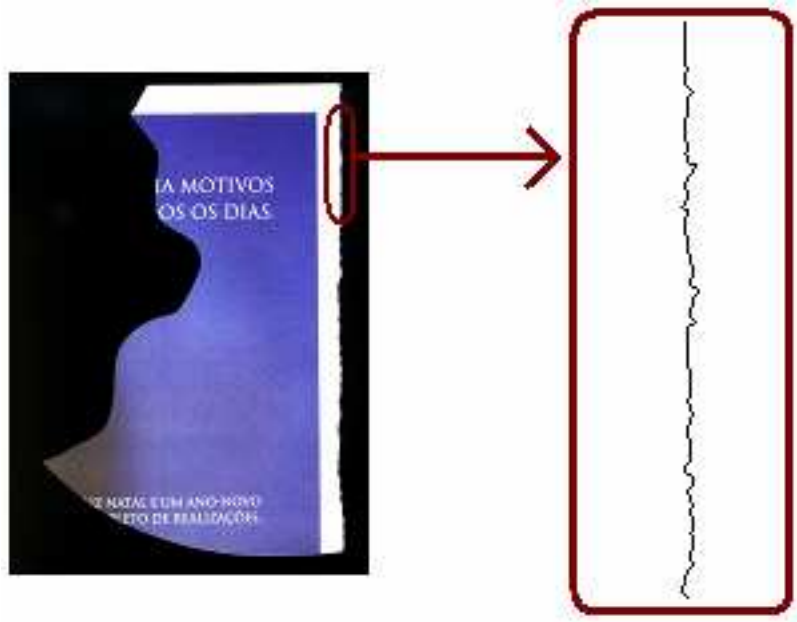


Figura 39 - Fragmento com irregularidades na borda.

Para minimizar os ruídos de borda dos fragmentos, duas técnicas foram avaliadas: cálculo de média móvel e técnica de reamostragem do contorno da imagem.

O cálculo de média móvel tem o objetivo de suavizar o contorno da imagem substituindo o pixel corrente pelo resultado da média das posições dos pixels anteriores e/ou adjacentes. Com os ruídos de borda suavizados, o processo de determinação de candidatos a parceiros seria facilitado, porém o cálculo de média móvel não altera a quantidade de pixels pertencentes ao contorno da imagem, causando distorções na imagem formada pela representação da nova cadeia de contornos encontrada devido ao alto nível de

ruídos apresentados pelas imagens. Devido a esta característica, a técnica de média móvel foi descartada do processo.

A técnica de reamostragem proposto por [GONZALEZ & WOODS, 2000], também possui o objetivo de retirar os ruídos e irregularidades encontradas nos contornos dos fragmentos. Esse processo é realizado submetendo a cadeia de pixels do contorno a uma reamostragem utilizando uma grade de pixels com espaçamento maior.

A figura 40 (a) mostra uma cadeia original contendo problemas com ruídos e irregularidades. Essa figura sendo submetida a uma reamostragem utilizando uma máscara com 5 pixels de espaçamento conforme figura 40 (b), resultará no contorno representado na figura 40 (c) e 40 (d). A técnica de reamostragem, além de realizar a suavização do contorno, diminui a quantidade de pixels na proporção da máscara de pixel utilizada para realizar a reamostragem.

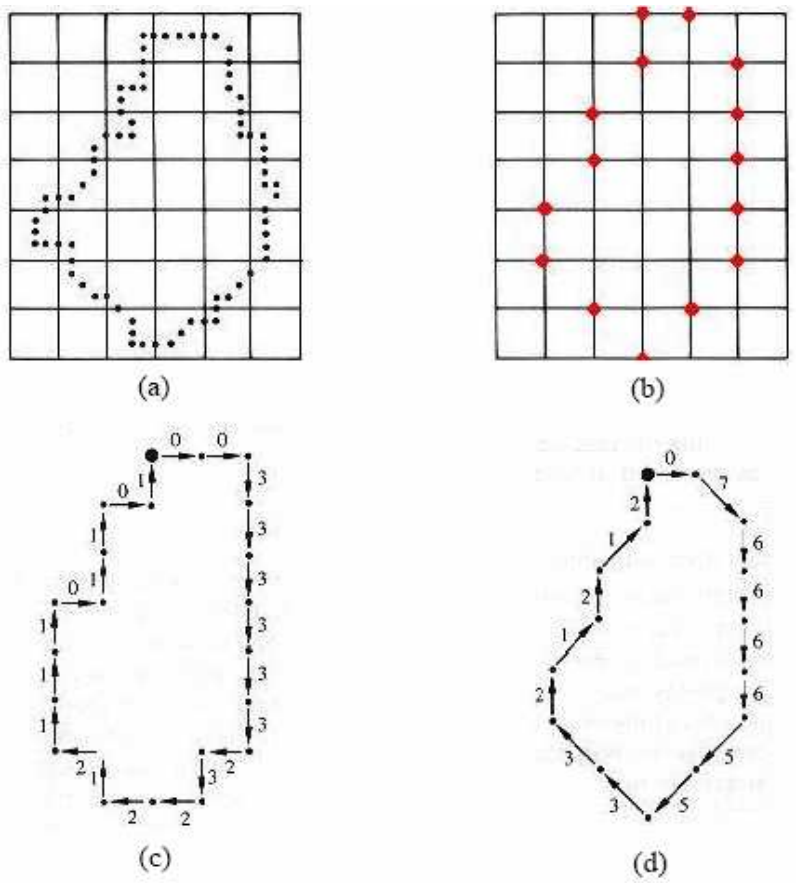


Figura 40 - (a) Grade e contorno; (b) Reamostragem; (c) Código de cadeia direcional de 4 segmentos; (d) Código de cadeia direcional de 8 segmentos [GONZALEZ & WOODS, 2000].

Esta técnica de suavização do contorno e diminuição da imagem foi avaliada como sendo positiva para o processo de reconstrução, sendo as novas cadeias geradas pela reamostragem submetidas aos passos seguintes do processo de análise de candidatos a parceiros.

4.3.2 INCLINAÇÃO AXIAL

Na figura 13, página 37, os fragmentos expostos como exemplo possuem a mesma inclinação axial, o que propicia uma análise adequada dos contornos. Assim, os valores das direções da cadeia de códigos de Freeman são equivalentes, o que possibilita a comparação e a verificação da região de combinação entre os fragmentos.

Porém, no processo de aquisição das imagens dos fragmentos de documentos mutilados, esse fato não é garantido. As imagens dos fragmentos na base de dados da PUCPR são obtidas através de scanner de mesa, porém através de qualquer meio de aquisição de imagens em formato digital, não há a certeza da exatidão do ângulo de inclinação da imagem a ser coletada.

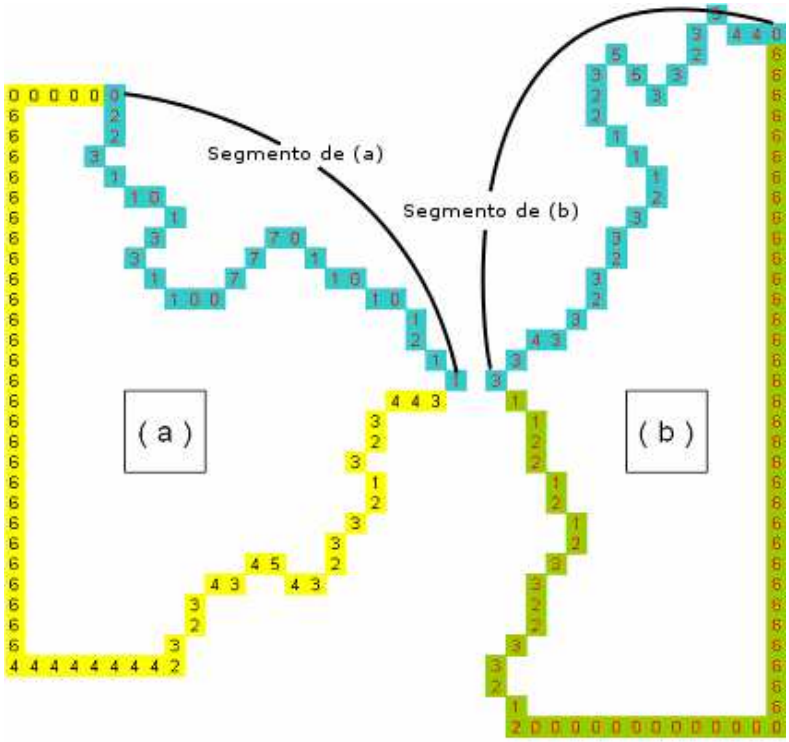


Figura 41 - Fragmentos com região de combinação adquiridos em defasagem axial.

Considerando que a aquisição é um processo manual e, portanto, não há exatidão plena no posicionamento dos fragmentos, inclusive pelo desconhecimento do formato original do documento, a inclinação axial dos fragmentos torna-se um problema a ser analisado.

A figura 41 ilustra o fragmento (a) e o fragmento (b) da figura 6. A região destacada nos dois fragmentos forma a região de combinação. Fragmento (a) sentido horário e fragmento (b) sentido anti-horário. Porém o fragmento (b) foi adquirido defasado em 90º em relação ao fragmento (a). Observando agora as

duas cadeias formadas pela região de combinação, percebe-se que não há mais equivalência. Submetendo essas novas duas cadeias ao algoritmo de LCS, conforme figura 42, o resultado é diferente do resultado obtido na seção 2.3, figura 11.

(b) Sentido Anti-Horário		(a) Sentido Horário																									
		0	2	2	3	1	1	0	1	3	3	1	1	0	0	7	7	7	0	1	1	0	1	0	1	2	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	0
0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	0	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	0	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3
3	0	1	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	0	1	2	2	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	0	1	2	2	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	0	1	2	2	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
3	0	1	2	2	3	3	3	3	3	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
2	0	1	2	3	3	3	3	3	3	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	6	6	6
2	0	1	2	3	3	3	3	3	3	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	6	6	6
1	0	1	2	3	3	4	4	4	4	4	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	7	7
1	0	1	2	3	3	4	5	5	5	5	5	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	8
1	0	1	2	3	3	4	5	5	6	6	6	6	7	7	7	7	7	7	7	8	8	8	8	8	8	8	8	8
2	0	1	2	3	3	4	5	5	6	6	6	6	7	7	7	7	7	7	7	8	8	8	8	8	8	9	9	9
3	0	1	2	3	4	4	5	5	6	7	7	7	7	7	7	7	7	7	7	8	8	8	8	8	8	9	9	9
3	0	1	2	3	4	4	5	5	6	7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	9	9	9
2	0	1	2	3	4	4	5	5	6	7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	9	9	9
3	0	1	2	3	4	4	5	5	6	7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	9	9	9
2	0	1	2	3	4	4	5	5	6	7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	9	9	9
3	0	1	2	3	4	4	5	5	6	7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	9	9	9
3	0	1	2	3	4	4	5	5	6	7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	9	9	9
4	0	1	2	3	4	4	5	5	6	7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	9	9	9
3	0	1	2	3	4	4	5	5	6	7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	9	9	9
3	0	1	2	3	4	4	5	5	6	7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	9	9	9

Figura 42 - Matriz LCS dos fragmentos (a) e (b) com diferenças de inclinação axial.

Na figura 43 estão os resultados das cadeias de *backtracking*. O símbolo "_" representa a penalidade.

Backtracking	023331112
Backtracking Horizontal com penalidade.	02_3___3311_____1____2__
Backtracking vertical com penalidade.	0___233__3__1112_____

Figura 43 - Resultado da cadeia de *backtracking*.

Verificando os resultados obtidos na tentativa de combinação entre os fragmentos com as cadeias de combinação em diferentes inclinações axiais, percebe-se que a inclinação axial é um fator que compromete totalmente a análise de combinação de bordas. Principalmente na tentativa de combinação utilizando diretamente as cadeias de códigos de Freeman originais.

A partir desta análise, percebe-se que não é possível realizar a combinação de cadeias de direções se as mesmas não estiverem representadas em cadeias de direções compatíveis entre os fragmentos.

Para tornar as cadeias compatíveis e passíveis de comparação, todas as cadeias que formam os contornos dos fragmentos, serão submetidas a um pré-tratamento. A função do pré-tratamento é avaliar a cadeia do contorno e construir uma nova cadeia que represente quantitativamente as alterações de direção, pixel a pixel, que cada elemento da cadeia possui. A essa nova cadeia formada dar-se-á o nome de cadeia de complemento.

A formação da cadeia de complemento será realizada analisando todos os valores de cada cadeia original, substituindo o elemento corrente analisado conforme a equação 4.

$$E_i = \begin{cases} E_{i+1} - E_i & se \quad E_{i+1} >= E_i \\ E_i = \{E_{i+1} - E_i + 8 & se \quad E_{i+1} < E_i \end{cases} \tag{4}$$

Sendo:

- E_i o elemento encontrado no índice i da cadeia de código de Freeman.

A figura 44 demonstra o resultado da *cadeia de complemento* para os fragmentos (a) e (b). As cadeias (ac) e (bc) são as respectivas cadeias de complemento. As cadeias de complemento não representam o contorno dos fragmentos.

(a)	0	2	2	3	1	1	0	1	3	3	1	1	0	0	7	7	7	0	1	1	0	1	0	1	2	1	1
(ac)	2	0	1	6	0	7	1	2	0	6	0	7	0	7	0	0	1	1	0	7	1	7	1	1	7	0	7

(b)	0	4	4	5	3	2	3	3	5	5	3	2	2	1	1	1	2	3	3	2	3	2	3	3	4	3	3
(bc)	4	0	1	6	7	1	0	2	0	6	7	0	7	0	0	1	1	0	7	1	7	1	0	1	7	0	5

Figura 44 - Cadeia de complemento resultante para os fragmentos expostos na Figura 21.

Observando as duas cadeias novas formadas, *cadeias de complemento*, percebe-se que há uma grande equivalência entre elas. Submetendo essas cadeias ao algoritmo de LCS, desprezando o primeiro e o último valor das seqüências, teremos o resultado conforme figura 45.

Observando a figura 45, o resultado da combinação entre as duas *cadeias de complemento,* é extremamente semelhante ao resultado obtido entre as cadeias originais que possuem angulação axial compatível, demonstrados na seção 2.3 figura 11. A única distorção apresentada ocorre no primeiro e no último valor da cadeia, devido à falta de elementos para realizar o cálculo do complemento. Em análises de cadeias completas, não apenas uma

porção do fragmento como exemplificado, este fato só ocorrerá no início das cadeias, uma vez que os valores dos elementos finais serão conhecidos.

(a) Sentido Horário

(b) Sentido Anti-Horário

		2	0	1	6	0	7	1	2	0	6	0	7	0	7	0	0	1	1	0	7	1	7	1	1	7	0	7
	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
4	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	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	0	0	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
6	0	0	1	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
7	0	0	1	2	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
1	0	0	1	2	3	3	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
0	0	0	1	2	3	4	4	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
2	0	1	1	2	3	4	4	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
0	0	1	2	2	3	4	4	5	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
6	0	1	2	2	3	4	4	5	6	7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
7	0	1	2	2	3	4	5	5	6	7	8	8	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
0	0	1	2	2	3	4	5	5	6	7	8	9	9	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
7	0	1	2	2	3	4	5	5	6	7	8	9	10	10	11	11	11	11	11	11	11	11	11	11	11	11	11	11
0	0	1	2	2	3	4	5	5	6	7	8	9	10	11	11	12	12	12	12	12	12	12	12	12	12	12	12	12
0	0	1	2	2	3	4	5	5	6	7	8	9	10	11	11	12	13	13	13	13	13	13	13	13	13	13	13	13
1	0	1	2	3	3	4	5	6	6	7	8	9	10	11	11	12	13	14	14	14	14	14	14	14	14	14	14	14
1	0	1	2	3	3	4	5	6	6	7	8	9	10	11	11	12	13	14	15	15	15	15	15	15	15	15	15	15
0	0	1	2	3	3	4	5	6	6	7	8	9	10	11	11	12	13	14	15	16	16	16	16	16	16	16	16	16
7	0	1	2	3	3	4	5	6	6	7	8	9	10	11	12	12	13	14	15	16	17	17	17	17	17	17	17	17
1	0	1	2	3	3	4	5	6	6	7	8	9	10	11	12	12	13	14	15	16	17	18	18	18	18	18	18	18
7	0	1	2	3	3	4	5	6	6	7	8	9	10	11	12	12	13	14	15	16	17	18	19	19	19	19	19	19
1	0	1	2	3	3	4	5	6	6	7	8	9	10	11	12	12	13	14	15	16	17	18	19	20	20	20	20	20
0	0	1	2	3	3	4	5	6	6	7	8	9	10	11	12	13	13	14	15	16	17	18	19	20	20	20	21	21
1	0	1	2	3	3	4	5	6	6	7	8	9	10	11	12	13	13	14	15	16	17	18	19	20	21	21	21	21
7	0	1	2	3	3	4	5	6	6	7	8	9	10	11	12	13	13	14	15	16	17	18	19	20	21	22	22	22
0	0	1	2	3	3	4	5	6	6	7	8	9	10	11	12	13	14	14	15	16	17	18	19	20	21	22	23	23
5	0	1	2	3	3	4	5	6	6	7	8	9	10	11	12	13	14	14	15	16	17	18	19	20	21	22	23	23

Figura 45 - Matriz LCS para as cadeias de complemento.

