3 SISTEMAS DE RECONHECIMENTO DE CARACTERES

Os sistemas de reconhecimento ótico de caracteres - OCR (*Optical Character Recognition*) são sistemas desenvolvidos para, de uma certa forma, reproduzir a capacidade humana de ler textos. Este capítulo abordará os diferentes tópicos relacionados com o assunto, apresentando uma classificação dos sistemas de reconhecimento de caracteres em geral (não somente os óticos) e descrevendo as principais características e algoritmos empregados em tais sistemas.

Mas afinal, o que é "reconhecer um texto"? Qual o tipo de classificação que é realizado? Um OCR recebe um texto sob a forma de uma imagem digital e através de um processo de reconhecimento e classificação dos padrões encontrados, que representam os caracteres, obtêm como resultado um código que representa cada caractere. Podem ser usados como códigos de identificação dos caracteres o ASCII, EBCDIC ou outro código similar.

3.1 Origens e Evolução

As origens da tentativa de simular a leitura humana datam de 1870, quando foi inventado por Carey o *scanner* de retina. A partir da evolução dos dispositivos capazes de captar e traduzir imagens em sinais que podiam ser medidos, novos horizontes foram descobertos. Entretanto, a primeira tentativa bem sucedida de reconhecimento de caracteres foi realizada somente em 1900 pelo cientista russo Tyutin, que desenvolveu um sistema que visava o auxílio de deficientes visuais [MAN 86].

A partir do desenvolvimento do computador digital nos anos 40, esta área de estudos ganhou um grande impulso, pois até o momento os dispositivos usados eram muito complexos e tinham que ser desenvolvidos unicamente com o propósito de reconhecer caracteres.

Os computadores digitais tornaram mais viáveis os estudos a respeito dos sistemas de reconhecimento de caracteres. Já na década de 50, são encontradas uma série de publicações e relatos a respeito de tais sistemas, como em Dineen (1955) e Bomba (1959), conforme citação em [HAR 72]. Nos anos 60, grandes avanços continuaram sendo realizados e esta área de estudos da computação começou a se consolidar [STE 61].

A partir desta época o crescimento da área de estudos do reconhecimento de caracteres foi muito veloz, principalmente devido às inovações tecnológicas que permitiram a utilização de dispositivos mais sofisticados, precisos e acessíveis em termos de custos, tanto para a aquisição como para o processamento dos dados. Um grande número de pesquisas e trabalhos foram desenvolvidos e publicados a respeito do assunto, possibilitando uma maior difusão e evolução da área.

Atualmente, uma grande quantidade de patentes e publicações específicas sobre sistemas OCR tem sido feitas, sendo que tais sistemas se tornaram disponíveis, tanto para aplicações comerciais de pequeno, médio ou grande porte, como para aplicações domésticas. Os objetivos atuais continuam sendo a busca de uma redução dos custos e o aumento da eficiência, bem como da flexibilidade de tais sistemas, sendo que ainda há muito a ser feito nesta área no futuro.

Mas por que este objetivo de desenvolver sistemas de reconhecimento de caracteres tem sido tão duramente buscado e gasto tantos esforços dos cientistas a procura de melhores soluções? Basicamente isto é devido às inúmeras aplicações possíveis de tais sistemas, tais como:

- Aquisição de dados numéricos comerciais[SUE 80];
- Compactação de dados [PRA 80];
- Leitura automática de formulários [DAI 90];
- Identificação de endereçamento postal [SCH 78];
- Reconhecimento de partituras musicais [SUE 80];
- Leitura de cheques bancários [HAR 72];
- Sistemas de aquisição de textos para a tradução automática [HAR 72];

- Interface homem x máquina mais natural e sem necessidade de recodificação dos dados (sujeitos a um menor número de erros, pois não é necessário redigitar os textos, introduzindo novos erros) [HAR 72] e

- Auxílio a deficientes visuais, na leitura de textos (tradução e impressão automática em Braille) [HAR 72];

Esta lista de aplicações poderia ser estendida por muitos e muitos outros ítens ainda. Como pode ser visto, o número de aplicações é bem amplo, mas além de todas estas possibilidades de utilização dos sistemas de reconhecimento de caracteres, ainda há um outro fator muito importante: o grande desafio de se conseguir replicar em uma máquina as funções humanas.

3.2 Tipos de Sistemas de Reconhecimento de Caracteres

Os sistemas de reconhecimento de caracteres podem ser desenvolvidos utilizando-se diferentes procedimentos tanto de aquisição de dados como de processamento de informações além de serem orientados para diferentes tipos de aplicações. O esquema da figura 3.1 apresenta uma classificação dos sistemas de reconhecimento de caracteres, em relação aos ítens citados acima. Este esquema é baseado em [MAN 86], com o acréscimo de mais alguns ítens.

Os diferentes sistemas de reconhecimento de caracteres podem ser classificados primeiramente quanto ao tipo de mecanismo utilizado na aquisição dos textos a serem reconhecidos. Nesta categoria, encontram-se os sistemas magnéticos (onde o texto é escrito com uma tinta magnética especial), os sistemas mecânicos e os sistemas óticos. Este último será o alvo dos estudos deste trabalho. Todos estes sistemas podem utilizar um processamento realizado a nível de sinais analógicos ou digitais. Os sistemas mais antigos, devido à tecnologia disponível, usavam mecanismos magnéticos, mecânicos ou óticos, baseados em dispositivos de tratamento de informações analógicas.

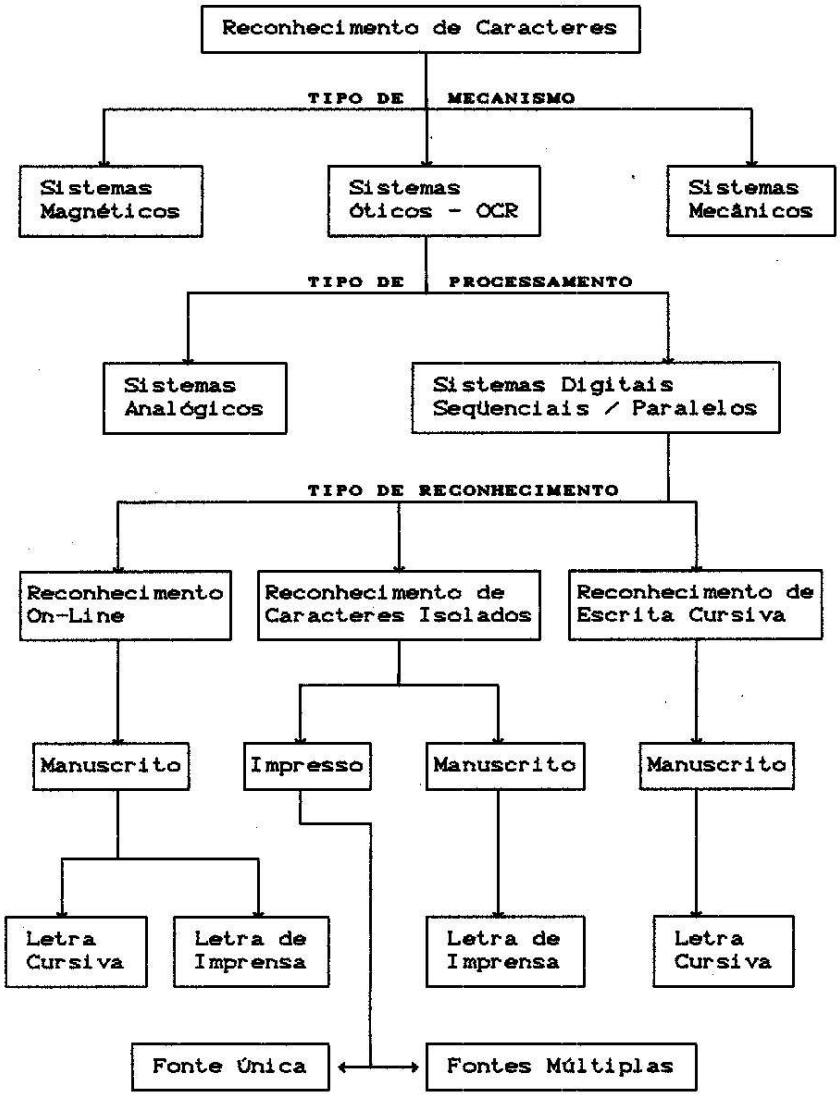


Figura 3.1 - Classificação dos sistemas de reconhecimento de caracteres

Um exemplo destes sistemas analógicos é descrito em [HAR 72], onde foi empregada uma técnica que trabalha com os campos de distribuição potencial. Nesta técnica, uma representação elétrica do caractere é colocada em uma matriz de resistores na qual um campo de potencial uniforme foi aplicado. A sobreposição da representação do padrão distorce o campo, de tal forma que a medida da distribuição potencial resultante nas extremidades da matriz, permite a obtenção de uma classificação única do padrão.

Os sistemas mecânicos, magnéticos e óticos que atuam sobre dados analógicos continuam a ser investigados até hoje, obtendo inclusive resultados muito promissores devido à possibilidade de exploração do paralelismo inerente a alguns destes processos. Infelizmente, tais sistemas são do tipo dedicado e exigem um custo muito alto em seu desenvolvimento.

Este estudo será restrito aos sistemas de OCR baseados no processamento digital, seja em sistemas sequenciais ou paralelos. Um dos motivos desta escolha é o fato de tais sistemas serem de uso geral, e desta forma seu custo pode ser reduzido em função da utilização do equipamento para outros fins.

Dentro dos sistemas OCR, são encontradas duas classes importantes, aqueles que são baseados no processamento sequencial das informações e os baseados no processamento paralelo. Este tipo de aspecto será abordado em maiores detalhes posteriormente. Entretanto, a principal distinção a ser feita em relação aos sistemas de OCR é em relação ao tipo de texto utilizado no reconhecimento. Existem três categorias principais: sistemas de reconhecimento *on-line*, sistemas de reconhecimento de caracteres isolados e sistemas de reconhecimento de escrita cursiva, onde estes dois últimos são realizados *off-line*.

Os sistemas de reconhecimento *on-line*, possuem características próprias que os distingüem dos demais, entre elas, a forma como os caracteres são entrados no sistema.

Nestes sistemas, utiliza-se usualmente uma mesa digitalizadora ou uma *light-pen* para fazer a aquisição unicamente de textos manuscritos. Devido à maneira como os caracteres são fornecidos, os dados disponíveis possuem características próprias. Estes dados são referentes a trajetória do traçado e não mais uma matriz de pontos, estando sujeito apenas a erros de precisão de coordenadas sem sofrer de problemas típicos da digitalização de imagens (inversão de pontos devido a interferências no processo de aquisição). Como resultado desta forma de entrada de dados, a descrição dos caracteres já possuirá uma representação "esqueletizada", apresentando apenas o contorno do caractere. Além da informação da forma do caractere que é fornecida por estes dispositivos de entrada, tais dispositivos também fornecem um dado extra referente à ordem do traçado (descrição temporal), podendo também contribuir para a classificação do mesmo.

