GERAÇÃO DE TEXTOS A PARTIR DE MATERIAL IMPRESSO

A tecnologia atual permite que sejam gerados textos em formato digital a partir de material originalmente impresso. O objetivo principal de tais procedimentos é o de permitir que documentos originalmente impressos, não disponíveis em formato digital, possam ser editados, corrigidos, reformatados, armazenados em mídia digital e até distribuídos entre leitores com interesses afins. O processo de conversão exige tanto equipamentos quanto programas específicos, que podem exigir mais conhecimento do utilizador conforme o grau de dificuldade na sua utilização. Deve-se considerar, ainda, que o processo de conversão não é isento de erros, nem é 100% correto. Em alguns casos, pode ser necessário digitalizar novamente o documento e tentar outros ajustes na imagem.

O equipamento necessário inicial exige, certamente, um computador executando um sistema operacional gráfico. No nosso exemplo, estaremos utilizando o MS Windows XP e um computador padrão PC-IBM com a configuração recomendada pela Microsoft para esse sistema. Além disso, precisamos de:

1. um digitalizador de imagens, ou Scanner.
2. um programa para edição de imagens digitadas no computador.
3. um programa para reconhecimento ótico de caracteres, ou OCR[1]

Os dois últimos itens podem estar disponíveis no pacote de programas que acompanha o scanner, geralmente no próprio CD com os programas de instalação.

AS ETAPAS DA CONVERSÃO

As etapas da conversão de imagens em texto implicam, inicialmente, na identificação do texto a ser convertido. Portanto, de posse do material impresso, do equipamento e dos programas acima, temos:

SCANNER	Conversão do conteúdo do material impresso em imagem digitalizada
Editor de imagens	Tratamento da imagem digitalizada (se necessário) para facilitar leitura pelo software OCR
OCR	conversão final da imagem gerada em texto compatível com editores de texto conhecidos (World, etc.)

Em muitos casos, o próprio OCR inclui ferramentas tanto para acionamento do scanner (digitalização da imagem) quanto edição do arquivo obtido, mas o controle do

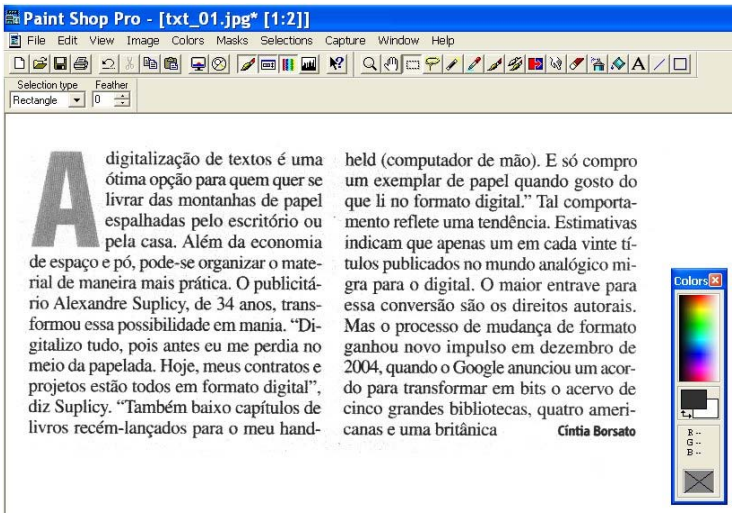
[1] OCR – Optical Character Recognition, ou Reconhecimento Ótico de Caracteres

processo pelo operador pode ser mais limitado do que aquele oferecido pelas etapas em separado.

SOFTWARE DE EDIÇÃO DE IMAGENS

Existem diversos programas para edição de imagem, e mesmo o Windows oferece um editor básico, o Paint. Um outro programa de edição de imagens comum, mas que oferece os recursos necessários para preparação de imagens para os programas OCR é o Paint Shop Pro. É necessário pesquisar o CD que acompanha o scanner pois ele certamente conterá um editor de imagens adequado para captura genérica.

A figura a seguir apresenta uma imagem da interface do programa Paint Shop Pro com o resultado do uso de um scanner sobre um texto impresso.



SCANNER – IMAGEM DIGITALIZADA NO COMPUTADOR

A finalidade básica de um scanner convencional é a de converter uma imagem regular, a partir de uma imagem arbitrária, em um arquivo de computador que pode ser armazenado, transferido e utilizado futuramente. Se a imagem gerada contiver texto impresso, ela será armazenada da mesma forma, como se fosse uma 'fotografia digital' do documento impresso. Esse tipo de informação não é adequado para armazenamento de textos, pois ocupa muito espaço em disco e é de difícil edição, ficando quase impossível a alteração de suas características visuais (tipo de fonte, tradução, etc.). Uma vez convertido em imagem digital, o documento original pode ser arquivado e sua imagem no computador pode ser tratada, ou seja, 'melhorada', de forma a poder ser convertida em texto.