A figura 46 demonstra a matriz de *backtracking* gerada para as cadeias de complemento*.* A semelhança entre a matriz de *backtracking* das cadeias originais apresentada na seção 2.3 figura 12, são evidentes. Os pontos de início e fim das cadeias que serão desconsiderados. Pontos em diagonal destacam a combinação entre as cadeias, pontos em seqüência horizontal e vertical destacam as penalidades.

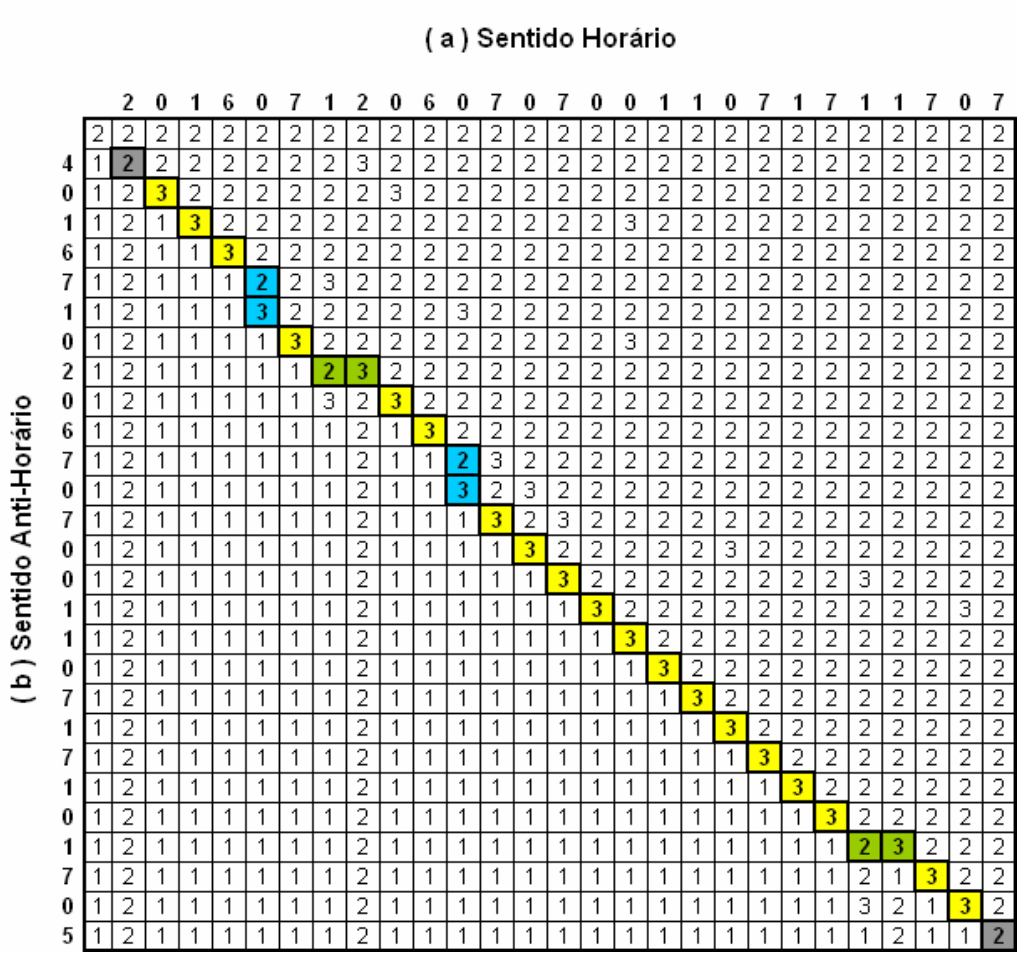


Figura 46 - Matriz *backtracking* para as cadeias de complemento.

Com a formação das cadeias de complementos, os problemas com inclinação axial dos fragmentos serão minimizados, possibilitando a aplicação do algoritmo de LCS para análise dos candidatos a parceiros diretamente sobre as cadeias de códigos de Freeman.

4.3.3 CADEIAS LONGAS E CADEIAS COM TENDÊNCIAS RETILÍNEAS

Mesmo depois de submetidas às cadeias aos processos de suavização das bordas proposto na subseção 4.3.1, as cadeias ainda podem ser extensas o suficiente a ponto de dificultar a análise de candidatos a parceiros.

Um dos problemas que causam grandes anomalias no processo é a presença de cadeias de bordas externas, cadeias dos contornos dos

documentos que, na sua maioria, apresentam regiões com tendências retilíneas. Essas cadeias são encontradas em todos os fragmentos que possuem partes do seu contorno que compõem o contorno total do documento mutilado.

A figura 47 mostra dois fragmentos de um mesmo documento mutilado constante na base de dados da PUCPR.

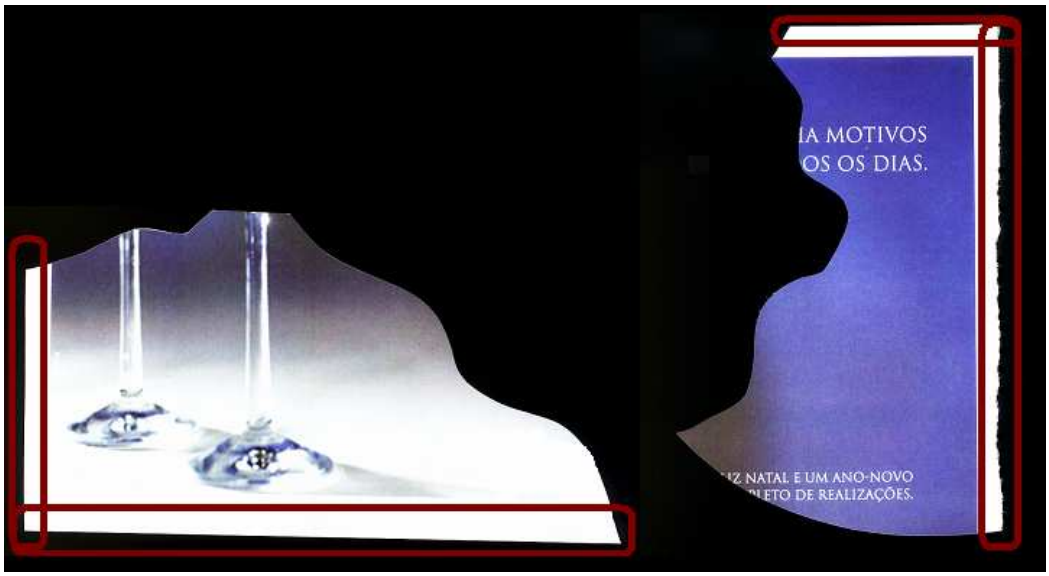


Figura 47 - Fragmentos de um documento mutilado.

Os dois fragmentos não possuem regiões de combinação, porém o contorno externo de ambas as imagens, destacados, quando submetidos ao algoritmo de LCS, resultarão em combinação, ou seja, falso positivo. Esse fato ocorre devido às cadeias destacadas de ambos os fragmentos possuírem subcadeias semelhantes.

Para que os contornos dos fragmentos analisados não influenciem negativamente no processo de análise de candidatos a parceiros, é necessário que estas cadeias sejam detectadas e removidas do processo.

Para diminuir à extensão das cadeias que serão analisadas, o contorno será dividido em segmentos menores, e a análise será realizada em todos os segmentos. A segmentação será realizada a partir de pontos identificados através de algoritmos de aproximação poligonal. Os pontos obtidos através de aproximação poligonal têm a tendência de mostrar os pontos onde ocorrem às mudanças de direções significativas nos contornos dos fragmentos.

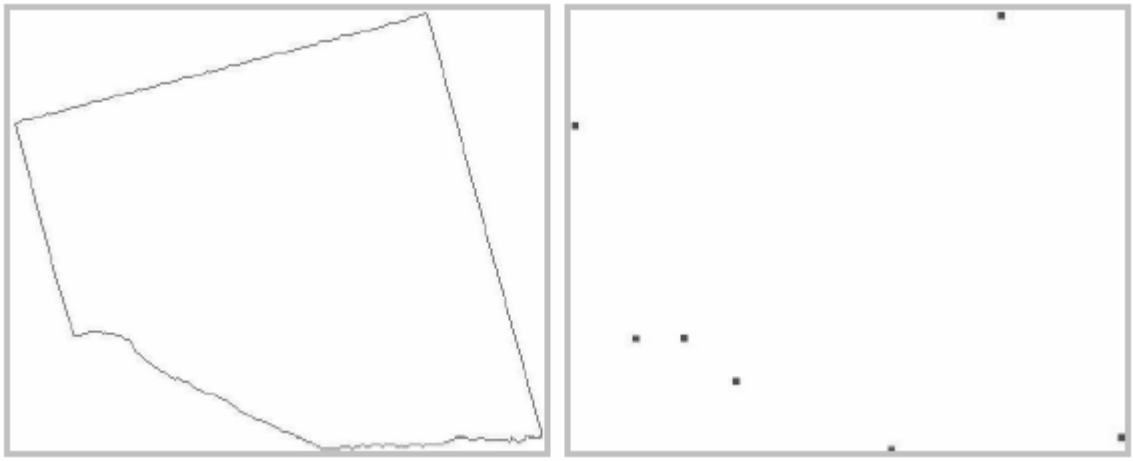


Figura 48 - (b) Resultado do algoritmo de aproximação poligonal Douglas e Peucker aplicado em (a) [SOLANA, 2005].

Na figura 48, temos um exemplo de pontos destacados através de aproximação poligonal. Os pontos identificados determinam o início e o fim de cada segmento. Os pontos destacados formam sete segmentos da cadeia original, com quantidades menores de elementos, facilitando o processo de análise do contorno.

Para a realização do descarte de contornos dos segmentos de fragmentos que apresentam tendências retilíneas, após o processo de segmentação, o segmento é submetido a um processo de verificação que analisa se o segmento deve ou não ser descartado.

A primeira tentativa para realizar o descarte de segmentos retilíneos foi através do cálculo da distância entre os pontos da extremidade do segmento em confronto ao número de pixel existente no segmento, ou seja, o comprimento da aresta do segmento e a quantidade de pixels existentes que compõem a aresta.

O cálculo da distância entre dois pontos é realizado através da equação 5.

$$d(P,Q) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (5)$$

Sendo:

- (x_1, y_1) é o ponto inicial do segmento;
- (x_2, y_2) é o ponto final do segmento;

Porém fazendo este confronto entre a distância entre os pontos da extremidade e a quantidade de pontos se aplica apenas em segmentos onde os pixels estejam alinhados na posição vertical ou horizontal e não se aplica em situações onde os pixels do segmento estejam na posição diagonal. Quando os pixels estão na posição vertical ou horizontal, o valor da distância é calculado de acordo com a equação 6.

$$d = N_p * L \quad (6)$$

Sendo:

- *d* = Distância;
- *Np* é o número de pontos do segmento.
- *L* é o comprimento do pixel. Utilizado valor 1 por padrão.

Para o cálculo para os segmentos em que os pixels estejam em sua totalidade na diagonal é utilizada a equação 7.

$$d = N_p * L * \sqrt{2} \tag{7}$$

A diferença nas fórmulas aparece devido ao comprimento da aresta de uma figura quadrada é diferente do comprimento da diagonal dessa mesma figura que é dada pelo comprimento da aresta multiplicado pela raiz quadrada de 2, conforme figura 49.

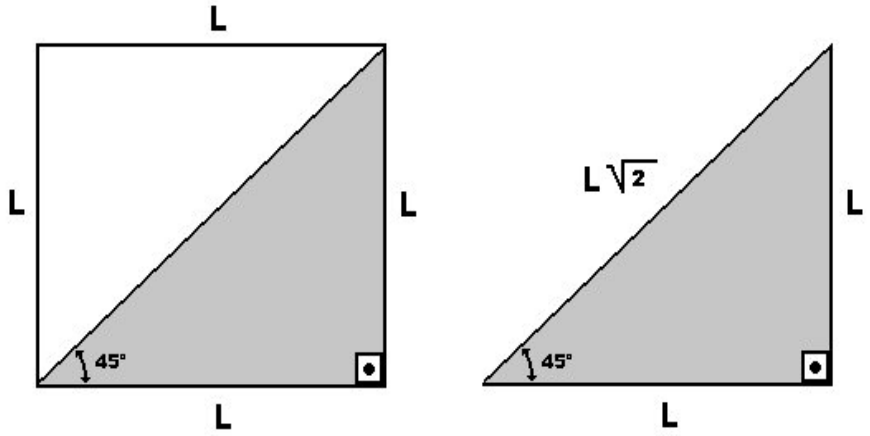


Figura 49 – Diferença do cálculo de distância: Horizontal / Diagonal.

Utilizando o valor da lateral do pixel como sendo de valor 1, o valor da sua diagonal será 1√2, que pode ser considerado aproximadamente 1,41, ou seja a diferença do valor calculado da distância para os valores em pixel, mesmo em valores corretos, podem variar em 41% o seu comprimento. Dessa forma essa técnica não produz resultados conclusivos se o segmento deve ou não ser retirado na análise.

A segunda tentativa trouxe resultados positivos quanto ao descarte de segmentos retilíneos. A técnica analisa a variância da distância dos pontos que compõem o segmento até a reta formada pelos pontos da extremidade. Caso a variância das distâncias esteja abaixo de um limiar especificado, o segmento é considerado retilíneo e o segmento então é descartado do processo de reconstrução.

Para realizar esta verificação, foram utilizadas a fórmula geral da reta e a fórmula da distância entre ponto e reta.

Através dos pontos de extremidade do segmento, foi encontrada a fórmula geral da reta, a qual passa pelos 2 pontos, conforme equação 8.

$$ax + by + c = 0 \tag{8}$$

Sendo o valor de "*a"* substituído pelo valor da coordenada *y* do primeiro ponto menos a coordenada *y* do segundo ponto, ou seja, *a = (Yb – Ya),* o valor de *b* substituído pelo valor da coordenada *x* do ponto *Q* menos a coordenada *x* do ponto *P*, ou seja, *b = (Xb – Xa)* e o valor de *c = Xa*Yb – Xb*Ya.*

Para cada pixel que compõe o segmento, foi calculada a distância entre este pixel e a reta formada pelos pontos da extremidade de acordo com a fórmula de distância entre ponto, conforme a equação 9.