Outra característica importante destes sistemas é o fato de ser inviável um treinamento prévio para uma grande quantidade de padrões exemplos, uma vez que o usuário tem que entrar manualmente com cada um destes padrões. Portanto os sistemas de OCR *on-line* devem normalmente já terem um conhecimento prévio embutido, não necessitando de uma fase de treinamento.

Os sistemas de reconhecimento *on-line* podem aceitar taxas de reconhecimento inferiores aos demais sistemas OCR. Isto se deve ao fato de tais sistemas serem interativos, permitindo um *feedback* de/para o usuário, de forma que ele se adapte, acostumando-se com as características e limitações do sistema, e desta forma incrementando posteriormente a taxa de reconhecimento. A taxa inicial de classificação correta de caracteres aceitável para sistemas *on-line* é de cerca de 95% de acertos. Este tipo de sistemas vem sofrendo um crescimento na sua utilização na atualidade e hoje já se encontram alguns equipamentos e aplicações nas quais o teclado foi totalmente substituído por entradas textuais realizadas através de um *tablet* [PCW 91].

Os sistemas de reconhecimento de caracteres isolados são os de maior utilização na atualidade, o que é devido provavelmente: à implementação mais simples, às taxas de reconhecimento correto mais altas e também devido à sua grande aplicabilidade. Este tipo de sistemas visa o reconhecimento de caracteres que estão representados no texto isoladamente, ou seja, textos impressos com letras não contíguas ou manuscritos com uso de letras de imprensa.

O último tipo de sistema OCR é o de reconhecimento de escrita cursiva, o qual é considerado o mais complexo dos três tipos de sistemas, em termos de dificuldade de implementação e de obtenção de taxas razoáveis de reconhecimento correto. Neste tipo de sistema, um dos maiores problemas (que não é tão grave nos outros) é o da separação dos caracteres, de forma a identificar cada um separadamente. Esta tarefa pode se tornar extremamente complexa e sujeita a muitos problemas. Um exemplo deste tipo de aplicação pode ser encontrado em [TAP 82], onde é aplicada uma técnica de "*elastic match*". Os sistemas de reconhecimento de caracteres escritos com letra cursiva não serão abordados neste trabalho, pois possuem características completamente diferentes dos outros sistemas, o que iria requerer todo um estudo específico para este caso.

Em relação a nomenclatura, costuma-se denominar de sistema OPCR (*Optical Printed Character Recognition*), os sistemas para reconhecimento de caracteres impressos, e, sistemas OHCR (*Optical Handwritten Character Recognition*), os sistemas para reconhecimento de caracteres manuscritos. No caso do termo OHCR, este é utilizado tanto para sistemas do tipo *handprint* (letra de imprensa), como para sistemas do tipo *handscript* ou *cursive script* (letra cursiva). Neste trabalho serão utilizados seguidamente os termos OPCR e OHCR, mas como não serão abordados os sistemas de reconhecimento de escrita cursiva, subentende-se que os sistemas OHCR referenciados são aqueles onde a escrita é do tipo *handprint* apenas.

Tem-se como exemplos de sistemas de reconhecimento de escrita do tipo OPCR, os sistemas descritos em [MAR 88, MAR 90b, SIL 90, WON 82, BRI 82 e HAR 72], e de sistemas OHCR, em [TAP 82, DAI 90, GON 90, SCH 78 e SUE 80].

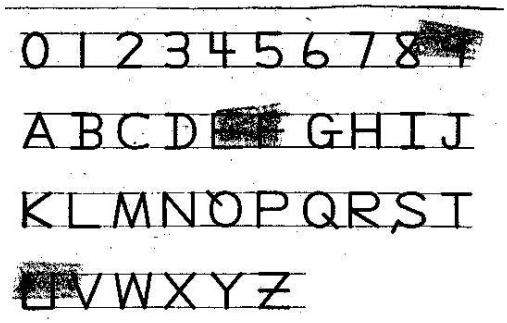
Na classificação dos sistemas OCR, ainda podem ser levados em consideração mais alguns aspectos tais como:

1) Limitações quanto à abrangência do conjunto de classificação: São encontrados diferentes tipos de sistemas OPCR/OHCR, onde pode ser limitada a quantidade de símbolos (caracteres) diferentes a serem reconhecidos, criando assim reconhecedores para classes como: números, caracteres alfanuméricos (alfabeto latino, grego, katakana, etc), conjunto de símbolos completo, como por exemplo o utilizado na linguagem FORTRAN, ou, os caracteres imprimíveis dos códigos ASCII, BRASCII, EBCDIC, etc.

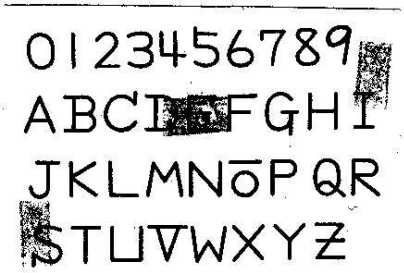
2) Limitações quanto ao número de estilos de escrita: Os sistemas OPCR podem ser limitados em relação ao número de fontes distintas de caracteres que o reconhecedor identifica simultaneamente. Pode-se ter reconhecedores apenas para um dos seguintes estilos: *sans-serif*, *pica*, *elite*, *itálico*, *negrito (boldface)*; ou para um conjunto formado pela união de um grupo destes estilos. Neste tipo de classificação, é possível distingüir principalmente entre textos datilografados (com uma única fonte usualmente), ou textos estilizados impressos (com várias fontes usualmente).

3) Limitações quanto à forma de escrita: Esta limitação em sistemas OHCR equivale à limitação em termos de estilos para os sistemas OPCR. É muito comum se encontrar na literatura referências a sistemas de OHCR onde os usuários são previamente treinados para escrever de um certo modo, alterando a sua forma natural de escrita. Desta maneira, consegue-se restringir o número de estilos diferentes de escrita existentes entre os diversos candidatos ao uso do sistema. Uma outra limitação ao estilo é o uso de formulários, onde o usuário fica restrito ao preenchimento dos caracteres dentro de campos de tamanho e posições fixas.

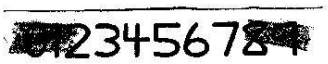
As restrições quanto à forma do caractere, que podem ser impostas de forma a facilitar e melhorar a eficácia do reconhecimento nos sistemas OCR, têm sido muito aconselhadas, principalmente pelos orgãos responsáveis pela normatização de diferentes países: OCR Committe of ANSI, Japanese OCR Committe e OCR Committe of Canada (figura 3.2). Baseado neste tipo de restrições, foram definidos diversos padrões de definição de conjuntos de caracteres, tanto para sistemas OPCR (OCR-A - Americano, OCR-B - Europeu), como para sistemas de OHCR (Conjunto de caracteres para escrita manual ANSI 1974) [SUE 80].



(A) Conjunto Padrão de Caracteres Manuscritos - OCR ANSI



(B) Conjunto Padrão de Caracteres Manuscritos - OCR Japão



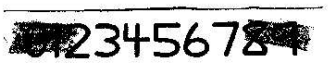
(C) Conjunto Padrão de Numerais para Sistemas OCR - Canadá

Figura 3.2 - Exemplos de padronização de conjuntos de caracteres para sistemas de OCR

O sistema N^2OCR, que é a base do estudo deste trabalho, está orientado principalmente ao reconhecimento de caracteres impressos isolados, onde pode ser considerada a utilização de múltiplas fontes de texto. Os conceitos apresentados em relação ao sistema N^2OCR podem ser facilmente estendidos para sistemas OHCR (orientados também ao reconhecimento de caracteres isolados), independentemente da entrada ser *on-line* ou *off-line*. Esta possibilidade de extensão do sistema será abordada no penúltimo capítulo. O único tipo de reconhecimento que não será abordado é o de escrita cursiva.

3.3 Descrição dos Sistemas OCR

Existem diferentes etapas de processamento das informações visando o reconhecimento de textos por um sistema de OCR. Estas etapas se estendem desde a captura da imagem do texto até a obtenção final de uma identificação deste. Em um sistema de OCR não se deve considerar apenas a parte referente às técnicas empregadas na resolução do problema de classificação, mas um item de extrema importância para sistemas de reconhecimento de caracteres é também a medição e avaliação dos resultados obtidos por estes.



(C) Conjunto Padrão de Numerais para Sistemas OCR - Canadá

Figura 3.2 - Exemplos de padronização de conjuntos de caracteres para sistemas de OCR

O sistema N^2OCR, que é a base do estudo deste trabalho, está orientado principalmente ao reconhecimento de caracteres impressos isolados, onde pode ser considerada a utilização de múltiplas fontes de texto. Os conceitos apresentados em relação ao sistema N^2OCR podem ser facilmente estendidos para sistemas OHCR (orientados também ao reconhecimento de caracteres isolados), independentemente da entrada ser *on-line* ou *off-line*. Esta possibilidade de extensão do sistema será abordada no penúltimo capítulo. O único tipo de reconhecimento que não será abordado é o de escrita cursiva.

3.3 Descrição dos Sistemas OCR

Existem diferentes etapas de processamento das informações visando o reconhecimento de textos por um sistema de OCR. Estas etapas se estendem desde a captura da imagem do texto até a obtenção final de uma identificação deste. Em um sistema de OCR não se deve considerar apenas a parte referente às técnicas empregadas na resolução do problema de classificação, mas um item de extrema importância para sistemas de reconhecimento de caracteres é também a medição e avaliação dos resultados obtidos por estes.

3.3.1 Etapas de Processamento

Em geral, os sistemas de OCR possuem implementadas as seguintes funções [SUE 80]: aquisição da imagem do texto, tratamento da imagem, localização e separação dos caracteres, pré-processamento dos padrões, extração de atributos, reconhecimento/classificação dos padrões e pós-processamento. A figura 3.3 apresenta um esquema representando as etapas citadas acima.

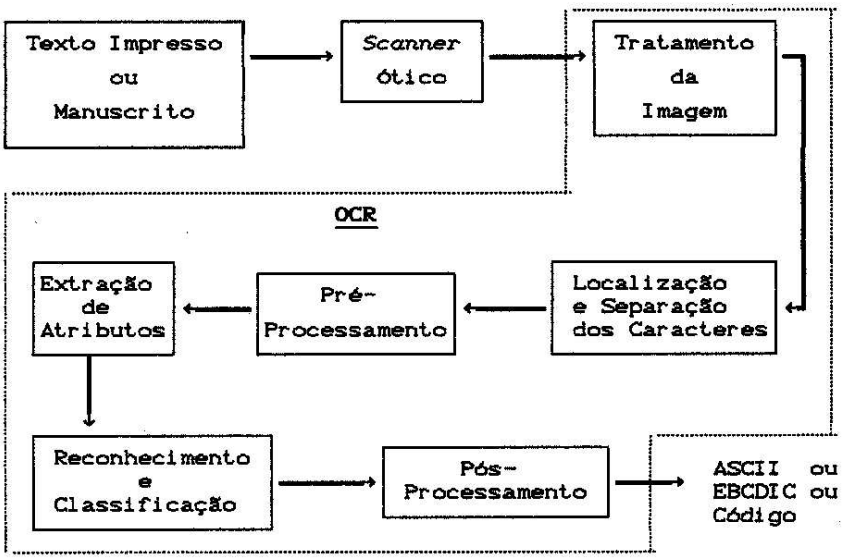


Figura 3.3 - Esquema das etapas de processamento de um sistema OCR

As etapas de tratamento da imagem, separação dos caracteres (segmentação), pré-processamento e extração de atributos, representadas na figura 3.3, são denominadas por alguns autores simplesmente de pré-processamento. Neste capítulo serão abordados os itens relacionados a partir da separação dos caracteres, visto que o capítulo 2 já abordou os itens referentes à aquisição e tratamento das imagens.