Ainda sobre os recursos do Scanner, as recomendações básicas para conversão de imagem em texto são:

- Gerar arquivos com imagens em tom de cinza; (a partir do editor de imagens)

- Manter a resolução entre 300 e 400 dpi[2];
- Verificar se o texto da imagem gerada é de fácil leitura[3];
- Eliminar elementos que possam interferir na leitura (figuras, símbolos, etc.);
- Salvar o arquivo com formato compatível com o OCR (PCX, TIFF, GIF, etc.);
- Utilizar o OCR e verificar se o arquivo de imagem gerado é facilmente convertido em um arquivo texto; caso contrário, tente novos ajustes ou digitalize com maior resolução;
- O arquivo texto gerado pelo OCR deverá ser salvo em um formato padrão (geralmente RTF[4]) e editado em um editor específico (tipo Word, do MS Office).

UTILIZANDO O SCANNER

Cada fabricante utiliza um software específico de controle (interface) para cada modelo de scanner que é colocado no mercado. São poucos os casos de diferentes scanners que utilizam o mesmo software de controle, mesmo que sejam do mesmo fabricante. Isso se deve principalmente ao fato de que diferentes scanners apresentam diferentes recursos e funcionalidades.

A interface do scanner Genius Color Page Vivid 1200XE pode ser vista abaixo.

Já o Scanner TCE S510 pode ser utilizado no Windows XP com a interface do Scanner AVISION modelo AV630C, apresentada abaixo.

[2] dpi = dots per inch square, ou pontos por polegada quadrada
[3] Se o texto pode ser lido facilmente pelo utilizador, a chance do OCR conseguir gerar um arquivo é maior
[4] RTF- Rich Text Format, formato criado pela Apple Computers, pode ser lido pelo Word

Em ambas as interfaces é possível:

- Digitalização de imagem colorida, em tom de cinza ou preto-e-branco;
- Resolução da imagem (disponível na janela Propriedades Avançadas da interface Genius);
- Ajuste de brilho e de contraste (disponível na janela Propriedades Avançadas da interface Genius)
- Visualização (Preview) e seleção de partes da imagem antes da digitalização final

A seguir podem-se visualizar três imagens. A primeira foi gerada com cores a partir de um original impresso. A segunda teve os elementos desnecessários removidos pelo programa de edição de imagens complementar. A terceira é uma representação gráfica de parte do texto que foi gerado pelo OCR Fine Reader. A imagem representa o texto gerado sem edição.


FAÇA VOCÊ

SCANNER + COMPUTADOR

COMO SALVAR

ONDE GUARDAR

MANDE FAZER

LEIA NA WEB


SCANNER + COMPUTADOR

COMO SALVAR

ONDE GUARDAR

MANDE FAZER

LEIA NA WEB


SCANNER + COMPUTADOR

COMO SALVAR

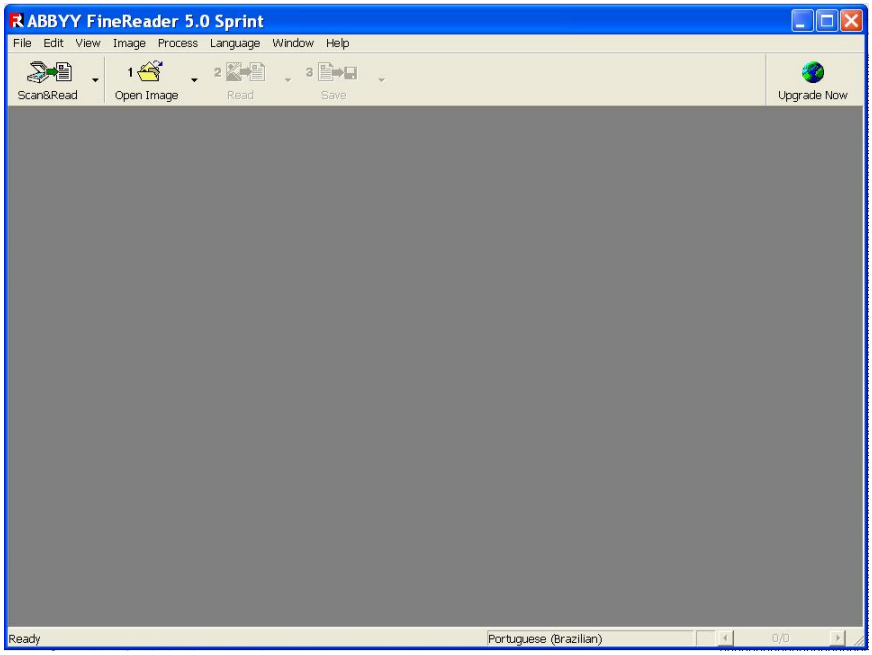
ONDE GUARDAR

MANDE FAZER

Uma vez que a imagem é gerada e adequada para leitura com o OCR, deve-se salvar no formato de arquivo adequado.