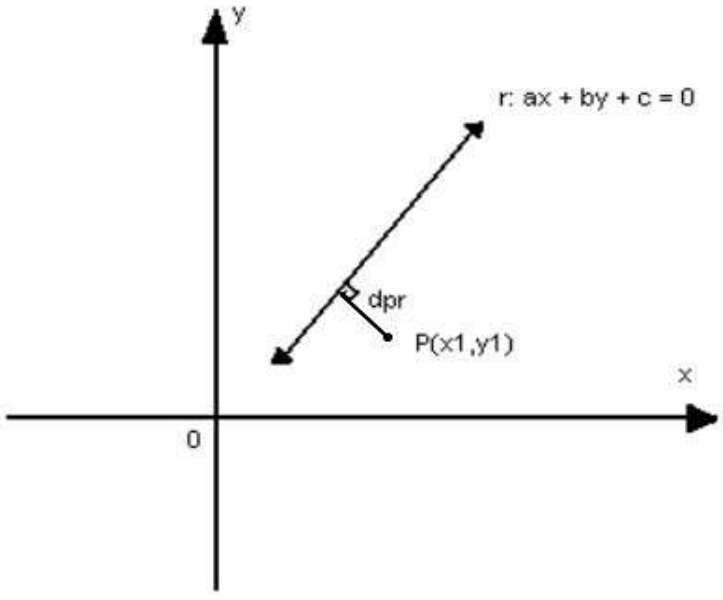


Figura 50 - Distância entre ponto e reta.

$$d_{pr} = \left| a_{x1} + b_{y1} + c \right| / \sqrt{(a^2 + b^2)} \quad (9)$$

Todos os valores são armazenados e por fim se calcula a média aritmética das distâncias. A média aritmética é utilizada para calcular a variância das distâncias. O cálculo utilizado para calcular a variância é dado pela equação 10:

$$V = \frac{1}{Nv} * \sum (d_a - média)^2 \quad (10)$$

Sendo:

- *V* é a variância;
- N_v é o número de elementos;
- d_a é o valor atual da distância;
- Média é a média aritmética das distâncias.

Utilizando a medida da variância, foi possível verificar em experimentos que quando o valor da variância das distâncias permaneceu abaixo de 0,4, o segmento tinha tendências retilíneas, logo é retirado do processo de reconstrução. Com a variância acima de 0,4, o segmento não é considerado retilíneo e não é retirado do processo de reconstrução. Em segmentos curtos, com poucas quantidades de pixels é passível de ocorrer o descarte do fragmento, mesmo não sendo fragmentos externos do documento, o que pode prejudicar o processo de reconstrução.

4.3.4 Análise de combinação do contorno

Para a realização da análise e combinação do contorno dos fragmentos, e apontar candidatos a parceiros no processo de reconstrução, as cadeias de

códigos de Freeman de todos os fragmentos foram submetidas aos processos descritos anteriormente.

Dessa forma, o método para realizar a verificação dos candidatos a parceiros nos fragmentos dos documentos mutilados, utilizando a análise do contorno, é composto de dez etapas:

a) Aquisição e leitura das cadeias de códigos de Freeman representativas dos contornos dos fragmentos

Esse procedimento tem a finalidade de realizar a leitura das cadeias dos pixels que formam o contorno dos fragmentos das imagens digitalizadas.

As cadeias de códigos de Freeman são lidas em arquivos gerados a partir dos fragmentos, segundo [SOLANA, 2005].

b) Rotulação e adequação das cadeias de códigos de Freeman em sentido horário e anti-horário do contorno.

Conforme exposto na subseção 2.2.1, as cadeias de códigos de Freeman só realizam combinação *(matching)* realizando o confronto entre cadeias em sentidos opostos de aquisição, cadeias horárias e anti-horárias. Sendo assim após a leitura das cadeias de Freeman no sentido horário, é realizado um processo de conversão desta cadeia para representar o fragmento também no sentido anti-horário.

c) Tratamento de irregularidades nas bordas.

Conforme técnicas expostas na subseção 4.3.1, os fragmentos são submetidos ao processo de reamostragem da imagem, através da técnica proposta por [GONZALEZ & WOODS, 2000].

A técnica de reamostragem é utilizada para diminuir o tamanho das imagens e conseguir uma cadeia menor de pixels para representar a borda dos fragmentos. Na figura 51, temos o exemplo de um fragmento de documento mutilado onde as imagens enumeradas de 1 a 10 sofreram o processo de reduções. Os números em cada imagem demonstram a largura da grade em pixels utilizada pelo algoritmo para realizar a reamostragem da imagem. O valor da grade em pixels representa exatamente a redução da imagem em comparação com a imagem original. Quando temos uma grade de 2 pixels, a imagem é reduzida à metade, quando a grade é de 3 pixels, a imagem é reduzida à um terço, etc.

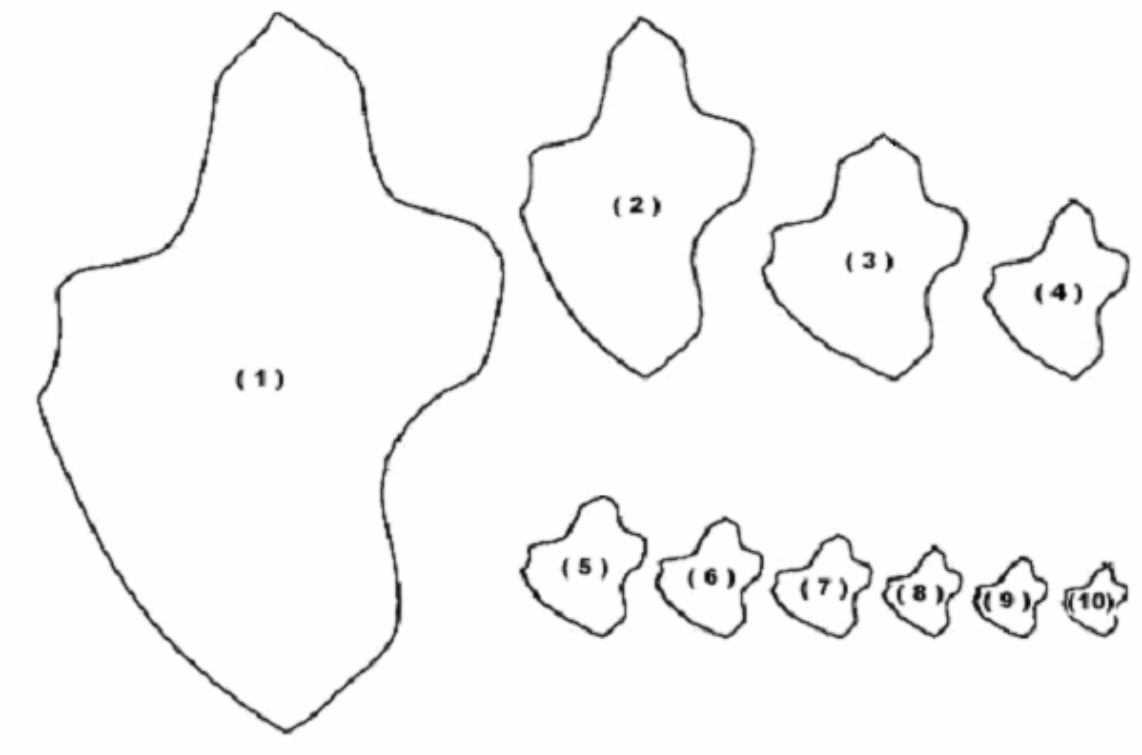
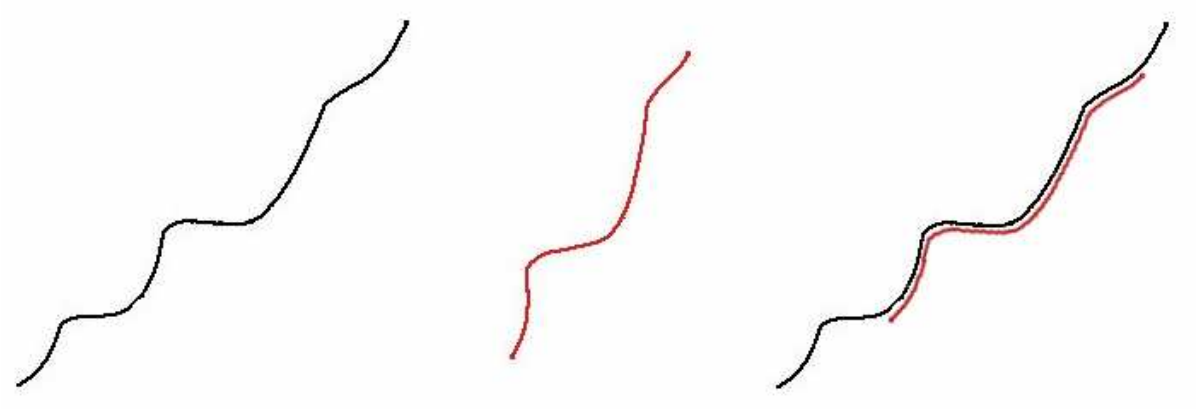


Figura 51 - Fragmento 2 do documento 1 da base de dados de documentos mutilados da PUCPR.

d) Análise e remoção das cadeias de fragmentos que possuem bordas externas com tendências retilíneas conforme técnica exposta na subseção 4.3.3.
e) Correção da inclinação axial das cadeias. Criação da cadeia de complemento. Conforme técnica exposta no item 4.3.2.
f) Segmentação das cadeias gerando cadeias menores. Conforme técnicas expostas na subseção 4.3.3.
g) Submeter as cadeias segmentadas ao algoritmo de LCS e armazenar características de combinação (*matching*), falsos positivos e penalidades resultantes do processamento.
h) Analisar características e resultados obtidos entre todas as cadeias segmentadas e definir os candidatos a parceiros para cada segmento.
i) Apontar os segmentos considerados vizinhos no processo de remontagem dos documentos mutilados.
j) Análise e estatística de resultados, positivos e falsos positivos, no processo de reconstrução.

4.3.5 PROBLEMAS IDENTIFICADOS

A figura 52 ilustra dois segmentos de contorno que possuem encaixes entre si extraídos através da cadeia de códigos de Freeman. Esses segmentos são submetidos ao algoritmo de LCS para analisar a combinação existente entre eles.



(a) (b) (c)

Figura 52 - Segmento de contorno: (a) Primeiro segmento de contorno do fragmento 1 do documento 1; (b) Último segmento de contorno do fragmento 2 do documento 1; (c) Encaixe manual entre os segmentos.

Na figura 52 (c), o encaixe entre os segmentos foi realizado manualmente para ilustrar que existe o encaixe, e que visualmente é possível identificá-lo. A tabela 5 mostra as cadeias de códigos representativos dos pontos que compõem o contorno dos fragmentos expostos na figura 52.

Tabela 5 - Cadeia de código de Freeman e cadeia de Complemento representativo dos segmentos expostos na figura 52. Na cadeia de complemento, está destacada a seqüência de combinação.

Segmento	Cadeia de códigos de Freeman	Cadeia de Complemento
(a)	00110110111111121112121212121221 22122121101010001000000000000000 10000001001010111021111121121212 12121221221222122221222212221111 01010001000071070000000000600106 00000000000000010160110101111121 0 21111211211211212112121212121212 12122121212121221212212212211101 1 10110101101010101101101111111112 11211212121212112121	01071071000***00017001717171717*** ***1071071071***70717170017000000000 00000170000017017171007270000170 17171717171071071007100071000710 07000717170017000727710000000006 20176200000000000001715210717100 00177270001701701701717017171717 17171717107171717171071710717107 00710071071710717171710710710000 00001701701717171717017171 7
(b)	24211212121212212212212221222222221 22222222242222222222222112101110 10100100000100001000100001000100 01010101111211121212121212212212 21222122221222122221222212222122 22212222222212222222212211211121 1 11111111111121121 12	26701717171710710710071000000071 00000000260000000000007017710071 71701700001700017001700017001700 1717171***00017001717171717107 1*** ***071071***00710007100710007100071 07100007100000071000000710701700 17000000000000017017017

Apesar de haver combinação, conforme a tabela 5, entre as cadeias que compõem as arestas, a quantidade de combinação é pequena, apenas 27 pontos. Estas mesmas arestas submetidas à análise de outras arestas dos mesmos fragmentos ou de outros fragmentos acarretam em quantidades de combinações semelhantes, conforme tabela 6.

Tabela 6 - Exemplos de resultado no processo de combinação de segmentos em pontos.

Fragmentos / Segmentos	**Fragmento 1, Segmento 0**	**Resultado**
Fragmento 2, segmento 5	27	Positivo
Fragmento 3, segmento 0	32	Falso-positivo
Fragmento 3, segmento 1	46	Falso-positivo
Fragmento 3, segmento 3	27	Falso-positivo
Fragmento 3, segmento 4	24	Falso-positivo
Fragmento 3, segmento 5	64	Falso-positivo
Fragmento 4, segmento 0	52	Falso-positivo
Fragmento 4, segmento 1	40	Falso-positivo
Fragmento 4, segmento 2	45	Falso-positivo
Fragmento 4, segmento 3	30	Falso-positivo
Fragmento 5, segmento 1	49	Falso-positivo
Fragmento 6, segmento 0	33	Falso-positivo
Fragmento 6, segmento 4	29	Falso-positivo
	Pontos coincidentes	**Resultado esperado**

Assim como mostra a tabela 6, os resultados obtidos entre as combinações dos segmentos dos fragmentos para todos os contornos de fragmentos resultaram em valores de combinação inconclusivos. Vários segmentos obtiveram valores de combinação aproximados, tanto para segmentos que possuem combinação quanto para segmentos que não possuem. A proximidade entre os valores de combinação e não combinação resultante do processo de programação dinâmica mostra que a utilização desse método não é efetiva.

Quatro principais problemas foram identificados durante o processo de combinação, fazendo com que a combinação através das cadeias não fosse possível. Os problemas principais identificados são: inclinação axial e a

distorção na cadeia de complemento; segmentação dos contornos em pontos não controlados; não combinação mesmo entre seqüências que possuem combinação. Na seqüência esses problemas serão detalhados.

4.3.5.1 DISTORÇÃO EXISTENTE NA CADEIA DE COMPLEMENTO E INCLINAÇÃO AXIAL

Conforme exposto no item 4.3.2, a técnica da cadeia de complemento tem a finalidade de eliminar a problemática da inclinação axial diferente entre os fragmentos. Porém, a cadeia de complemento não corrige os problemas por completo, uma vez que, para realizar a rotação de uma imagem os comprimentos das arestas devem ser recalculados e redesenhados. A figura 53 mostra a distorção que ocorre nas arestas utilizando a cadeia de complemento.

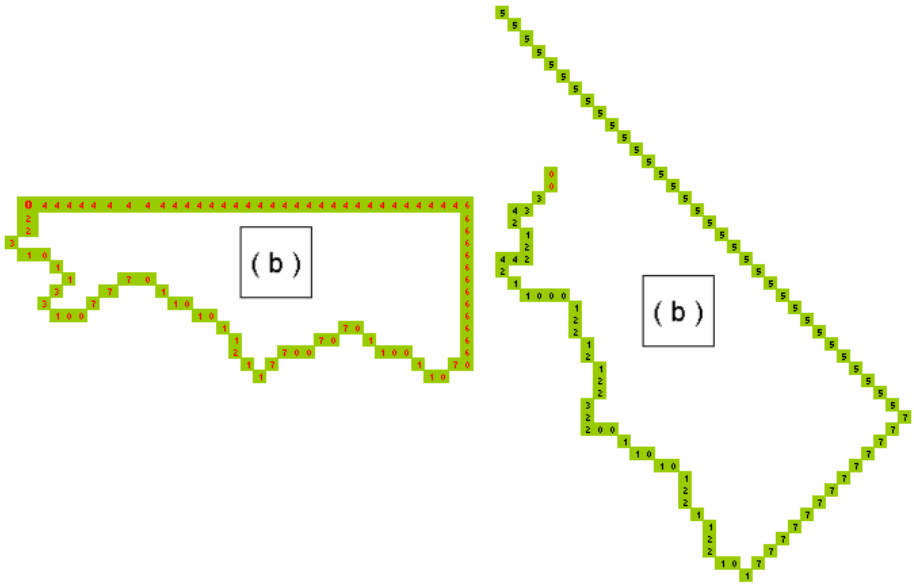


Figura 53 - Exemplo do fragmento (b) da figura 5 submetido à técnica da cadeia de complemento.