A separação dos caracteres, ou segmentação, é uma importante etapa do trabalho de reconhecimento, envolvendo a separação individual dos caracteres do resto do conteúdo da imagem digitalizada. Numa primeira etapa da segmentação torna-se importante realizar a diferenciação e separação entre imagens gráficas (desenhos, fotografias, etc) das imagens textuais. Após, as áreas que foram identificadas como sendo áreas cujo conteúdo são textos sofrem um processamento visando isolar cada caractere em uma pequena matriz de pontos. Somente a partir do momento em que for obtida uma matriz representando apenas um caractere é que este poderá ser então processado visando a sua classificação.

Para a separação dos caracteres, o algoritmo tem que levar em consideração diversos fatores, entre eles: o espaçamento entre caracteres e tamanho do caractere (fixo ou variável), a utilização de uma grade auxiliar (no caso de formulários com espaçamento bem definido), e o tipo de texto a ser reconhecido, onde podem haver caracteres que se tocam ou caracteres com descontinuidades (características muito importantes para o módulo de segmentação). Esta etapa de processamento é até hoje uma das etapas mais críticas, sendo um problema ainda não completamente solucionado [CAS 82]. Portanto, a etapa de segmentação merece uma destacada importância dentro do processo global de reconhecimento. O pré-processamento realizado após a segmentação dos caracteres consiste basicamente em realizar alguma correção em termos de distorções do caractere. O principal objetivo é normalmente o de ajustar o tamanho da matriz resultante da segmentação dos caracteres, de forma que esta tenha um tamanho compatível com o tipo de "padrões exemplos" das classes a serem identificadas. Logo, o pré-processamento envolve em primeiro lugar o ajuste de escala (normalização de tamanho), podendo incluir outros procedimentos como: correção de distorções, ajuste de posição, correção de inclinações e rotações, e até mesmo um segundo nível de tratamento de imagens. A diferença do pré-processamento para o tratamento de imagens é que neste o tratamento é realizado a nível de um único padrão e não mais sobre a imagem inteira.

Após terem sido isolados os caracteres, e preferencialmente com a realização de um ajuste de tamanho e posição pré-definidos, pode-se então realizar uma etapa de extração de atributos. Esta etapa é opcional e dependerá exclusivamente do tipo de técnica empregada no reconhecimento e classificação dos padrões. A recodificação de um padrão através de seus atributos visa ressaltar algumas características dos padrões, de forma a facilitar a sua identificação. Por exemplo, pode-se processar um padrão correspondente a um caractere, de forma a extrair segmentos de retas e curvas presentes neste. Usando esta descrição do caractere através deste novo conjunto de atributos, pode ser implementado um algoritmo de classificação cuja função de decisão leve em consideração a presença, ausência ou inter-relações entre estes novos atributos.

O reconhecimento consiste na aplicação de uma técnica para reconhecimento de padrões (como as citadas no item 2.4) do qual resulta uma classificação deste em uma das classes pré-definidas. O reconhecimento dependerá muito da técnica empregada e dos atributos levados em consideração, sendo que no próximo item serão analisadas as principais técnicas empregadas no caso específico do reconhecimento de padrões, onde estes padrões são caracteres.

Uma etapa final a ser realizada no reconhecimento de caracteres é o emprego de técnicas de validação do texto reconhecido. Nesta etapa, realiza-se um pós-processamento, baseado no contexto, visando corrigir alguns erros realizados durante a fase de reconhecimento. Este assunto será abordado posteriormente.

Todas as etapas citadas acima, compreendem um sistema de OCR o mais completo possível, sendo que, dependendo do tipo de sistema de OCR, algumas etapas podem ser omitidas ou podem sofrer algumas modificações.

3.3.2 Técnicas de Reconhecimento

Existem diferentes técnicas empregadas no reconhecimento de padrões, mas estas podem ser classificadas em três grandes categorias, técnicas baseadas em: atributos globais, distribuição de pontos e atributos geométricos e topológicos [SUE 80].

3.3.2.1 Atributos Globais

Nesta categoria, os atributos (*features*) são extraídos de cada ponto interior ao retângulo que circunscreve a matriz do caractere, a qual também pode ser denominada de *frame*. Os atributos empregados não refletem nenhuma propriedade local, geométrica ou topológica do próprio traçado. Algumas destas técnicas foram desenvolvidas originalmente para reconhecer apenas caracteres impressos.

3.3.2.1.1 Casamento de Padrões e Correlações

Esta técnica toma os atributos do estado (preto ou branco) de todos os pontos pertencentes ao *frame*. Ela mede simplesmente a similaridade entre o caractere de entrada e as referências armazenadas, pelo casamento e obtenção da correlação de pontos ou grupos de pontos do *frame*. Exemplos de emprego desta técnica são encontrados em [MAR 90a e KOZ 71].

3.3.2.1.2 Transformações e Expansões em Séries

Um problema vindo do uso dos próprios pontos como atributos é a alta dimensão do vetor de atributos usado no casamento de padrões. Para reduzir esta dimensão e extrair atributos invariantes a deformações globais (como por exemplo, rotação e translação), faz-se transformações ou expansões em séries. Transformações como de Fourier, Walsh, Haar e Hadamard, bem como expansões em séries de Karhunen-Loeve tem sido exploradas. No capítulo 2, já foi abordado um exemplo de uso da transformada de Fourier para tratamento de padrões/sinais.

3.3.2.2 Distribuição de Pontos

É uma outra forma de reduzir a dimensionalidade dos conjuntos de atributos, onde estes são derivados a partir das distribuições estatísticas dos pontos. Diferentes tipos de distribuições tem sido utilizadas, correspondendo a diferentes técnicas de reconhecimento.

3.3.2.2.1 Zoneamento

O *frame* que contém o caractere é subdividido em várias zonas sobrepostas ou não, e a densidade de pontos nestas diferentes regiões formarão os atributos. Este método é usado comumente em reconhecimento de caracteres manuscritos em letra de imprensa. Este tipo de procedimento é abordado em [SIL 90 e HUS 72].

3.3.2.2.2 Momentos

Os momentos dos pontos do caractere em relação a um centro escolhido, por exemplo, o centro de gravidade do caractere, ou de um sistema de coordenadas escolhido, são usados como atributos.

3.3.2.2.3 N-Uplas

A ocorrência de elementos pretos ou brancos ou a ocorrência da junção destes pontos são usadas como atributos. Um exemplo desta técnica é encontrado em [ULL 69].

3.3.2.2.4 Characteristic Loci

Para cada ponto pertencente ao interior do caractere, geram-se vetores horizontais e verticais. O número de vezes (um máximo de dois é usado na ordem, para limitar a dimensão do atributo) que um segmento de linha é interseccionado por estes vetores é usado como atributo. Um exemplo do emprego desta técnica é encontrado em [SIL 90].

3.3.2.2.5 Cruzamentos e Distâncias

Atributos são obtidos medindo-se o número de vezes que segmentos de linha são atravessados por estes vetores em direções específicas ou medindo-se as distâncias dos elementos ou segmentos de linha em relação a uma dada borda, tal como as bordas do *frame* que contém o caractere.

3.3.2.3 Atributos Geométricos e Topológicos

Esta técnica é baseada na extração de atributos que descrevem a geometria ou topologia de interesse no traçado do caractere. Estes atributos podem representar propriedades locais ou globais do caractere. Esta é de longe a mais popular técnica estudada pelos pesquisadores, e alguns exemplos de atributos extraídos são descritos abaixo:

1) Retas e Curvas em Várias Direções: É realizada a identificação de elementos como

$$-,\ |,\ \backslash,\ /,\ \subset,\ \supset,\ \cup,\ \cap$$

2) Pontos Finais, Intersecção de Segmentos de Reta e Laços: É realizada a identificação de elementos como

Todas as técnicas acima visam facilitar e melhorar a eficácia do reconhecimento de caracteres, através de uma melhor seleção dos atributos que representam este. Estas técnicas listadas acima servem para compor o vetor de atributos de um padrão, podendo ser a própria matriz de pontos do caractere, ou uma nova codificação obtida a partir desta, ou ainda podendo ser resultante da composição de dois ou mais conjuntos de atributos diferentes.

Estes variados métodos influenciarão diretamente a capacidade do sistema em ser imune a problemas de deformação, ruído e outras perturbações. Este aspecto será analisado no item seguinte.

3.3.3 Avaliação do Reconhecimento

A comparação do desempenho entre os diversos sistema de OCR é uma tarefa muito difícil e provavelmente os resutados obtidos nesta comparação não sejam muito confiáveis. Para cada algoritmo candidato a avaliação, sempre haverá provavelmente um conjunto de padrões de teste para o qual a avaliação será superior a de um outro algoritmo rival [NAG 83]. Baseado nestas constatações, foi desenvolvida uma série de métodos e parâmetros a serem utilizados, visando no mínimo uma melhor caracterização dos algoritmos e sistemas de OCR em geral.

Uma das formas de realizar a avaliação destes sistemas foi através da criação de bancos de dados padrões para análise de desempenho, contendo amostras variadas de padrões a serem reconhecidos por um novo algoritmo submetido à avaliação. Alguns destes bancos de amostras de padrões de caracteres se tornaram muito difundidos e utilizados na avaliação de diferentes sistemas. Dois destes bancos de dados muito conhecidos são: o de *Highleyman* e o de *Munson*. Estes bancos de dados estão disponíveis junto ao IEEE e fornecem um conjunto de caracteres manuscritos em letra de imprensa. Os dados de *Munson* possuem 12760 amostras de 46 tipos diferentes de caracteres, com uma resolução de 24 x 24 pixels, e os dados de *Highleyman* possuem 1800 amostras de letras e 500 de numerais, com um total de 36 tipos diferentes de caracteres, com uma resolução de 12 x 12 pixels. Esta resolução (12x12) infelizmente não é a mais adequada, uma vez que para o reconhecimento de caracteres manuscritos é aconselhada uma matriz com dimensões em torno de 20 x 25 pixels [SUE 80]. O maior problema referente a tais bancos de dados é a divulgação de sua existência e as formas de acesso aos mesmos. Infelizmente, para a realização deste trabalho não foi possível a utilização destas bases de dados de padrões de caracteres, devido primeiramente ao fato de serem orientados a sistemas de reconhecimento de caracteres manuscritos e, em segundo, porque não foi possível até o momento ter acesso a eles.

Espera-se, num futuro próximo, que seja possível o acesso a estas e a outras bases de dados de testes, visando uma melhor validação do trabalho desenvolvido.