UTILIZANDO O OCR

O programa denominado OCR executa uma tarefa comum para o ser humano, mas que pode atingir graus de complexidade no sistema computadorizado não facilmente compreensíveis para aqueles que não lidam diretamente com seu desenvolvimento e criação. Um OCR deveria ser capaz de converter o conteúdo de uma imagem armazenada em um arquivo em texto digitado, compatível com os editores de texto disponíveis no mercado.



Como existem diversos tipos de OCR no mercado, não será possível gerar uma referência global e 100% abrangente. Utilizaremos o FineReader, um versátil OCR, utilizado nos exemplos contidos nesse texto.

Ao ser iniciado, o Fine Reader apresenta a interface apresentada ao lado para utilização.

Essencialmente, um OCR permite que um arquivo de imagem (contendo informações que possam ser “lidas”) seja aberto através de uma seqüência comum do Windows: Arquivo, Abrir Imagem (File, Open Image, se for em Inglês).

Esse ícone permite acesso direto para abrir arquivos do tipo imagem. Os arquivos compatíveis serão mostrados e, após selecionado e aberto um deles, o FineReder procede a uma rápida verificação.

Após aberto o arquivo, esse ícone fica ativo e permite o início do processo e leitura da imagem, para geração do arquivo texto.

Esse ícone indica que o arquivo de imagem já foi convertido para texto digitado e que pode ser salvo adequadamente em mídia (disquete, disco rígido, pendrive, etc.).

No exemplo anterior, o texto resultante, depois de editado, pode ter a seguinte forma:

A digitalização de textos é uma ótima opção para quem quer se livrar das montanhas de papel espalhadas pelo escritório ou pela casa. Além da economia de espaço e pó, pode-se organizar o material de maneira mais prática. O publicitário Alexandre Suplicy, de 34 anos, transformou essa possibilidade em mania. "Digitalizo tudo, pois antes eu me perdia no meio da papelada. Hoje, meus contratos e projetos estão todos em formato digital", diz Suplicy. 'Também baixo capítulos de livros recém-lançados para o meu handheld (computador de mão). E só compro um exemplar de papel quando gosto do que li no formato digital." Tal comportamento reflete uma tendência. Estimativas indicam que apenas um em cada vinte títulos publicados no mundo analógico migra para o digital. O maior entrave para essa conversão são os direitos autorais. Mas o processo de mudança de formato ganhou novo impulso em dezembro de 2004, quando o Google anunciou um acordo para transformar em bits o acervo de cinco grandes bibliotecas, quatro americanas e uma britânica **Cíntia Borsato**

SCANNER + COMPUTADOR

O scanner portátil é a melhor opção para copiar trechos de textos. Em casa ou no escritório, podem ser usados os scanners convencionais ou os multifuncionais (reúnem scanner, copiadora e impressora num só aparelho). Para que a conversão não perca qualidade, esses equipamentos devem ter resolução de pelo menos 300 DPIs (dots per inch, ou pontos por polegada).

COMO SALVAR

O arquivo de texto pode ser guardado no computador em, basicamente, dois formatos. Um deles é o PDF, que ocupa pouco espaço e é compatível com qualquer modelo de PC. O ponto negativo: uma vez em PDF, o arquivo não pode ser alterado. Outra alternativa é o OCR (Optical Character Recognition, ou Reconhecedor Óptico de Caracteres), software que permite anotações e variações no texto.

MANDE FAZER

Não são muitas as empresas que prestam pequenos serviços desse tipo. Alguns sites: www.alphaqraphics.com.br. ou www.powerhrasll.com.br.

LEIA NA WEB

Endereços na internet permitem a leitura total ou parcial de livros digrtalizados. Três opções: www.dominiopublico.gov.br; www.books.Qooole.com.br e www.bn.br.

ONDE GUARDAR

Além do computador, é possível guardar informações na web. Há a opção de blogs com senha (a maior parte dos provedores oferece esse serviço). Os textos também podem ser guardados em CDs e DVDs, em cópias de segurança (back-up). Nesse caso, os limites de armazenamento variam: um CD comporta cerca de 15 000 páginas de texto em preto-e-branco, enquanto um DVD guarda 100 000 páginas em preto-e-branco.

eBook: a Sony lançou no início de 2006 o Reader, uma base para a leitura de livros em formato digital, que pode guardar até oitenta títulos. Custa 360 dólares. A empresa não vende o produto no Brasil.