A figura 53 mostra que, apesar da cadeia de complemento transformar as cadeias originais de Freeman em valores comparáveis, o problema da inclinação axial ainda permanece. O problema da inclinação axial deve ser

resolvido realizando as transformações geométricas de rotação onde todos os pontos são recalculados baseados no comprimento das arestas.

Com o problema da inclinação axial não resolvido, o processo de combinação dos pontos dos contornos fica comprometido.

4.3.5.2 SEGMENTAÇÃO DO CONTORNO EM PONTOS NÃO CONTROLADOS

O processo de segmentação dos contornos utiliza, como vértices para a separação, pontos que foram encontrados através do processo de aproximação poligonal. Não há como definir os pontos de início e fim dos fragmentos que possuem encaixes com outros fragmentos precisamente, esse é objetivo de todo o processo de reconstrução. Dessa forma, consegue-se a segmentação em pontos onde ocorrem mudanças significativas de direção nos contornos, porém não nos pontos de encaixes.

Como ocorre na figura 52, verifica-se que as seqüências possuem combinação, mas o tamanho das seqüências são extremamente diferentes, pois no mesmo segmento, pode-se ter porções de encaixes em diferentes arestas de demais fragmentos, resultando em um processo falho.

4.3.5.3 NÃO COMBINAÇÃO DE ARESTAS SEMELHANTES

Apesar das dificuldades citadas acima serem importantes, o maior problema identificado no processo de combinação ponto a ponto, está na própria não combinação de seqüências que possuem combinações claras entre si.

Esse problema é facilmente verificado no exemplo da figura 52 e no resultado da combinação exposta na tabela 5. Visivelmente os contornos

possuem encaixes, inclusive o segmento do fragmento 2 possui combinação em toda a sua extensão com o fragmento 1. Porém, se observarmos os resultados de encaixes encontrados na tabela 5, apenas 27 pontos entre as cadeias tiveram combinação, sendo que as cadeias possuem respectivamente 309 e 211 pontos, refletindo uma combinação de apenas 9% e 13%, onde deveria ser 68% e 100% respectivamente.

4.3.5.4 CONCLUSÃO

Neste tópico foi exposto o processo de reconstrução baseado em segmentação dos contornos dos fragmentos formando cadeias menores de códigos de Freeman. São esses segmentos que formam as características do contorno do fragmento e que são submetidas ao processo de programação dinâmica.

Esse processo demonstrou não ser eficiente para a reconstrução digital de documentos, apesar dos ensaios iniciais apontarem ao contrário. Não foi possível utilizar a seqüência de pontos do contorno como características dos fragmentos a serem reconhecidas.

No item 4.4 será descrita a metodologia baseada em extração de características geométricas dos contornos dos fragmentos em que os resultados foram promissores.

4.4 METODOLOGIA BASEADA NAS CARACTERÍSTICAS GEOMÉTRICAS DO CONTORNO

4.4.1 EXTRAÇÃO DE CARACTERÍSTICAS

Após o processo de aquisição das imagens dos fragmentos, realiza-se o processo de extração de características dos contornos. A extração de características é realizada semelhantemente ao proposto por [SOLANA, 2005].

Inicialmente o contorno dos fragmentos é submetido ao processo de aproximação poligonal. Os pontos obtidos através de aproximação poligonal têm a tendência de mostrar onde ocorrem as mudanças de direções significativas nos contornos dos fragmentos. Ver exemplo na figura 54.

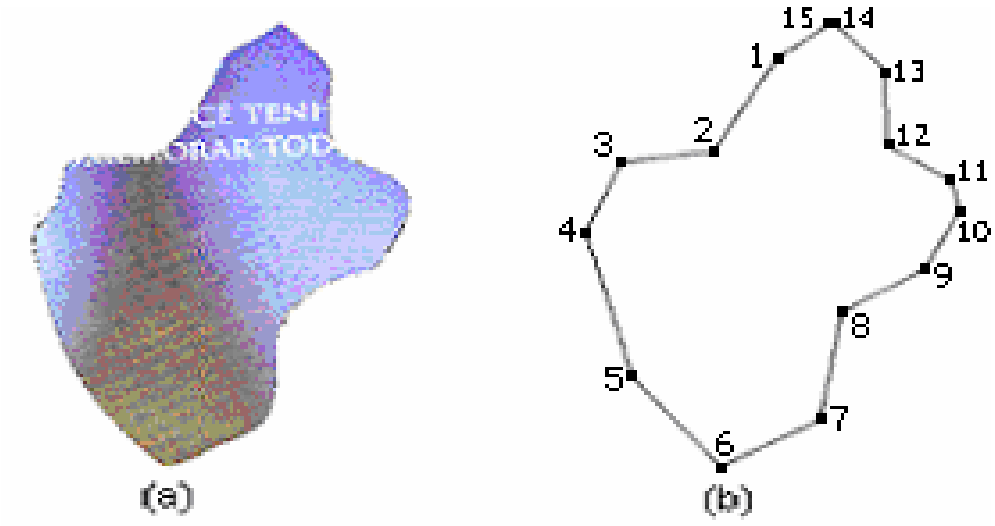


Figura 54 - (a) Fragmento original; (b) Contorno do fragmento submetido ao processo de aproximação poligonal.

O algoritmo de aproximação poligonal utilizado no processo de reconstrução é o algoritmo proposto por [DOUGLAS & PEUCKER, 1973]. Este algoritmo já foi alvo de estudos avançados e é reconhecido por melhor preservar as características do polígono original [SOLANA, 2005].

As características coletadas do contorno da imagem são compostas pelo trio formado por dois segmentos de reta consecutivos e o ângulo externo à imagem formado por eles. O cálculo do comprimento das arestas é calculado seguindo o teorema de Pitágoras, e o ângulo formado pelas arestas é calculado através da Lei dos Cossenos, equação 11, onde *a* é a distância entre o ponto B e C, *b é* a distância entre os pontos A e C e *c* é a distancia entre os pontos A e B.

$$\cos\alpha = \frac{a^2 - b^2 - c^2}{2 * b * c} \tag{11}$$

Conforme a figura 55, o ângulo formado pelos pontos *1, 2 e 3,* correspondente ao índice 2 da tabela 7, é um ângulo convexo ao polígono, resultante na área A1, diferentemente do ângulo formado pelos pontos *12, 13 e 14,* resultante na área A2, correspondente ao índice 13 da tabela 7, que se configura como um ângulo côncavo.

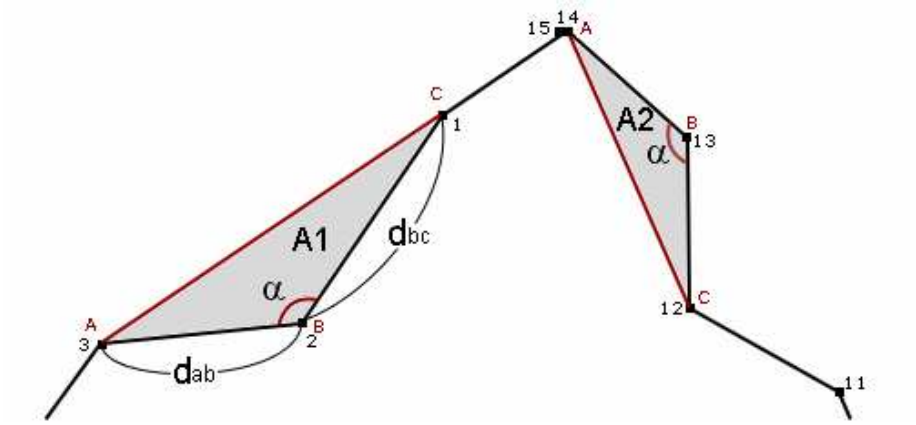


Figura 55 - Vértices de extração de características.

Para verificar se o ângulo é côncavo ou convexo ao polígono, calcula-se a interseção da área formada pelos pontos em análise, ou seja, a área A1 e A2, etc, com a área formada pelo polígono do fragmento. Se a interseção for 100%,

o ângulo é interno, nesse caso utiliza-se o ângulo suplementar ao ângulo encontrado como característica. Se a interseção for 0%, o ângulo é externo, nesse caso utiliza-se o próprio ângulo como característica.

A tabela 6 mostra o resultado da extração de características da borda para cada ponto do polígono exposto na figura 54 (b).

Após a extração das características de cada fragmento, é iniciado o processo de combinação (*matching*) entre as características dos fragmentos. A combinação de características é o processo mais importante para a realização da reconstrução dos documentos. É nessa fase que são gerados os valores de encaixes entre os fragmentos e onde são descartadas várias características e junções.

Tabela 7 - Características extraídas do fragmento da figura 54 (b).

Vértices	Ângulo	Distância d_{ab}	Distância d_{bc}
1	200º	80	160
2	130º	160	110
3	240º	110	110
4	220º	110	220
5	200º	220	160
6	263º	160	130
7	260º	130	160
8	130º	160	120
9	210º	120	93
10	220º	93	43
11	220º	43	90
12	130º	90	110
13	220º	110	89
14	0º	89	1,4
15	190º	1,4	80

Porém, na execução do algoritmo de aproximação poligonal, utiliza-se um parâmetro de erro, que tem a função de determinar o quão próximo e o

quão fiel à imagem gerada resultante será em relação à imagem original dos fragmentos.

Durante os experimentos foi constatado que para todos os documentos, o valor utilizado como erro para o algoritmo de aproximação poligonal deve ser customizável, ou seja, para cada documento existe um valor de erro que melhor representa as tendências do contorno resultando em uma melhor reconstrução final. Diferentemente do proposto por [SOLANA, 2005], não utilizamos valores fixos de erro na aproximação poligonal, mas sim se verifica no processo qual o melhor valor a ser utilizado individualmente para cada documento analisado.

A detecção do valor de erro na aproximação poligonal é feita analisando o número de combinações finais conseguidas pelo processo e pelo valor final da nota empregada à reconstrução para cada documento. Nos experimentos a faixa de valores de erro utilizada foi de 2 a 30.

O processo de combinação de características é realizado em duas etapas principais: Combinação de características utilizando algoritmo de programação dinâmica e posteriormente o processo de descarte e seleção de combinações.

4.4.2 Programação dinâmica e combinação de características

Para realizar a combinação de características, será utilizado um algoritmo LCS, algoritmo já explorado na seção 2.3. Este algoritmo tem o objetivo de apontar a maior subseqüência comum encontrada comparando a seqüência de duas cadeias. Esta técnica será utilizada para encontrar as

cadeias de seqüências comuns entre as características extraídas do contorno dos fragmentos mutilados, com o intuito de se avaliar as condições para considerar os fragmentos como candidatos a parceiros no processo de reconstrução.

Após a criação da matriz de LCS inicial, os restantes dos valores são completados seguindo a definição exposta no item 2.3, equação 1. Porém com os parâmetros diferenciados, definidos abaixo:

- *S* = 1, se o valor de *i* na seqüência (a) for igual ao valor de *j* na seqüência (b). Ou seja, combinação (match).
- *S* = -100, se o valor de *i* na seqüência (a) for diferente ao valor de *j* na seqüência (b). Ou seja, sem combinação (*mismatch*).
- *P* = -100. Valor de penalidade.

Nos experimentos realizados foram utilizados valores 1 para combinação, -100 para não combinação e valor -100 para penalidade. O valor alto de penalidade e não combinação causará uma quebra na seqüência nos pontos onde não ocorre a combinação *(matching)* das características. Dessa foram conseguimos o melhor alinhamento local entre as cadeias de características, sendo estas contínuas não permitindo que penalidades componham a seqüência final de reconhecimento.

Essa mesma técnica é aplicada para realizar a correspondência entre as seqüências de características encontradas em cada fragmento que compõe o documento original. Apenas altera-se o valor de *S* que passa a ser:

- *S* = 1, se o valor de cada segmento de reta da característica de *i* tiver o mesmo comprimento dos segmentos de reta da característica de *j* com tolerância de 10 pontos e a soma dos ângulos formados pelos segmentos for 360º com tolerância de 5º. Os valores de tolerância foram obtidos experimentalmente.
- *S* = -100, para qualquer outro caso.

Cada par de fragmentos que é submetido ao algoritmo de LCS resulta em um conjunto de relações. Cada relação entre os fragmentos é composta por um índice, pela nota representativa da relação e pelos índices das características de cada fragmento onde ocorreu a combinação. Como o objetivo do algoritmo de LCS é analisar diversos batimentos seqüencialmente, a nota para cada relação será a nota total de todas as relações que forem combinadas seqüencialmente. A nota é calculada de acordo com a equação 12.

$$\sum (A_i + B_j) \tag{12}$$

Sendo:

- A_i a soma dos comprimentos das arestas que compõem o vértice da característica extraída da primeira lista de características de fragmento submetido ao processo de LCS.
- B_j a soma dos comprimentos das arestas que compõem o vértice da característica extraída da segunda lista de características de fragmento submetido ao processo de LCS.

Dessa forma, caso 2 fragmentos tenham 5 pontos coincidentes, serão 6 arestas em cada fragmento totalizando 12 arestas. A nota total será a soma do comprimento dessas 12 arestas.

Esse procedimento é realizado para todos os fragmentos existentes para remontar o documento. A quantidade de execuções é dada pela análise combinatória de acordo com a equação 13:

$$C\,{}^{F}_{2} = \frac{F!}{2!*(F-2)!} \tag{13}$$

Sendo F o número de fragmentos do documento a ser remontado.

4.4.3 ANÁLISE E DESCARTE DE COMBINAÇÕES

Após o procedimento de combinação, todas as relações encontradas entre os fragmentos são adicionadas a uma mesma lista de relações. A cada inserção de uma nova relação na tabela, é iniciada uma análise para verificar a existência de outras relações que tenham características coincidentes, ou seja, o mesmo vértice encaixado em mais de um fragmento. Nesse caso, o índice da relação que possui a menor nota é adicionado a uma lista interna de índices a serem removidos na relação que possui a maior nota. A tabela 8 mostra um exemplo da lista de relações encontradas para um documento da base de dados.

Tabela 8 - Relações de características para um documento da base de dados.

Índice	Fragmento A	Fragmento B	Índice Caract. A	Índice Caract. B	Nota	Índice para Remoção
1	1	2	8	0	1438	
1	1	2	7	1	1438	
1	1	2	6	2	1438	
2	1	3	3	2	407	
3	1	4	4	0	1104	
3	1	4	3	1	1104	2, 7, 10
3	1	4	2	2	1104	10
6	2	3	13	2	1684	2, 7, 10
6	2	3	12	3	1684	
6	2	3	11	4	1684	
6	2	3	10	5	1684	
6	2	3	9	6	1684	
7	2	4	13	1	388	
9	2	6	7	6	436	
10	3	4	3	1	800	7
10	3	4	2	2	800	2
13	4	5	11	0	1220	
13	4	5	10	1	1220	
14	4	6	7	0	871	
14	4	6	6	1	871	
15	5	6	4	8	1250	
15	5	6	3	9	1250	

A coluna índice da tabela 8 refere-se às comparações realizadas entre as cadeias de características de cada fragmento, de acordo com a equação 13. Nesse exemplo temos um documento com seis fragmentos totalizando 15 comparações, ou seja, o fragmento 1 com o 2, o 1 com o 3, 1 com 4, etc. Portanto o índice varia de 1 à 15.

Na coluna fragmento A e fragmento B, temos os índices dos fragmentos que estão sendo comparadas, na primeira linha, por exemplo, temos o fragmento rotulado como 1 sendo comparado com o fragmento rotulado como 2. Nessa comparação, a característica de índice 8 do fragmento 1 teve correspondência com a característica de índice 0 do fragmento 2.

Nas próximas duas linhas da tabela, também há mais duas combinações entre as características do fragmento 1 com o fragmento 2, totalizando 3 combinações. Estas três combinações possuem a nota de 1438, conforme a equação 12.