Os principais parâmetros a serem tomados como referência são os índices de avaliação de desempenho de sistemas de OCR. Estes índices consistem de três indicadores: taxa de caracteres reconhecidos corretamente, taxa de caracteres substituídos (classificações incorretas) e taxa de rejeição (caracteres não identificados) [MAR 90c]. Estas taxas são as mesmas que já haviam sido relacionadas no item 2.4 referente à avaliação de sistemas de reconhecimento de padrões.

O objetivo deste texto não é o de estabelecer valores fixos a serem atingidos por estas taxas de avaliação de desempenho. Isto não é aconselhável, pois estes índices de desempenho dependem de uma série de fatores próprios de cada aplicação. O que será feito é uma apresentação com o objetivo de orientar o leitor em termos de parâmetros e itens a serem considerados, de forma a se ter uma noção do desempenho esperado para um dado sistema alvo.

Quando se trabalha com sistemas OCR, é importante se considerar as taxas referentes ao seu desempenho para que seja possível determinar a viabilidade, em termos econômicos, da aplicação prática do sistema e em relação ao objetivo de uso destes sistemas como forma alternativa de aquisição de informações pelo computador. Para que seja possível fazer este tipo de análise é importante se tomar conhecimento de alguns dados de referência muito importantes. O primeiro se refere a capacidade do ser humano em reconhecer caracteres, onde a taxa de erros na identificação dos caracteres, na ausência de contexto, é de 4% aproximadamente. Desta forma, um sistema OCR, que não se utiliza de um pós-processamento para avaliação de contexto, onde é buscado um desempenho similar ao humano, deve atigir taxas de 96% de reconhecimento correto de caracteres. Obviamente que taxas superiores a esta devem ser sempre buscadas.

Um outro dado importante é a taxa de substituição de caracteres em textos datilografados que, antes de sofrerem revisão, se situa em torno de 0.1%, para dados que possuem um certo significado contextual [MAR 90c]. Esta última taxa indica que sistemas OCR com uma eficácia igual ou superior a esta são plenamente indicados para substituir o ser humano em tarefas de transcrição de dados por digitação.

Mas então quais são as taxas aconselháveis de desempenho para um sistema OCR? É muito difícel de estimar, e inclusive são encontradas contradições entre diferentes autores. Segundo [MAR 90c], os sistemas comerciais possuem tradicionalmente taxas de rejeição mais substituição em torno de 10:1 ou 100:1. Como pode ser visto, esta faixa é muito ampla, e isto é devido ao fato de que se deve levar em conta o tipo de aplicação do sistema OCR. Um sistema de OPCR terá uma certa taxa de reconhecimento aceitável, e um sistema de OHCR terá outra, dependendo ainda do tipo de dados originais.

A seguir, ver-se-á um exemplo de avaliação a ser seguido. Uma vez determinado o tipo de aplicação, como por exemplo, um sistema de entrada automática de dados para o computador através da aquisição de textos impressos; passa-se para a análise das taxas de erros e rejeições admissíveis. Neste sistema exemplo, uma taxa de substituições e rejeições em torno de 10:1 é inaceitável, pois em cada uma ou duas palavras de tamanho médio, haverá uma letra incorreta e o trabalho de correção será muito demorado. Neste caso seria preferível digitar todo o texto diretamente no computador.

Para dar uma noção das taxas de desempenho, foi reproduzida uma tabela de valores destas representando diferentes tipos de sistemas analisados em [SUE 80]. Esta tabela apresenta o desempenho de algoritmos referenciados na literatura, onde para cada caso estão listados um exemplo de baixa, mediana e alta eficácia. Todos os algoritmos listados na tabela 3.1 são referentes a sistemas OHCR.

Tabela 3.1 - Taxas de reconhecimento obtidas por diferentes sistemas OHCR

Tipo de Reconhecimento	Autor do Algoritmo	Desempenho		
		TA	TE	TR
Numerais	Neill, 1969	85 %		
	Spantersberg, 1974	92,80%	2,7 %	4,5%
	Masterson, 1962	99,79%		
Alfanumérico	Dydyk, 1970	65 %		
	Naylor, 1971	80,40%		
	Krause, 1973	97,10%	2,9 %	0 %
Símbolos Fortran	Caskey, 1973	91,80%		
	Mori, 1975	96 %	0,5 %	3,5%
	Fujimoto, 1976	99,74%	0,06%	0,2%
Caracteres Katakana	Masuda, 1972	94,4 %	0,2 %	5,4%
	Tomita, 1967	98,1 %		
	NEC Lab., 1974	98,4 %	0,6 %	1 %
	ETL Group, 1976	98 %	0,2 %	1,8%

Obs.: TA - Taxa de Acertos TE - Taxa de Erros
TR - Taxa de Rejeições

É interessante salientar que, entre os algoritmos relacionados na tabela 3.1, todos aqueles que obtiveram o melhor desempenho em seu respectivo grupo (Masterson, Krause, Fujimoto) fizeram uso de restrições na escrita dos caracteres. Foram utilizandos nestes testes um conjunto de caracteres OCR padronizado, com o treinamento prévio dos usuários para escrever da forma solicitada! Isto demonstra como uma tabela comparativa pode ocultar certos detalhes muito importantes. Outros comparativos podem ser também encontrados em [NAG 82].

Além das taxas de avaliação de desempenho, descritas anteriormente, existem outros fatores muito importantes:

- Sensibilidade à deformação da imagem (será abordada no item 3.3.4);
- Implementação prática da técnica;
- Custo do sistema e

- Características da aplicação: tipo de caracteres, restrições em relação aos documentos digitalizados, volume de dados envolvidos e qualidade do texto ou do meio utilizado.

Quanto à implementação prática do sistema, alguns fatores importantes são: a possibilidade de obtenção automática de máscaras para casamento de padrões, a velocidade de reconhecimento em termos do número de caracteres reconhecidos por segundo, a complexidade e viabilidade da implementação prática do sistema e a independência da técnica em relação à utilização de técnicas suplementares. Outro aspecto muito importante em relação à implementação, é o grau de paralelismo empregado no sistema, o qual influi drasticamente na velocidade de reconhecimento do sistema.

Em relação ao custo de sistemas OCR comerciais, é possível encontrá-los com preços que variam de US$100 (sistemas usados junto a *handy scanners*, com custo aproximado de US$250) até sistemas de US$4000 (sistemas acoplados a leitores automáticos de páginas, com custo aproximado de US$1500). As velocidade de reconhecimento destes sistemas variam de 0,1 a 2,0 caracteres por segundo, operando com um conjunto de caracteres para reconhecimento de 50 a 3.000 símbolos diferentes [MAR 90c]. Existem também sistemas dedicados de alto desempenho que possuem taxas de até 40.000 caracteres reconhecidos por segundo e com custos bem mais elevados, em torno de US$1.000.000 [HAR 72].

3.3.4 Principais Dificuldades

Existem diversos elementos que dificultam a tarefa de reconhecimento de caracteres, influenciando diretamente no desempenho dos sistemas OCR. Os principais problemas enfrentados no reconhecimento de caracteres são:

- Deformações dos caracteres a serem reconhecidos;
- Problemas de segmentação e separação dos caracteres;
- Abrangência e tipo dos caracteres a reconhecer;
- Desempenho do sistema independente da quantidade e qualidade da massa de dados manipulada e
- Capacidade de adaptação a novos caracteres.

As deformações possíveis de ocorrer em caracteres são as seguintes: ruído, distorção, translação, rotação, mudança de escala e variação de estilo. O ruído consiste da inversão de pontos do padrão da imagem, ocorrendo normalmente na etapa de digitalização (possivelmente associado ao dispositivo de aquisição), gerando linhas disconectas, espaçamentos inexistentes, preenchimento de espaços, união de caracteres e surgimento de pontos esparsos. Um outro tipo de "ruído" muito comum é o proveniente do meio onde está escrito o texto a ser digitalizado. Isto pode ser facilmente reparado em reproduções de um original através de uma fotocopiadora. Quanto maior o número de reproduções do original, quanto mais for manipulado este material, e quanto pior for a qualidade do papel, mais a imagem digital capturada estará sujeita a conter *spots* (pontos de sujeira) no texto, que dificultarão o processo de reconhecimento.

As distorções são variações locais, arredondamento de bordas, dilatação ou compressão do padrão, torções, e todo o tipo de deformação resultante de uma perturbação do processo de digitalização. Estas distorções ocorrem muito em dispositivos tipo *handy scanner*, onde o usuário é o responsável pelo deslocamento do dispositivo de varredura. Pode ocorrer que ao realizar esta tarefa, o usuário faça pequenos deslocamentos indevidos na direção de varredura, causando assim o surgimento de distorções. Outra causa de distorções pode ser a digitalização de uma folha com dobras ou ondulações ou ainda a ocorrência de uma variação indevida da velocidade de varredura para aquisição da imagem.

Ao fazer a aquisição do texto, é difícil a obtenção de um referencial extremamente preciso, de forma a que o texto seja posicionado exatamente no ponto desejado e com a orientação de forma que as linhas do texto estejam colocadas perfeitamente na horizontal em relação ao dispositivo de aquisição. Devido aos problemas de posicionamento e orientação, os caracteres muitas vezes se encontrarão levemente rotacionados ou transladados, dificultando a sua identificação.

Além da translação e rotação, os caracteres estão sujeitos a problemas de distorções que podem provocar uma alteração no tamanho destes. A alteração de tamanho também pode ser resultante do afastamento do dispositivo de aquisição das imagens em relação ao objeto alvo, resultando uma redução ou ampliação dos padrões contidos na cena, e conseqüentemente, necessitando um ajuste de escala. Em certos casos, como no casamento de padrões onde são levados em consideração os próprios pontos do padrão, este tipo de perturbação terá que ser corrigido.

A variação de estilo pode ser considerada como uma distorção, pois causa a deformação do padrão, que assume uma nova forma, diferente daquela para a qual o sistema foi treinado. Quando ocorrem variações de estilo em um mesmo texto, o sistema de reconhecimento tem que ser capaz de identificar diferentes representações para um mesmo caractere.

Cada uma das técnicas apresentadas no item 3.3.2 terá um comportamento diferente diante da presença das várias deformações possíveis dos caracteres. A tabela 3.2 apresenta um comparativo destas técnicas, em relação ao tipo de distorção que pode ser enfrentado [SUE 80].

Tabela 3.2 - Comportamento das técnicas de reconhecimento

CRITÉRIO →	Tolerância a deformação da imagem				
MÉTODOS ↓	Ruído	Distorção	Variação de Estilo	Translação	Rotação
1. Atributos Globais					
Casamento de Padrões	+	□	-	-	-
Transform.	-	+	+	+	+
2. Distrib. de Pontos					
Zoneamento	-	□	-	-	□
Momentos	□	□	-	+	+
N-Uplas	□	-	□	-	□
Characteris. Loci	-	+	+	+	□
Cruzamento e distância	-	+	+	+	□
3. Atributos Geométric. e Topológ.	-	+	+	+	□

+ Alta Tolerância □ Média - Baixa Tolerância

Os problemas de segmentação típicos são a separação de textos e gráficos e o problema dos caracteres que se tocam - *touching characters*. A separação de textos e gráficos pode ser resolvida de uma maneira eficiente, como será demonstrado no item 3.5.1.1, o que não ocorre com os caracteres que se tocam. A união de dois caracteres pode ser causada por deformações dos caracteres ou inclusive ser resultante de características inerentes ao estilo da letra. Esta união é bastante problemática, sendo inclusive um dos motivos porque a segmentação de caracteres ainda é um problema não completamente resolvido até hoje. Uma solução adotada em relação ao problema dos *touching characters* consiste em simplesmente considerar o par de caracteres unidos, como uma classe a mais a ser reconhecida, e desta forma contornar o problema, sem na realidade resolvê-lo.

