Capítulo 3

Base de Dados

Este Capítulo descreve o processo de formação da base de dados de laboratório de imagens de cheques bancários brasileiros. Descreve-se a metodologia empregada no estabelecimento da base, coleta, aquisição digital e armazenamento das imagens. A base de dados de imagens de cheques bancários brasileiros constitui CD-ROM's e encontra-se disponíveis no Laboratório de Análise e Reconhecimento de Imagens de Documentos - LARDOC da PUC-PR.

3.1. Bases de Dados

Existem diversas bases de dados de cheques bancários em nível internacional (CENPARMI, CEDAR, NIST,SRTP, ETL, ESSEX, IRONOF), porém as mesmas não se enquadram nas especificações necessárias para atender os objetivos do presente trabalho. Algumas bases apresentam dados isolados que não representam os dados encontrados na realidade dos cheques brasileiros. Outras bases apresentam dados reais, porém em quantidades insuficientes para testes em processos de reconhecimento automático. Outras, ainda, apresentam somente imagens binárias, estando o usuário limitado ao processo de aquisição destas imagens, bem como aos processos e algoritmos de limiarização ou *thresholding* aplicados. Além do fato, de que as bases de dados que possuem os valores numéricos escritos por extenso e as datas, apresentam palavras em língua inglesa ou francesa, não sendo compatível com o estudo de palavras da língua portuguesa.

Deste modo, observa-se que as bases de dados internacionalmente relevantes possuem uma grande quantidade de imagens armazenadas, permitindo uma grande variedade de trabalhos de pesquisa relativos a diferentes etapas do processo de

reconhecimento de documentos manuscritos. A seguir estão apresentadas resumidamente as bases de dados reconhecidas em nível internacional:

- Centre d'Etudes en Reconnaissance des Formes et Intelligence Artificielle (CENERFIA/CENPARMI) - Montreal, Canadá: composta por 2500 cheques manuscritos preenchidos em inglês provenientes de 800 escritores e 1900 cheques manuscritos preenchidos em francês provenientes de 600 escritores. E, ainda, códigos de endereçamento postal (CEP ou *ZIP code*), com aproximadamente 17.000 dígitos obtidos a partir dos códigos de endereçamento postal, adicionalmente a um conjunto de 20.000 números manuscritos sem restrições coletados de 500 estudantes da Concordia University [GUILLEVIC, 1995] e [CÔTE, 1997];
- Center of Excellence in Document Analysis Recognition (CEDAR) - Buffalo, USA: possui dados de palavras manuscritas relativas a imagens de 5632 nomes de cidades, 4938 nomes de estados e 9454 códigos de endereçamento postal (CEP ou *ZIP code*), caracteres alfabéticos e numéricos mixados e dígitos extraídos de imagens digitais de endereços manuscritos [MOHAMED, 1995] e [KIM, 1996];
- United State National Institute of Standards and Technology (NIST) - Maryland, USA: as bases de dados correlatas com a área de reconhecimento de manuscritos são as de número 1, 3 e 7. Estas bases contém números e textos manuscritos (cada texto de um mesmo escritor), compondo um total de 402.953 dígitos, 220.304 caracteres manuscritos em caixa alta e 190.998 caracteres manuscritos minúsculos [HULL & FERINCH, 1993], [HULL, 1994] e [MONTOLIU, 1995];
- Service de Recherche Technique de la Poste (SRTP) - Paris, França, ligado ao Serviço de Correios da França (que também presta serviços bancários): possui aproximadamente 50.000 imagens de endereços extraídos de envelopes e cheques postais, podendo ser considerada a maior base de dados de imagens extraída de situações reais [GILLOUX et al., 1993], [EL YACOUBI, 1996], [AVILA, 1996] e [GRANDIDIER et al., 2000];
- Electrotechnical Laboratory in Japan (ETL) - Japão: a base de dados como um todo é sub-dividida em 9 sub-bases, sendo as bases ETL1 (141.319 imagens) e ETL6 (157.662 imagens) relacionados aos trabalhos de

reconhecimento de dados alfa-numéricos e a base ETL9 relacionada com os trabalhos de reconhecimento de caracteres [HULL & FERINCH, 1993];