Os índices que não constam na tabela, como, por exemplo, os índices 4 e 5, são casos em que não houveram nenhuma correspondência entre nenhuma características dos vetores analisados.

Na seqüência, analisando a tabela 8, os índices de relação 2, 7 e 10 serão removidos, pois apresentam pontos em coincidência com outras relações que possuem nota mais alta, sendo assim tratados como falsos positivos.

No caso do índice 2 da tabela, foram analisados os fragmentos rotulados como 1 e 3. Houve uma correspondência entre a característica de índice 3 do fragmento 1 com a característica de índice 2 do fragmento 3, com uma nota de 407.

Porém na comparação de índice 3, o mesmo fragmento 1 foi comparado com o fragmento 4, e a mesma característica de índice 3 do fragmento 1 houve correspondência com a característica de índice 1 do fragmento 4, com uma nota de 1104 que é maior do que combinação anterior de 407. Sendo assim a combinação encontrada de índice 2, com nota de 407 é considerada como falso-positiva e por esse motivo ela é retirada da tabela 8.

As relações restantes formam um grafo representativo dos encaixes entre os fragmentos do documento mutilado, representado na figura 56.

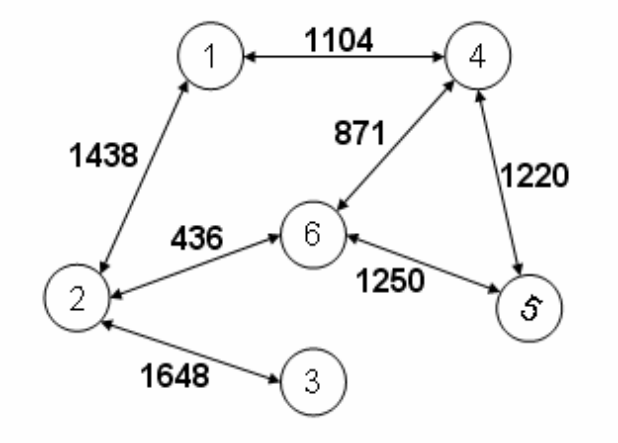


Figura 56 - Grafo de fragmentos parceiros.

Após realizar o processo de combinação de características e encontrar o grafo de encaixes dos fragmentos já é possível verificar a taxa de remontagem dos documentos.

Porém a análise baseada apenas na rotulação, desconsiderando a composição gráfica da remontagem, acaba por descartar um grande aliado contra falsos positivos. Com a análise gráfica da remontagem é possível verificar áreas de remontagem que acabam em oclusões, fato que não pode ocorrer no processo de reconstrução.

Para realizar a reconstrução gráfica dos documentos mutilados a partir do grafo de relacionamentos, é necessário realizar três operações principais: sequenciamento de reconstrução, translação e a rotação dos fragmentos.

4.4.4 SEQUENCIAMENTO DE RECONSTRUÇÃO

Devido às diversas ligações cíclicas existentes entre os nós de um grafo, é necessário primeiramente realizar um processo para analisar qual fragmento será transladado e rotado em relação aos seus candidatos a parceiros, garantindo que um fragmento seja rotado e transladado apenas uma vez durante o processo.

A técnica utilizada para conseguir o sequenciamento dos fragmentos a serem reconstruídos graficamente é baseada na transformação do grafo de candidatos a parceiros em uma árvore geradora mínima. Uma árvore geradora mínima, na teoria dos grafos, é um grafo onde cada vértice é visitado apenas uma vez, não havendo ciclos entre os nós.

Para realizar a transformação do grafo em uma árvore geradora mínima e o sequenciamento da reconstrução, é utilizado o algoritmo de Prim, descrito na seção 2.5, com algumas alterações.

- Ao invés de escolher um vértice inicial aleatoriamente, sempre iniciamos a árvore resultante com um dos vértices que possui a maior nota referente ao processo de combinação.
- Na seqüência, as arestas que são adicionadas à árvore resultante são as arestas que possuem o maior peso, diferente do algoritmo original que escolhe o menor peso. Escolher o maior peso significa escolher a

seqüência de encaixe dos fragmentos de acordo com a maior probabilidade de encaixes corretos.

- Antes de adicionar os fragmentos na árvore resultante, é realizado o processo de rotação e translação do fragmento que está sendo incluído para o ponto de remontagem. Após este processo, é verificado se existe sobreposição da área do polígono formado entre os fragmentos já existentes na árvore de retorno e o polígono formado pelo novo fragmento. Caso exista uma sobreposição de área superior ao limiar de 3%, valor determinado experimentalmente, significa que as imagens não podem ser encaixadas. Nesse caso, o fragmento é removido do processo de reconstrução. Esse passo é de extrema importância para retirar falsos positivos ainda remanescentes do processo de rotulação. A figura 57 mostra um exemplo de encaixe entre fragmentos, porém descartada por haver oclusão entre os fragmentos.

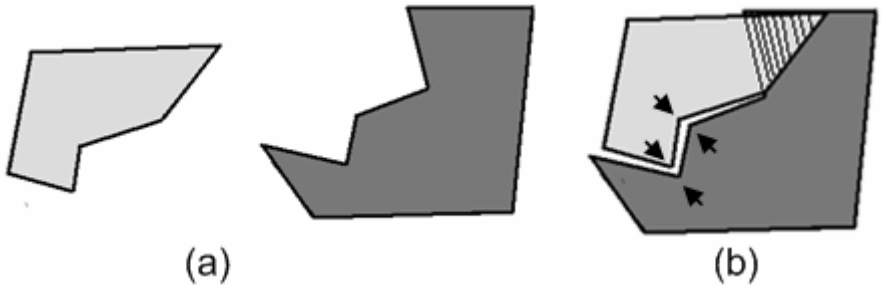


Figura 57 - (a) Fragmentos candidatos; (b) Fragmento encaixado, porém com região em sobreposição.

Na figura 58 está o resultado do grafo exposto na figura 56 após a execução do algoritmo de Prim. A seqüência de nós representada na árvore será a seqüência utilizada para a reconstrução do documento mutilado. As setas definem a orientação de qual fragmento deve rotar e transladar em relação ao seu candidato a parceiro, porém servem apenas como orientação

no momento da reconstrução visual. Caso um vértice aleatório inicial seja escolhido, não haverá problemas. Devido à criação da árvore geradora mínima, o caminho a ser seguido no grafo será sempre sem ciclos ou repetições de nós.

Após a criação da árvore de Prim, várias relações são excluídas, algumas por serem consideradas falso-positivas, outras por realizarem ciclos no processo de reconstrução e também serão removidas as relações com notas menores que possuírem mais de um encaixe com o mesmo parceiro. Para este caso, as relações de índice número 9, 14 e 2 foram retiradas da lista de relações dispostas na tabela 8, formando o grafo da figura 58.

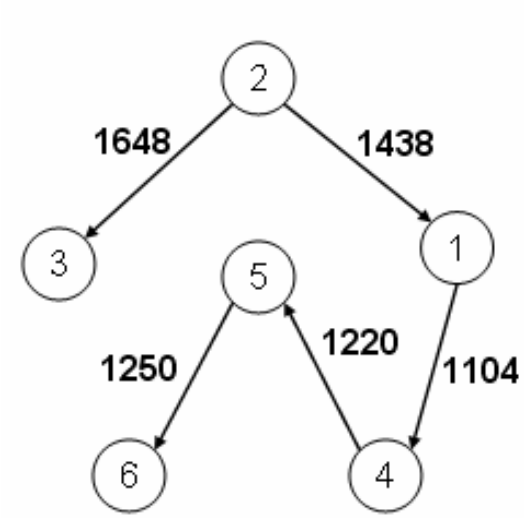


Figura 58 - Resultado do algoritmo de Prim aplicado ao grafo da figura 56.

Após o sequenciamento dos fragmentos, inicia-se o processo de rotação e translação para realizar a reconstrução gráfica do documento.

4.4.5 ROTAÇÃO E TRANSLAÇÃO DOS FRAGMENTOS

O último processo realizado para a reconstrução do documento é a rotação e a translação dos fragmentos para a posição correta de remontagem. Esse procedimento faz-se necessário devido ao processo de aquisição das

imagens dos fragmentos ser realizado de forma aleatória e com inclinações axiais diferentes.

Após a obtenção da árvore gerada pelo algoritmo de Prim, a tabela de relações possui apenas as relações que serão utilizadas para a realização dos encaixes. Através das relações sabe-se em qual parte da borda os fragmentos se encaixam, uma vez que na relação está contida a característica de combinação, ou seja, a seqüência dos pontos de combinação.

Com esses dados, caso exista mais de um ponto seqüencial de combinação, escolhe-se por definição o segundo ponto de combinação. Não há um motivo específico para a escolha do segundo ponto, caso seja escolhido qualquer outro ponto encontrado na combinação, o resultado deve ser semelhante. Em trabalhos futuros pode-se realizar um processamento melhor no momento da escolha dos pontos e dos valores para rotação e translação com o objetivo de alcançar o melhor encaixe global.

Na seqüência é realizada a translação do mesmo ponto de encaixe do fragmento B em relação ao fragmento A. Qualquer ponto em um plano pode ser transladado simplesmente adicionando um valor inteiro a cada uma de suas coordenadas. Assim os pontos do fragmento B são transladados conforme a equação 14.

$$X_{Fbf} = X_{Fbi} + (X_{Fai} - X_{Fbi})$$
$$Y_{Fbf} = Y_{Fbi} + (Y_{Fai} - Y_{Fbi}) \quad (14)$$

Sendo:

- X_{Fbf} e Y_{Fbf} as coordenadas do ponto final de translação do fragmento B.
- X_{Fbi} e Y_{Fbi} as coordenadas do ponto inicial de translação do fragmento B.
- X_{Fai} e Y_{Fai} as coordenadas do ponto do fragmento A.

Após o processo de translação, inicia-se o processo de rotação do fragmento. Para realizar a rotação do fragmento, inicialmente precisamos encontrar o ângulo correto de rotação.

Para encontrar o ângulo de rotação, novamente é aplicada a Lei dos Cossenos, equação 11, utilizando como os vértices do triângulo o ponto de combinação dos dois fragmentos, ponto *P2*, o ponto adjacente ao da combinação para o fragmento A, ponto *P1*, e o ponto anterior ao ponto de combinação para o fragmento B, ponto *P3*, depois da translação, conforme a figura 59.

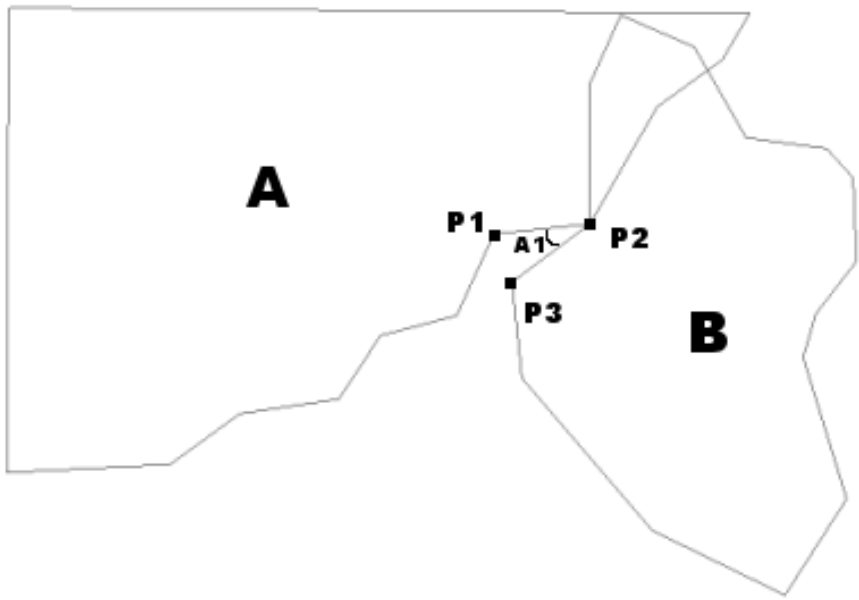


Figura 59 - Fragmentos parceiros transladados no ponto de encaixe; P2 Ponto de encaixe escolhido; P1 Ponto adjacente do fragmento A; P3 Ponto anterior do fragmento B.

Depois de encontrado o ângulo de defasagem entre os fragmentos, todos os pontos do fragmento B são rotados de acordo com a equação de rotação de pontos, equação 15.

$$X_f = X_i * \cos(\alpha) + Y_i * sen(\alpha)$$
$$Y_f = Y_i * \cos(\alpha) + X_i * sen(\alpha) \quad (15)$$

O processo de reconstrução do documento termina após todos os fragmentos do documento terem sido rotados e encaixados.

Todas as informações de translação e rotação aplicadas aos fragmentos são armazenadas e posteriormente aplicadas aos fragmentos originais digitalizados, no intuito de realizar a remontagem digital do documento propriamente dito.

A figura 60 demonstra o esquema geral da metodologia proposta para a reconstrução de documentos mutilados.

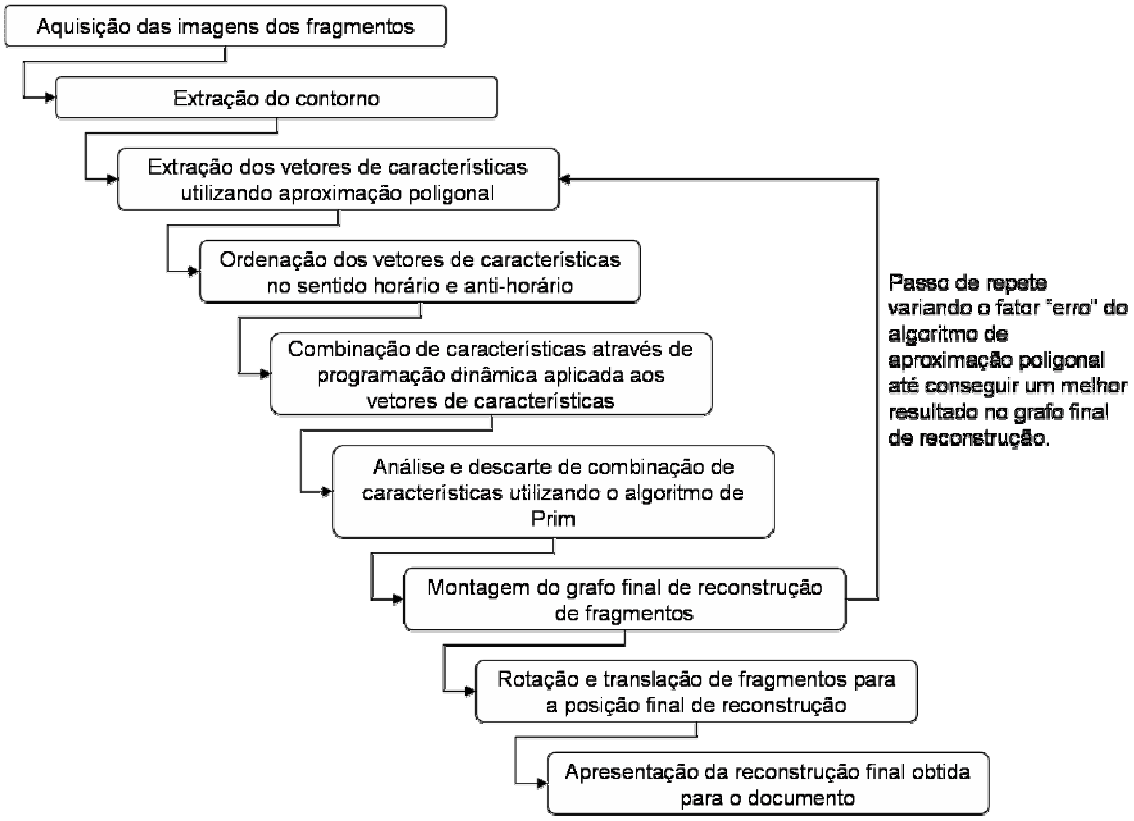


Figura 60 – Esquema geral da metodologia proposta.