A abrangência e o tipo de caracteres a serem reconhecidos são pontos que podem aumentar bastante a complexidade de um sistema OCR. Um sistema de reconhecimento pode ser orientado apenas para o reconhecimento de numerais, o que facilita muito o processamento, uma vez que o conjunto de classes é bastante reduzido. Imagine-se agora um sistema para o reconhecimento de múltiplas fontes, para todos os caracteres ASCII imprimíveis... O número de classes é extremamente grande, possibilitando confusões entre uma classe e outra e, como se isto não bastasse, ainda há o problema das múltiplas fontes implicando vários padrões de representação para uma mesma classe. A figura 3.4 apresenta uma classificação em termos de complexidade dos sistemas OCR, em relação à abrangência do reconhecimento realizado por estes sistemas.

Menor Complexidade

Numerais	Datilografado (1 fonte)
Letras Maiúsculas (A - Z)	Impresso (+ de 1 fonte)
Todas as Letras (a-z,A-Z)	Manuscrito - Letras Isoladas
Letras e Números	Manuscrito - Escrita Cursiva
Código ASCII Completo	
Caracteres Katakana	Documento Original
Vários Alfabetos (Latino, Grego, Japonês)	Reprodução do Original (Fotocópia)

Maior Complexidade

Figura 3.4 - Complexidade dos sistemas OCR

Em relação ao número de caracteres diferentes a serem reconhecidos, fica fácil demonstrar os problemas resultantes deste aumento, bastando listar as subtituições de caracteres mais comuns em sistemas de OCR. Nota-se que só o fato de trabalhar apenas com letras ou apenas com números já possibilita uma grande redução neste tipo de erros.

Mesmo em sistemas restritos a conjuntos menores, certos erros comuns podem ocorrer, como é o caso da confusão entre letras minúsculas e maiúsculas (s, S; c, C; u, U; etc), e outros erros como os apresentados na tabela 3.3, onde a melhor solução para uma classificação correta é o emprego de correções baseadas no contexto.

Tabela 3.3 - Erros de substituição mais comuns

Troca	Motivo da confusão / O que fazer
4 ⟷ A	4 é similar o A com uma falha na base Solução : Escrita do 4 com topo aberto
2 ⟷ Z	Podem possuir a mesma forma Solução : Escrita do Z com um risco central
1 ⟷ I	Diferença muito pequena apenas em cima Solução : Escrever 1 sem risco em baixo e o I com o uso de 2 riscos nas extremidades
0 ⟷ O	Não há nenhuma diferença (apenas o contexto) Solução : Escrever o zero cortado Ø
7 ⟷ Z	O 7 e Z só se diferenciam pelo risco inferior Solução : Escrever o 7 sem risco central e o Z com o uso deste risco
Outros Erros Comuns	2 ↔ 7 ↔ Z ; U ↔ V ; C ↔ L ; *a* ↔ *d* n ↔ h ; L ↔ I ↔ 1 ; S ↔ 5 ; G ↔ 6
SOLUCAO GERAL	ACONSELHA-SE CONSIDERAR O CONTEXTO

Quanto ao desempenho dos sistemas de OCR, este é um problema difícil de ser resolvido, pois para um melhor reconhecimento há um grande acréscimo no número de informações manipuladas. No caso de sistemas *on-line*, existe uma restrição em relação ao tempo de processamento um pouco mais rígida, uma vez que não deve ser prejudicada a "taxa natural" de entrada de caracteres. Em sistemas *off-line* esta taxa de aquisição de dados pode ser controlada, mas mesmo assim o sistema deve ter um desempenho razoável, de forma a satisfazer o usuário no tipo de aplicação para o qual está voltado. Uma das formas de acelerar o processamento é através da execução de tarefas em paralelo, otimizando o processamento.

Este trabalho apresenta uma proposta que permite a aceleração do reconhecimento através do emprego de paralelismo. Este assunto será abordado em maiores detalhes nos capítulos 5 e 6.

Por fim, uma última dificuldade dos sistemas de reconhecimento de caracteres é a implementação de uma capacidade de adaptação ou aprendizado. Conforme foi apresentado, existe um conjunto enorme de tipos e estilos de caracteres e, mesmo se fosse feita uma restrição apenas a caracteres impressos, ainda assim o número de fontes para impressão seria muito grande. Uma solução para este problema, que gera uma dificuldade a mais, é a criação de um sistema com capacidade de adaptação, onde cada usuário realiza um treinamento prévio específico para o tipo de caracteres que ele deseja reconhecer. Esta fase inicial de adaptação permite que o sistema aprenda a reconhecer melhor os caracteres que lhe serão posteriormente apresentados, obtendo assim taxas melhores de reconhecimento. Um sistema adaptativo também fornece uma maior flexibilidade ao OCR, uma vez que este não precisa mais ser desenvolvido para o reconhecimento de um conjunto específico de caracteres. Este conjunto, que deveria ser o mais amplo possível para que o sistema tivesse um bom desempenho de reconhecimento, aumenta muito a complexidade do sistema, o tempo de resposta e o número de erros. Sistemas adaptativos podem aprender um conjunto mais restrito de caracteres a serem reconhecidos, específicos para cada aplicação, e desta forma possibilitam ainda uma redução quanto ao número de erros que podem ocorrer.

Uma vez que as características e dificuldades de implementação de sistemas OCR já foram apresentadas, serão abordados agora os algoritmos desenvolvidos para solucionar os principais problemas referentes ao assunto, bem como as técnicas necessárias para que o objetivo final de reconhecer caracteres seja alcançado.

3.4 Análise de Algoritmos

Este item irá resumir as técnicas consideradas de maior importância, na implementação de um sistema OCR, e mais especificamente aquelas que foram utilizadas no desenvolvimento do sistema N^2OCR.

3.4.1 Segmentação

Em relação à segmentação, serão abordados dois aspectos, separação entre textos e gráficos, e separação de caracteres entre si.

3.4.1.1 Separação de Textos e Gráficos

Um documento impresso normalmente é constituído de duas classes bem distintas de informações: os gráficos e os textos. Para realizar o reconhecimento de caracteres é importante que o sistema de OCR tenha a capacidade de distingüir entre as regiões de texto e as regiões de gráficos. O método que será apresentado é o *Run-Length Smoothing*, o qual permite a realização da tarefa descrita acima [WON 82].

Os documentos digitalizados, utilizados junto a esta técnica, são imagens do tipo B&W com figuras, desenhos e fotografias. As fotografias são representadas através de *halftoning*. O procedimento empregado na separação consiste basicamente de dois passos: o primeiro é um procedimento que subdivide a área global do documento em regiões (blocos), cada uma delas podendo conter apenas um tipo de dados (texto, figuras, fotografias, etc). A seguir, alguns atributos básicos destes blocos são calculados, permitindo que seja realizada uma classificação destes. Cada bloco será classificado de acordo com uma avaliação a respeito de dados estatísticos sobre seus atributos.

O algoritmo básico utilizado, o RLSA - *Run Length Smoothing Algorithm*, tem como objetivo original a detecção de longas linhas horizontais e verticais. Este algoritmo foi estendido para obter um *bit-map* com áreas bem definidas, representando os blocos que contém os diversos tipos de dados. O RLSA básico é aplicado sobre uma seqüência binária, na qual pixels brancos são representados por 0's e pixels pretos são representados por 1's. O algoritmo transforma uma seqüência binária *X* em uma seqüência de saída *Y* de acordo com as seguintes regras:

- 0's em *X* são trocados para 1's em *Y*, se o número de 0's adjacentes for menor ou igual a um limiar *C* pré-definido.
- 1's em *X* são mantidos inalterados em *Y*.

Por exemplo, com C=4, a seqüência *X* é transformada em uma seqüência *Y* como segue:

X: 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0
Y: 1 1 1 1 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 1 1 1 1 1

Quando o RLSA é aplicado sobre matrizes de pontos, ele possui o efeito de juntar as áreas pretas, que são separadas por menos que *C* pixels. A escolha apropriada de *C* faz com que sejam unidos todos os pontos de regiões de um tipo de dados único. O grau de união dos pontos dependerá do valor de *C*, da distribuição dos pontos pretos e brancos no documento (característica para cada tipo de dado) e da resolução de digitalização. O valor de *C* pode ser estimado experimentalmente, de acordo com as características de resolução e do tipo de dados presentes nos documentos.

O RLSA é aplicado linha a linha, coluna a coluna em um documento, obtendo dois distintos *bit-maps*. Por causa do fato de se ter uma variação independente do espaçamento horizontal e vertical dos diferentes componetes presentes nos documentos, dois valores diferentes de *C* podem ser usados (C_H e C_V).

A figura 3.5 [WON 82] apresenta um documento (a), e o resultado da aplicação do RLSA na direção horizontal (b) e na vertical (c). Os dois bit-maps são combinados através de uma operação lógica *AND*, realizada ponto a ponto, obtendo a imagem representada na figura 3.5 (d).

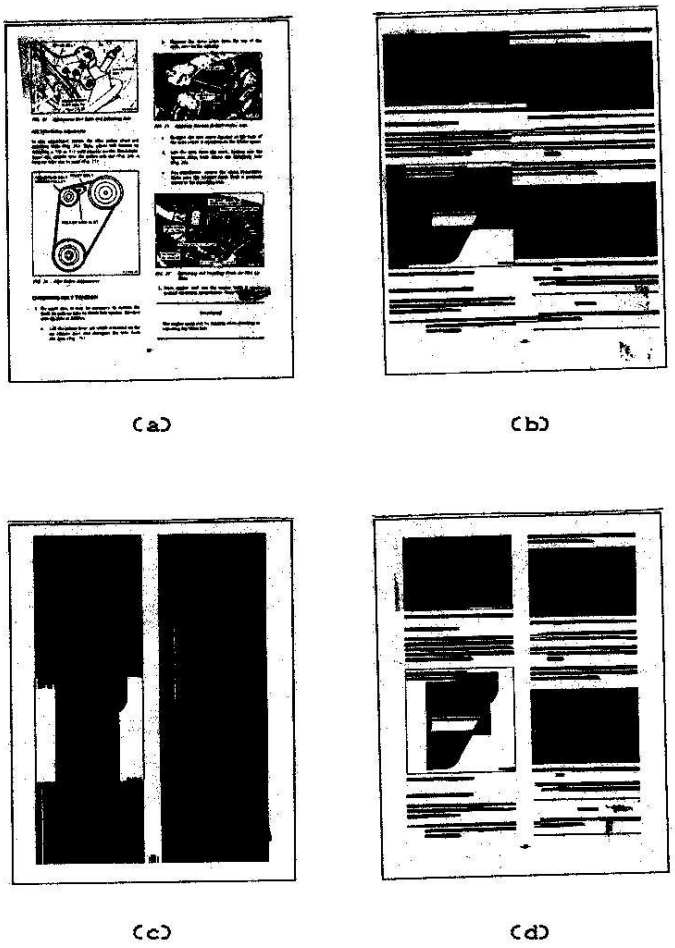


Figura 3.5 - Etapas do processamento do algoritmo RLSA

Os blocos apresentados na figura 3.5 (d) podem ser localizados e classificados de acordo com o seu conteúdo. Os blocos são isolados através de um algoritmo de identificação de regiões contínuas de pontos pretos, atuando de forma similar a do algoritmo que será descrito na seção 3.5.1.2. Para que seja possível a identificação realizada no passo seguinte, é necessário obter as seguintes medidas para cada um dos blocos:

- Número total de pontos pretos no bloco segmentado;
- Número total de pontos pretos da imagem original referentes aos pontos contidos no bloco obtido;
- Tamanho do bloco, comprimento na horizontal e vertical e
- Transições de pontos brancos para pretos da imagem original para a resultante do processamento.