- University of Essex (ESSEX) - Inglaterra: base de dados criada a partir da preenchimento à mão de envelopes padrões na Inglaterra [HULL & FERINCH, 1993];
- IReste ON/OFF Dual Database (IRONOFF) - Nantes, França: base de dados manuscrita com dupla entrada de dados (on-line e off-line) constituída por 32.000 caracteres isolados e 50.000 palavras cursivas [VIARD-GAUDIN et al., 1999];
- Empresa de Correios Brasileiros (ECT) - Brasil: possui 50.000 imagens de envelopes postais, sub-divididas em 4 sub-bases de 2 x 20.000 imagens e 2 x 5.000 imagens, com o objetivo do estudo da triagem automática de envelopes.

Como iniciativa brasileira, com cheques bancários brasileiros, pode-se citar a base de dados do Laboratório de Reconhecimento de Padrões e Rede de Computadores (LRPRC) da Universidade Estadual de Campinas, que possui 620 cheques bancários reais provenientes de 62 escritores [KOERICH, 1997].

A Tabela 3.1 apresenta um resumo das caraterísticas das bases de dados já existentes e alguns trabalhos que fizeram uso das mesmas. Observa-se que as bases de dados possuem uma grande quantidade de imagens armazenadas, tanto para estudos de endereços (nomes de cidades, estados e códigos de endereçamento postal), bem como de cheques bancários e postais (valor numérico e por extenso).

Tabela 3.1: Características e aplicações de trabalhos com bases de dados existentes

Côte, M. (1997) – [CÔTE, 1997]						
BASE de DADOS	**Cheques**		**Escritores**		**Palavras**	
Tipo da Base	Inglês	Francês	Inglês	Francês	Inglês	Francês
CENPARMI	2500	1900	800	600	7837	7135
Aplicação: Sistema PERCEPTO	-	-	-	-	184(A) 2929(T)	-
Guillevic, D. (1995) – [GUILLEVIC, 1995]						
Tese Doutorado	2500	1900	800	600	5322(A) 2515(T)	-

Tabela 3.1: Características e aplicações de trabalhos com bases de dados existentes(continuação)

Mohamed, M. (1995) – [MOHAMED, 1995]										
BASE de Dados	**Cidades**		**Estados**		**ZIP code**		**Caracteres**		**Dígitos**	
Tipo da Base	A	T	A	T	A	T	A	T	A	T
CEDAR / CD-ROM 1	5100	532	4465	470	9019	435	19145	2183	5802	707

Montoliu, L. (1995) – [MONTOLIU, 1995]						
Base de Dados	**Cheques**		**Números**		**Palavras**	
Tipo da Base	A	T	A	T	A	T
NIST	----		223125 imagens		----	
LIX	1720	1585	10010	9259	16819	15538

A = treinamento e T = testes.

Deste modo, optou-se pela criação de bases de dados de laboratório de cheques bancários brasileiros para atender as necessidades do presente estudo: reconhecimento de palavras extenso manuscrito.

3.2. Criação da Base de Dados de Laboratório de Cheques Bancários Brasileiros

A criação de uma base de dados, seja essa, de laboratório ou real, não constitui uma tarefa fácil. Pode-se dizer, que esta etapa é a mais trabalhosa e depende do esforço e colaboração de todo o grupo envolvido com atividades de pesquisa nesta área. Portanto, as deficiências e bons exemplos existentes em outras bases permitiram estabelecer os requisitos necessários da base a ser gerada com as finalidades já apresentadas.

Bases de dados reais, aquelas formadas com imagens de cheques reais, encontram uma grande barreira quando do fornecimento dos cheques reais, ou seja, das suas respectivas imagens pelas instituições bancárias. Por razões de proteção e segurança, no que diz respeito tanto aos clientes como das próprias instituições, os centros de pesquisas não conseguem obter os reais documentos bancários. A maioria dos estudos com cheques bancários reais brasileiros apresenta um número reduzido de exemplares.