4.4.6 PROBLEMAS IDENTIFICADOS

Durante os experimentos, foram identificados alguns problemas e algumas situações onde o algoritmo proposto não conseguiu atuar de forma efetiva.

O algoritmo atual necessita que sejam encontrados arestas e vértices em comum entre os fragmentos para realizar as junções necessárias. Porém em mutilações, cortes, em que as arestas tenham tendências circulares grandes, o algoritmo de aproximação poligonal acaba por encontrar pontos que não são correspondentes entre os fragmentos.

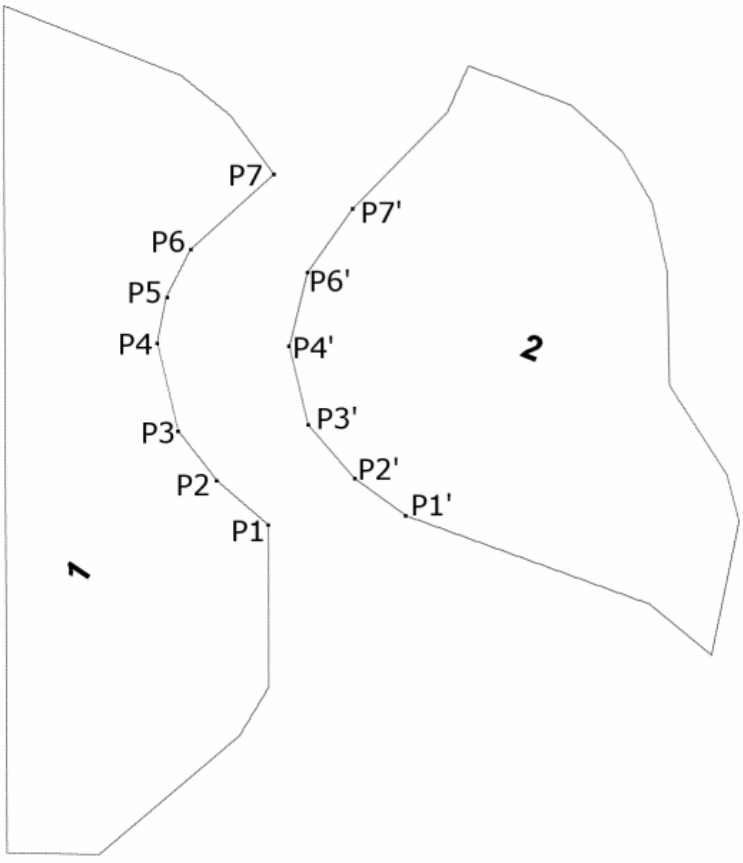


Figura 61 - Ponto de junção entre fragmentos com contornos curvilíneos.

Na figura 61, o fragmento 1 encaixa com o fragmento 2 nos pontos P indicados. Mesmo utilizando o mesmo valor de erro no processo de

aproximação poligonal, as arestas e ângulos não possuem os mesmos valores incluindo as tolerâncias, inclusive o ponto P5 do fragmento 1 não possui correspondente no fragmento 2. Este é um exemplo de problemas encontrados em documentos que possuem contornos curvilíneos bem definidos aos quais o algoritmo de aproximação poligonal não atua de forma igual.

Outra limitação no processo de análise de candidatos a parceiros realizado pelo algoritmo atual, é a comparação de dois fragmentos a cada momento, não realizando o processo de verificação de candidatos após 2 ou mais fragmentos já serem considerados parceiros.

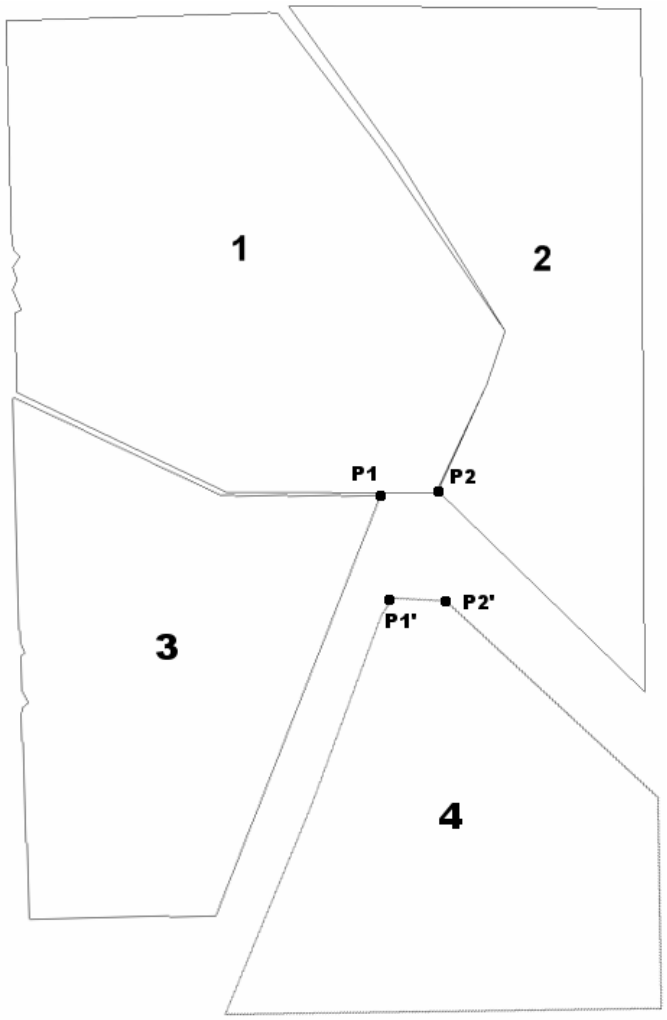
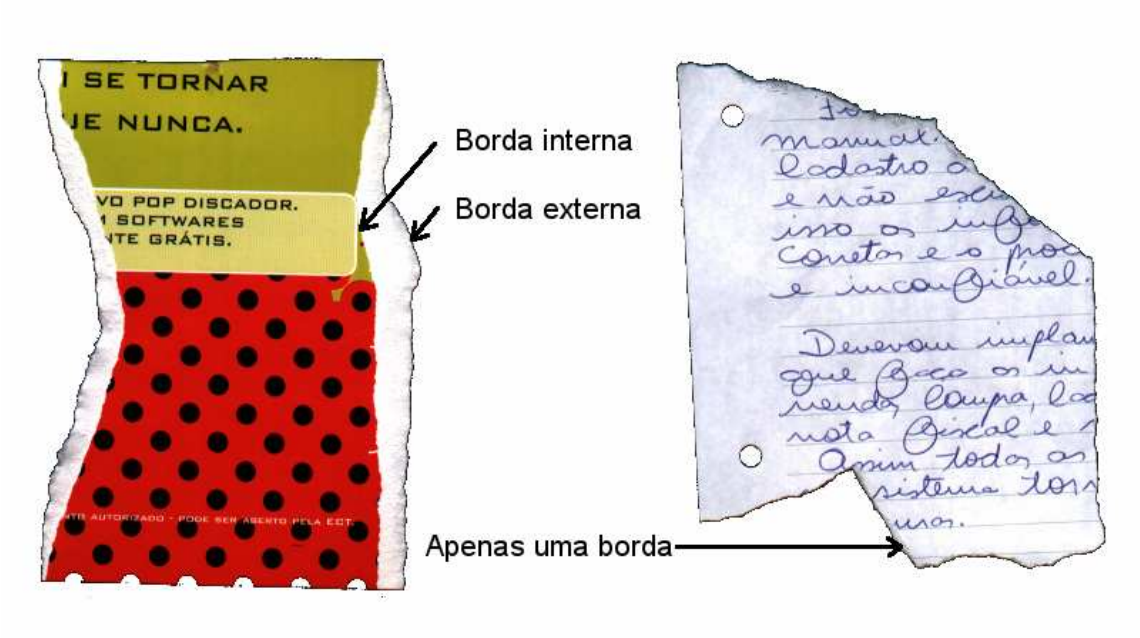


Figura 62 - Exemplo de possível encaixe entre os fragmentos utilizando processo de combinação com convergência dos fragmentos.

Dessa forma, conforme a figura 62, fragmentos que possuem ângulos e arestas que formam divisas com mais de um fragmento distinto, acaba por não ser detectado como parceiro no momento da reconstrução. Na figura 62, o fragmento de número 4 não foi adicionado à reconstrução final do documento por não ter nenhum ângulo coincidente com os demais fragmentos. Porém se observarmos o ponto P1, formado pela junção dos fragmentos 1 e 3, e P2, formado pela junção do fragmento 1 e 2, é evidente que existe uma correlação com os pontos P1’ e P2’ do fragmentos 4. Processamentos com técnicas de convergência poderiam ser aplicados nas fases de reconstrução para detectar essa característica e melhorar a qualidade do processo.

Em casos de documentos rasgados à mão, conforme exposto por [SOLANA, 2005], pode ocorrer o fenômeno de criação de bordas duplas nos fragmentos, conforme mostra a figura 63.



(a) (b)

Figura 63 - Exemplo de fragmentos de documentos rasgados a mão; (a) Com borda dupla, borda interna e borda externa; (b) Com apenas uma borda.

Em casos de documentos rasgados a mão semelhante à figura 62 (b), que não há borda dupla, ou quando a faixa entre as bordas é desprezível, o algoritmo atua sem problemas, porém nos casos como exposto na figura 62 (a) a detecção de candidatos a parceiro é dificultada devido ao algoritmo de extração do contorno não atuar em bordas duplas. A solução para este problema continua em aberto.

4.5 CONCLUSÃO

Neste capítulo foram apresentadas duas metodologias para a reconstrução digital de documentos mutilados.

Na primeira metodologia foi exposto o processo de reconstrução baseado em segmentação dos contornos dos fragmentos formando cadeias menores de códigos de Freeman. São esses segmentos que formam as características do contorno do fragmento e que são submetidas ao processo de programação dinâmica.

Esse processo demonstrou não ser eficiente para a reconstrução digital de documentos, apesar dos ensaios iniciais apontarem ao contrário. De acordo com os problemas identificados, não foi possível utilizar a seqüência de pontos do contorno como características dos fragmentos a serem reconhecidas, sendo esta metodologia descartada.

Na segunda metodologia foi exposto o processo de reconstrução digital de documentos baseado na extração das características geométricas do contorno na imagem. As características são aplicadas ao algoritmo de LCS e

posterior a uma modificação do algoritmo de Prim para retirar falsos candidatos e alinhar o processo de reconstrução dos fragmentos.

Esta metodologia demonstrou ser eficiente para a reconstrução de documentos, incluindo a reconstrução visual do documento.

No capítulo 5 serão demonstrados os resultados obtidos na aplicação dessa metodologia.

Capítulo 5

RESULTADOS OBTIDOS

5.1 INTRODUÇÃO

Este capítulo será dividido em 2 partes: na primeira parte serão expostos os resultados obtidos utilizando a metodologia proposta no capítulo 4 aplicada à base de dados de documentos mutilados da PUCPR comparando o resultado com os resultados anteriores alcançados por Solana [SOLANA, 2005]; na segunda etapa serão expostos alguns resultados de documentos reconstruídos ao final do processo, incluindo documentos que foram reconstruídos integralmente e documentos que foram reconstruídos parcialmente.

Durante o processo de pesquisa e análise de reconstrução observa-se que a utilização do algoritmo modificado de Prim para a verificação e geração da seqüência correta de remontagem possui grande influencia nos resultados finais, sendo tão importante quanto o próprio algoritmo de programação dinâmica, o qual tem a finalidade de apontar os candidatos a parceiros. O algoritmo de Prim auxiliou tanto no aumento na taxa de candidatos parceiros corretos, quanto no processo de reconstrução e apresentação visual do documento.

5.2 CLASSIFICAÇÃO DOS CANDIDATOS A PARCEIROS

Na subseção 3.3.3 estão expostos os resultados alcançados por [SOLANA, 2005] divididos em 3 grupos: classificação com repetição de candidatos a parceiros; classificação sem repetição de candidatos a parceiros; classificação com convergência dos fragmentos.

Classificação com repetição de candidatos a parceiros significa que um mesmo fragmento que já tenha encontrado um potencial parceiro será submetido à análise e verificação da possibilidade de ocorrer parcerias com os demais fragmentos que ainda restam do documento. Utilizando esta técnica temos a vantagem de analisar se um fragmento possui mais de 1 candidato a parceiro, porém teremos a desvantagem de encontrar um número maior de falsos candidatos.

No trabalho realizado por Solana [SOLANA, 2005], a definição da escolha de um candidato a parceiro em detrimento de outros, utiliza apenas como parâmetro um valor calculado baseado no perímetro de combinação, representando uma nota para a parceria. Porém não se colocava a atenção sobre o resultado gráfico daquela parceria.

Nesta pesquisa, realizamos o processo com a repetição de candidatos a parceiros, porém com a verificação de oclusão entre os fragmentos para realizar o descarte de falsos candidatos. A comparação dos resultados obtidos por [SOLANA, 2005] está exposto na tabela 9, e pôde ser realizado devido à utilização da mesma base de dados de documentos mutilados para os experimentos, conforme seção 4.2.

Tabela 9 - Comparação do método proposto em relação ao método proposto por [SOLANA, 2005], com repetição de candidatos a parceiros.

Pesquisa	Tolerância aproximação poligonal	Número de Documentos	Erros durante o processo	Falsos candidatos	Candidatos corretos
Solana, 2005	Baixa	81%	15%	34%	51%
Solana, 2005	Média	81%	17%	40%	43%
Método proposto	Com variação da tolerância	81%	11%	14%	75%

Conforme os resultados da tabela 9, podemos avaliar que a metodologia proposta possui um ganho considerável em relação à metodologia proposta por [SOLANA, 2005]. Houve uma redução em ao menos 4% nos erros durante o processo, uma grande redução de falsos candidatos a parceiros ao menos em 20% e ainda um ganho em 24% na combinação correta de candidatos a parceiros.

A coluna de falsos positivos representa fragmentos que possuem boas características para serem considerados parceiros, porém claramente não são, estando assim encaixados erroneamente. A coluna de erros durante o processo representa os fragmentos que durante o processo de LCS não foram encontrados combinações. Um problema comum de erro é o caso de um fragmento não possuir características suficientes devido a cortes lineares, que por sua vez, venha a não produzir vértices e ângulos para serem utilizados como características.

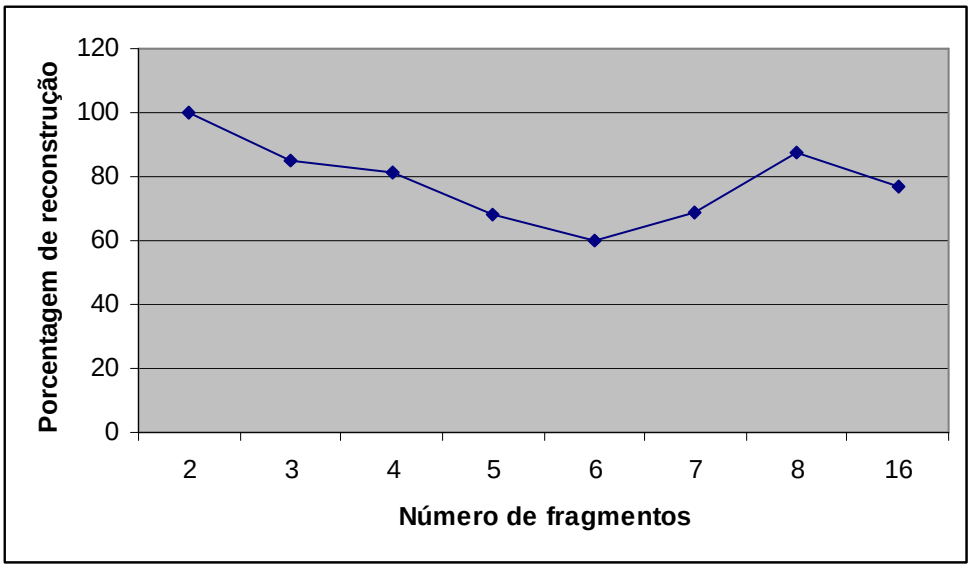


Figura 64 - Porcentagem de reconstrução por número de fragmentos.