De posse destes dados, pode-se então caracterizar o tipo de conteúdo de cada bloco. Segundo [WON 82], cada tipo de dado da imagem (texto, figuras, fotografias, etc), possui um tipo específico de formação, que resulta em valores característico destas medidas. Este valores podem ser determinados através da realização de experimentos onde são recolhidos dados estatísticos a respeito de cada tipo de dado. Por exemplo, é possível determinar a presença de textos com caracteres de tamanho entre 10 x 10 e 20 x 20 pixels, examinando a altura dos blocos. Uma altura entre 10 e 20 pixels indica uma boa possibilidade de que o conteúdo do bloco seja um texto. Analisando o conjunto de todas as informações extraídas a respeito dos blocos, pode ser obtida uma classificação perfeita de cada tipo de bloco. A figura 3.6 apresenta o resultado da aplicação desta técnica, para o caso dos dados obtidos a partir do documento da figura 3.5.

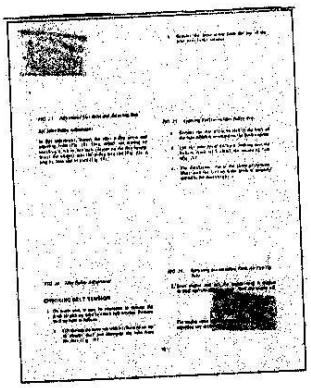


Figura 3.6 - Separação dos textos do documento da figura 3.5

Além desta técnica, existem outras que são normalmente baseadas na identificação prévia da altura e largura dos caracteres [WON 82]. Estas outras técnicas são bastante simples e podem levar a alguns erros na classificação dos elementos que compõem o documento, devido ao fato de considerarem apenas informações como a altura e a largura dos elementos extraídos. O algoritmo RLSA certamente obtém melhores resultados que estes métodos, pois este utiliza uma técnica que considera a altura e largura dos elementos constituintes de cada bloco extraído, além de outras informações complementares.

3.4.1.2 Separação de Caracteres

Depois da separação dos gráficos e textos, pode-se então realizar a separação dos caracteres um a um, de forma a obter pequenas matrizes de pontos que serão fornecidas para o processo de classificação. Existem basicamente dois tipos de processos para a separação de caracteres: separação de caracteres escritos com espaçamento fixo e separação de caracteres escritos com espaçamento variável.

Um texto pode ser escrito usando-se um espaçamento fixo entre caracteres, ou seja, todos os caracteres terão um mesmo tamanho. Este tipo de textos pode ser encontrado em impressoras matriciais, onde a impressão é realizada em modo textual sem uso de recursos gráficos.

Outra possibilidade de se ter um espaçamento fixo é através do preenchimento de um formulário, onde a pessoa que escreve está restrita a pequenos quadros, e em cada quadro deve ser escrito apenas um caractere. A separação dos caracteres neste tipo de textos é simples. Basta que seja localizado um primeiro caractere de uma linha e, a partir deste caractere, quebra-se toda a imagem em pequenos blocos de $M \times N$ pontos, sendo os valores de M e N indicados pelo usuário, ou estimados automaticamente pelo sistema a partir do primeiro caractere identificado. Os valores M e N indicam o tamanho exato de um caractere, em relação a sua altura e largura. Estes valores devem ser considerados com uma pequena margem de erro, de um ou dois pixels em ambas as direções, devido a problemas referentes a possíveis deslocamentos indevidos na escrita do texto ou até a pequenos deslocamento ocorridos na digitalização da imagem.

No caso da digitalização de textos obtidos através da impressão em uma impressora matricial, há o problema dos caracteres não serem formados por linhas e curvas contínuas, mas sim por um conjunto de pequenos pontos, que normalmente se encontram espaçados entre um e o outro. Para estes caracteres serem isolados, pode ser feito uso do fato de terem um espaçamento fixo, ou também através do uso de uma operação de *smoothing*, que permita a união entre os pontos adjacentes. O algoritmo de *smoothing* (suavização) permitirá uma aplicação posterior do algortimo de separação por continuidade, como será demonstrado logo a seguir. O problema da aplicação de um algoritmo de *smoothing* é que pode ser causada não somente a união dos pontos internos de um caractere, mas também dos pontos pertencentes a dois caracteres. Esta união entre dois caracteres (*touching characters*) irá provocar um erro mais difícil ainda de ser tratado. Como pode ser visto, tais algoritmos exigem que se tenha a possibilidade de realizar ajustes mínimos em relação a parâmetros de funcionamento, tais como o tamanho estimado do caractere, o grau de *smoothing* aplicado na imagem, o tipo de espaçamento, etc.

Isto leva à utilização de sistemas parametrizáveis, o que é feito normalmente através do uso de um arquivo de configuração, a exemplo do utilizado no sistema N^2OCR.

Será abordado em maiores detalhes, neste trabalho, o reconhecimento de caracteres com espaçamento variável, pois estes constituem a maioria dos textos impressos ou manuscritos que se manipula. Além disto, o algoritmo de separação de caracteres com espaçamento fixo é trivial, bastando quebrar a imagem em pequenos pedaços, a cada $M \times N$ pixels.

A primeira etapa na separação de caracteres com espaçamento variável é a localização do início da linha de texto. O caractere é localizado buscando-se um pixel com a cor de escrita deste, a partir do topo (canto superior esquerdo) da matriz de pontos que compõe a imagem. Ao ser localizado um ponto que pertence a um caractere, pode então ser realizada a sua extração. A extração de um caractere segue o princípio de que cada caractere é composto por um conjunto de pontos unidos entre si (contínuos). Após serem extraídos os pontos que estão unidos entre si e que formam um caractere, segue-se na direção horizontal em busca de mais um ponto da cor dos caracteres, que ao ser localizado é extraído da mesma forma que o anterior. Este processo se repetirá até que se chegue ao limite extremo direito da imagem, sendo após iniciada uma busca de uma nova linha inferior a que foi identificada e que sofreu a separação. O algoritmo da figura 3.7 apresenta de uma forma global o processo de separação de caracteres com espaçamento variável.

```
Início
  X=0; Y=0;
  Enquanto Não Terminar a Imagem (X < Xmax, Y < Ymax)
  [ Acha_Início_da_Linha (X,Y);
  [ Enquanto Não Terminar a Linha
  [   [ Isola_Letra (X,Y);
  [   [ Se Em_Treinamento Então Rotula_Letra;
  [   [ Salva_Letra;
  [   [ Avança_Próxima_Letra (X,Y);
  [ Avança_Próxima_Linha (X,Y);
Fim
```

Figura 3.7 - Algoritmo geral de separação de caracteres

A rotina de "Acha_Início_da_Linha" consiste em dada uma coordenada Y de base para o início da procura, e sendo a coordenada X considerada igual a zero, faz-se uma varredura da imagem, da esquerda para a direita, de cima para baixo em busca de um pixel com a cor de escrita dos caracteres. Como pode ocorrer de uma letra do meio da linha aparecer com pontos mais altos em relação às demais, a varredura ocorre simultaneamente na horizontal e na vertical de acordo com valores de deslocamento previamente especificados.

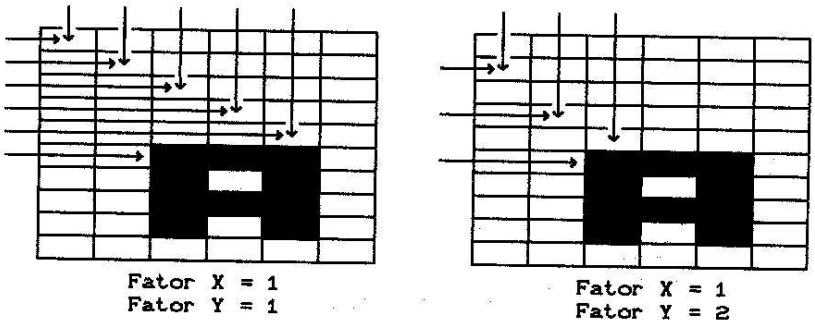
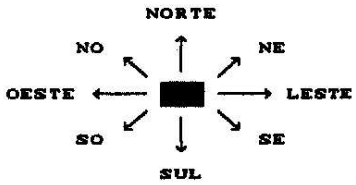


Figura 3.8 - Localização do início de uma linha

A tarefa de identificar o início de uma linha leva em consideração dois fatores configuráveis (fator x e fator y, conforme a figura 3.8) que se encaixam dentro do conceito de um sistema totalmente configurável, de acordo com o tipo de documento analisado. Isto é necessário devido ao fato de os documentos apresentarem diferentes organizações e distribuições do texto em seu interior. Além destes dois fatores podem ser definidos outros, como por exemplo, um fator máximo para o incremento em X e Y. Caso seja importante manter a formatação do texto (o que normalmente é necessário), deve ser possível ao sistema detectar também quando há a presença de espaços em branco ou linhas em branco no texto. Novamente será necessária a definição de valores máximos configuráveis para que se considere X pixels em branco, como sendo um espaço em branco, e Y pixels em branco, como sendo uma linha em branco. O valor de X e Y deve ser um pouco superior ao maior tamanho de um caractere. Todos estes fatores que foram apresentados aqui serão detalhados posteriormente no capítulo que descreve o sistema N^2OCR.