Por todas estas razões, as pesquisas na área de reconhecimento de palavras manuscritas em cheques bancários dependem das bases de dados fornecidas pelas instituições responsáveis pelo seu manuseio e tratamento ou, na inexistência dessas

bases, trabalham com a criação de bases de laboratório, atendendo então, as necessidades peculiares do processo de reconhecimento de manuscritos.

A seguir apresenta-se cada uma das etapas da criação da base de dados, sendo estas as seguintes [FREITAS et al., 2000a]:

- critérios para o estabelecimento da base de dados,
- vocabulário de cheques bancários,
- geração dos valores dos cheques bancários,
- estatística prévia da base de dados,
- processo de coleta dos cheques,
- armazenamento e recuperação das informações,
- características da base de dados de treinamento.

3.2.1. Critérios de Estabelecimento da Base de Dados

As bases de dados de laboratório estabelecidas atendem aos seguintes critérios:

- valor mínimo de 0,01 (um centavo);
- valor máximo de 999.999,99 (novecentos e noventa e nove mil, novecentos e noventa e nove reais e noventa e nove centavos);
- garantia de repetição de no mínimo 20 vezes do mesmo dígito em cada posição do domínio do conjunto de valores numéricos e
- existência das palavras: real, reais, centavo e centavos.

Visando o reconhecimento as palavras manuscritas do extenso dos cheques bancários divide-se a base de dados original em 3 bases, denominadas como a seguir:

- BASE 1 - Base de Treinamento (A),
- BASE 2 - Base de Validação (V),
- BASE 3 - Base de Teste (T).

A dimensão de cada uma das bases é de 1600 exemplares de cheques de laboratórios, dado que o número de 20 (vinte) repetições para cada uma das 8 (oito) possíveis (xxx.xxx,xx) do maior valor numérico. Portanto, o valor máximo de R$ 999.999,99 tem por justificativa a própria dimensão das bases de dados a serem coletadas e, ainda, que a inclusão de valores maiores, com cifras de milhão ou bilhão, acarreta somente a inclusão dos modelos destas palavras, sendo as palavras correspondentes aos valores numéricos tratados pelos mesmos modelos já definidos.

3.2.2. Vocabulário de Cheques Bancários

O vocabulário de cheques bancários estudado, considerando o domínio de valores a serem representados e reconhecidos está apresentado na Tabela 3.2. Este vocabulário tem por objetivo representar a maioria das possibilidades de escrita das palavras manuscritas encontradas em cheques bancários brasileiros. Não constituem objetos do presente estudo os erros ortográficos.

Tabela 3.2: Vocabulário de cheques bancários para o domínio estudado

ITEM	VALOR NUMÉRICO	EXTENSO CURSIVO
1	1	hum, um
2	2	dois
3	3	três, treis
4	4	quatro
5	5	cinco
6	6	seis
7	7	sete
8	8	oito
9	9	nove
10	10	dez
11	11	onze
12	12	doze
13	13	treze
14	14	quatorze, catorze
15	15	quinze
16	16	dezesseis
17	17	dezessete
18	18	dezoito
19	19	dezenove
20	20	vinte
21	30	trinta
22	40	quarenta
23	50	cinqüenta, cincoenta
24	60	sessenta
25	70	setenta
26	80	oitenta
27	90	noventa
28	100	cem, cento
29	200	duzentos
30	300	trezentos
31	400	quatrocentos
32	500	quinhentos
33	600	seiscentos
34	700	setecentos
35	800	oitocentos
36	900	novecentos
37	1000	mil
38	----	real/reais
39	----	centavo/centavos

Fazem parte do vocabulário as seguintes palavras:

- 37 palavras relativas aos valores numéricos,
- 02 palavras que representam a moeda corrente no Brasil: “real/reais” e “centavo/centavos”.

3.2.3. Geração dos Valores dos Cheques Bancários

A geração aleatória de valores para preenchimento dos cheques se fez necessária, uma vez que, ao se solicitar às pessoas que preenchessem os cheques sem orientação alguma, muitos valores do tipo “um real”, ou “cem reais” seriam efetivados.