O gráfico da figura 64 demonstra a porcentagem média de reconstrução dos documentos de acordo com a quantidade de fragmentos. Analisando o gráfico percebe-se que não há uma relação de degradação do método em relação à quantidade de fragmentos utilizado, mas sim os resultados piores estão entre 5 e 7 fragmentos. Analisando os documentos da base de dados, percebe-se que os documentos de 5 a 7 fragmentos possuem mais casos específicos de mutilação onde o método proposto não atua com total eficiência. Um dos problemas mais evidentes é o exposto na figura 62, onde os vértices da mutilação não estão disponíveis em apenas um fragmento.

Esse resultado é considerado positivo nesse contexto, pois se pode melhorar o processo para que seja mais eficiente nesses casos, e ainda projetar uma expectativa melhor avaliando o processo com um número maior de fragmentos para cada documento.

Na figura 65 está o gráfico de convergência dos fragmentos de documentos de acordo com a sua quantidade referente ao método proposto

por Solana. Percebe-se no método que à medida que o número de fragmentos aumenta, o desempenho do método diminui.

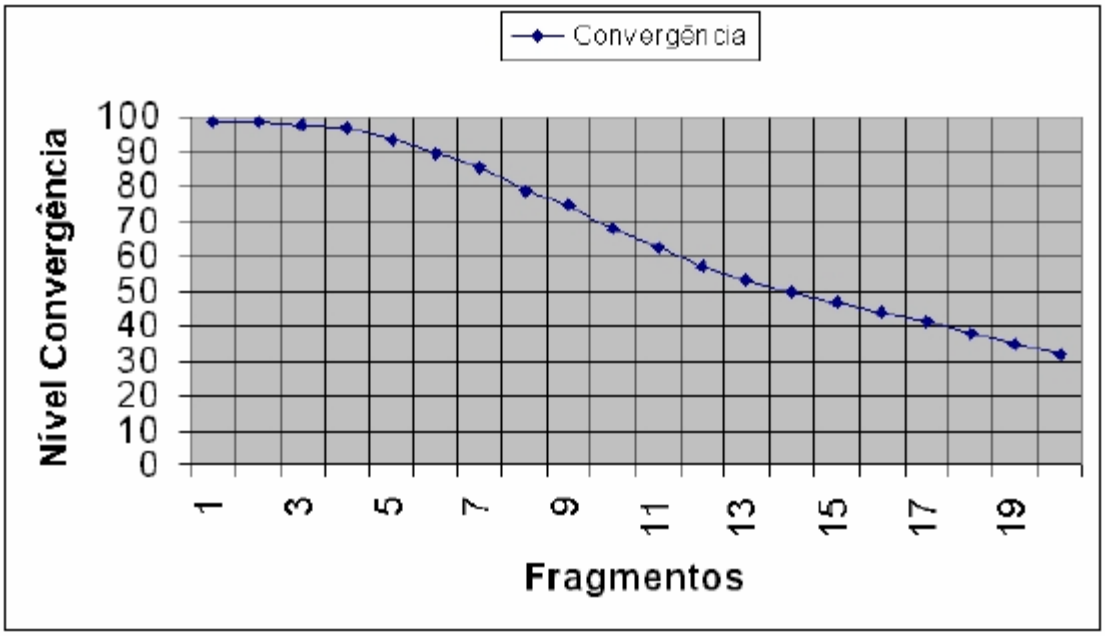


Figura 65 - Nível de convergência de acordo com a quantidade de fragmentos. Baixa tolerância [SOLANA, 2005].

Na figura 66, está o gráfico do tempo de processamento do método proposto em relação a quantidade de fragmentos.

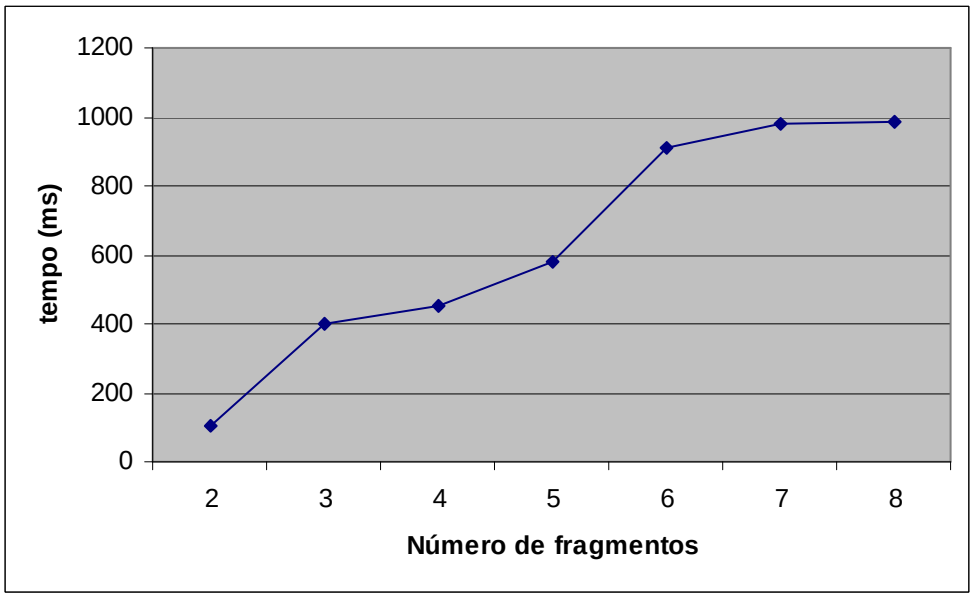
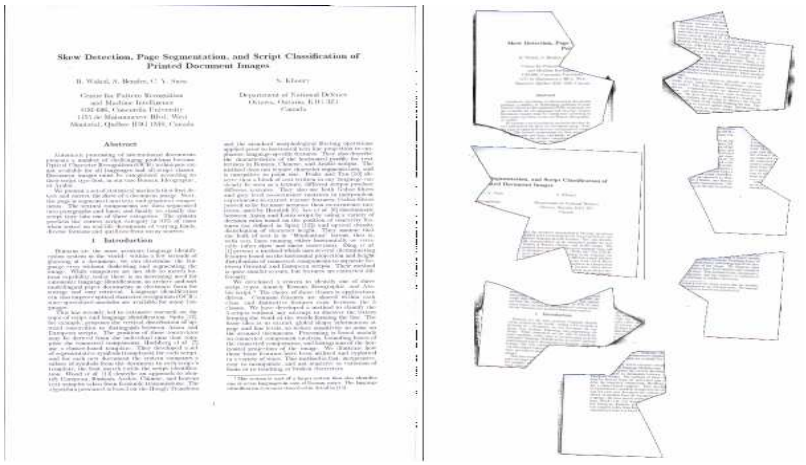


Figura 66 - Tempo médio de processamento por quantidade de fragmentos.

Além da quantidade de fragmentos, um fator que interfere diretamente no aumento do tempo de processamento é a quantidade de características extraídas de cada contorno de fragmento.

De acordo com o gráfico da figura 66 percebe-se que o aumento médio do tempo de processamento em relação ao número de fragmentos não é expressivo, o tempo aumenta proporcionalmente ao número de fragmentos, identificando que a utilização da técnica de programação dinâmica está sendo aplicada corretamente. O tempo médio geral de processamento considerando todos os documentos analisados é de aproximadamente 630ms, valor obtido utilizando um microcomputador desktop IBM-PC compatível, com velocidade de 3800Mhz e 1GB de memória ram.

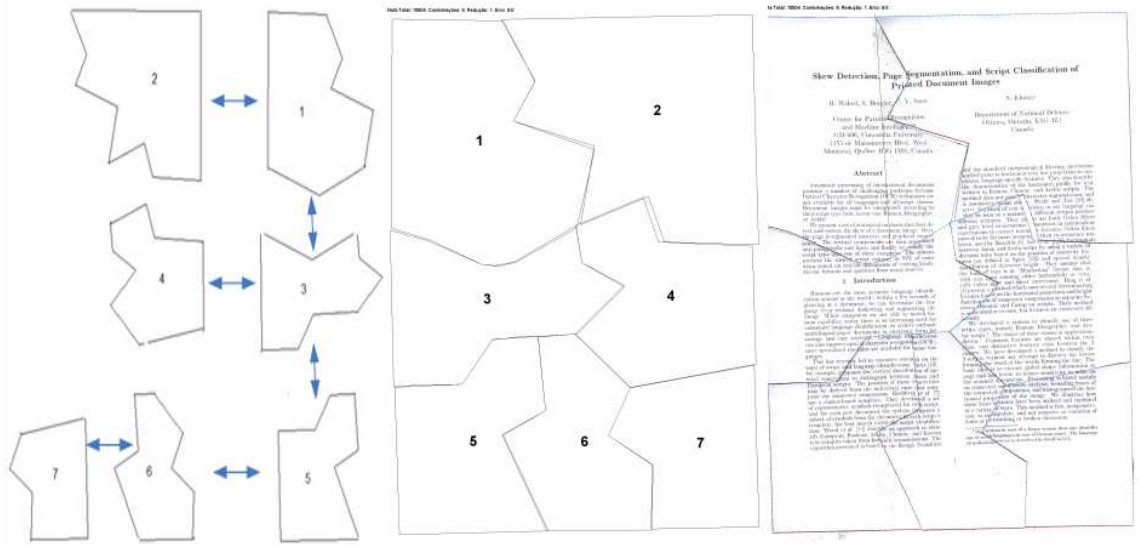
A figura 67 mostra um documento da base de imagens da PUCPR que foi submetido ao processo de reconstrução. A figura 67 (a) mostra o documento original e a figura 67 (b) mostra os fragmentos resultantes da mutilação.



(a) (b)

Figura 67 - (a) Documento original da base de imagens; (b) Fragmentos após a mutilação.

A figura 68 mostra o processo de remontagem do documento da figura 66. Todo o processo de recomposição do documento é automatizado.



(a) (b) (c)

Figura 68 - (a) Seqüência de remontagem dos fragmentos sem ciclos; (b) Polígonos remontados; (c) Documento original remontado.

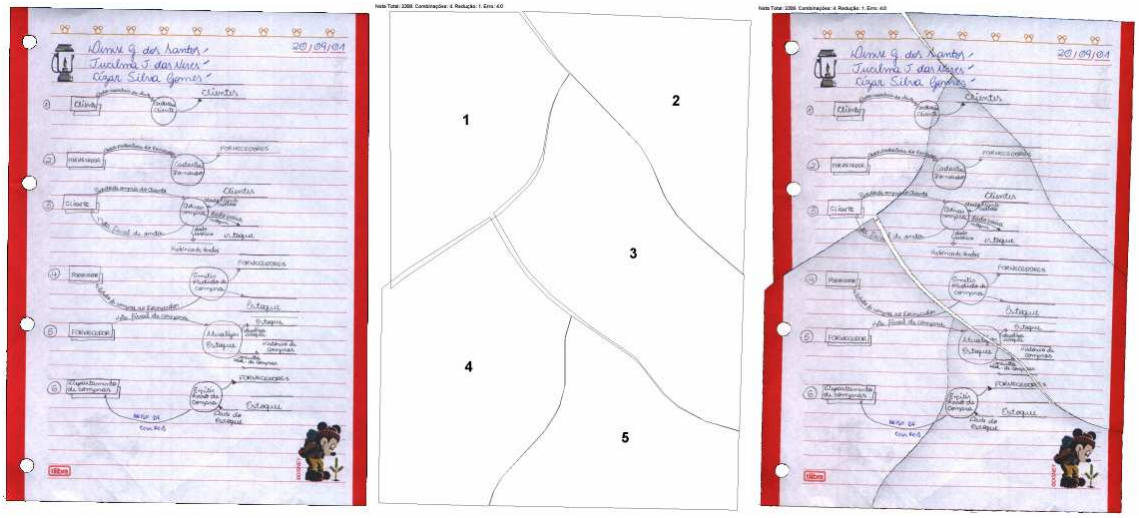


Figura 69 - Documento 41 da base de imagens reconstruído.

As figuras 69, 70 e 71 são exemplos de documentos constantes na base de dados de imagens que foram totalmente reconstruídos após serem submetidos ao processo de reconstrução.

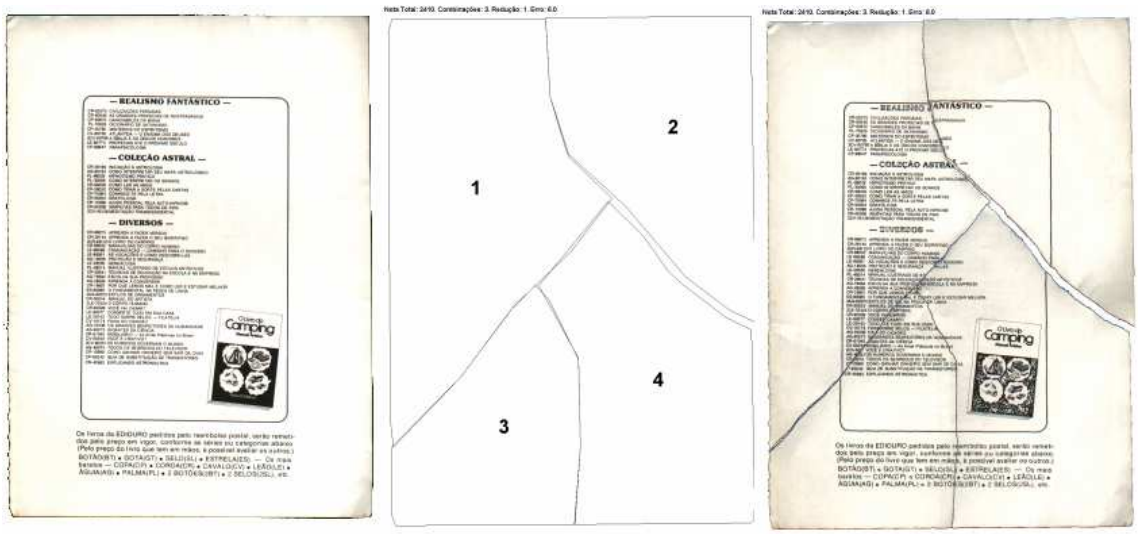


Figura 70 - Documento 96 da base de imagens reconstruído.

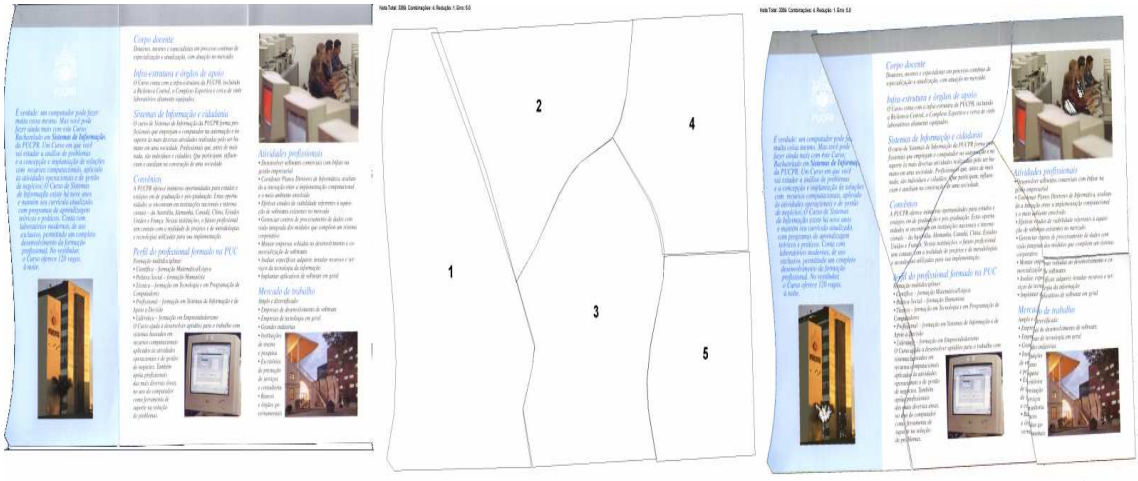


Figura 71 - Documento 3 da base de imagens reconstruído.