Uma vez identificada uma posição na qual se encontra um ponto pertencente a um caractere, é necessário realizar uma operação de extração do mesmo. Esta tarefa é realizada pela função "Isola_Letra", que recebe como parâmetro uma coordenada de um ponto pertencente à letra e devolve uma matriz de pixels contendo o caractere isolado. O algoritmo empregado para isolar todo o caractere se encontra listado na figura 3.9.

```
Início - Isola_Letra
  Se Pixel (X,Y) é da Cor do Caractere
  Então Faz
     [  Copia Pixel (X,Y) para o Buffer
        Apaga Pixel (X,Y) da Imagem
        Isola_Letra (X+1,Y)
        Isola_Letra (X,Y+1)
        Isola_Letra (X+1,Y+1)
        Isola_Letra (X-1,Y)
        Isola_Letra (X,Y-1)
        Isola_Letra (X-1,Y-1)
        Isola_Letra (X+1,Y-1)
        Isola_Letra (X-1,Y+1)
Fim
```



Pesquisa Recursiva

Figura 3.9 - Algoritmo para isolar um caractere

Existem dois problemas básicos que podem ocorrer neste processo de extração de um caractere: caracteres unidos e caracteres compostos por porções isoladas. O caso dos caracteres unidos ou *touching characters*, conforme já foi descrito, será tratado como um único padrão, ou seja, a extração considera ambos os caracteres unidos como sendo um único caractere. No caso de caracteres com porções isoladas, é necessário realizar algumas alterações no algoritmo básico de extração de forma que este consiga extrair corretamente este tipo de caracteres. Um exemplo típico deste tipo de problema são os caracteres com acentuação, onde o acento faz parte do caractere mas está separado do mesmo, ou casos em que, devido a um ruído na digitalização, o caractere ficou com porções desconexas (figura 3.10). A alteração do algoritmo consiste em se determinar um MBB (Menor retângulo envolvente - *Minimum Bounded Box*) do caractere extraído pelo processo descrito anteriormente, adicionar uma margem extra de pontos superiores e inferiores, e varrer todo o interior desta região extraindo qualquer pixel que ainda não tenha sido lido. A implementação desta função deve ser muito cuidadosa pois pode acabar provocando erros indesejáveis como os *touching characters*.

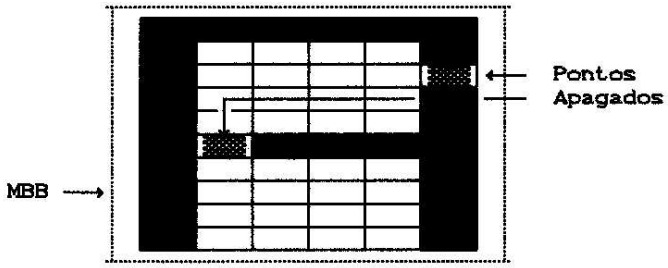


Figura 3.10 - Extração de um caractere desconexo

Ao isolar o caractere, alguns dados são guardados, como por exemplo, as coordenadas Y inferior e superior máximas que este ocupava na imagem. Baseado nestes valores, segue-se em direção à direita, verificando os valores dos pixels que se encontram dentro desta faixa de linhas, até que se encontre um novo pixel da cor do caractere. Neste ponto, realiza-se novamente a extração do caractere, repetindo o processo de busca do próximo e extração até que se chegue ao canto direito final da imagem (ou do bloco extraído do documento de acordo com o algoritmo descrito na seção 3.5.5.1).

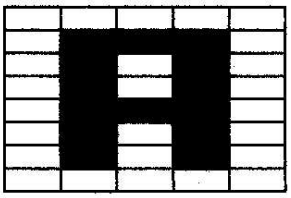
Do processo de separação de caracteres, resulta portanto um conjunto de pequenas matrizes contendo apenas um caractere cada uma (ou dois, caso haja caracteres que se tocam). Estas matrizes servirão como dados de entrada para o processo seguinte do reconhecimento.

3.4.2 Extração de Atributos

Serão apresentados nesta seção alguns dos métodos de extração de atributos de interesse para este trabalho descrevendo-se a sua implementação em maiores detalhes. Estas técnicas apresentadas, recebem denominações próprias dada pelo autor, mas todas elas podem ser classificadas (com nomes iguais ou não) de acordo com a classificação apresentada na seção 3.3.2.

3.4.2.1 Vizinhança Global

Consiste na contagem dos pixels do caractere que são vizinhos a outros pontos do mesmo, em relação a uma dada direção. A vizinhança pode ser imediata ou pode também ser considerada uma vizinhança em segundo grau (distância maior que 1).



Vizinhança Global Imediata

Norte = 10
Noroeste = 3
Nordeste = 3
Leste = 4

Sul, Oeste, Sudeste e Sudoeste são redundantes a estes dados

Figura 3.11 - Exemplo de vizinhança global

3.4.2.2 Vizinhança Local

Realização de uma codificação para cada tipo de subconjunto de pontos do padrão (usualmente 3 x 3 pontos). Após definidos os tipos de padrões possíveis, é realizada uma contagem da ocorrência de cada um destes padrões no caractere.

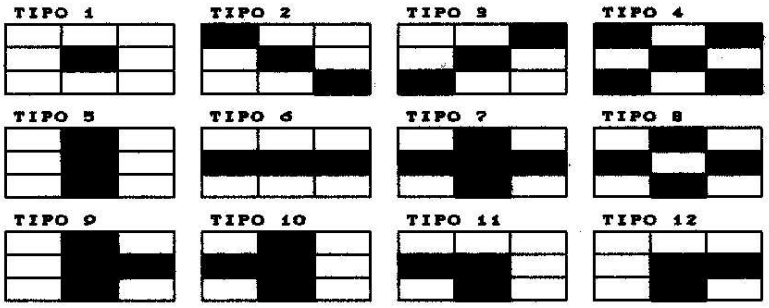


Figura 3.12 - Exemplos de padrões de vizinhança local

Para o caractere representado na figura 3.11, os valores obtidos utilizando-se as matrizes representadas na figura 3.12 são:

- Tipos 1 a 8 : 0 ocorrências
- Tipos 9 a 12 : 1 ocorrência

3.4.2.3 Projeção

Consiste em fazer a contagem do número de pixels pertencentes ao caractere, encontrados em uma certa direção (horizontal, vertical, diagonal esquerda e diagonal direita) de análise do padrão.

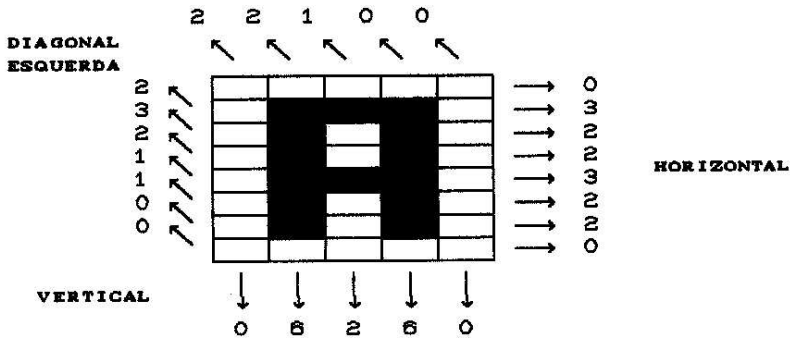


Figura 3.13 - Exemplo de cálculo de projeções

3.4.2.4 Alternância

É similar à projeção mas, ao invés de contar os pixels, conta-se o número de alternâncias de cores, ou seja, conta-se o número de cruzamentos por linhas. Para o caso do caractere representado na figura 3.13 serão obtidos os seguintes valores:

- Alternância Horizontal : 0, 2, 4, 4, 2, 4, 4, 0
- Alternância Vertical : 0, 2, 4, 4, 2, 0
- Alternância Diagonal Esquerda: 0, 0, 2, 2, 4, 2, 4, 4, 2, 2, 0, 0
- Alternância Diagonal Direita : 0, 0, 2, 2, 4, 2, 4, 4, 2, 2, 0, 0

3.4.2.5 Simetria

A partir da definição das projeções é possível definir se o padrão possui alguma simetria. A simetria tanto pode ser relativa, se for tomada em relação a uma linha imaginária central horizontal, vertical ou diagonal, como também pode ser absoluta, se for considerado o padrão como um todo. A figura 3.14 demonstra estas possibilidades.

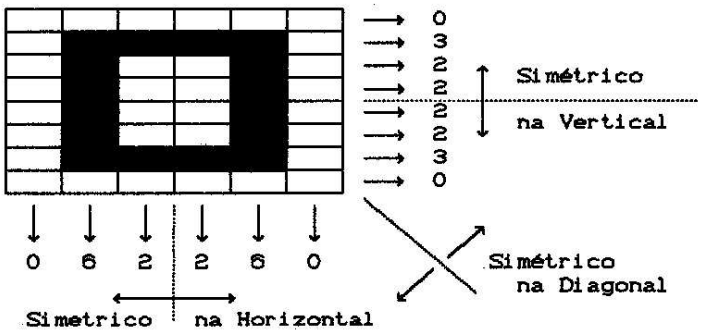


Figura 3.14 - Exemplo de cálculo de simetria

3.4.2.6 Densidade

Uma técnica muito empregada que obtém ótimos resultados consiste em calcular a densidade de pixels em sub-regiões do padrão (figura 3.15). Através desta técnica, é detectada a presença ou ausência de pontos em posições de relevada importância para a identificação dos caracteres.

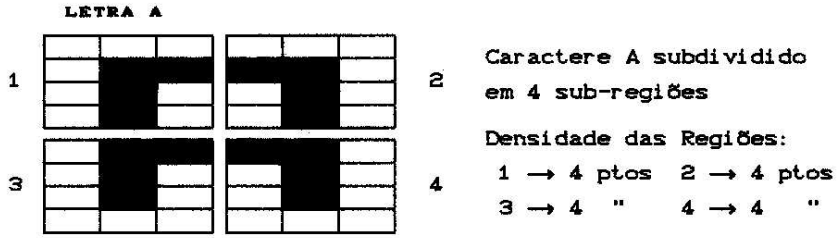


Figura 3.15 - Exemplo de cálculo da densidade de sub-regiões

3.4.3 Alteração de Escala e Posição

Os caracteres que foram isolados através da aplicação do algoritmo de segmentação apresentado na seção 3.5.1 tem diferentes tamanhos. Para que seja possível a utilização destes padrões junto a um sistema de classificação baseado no casamento de padrões, é necessária a implementação de um módulo para padronização de tamanho e posição dos caracteres.

Quanto ao ajuste de posição, todos os caracteres extraídos da imagem (com algumas poucas exceções) devem ser ajustados pelo seu canto superior esquerdo, o qual é tomado como referencial. Existem outras formas de ajuste de posição, como por exemplo, em relação ao centro deste, mas que não serão abordadas neste trabalho. Antes de realizar o ajuste de posição, é realizado um teste em relação ao tamanho do padrão extraído:

1) Se este tamanho for superior ao tamanho utilizado na classificação, então o ajuste será realizado pela extremidade superior esquerda seguindo-se uma alteração de redução do padrão.

2) Se este tamanho for igual ao tamanho utilizado na classificação, nenhuma alteração é feita;

3) Se o tamanho for menor que o tamanho utilizado na classificação, é feito um "ajuste especial" de posição, sem realizar uma alteração do tamanho real do caractere extraído. Este caso ocorre em caracteres como : ponto, vírgula, aspas, etc.

O "ajuste especial" de posição consiste em manter as características essenciais do caractere extraído, de forma a possibilitar a sua identificação. Caso seja feito sempre o ajuste pelo canto superior esquerdo, tanto a vírgula (",") como o apóstrofo ("'") serão identificados como um único caractere, não possuindo nada que os diferencie.

Nestes casos, o padrão possui um posicionamento onde deve ser mantida uma relação com a linha original em que ele estava inserido, para que seja feita uma classificação correta. Este processo deve ser realizado junto à separação dos caracteres, pois é nesta etapa que será possível identificar o posicionamento do caractere em relação à linha em que está contido. O processo é bem simples, em vez de ajustar a dimensão do caractere, cujo tamanho é sensivelmente inferior ao tamanho médio dos outros caracteres obtendo-se assim uma matriz de dimensões reduzidas, faz-se a extração do mesmo, mantendo a sua posição relativa a linha (altura), e preenchendo com pixels da cor de fundo o restante da matriz, de forma que esta tenha o mesmo tamanho dos caracteres "normais" da linha. Este tipo de procedimento está demonstrado na figura 3.16.

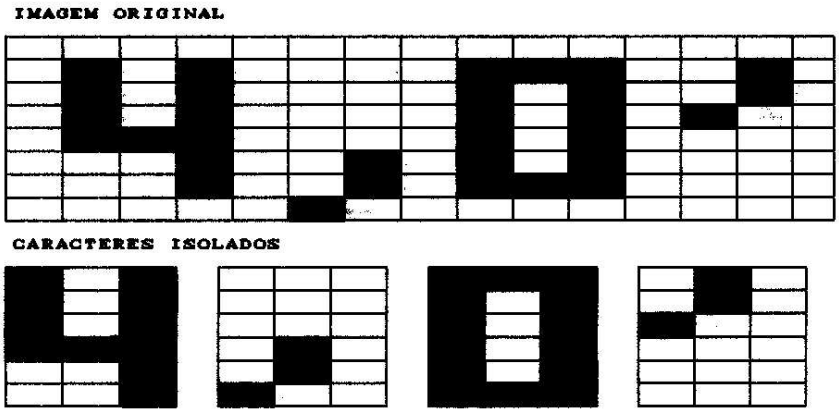
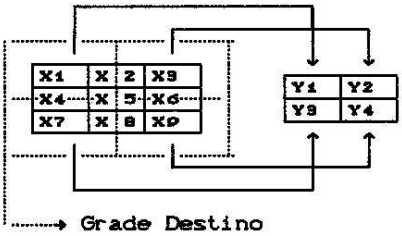


Figura 3.16 - Separação e posicionamento de caracteres

O ajuste de escala deve realizar uma conversão de um padrão de tamanho M x N para outro padrão J x K equivalente, onde M, N, J e K são números inteiros quaisquer. É interessante notar que J e K não precisam ser múltiplos exatos de M e N, o que implica na criação de um algoritmo um pouco mais complexo. A figura 3.17 apresenta um exemplo de mapeamento de um padrão 3 x 3 para outro 2 x 2.



$Y1 = 1.X1 + 0,5.X2 + 0,5.X4 + 0,25.X5$

$Y2 = 0,5.X2 + 1.X3 + 0,25.X5 + 0,5.X6$

$Y3 = 0,5.X4 + 0,25.X5 + 1.X7 + 0,5.X8$

$Y4 = 0,25.X5 + 0,5.X6 + 0,5.X8 + 1.X9$

Máximo = 1+0,5+0,5+0,25 = 2,25

Se Yi >= 50% do Máximo
Então Yi ← 1
Senão Yi ← 0

Figura 3.17 - Redução de um padrão 3 x 3 para 2 x 2

O algoritmo, apresentado na figura 3.18, que realiza a alteração da escala é uma generalização do processo descrito através da figura 3.17. O primeiro passo é a determinação de duas grades, uma original (grade da matriz do padrão) e uma destino (grade do padrão com a escala alterada). A sobreposição destas duas grades determinará a influência (ponderação) que cada pixel da grade origem terá sobre um pixel da grade destino. No caso de uma redução, existe um ou mais pixels da grade origem a influenciar a cor resultante do pixel correspondente na grade destino e, no caso de uma ampliação, existe um pixel da grade original a influenciar um ou mais pixels da grade destino. O cálculo da cor final dos pixels na grade destino ocorre da seguinte forma: se o somatório da multiplicação dos valores dos pixels da grade origem (referentes a um pixel da grade destino) pela sua influência for maior que 50% do valor máximo possível de ser atingido neste somatório, então o pixel terá o valor 1 (aceso) e, em caso contrário, terá o valor 0 (apagado). Esta técnica, na forma em que está descrita, deve ser empregada apenas em imagens binarizadas.

SE $\sum$ (Pixel_Origem · Influência_no_Pixel_Destino)
>= (0,5 · $\sum$ (1 · Influência_no_Pixel_Destino))
ENTÃO Pixel_Destino = 1
SENÃO Pixel_Destino = 0

Figura 3.18 - Algoritmo de alteração de escala

Este algoritmo produz resultados razoáveis que satisfazem o tipo de aplicação que está sendo abordado. É claro que algumas deformações podem ocorrer na aparência do caractere, mas uma vez que para um mesmo padrão, a mesma deformação será gerada, então este problema não afetará o reconhecimento de caracteres. Existem outros algoritmos de mudança de escala, como o apresentado em [CAS 82], que não provocam deformações, mas a sua implementação implica tempos de processamento muito altos, incompatíveis com o desempenho esperado de um sistema OCR.

3.5 Perspectivas

Os sistemas de OCR desenvolvidos até o presente momento possuem um bom desempenho, mas ainda há muito a ser feito, pequisado e melhorado, para que se tenha sistemas com uma capacidade de reconhecimento igual ou superior a dos seres humanos. Dentre os diferentes aspectos abordados neste capítulo alguns se destacam, indicando tendências das pesquisas atuais em sistemas OCR. São eles: a capacidade de adaptação, o processamento paralelo e o uso de contexto.

3.5.1 Adaptação

O ser humano possui uma grande capacidade de adaptação ao meio em que está inserido. Através do aprendizado, ele consegue se adaptar a novas situações. No caso específico do reconhecimento de caracteres, o ser humano possui uma grande facilidade para o aprendizado de diferentes conjuntos de símbolos, podendo desta forma ampliar o conjunto de elementos que ele reconhece. Além de aprender novos padrões, o ser humano também possui a capacidade de se "ajustar" a um tipo de escrita, acostumando-se a interpretá-la, mesmo que esta a princípio pareça ilegível.

Esta capacidade de aprendizado e adaptação tem levado muitos pesquisadores a buscar formas de incorporá-las aos sistemas de OCR. Com isto, tais sistemas se tornam mais flexíveis e possivelmente terão seu desempenho melhorado. Afinal nada mais lógico do que, na busca de um sistema que simule o comportamento humano de ler textos, se fazer a implementação de recursos similares aos que o ser humano utiliza para realizar este reconhecimento.

Estes sistemas dotados de aprendizado e adaptação podem ser implementados de diferentes formas, como por exemplo, a descrita em [MAR 89]. Uma das maneiras muito interessantes de dotar um sistema OCR com estas características é através do emprego de redes neurais, as quais serão apresentadas no capítulo 4. Um dos motivos da adoção das redes neurais como forma de implementação do sistema N^2OCR foi justamente pelo fato delas possibilitarem que o sistema tenha esta característica de aprendizado e adaptação.

3.5.2 Processamento Paralelo

Como já foi citado, o processamento paralelo possibilita o aumento do desempenho dos sistemas, através da execução simultânea de tarefas. No caso do casamento de padrões, a possibilidade de acelerar o processo de classificação através do uso de processamento paralelo fica muito clara. Uma vez que, neste processo, é necessário comparar um padrão não identificado com N outros padrões modelos, a fim de identificar qual é o mais parecido, pode-se realizar em paralelo a tarefa de comparação entre padrão desconhecido *versus* padrões modelo, já que não há uma dependência entre uma comparação e outra.

O processamento paralelo tem se tornado uma alternativa viável devido ao desenvolvimento atual da tecnologia, que possibilitou o surgimento de novos equipamentos baseados no processamento paralelo.

Em vista deste fato, nesta última década, muitas aplicações começaram a fazer uso de equipamentos que possibilitassem este tipo de processamento. Os sistemas OCR não são uma exceção, podendo também ser implementados em equipamentos que possuam uma arquitetura paralela, e que possibilitem uma aceleração do processamento.

As redes neurais são uma proposta de um novo modelo de computação, baseado no processamento paralelo. Este é mais um dos motivos que levaram à sua utilização junto ao sistema N^2OCR.

3.5.3 Uso de Contexto

A tarefa de reconhecimento pode ser desenvolvida de forma a permitir um pós-processamento. A idéia básica consiste de, na classificação, indicar não apenas a classe mais adequada à classificação do padrão, mas sim um conjunto de classes mais prováveis às quais o padrão pode pertencer. Este conjunto de classes possíveis serão avaliadas, conforme o contexto em que estão inseridas, de forma a selecionar a melhor destas classes.

O contexto também pode ser utilizado em uma fase de pré-processamento onde, antes mesmo de iniciar o reconhecimento do padrão, já se possui uma lista de prováveis candidatos ao reconhecimento. Com isto, é reduzido o espaço de soluções possíveis, facilitando o reconhecimento e diminuindo a possibilidade de ocorrência de erros. No entanto, a forma mais empregada em relação ao uso de contexto, é o pós-processamento.

Em resumo, o emprego do contexto para a correção do texto reconhecido em uma etapa de pós-processamento é uma importante etapa do reconhecimento, que atua em um nível mais alto do processamento das informações. A idéia de contexto consiste em se observar características gramaticais, sintáticas ou semânticas do texto em reconhecimento. Com o uso do contexto pode-se corrigir muitos erros, melhorando sensivelmente o índice de reconhecimentos corretos [HAR 72].

Sabe-se que o contexto é utilizado naturalmente pelo ser humano ao realizar uma tarefa como a leitura de um texto, sendo que já foi demonstrado que o ser humano causa cerca de 4% de erros quando realiza a leitura de um texto na ausência de contexto [SUE 80]. O homem é também capaz de remontar um texto, sem numerais e inserido em um contexto, mesmo que este possua 25% dos caracteres removidos randomicamente [HAR 72]. Deve-se salientar que a noção de contexto está intimamente ligada à língua utilizada na representação do texto (inglês, português, alemão, etc), e portanto estudos desta natureza devem levar em consideração aspectos lingüísticos.

Assim como o homem utiliza o contexto no reconhecimento de caracteres, é portanto interessante que um sistema de OCR siga este exemplo e também se utilize deste tipo de informações para melhorar o seu desempenho. O problema da implementação de um sistema OCR com uso de regras de validação do texto reconhecido através do contexto consiste em se determinar este conjunto de regras. Como já foi citado, estas regras estão intimamente ligadas a aspectos lingüísticos e, portanto, sua determinação requer amplos conhecimentos em relação a esta área de estudos. Um trabalho de estudo de aspectos lingüísticos visando a aplicação do contexto na correção de textos resulta em um estudo extenso e que escapa do escopo deste trabalho. Desta maneira, será considerada apenas a implementação do sistema N^2OCR de tal forma que este possa ser estendido por um módulo complementar para a validação por contexto. Este módulo não será estudado nem implementado na versão atual do sistema, sendo apenas descrita, no capítulo 5, a forma como pode ser adicionado ao sistema N^2OCR este módulo para validação do texto reconhecido através do contexto.