As figuras 72 e 73 são exemplos de documentos que foram parcialmente reconstruídos. Na figura 72, um fragmento não foi adicionado ao processo devido a não encontrar nenhum candidato a parceiro, nesse caso é incrementada a estatística de erro no processo.

Na figura 73, além de fragmentos que não foram adicionados ao processo, existe um fragmento, fragmento 4, que está erroneamente encaixado, considerado como falso-positivo.

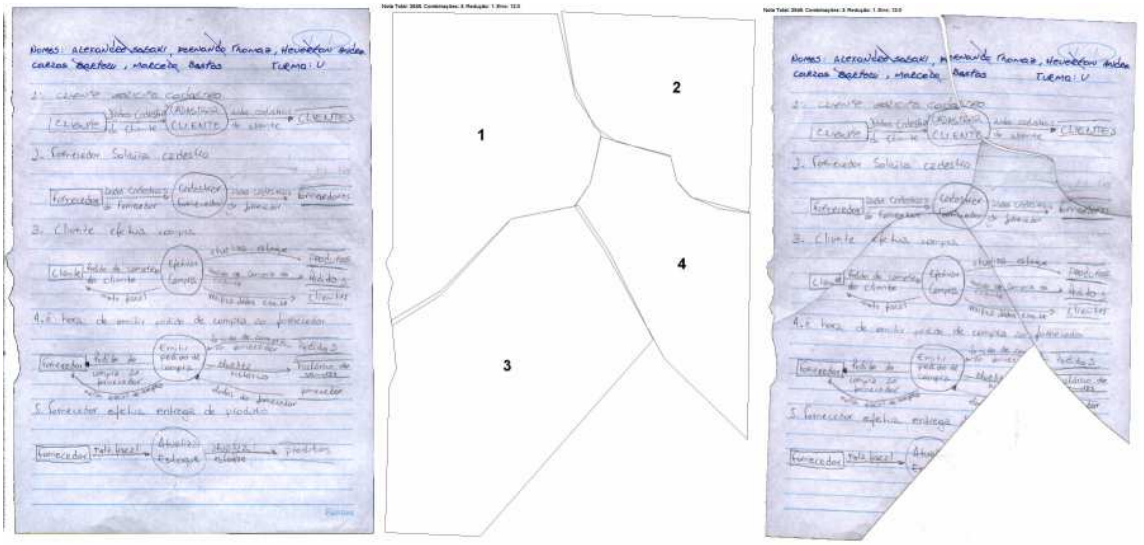


Figura 72 - Documento 38 da base de imagens parcialmente reconstruído.

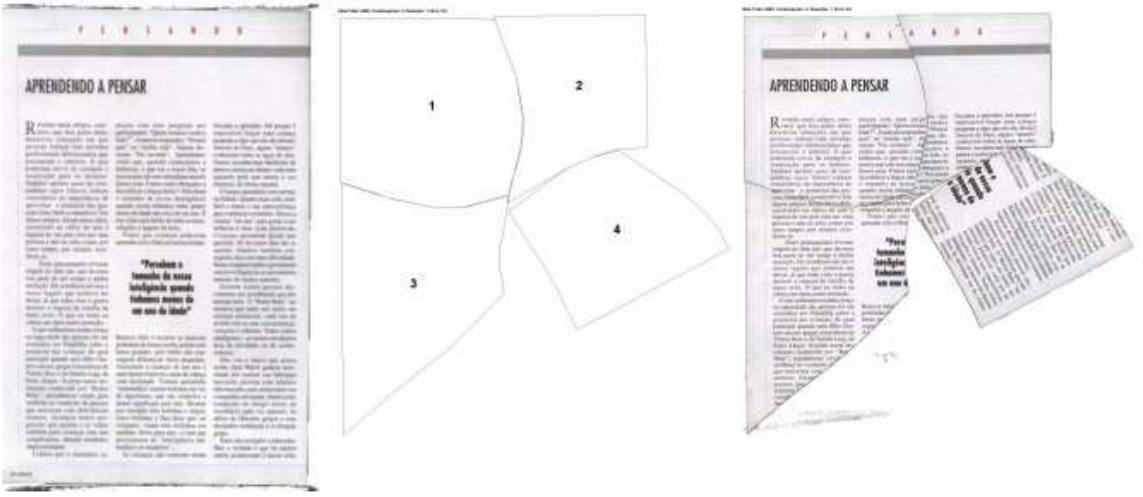


Figura 73 - Documento 62 da base de imagens parcialmente reconstruído e com falso candidato.

Conforme apresentado nas figuras 72 e 73, mesmo em casos de reconstrução parcial, a informação recuperada automaticamente pode diminuir consideravelmente o tempo necessário para realizar a atividade de reconstrução manualmente.

5.3 CONCLUSÃO

Este capítulo apresentou os resultados alcançados nos experimentos realizados através da metodologia proposta. Foram realizadas comparações e

análises estatísticas da metodologia com trabalhos já publicados, concluindo que a mesma é eficiente no processo de reconstrução de documentos mutilados.

Neste capítulo ainda foram apresentadas algumas imagens de documentos reconstruídos pelo processo, incluindo documentos totalmente reconstruídos e outros parcialmente reconstruídos, indicando que os resultados aqui alcançados podem ser melhorados, justificando a continuidade de pesquisas nesta área.

No próximo capítulo será exposta a conclusão deste trabalho.

Capítulo 6

CONCLUSÃO E TRABALHOS FUTUROS

A reconstrução de documentos mutilados não se configura como uma atividade trivial no processo de perícia documentoscópica. Problemas com o manuseio negligente, tempo e esforço dispensado nesta atividade e, principalmente, métodos destrutíveis de reconstrução, constituem uma forte razão para que se tenha um processo computacional de reconstrução digital de documentos mutilados.

O foco deste trabalho é auxiliar o processo de perícia, munido de métodos computacionais automatizados ou semi-automatizados, de forma a facilitar o processo trazendo mais agilidade e qualidade. Em conjunto com esta questão, os resultados apresentados por este trabalho, apontam novas possibilidades de pesquisa nesta área, gerando a continuidade e aperfeiçoamento dos métodos digitais de reconstrução de documentos mutilados.

As contribuições efetivas deste trabalho são:

- Aumento da taxa de reconstrução dos documentos da base de dados da PUCPR, em 24%, comparada ao trabalho pioneiro realizado por [SOLANA, 2005], elevando a taxa média de reconstrução para 75%.

- O resultados obtidos não são mais apresentados apenas no formato de rótulo, como na maioria das pesquisas apresentadas até o momento, mais sim apresenta a reconstrução visual final do documento, que já possui embutido o rótulo, a posição e a prévia do documento reconstruído.
- Diminuição dos recursos dispensados, principalmente tempo, na reconstrução final do documento para avaliação pericial.
- Processo e análise de combinação considerando características gráficas e detectando sobreposição entre fragmentos candidatos. Não se avalia apenas a característica para realizar a combinação, mais sim a composição global dos fragmentos diante de reconstrução gráfica do documento.
- Aumento médio de tempo de processamento proporcional à quantidade de fragmentos do documento.
- Baixo tempo de processamento e baixo nível de recursos consumidos, indicando que o método pode ser implementado em máquinas desktop’s convencionais.

Podemos destacar como possibilidades de melhorias em pesquisas e trabalhos futuros:

- Melhorar o processo e adicionar módulos para a realização de reconstrução com convergência dos fragmentos.
- Coletar características entre fragmentos já remontados durante o processo, ou seja, a partir da reconstrução de 2 ou mais fragmentos, recriar novos vetores de características.

- Coletar novas características para reforçar o descarte de falsos-candidatos, como cores e informações do contexto.
- Utilizar as cadeias secundárias encontradas no processo de programação dinâmica, ma matriz de LCS. No processo atual apenas a cadeia principal é utilizada.
- Detectar e separar os contornos com tendências curvilíneas e avaliar a utilização de técnicas de reconhecimento de curvas.
- Realizar experimentos com documentos que sejam digitalizados em maior escala.
- Utilizar outros algoritmos de aproximação poligonal e avaliar o impacto nos resultados atuais.
- Aumentar a base de dados com documentos que possuam mais fragmentos e em tamanhos menores.

REFERÊNCIAS BIBLIOGRÁFICAS

[BELLMAN, 1957] BELLMAN R. E, *Dynamic Programming*. Princeton University Press, 1957. Disponível em http://books.google.com. Acessado em 18/03/2007.

[BREUEL, 2001] BREUEL, T. M. *Segmentation of Handprinted Letter Strings using a Dynamic Program Algorithm.* Sixth International Conference on Document Analysis and Recognition (ICDAR'01) icdar pp. 0821, 2001.

[CAMPELLO, 2005] Campello, L. G. B, *As Provas e o Recurso à Ciência no Processo.* Revista da Faculdade de Direito de Campos, Ano VI, Nº 6 – junho de 2005.

[DOUGLAS & PEUCKER, 1973] DOUGLAS, D; PEUCKER T., *Algorithms for the reduction of the number of points required to represent a digitalized line or its caricature.* The Canadian Cartographer 10(2), pp 112 – 122, 1973.

[ECKERT, 1992] ECKERT, W. G, *Introduction to forensic sciences*, second edition, New York: Elsevier. 319p, 1992.

[ERNST & FLINCHBAUGH, 1989] ERNST, M. D e FLINCHBAUCH, B. E, *Image map correspondence using curve matchig.* Texas Intruments, Computer Science Center. Dallas, Texas. Março de 1989.

[FBI, 2004] Federal Bureau of Investigation. *FBI Laboratory 2003 Report.* FBI Laboratory Publication – www.fbi.gov/hq/lab/labannual03.pdf, Quantico, Virginia, 2004. Acessado em 10/10/2006.

[FBI, 2005] Federal Bureau of Investigation. *FBI Laboratory 2005 Report.* FBI Laboratory Publication – www.fbi.gov/hq/lab/labannual05.pdf, Quantico, Virginia, 2005. Acessado em 10/10/2006.

[FREEMAN, 1974] FREEMAN, H., *Computer processing of line-drawing images*. Computing Surveys (CSUR), v. 6 n. 1, pp 57-97, Março 1974.

[FREITAS, 2008] FREITAS, E. A. M. *Questões Legais da Digitalização de Documentos: evolução necessária para a redução do Custo Brasil.* Comitê de Direito da Tecnologia. AMCHAM – Brasil, São Paulo, julho de 2008.

[GANDINI, 2002] J. A D; SALOMÃO, Diana Paola da Silva et al. *A validade jurídica dos documentos digitais.* Jus Navigandi, Teresina, ano 6, n. 58, agosto de 2002. http://jus2.uol.com.br/doutrina/texto.asp?id=3165. Acessado em: 14/12/ 2008.

[GONZALEZ & WOODS, 2000] GONZALES, R.C e WOODS, R.E. *Processamento de imagens digitais*. Addison-Wesley Publishing Company, Inc, New York, 1992, Editora Edgard Blucher Ltda. São Paulo SP, 2000.

[GREENBERG, 2003] GREEBERG R. I, *Bounds on the Number of Longest Common Subsequences*. The Computing Research Repository cs.DS/0301034, 1-3. Department of Mathematical and Computer Sciences, Loyola University.

[KAMPEL & SABLATNIG, 2004] KAMPEL, M.; SABLATNIG, R. *On 3D Mosaicing of Rotationally Symmetric Ceramic Fragments.* 17th Internationa Conference on (ICPR'04). pp 265-268. Vienna University of Technology, Vienna, Áustria, 2004.

[KLAJNSEK, 2000] KLAJNASEK, G. *Merging a set of polygons with non-stable borders. 4th Central European seminar on computer graphics, May 1-3, 2000, Budmerice, Slovakia* Laboratory for Geometric Modeling and Multimedia Algorithms. Faculty of Eletrical Engineering and Computer Science; University of Maribor.

[KONG & KIMIA, 2001] KONG, W. E KIMIA, B. B., *On solving 2D and 3D puzzles using curving matching,* IEEE, Computer Society . Proceeding of Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, Dezembro 2001.

[KULESH & MEMON, 2003] S. N. Memon, *Automatic Reassembly of Document Fragments via Context Based Statistical Models.* 19th Annual Computer Security Applications Conference. p. 152. Departamente of Computer and Information Science. Polytechnic University. Brooklyn, USA.

[LEITÃO, 2000] LEITÃO, H. C. G., *Reconstrução automática de objetos fragmentados.* Tese de Doutorado de 21/10/1999, 138p. Instituto de Educação, Universidade Estadual de Campinas. Campinas, Brasil, 2000..

[LEITÃO & STOLFI, 2002] LEITÃO, H. C. G; STOLFI, J., *A multiscale method for the reassembly of two-dimensional fragmented objects.* IEEE Trans, Patter analisys and Machine Inteligence., vol 24, pp 1239-1251, 2002.

[PAPAODYSSEUS et al, 2002] PAPAODYSSEUS, C. T. Panagopoulos, M. Exarhos, C. Triantafillou, D. Fragoulis, and C. Doumas. *Contourshape based reconstruction of fragmented*. IEEE Trans. Signal Processing, vol. 50, pp 1277–1288, 2002.

[PATTERSSON, 2005] PETTERSSON, N. *Measuring Precision for Static and Dynamic Design Pattern recognition as a Function of Coverage.* Workshop

on Dynamic Analysis (WODA). pp 1-7 School of Mathematics and Systems Engineering. Växjö University. Växjö, Sweden. 2005.

[PIMENTA et al, 2008] PIMENTA, A. M.; JUSTINO. E. J. R.; OLIVEIRA, L. E. S. *Document Reconstruction Using Dynamic Programming.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009), Abril 19-24, pp1393-1396 2009. Taipei, Taiwan.

[PRANDTSTETTER, 2008] M. G. R. Raidl, *Combining Forces to Reconstruct Strip Shredded Text Documents.* 5th International Workshop on Hybrid Metaheuristics. pp 175-189. Malasia, Spain.

[PRIM, 1957] PRIM, R. C, *Shortest connection networks and some generalism*, Bell System Technical Journal, vol. 36, pp. 1389-1401, 1957.

[SCHMITKNECHT, 2004] SCHMITKNECHT, D.A, *Building FBI computer forensics capacity: one lab at a time.* Digital Investigation 2004. http://www.rcfl.gov/downloads/documents/DigitalInvestigator.pdf. Acessado em 11/10/2007.

[SMET, 2007] SMET, P. *Reconstruction of ripped-up documents using fragment stack analysis procedures,* Forensic Science Intern, vol. 176, pp. 124-136, 2008.

[SOLANA, 2005] SOLANA, C. D. O., *Reconstrução Digital de Documentos por Aproximação Poligonal*. Dissertação de Mestrado de 01/08/2005, 107p, no Programa de Pós Graduação em Informática Aplicada da Universidade Católica do Paraná PUCPR, Curitiba, Brasil, 2005.

[UKOVICH, 2008] A, G. Ramponi, H. Doulaverakis, Y. Kompatsiaris, M.G. Strintzis, *Shredded Document reconstruction using MPEG-7 Standard Descriptors.* IEEE International Symposium on Signal Processing and Information Technology, December 2004, pp 334-337. IPL – DEEI. Thessaloniki, Greece.

[YAO & SHAO, 2003] YAO, F.; SHAO, G., *A shape and image merging technique to solve jigsaw puzzles,* Patter Recognition Letters (PRL), Vol. 24, Nr. 12, August 2003, pp 1819-1835, 2003.

[WILLIS & COOPER] WILLIS A. R.; COOPER, D. B. *Computational reconstruction of ancient artifacts*. IEEE Signal Processing Magazine, vol. 25, pp 65–83, 2008.

[WOLFSON, 1990] WOLFSON, H. *On curve matching.* IEEE Trans. Pattern Anal. and Machine Intell., vol. 12, pp 483–489, 1990.