Deste modo, sabe-se que estudos de reconhecimento da parte manuscrita requerem do preenchimento de cheques uma distribuição da amostra de caracteres numéricos que possibilitem a uma distribuição da amostra de palavras dos valores por extenso adequada. A amostra de valores numéricos está diretamente ligada à amostra de palavras, pois sabe-se que o extenso tem sua origem no valor numérico.

A geração dos valores a serem preenchidos nos cheques de laboratório considera um gerador de números aleatórios com distribuição uniforme entre os dígitos [0,..,9] com igual probabilidade de ocorrência.

Controla-se o número de 20 (vinte) repetições de cada dígito, em cada posição possível do valor numérico, formatando-se o problema conforme Tabela 3.3, na forma R$ w3w2w1.y3y2y1,x2x1.

As posições foram denominadas w3, w2, w1, y3, y2, y1, x2 e x1 correspondendo ao valor numérico de oito dígitos. Na Tabela 3.3 a posição x1 foi fixada com o dígito **0**, repetidamente vinte vezes. Deve-se observar, também, que as posições w3, w2, w1, y3, y2, y1 foram fixadas com o dígito **0**, nas linhas de 1 a 6, formando um conjunto de valores específicos, os quais com a geração aleatória poderiam ser determinados. A exceção é a coluna y1, linha 6 que foi fixada com o dígito **1**, para que fossem gerados valores numéricos com a palavra “real”. Pois, somente, quando encontra-se o dígito **1** nesta posição é que existe a grafia da palavra “real”, em todos os demais casos a grafia é “reais”.

Fixou-se a posição x2, nas linhas de 1 a 6, com os dígitos 1,2,...,9, respectivamente, estabelecendo-se assim a possibilidade de grafia dos números de 10 a 99 centavos.

Tabela 3.3: Geração aleatória dos valores numéricos para a base de treinamento

CÓDIGO	**POSIÇÃO**								
	w3	**w2**	**w1**	**y3**	**y2**	**y1**	**,**	**x2**	**x1**
AE001	5	6	1	3	1	**0**	,	1	**0**
AE002	3	7	2	8	**0**	**0**	,	2	**0**
AE003	0	8	4	**0**	**0**	**0**	,	3	**0**
AE004	0	2	**0**	**0**	**0**	**0**	,	4	**0**
AE005	1	**0**	**0**	**0**	**0**	**0**	,	5	**0**
AE006	**0**	**0**	**0**	**0**	**0**	*1*	,	6	**0**
AE007	2	3	6	2	6	8	,	7	**0**
AE008	1	5	8	0	4	9	,	8	**0**
AE009	9	6	2	5	8	5	,	9	**0**
AE010	3	9	5	0	5	4	,	7	**0**
AE011	4	1	4	2	1	9	,	0	**0**
AE012	9	1	6	0	7	0	,	3	**0**
AE013	7	0	8	6	8	9	,	6	**0**
AE014	7	9	9	1	1	4	,	4	**0**
AE015	9	6	6	1	1	7	,	6	**0**
AE016	0	0	7	9	2	3	,	7	**0**
AE017	2	1	5	7	6	9	,	3	**0**
AE018	7	5	0	0	6	7	,	8	**0**
AE019	8	5	4	7	6	4	,	8	**0**
AE020	6	3	0	1	3	9	,	9	**0**

3.2.4. Estatística Prévia das Bases de Dados

Com base nos números aleatórios gerados para formação das 3 bases, efetuou-se uma estatística prévia sobre o total de dígitos, considerando o domínio [0,..,9]. A Tabela 3.4 apresenta o total de dígitos aleatórios gerados. Alguns exemplos de valores gerados aleatoriamente estão mostrados na Tabela 3.5.

Tabela 3.4: Total de dígitos aleatórios gerados

TOTAL de DÍGITOS	**BASE 1 (Treinamento)**	**BASE 2 (Validação)**	**BASE 3 (Teste)**
dígito 0	1.926	1.662	1.778
dígito 1	1.244	1.233	1.263
dígito 2	1.230	1.266	1.200
dígito 3	1.220	1.264	1.186
dígito 4	1.181	1.284	1.243
dígito 5	1.260	1.235	1.299
dígito 6	1.156	1.200	1.168
dígito 7	1.205	1.254	1.218
dígito 8	1.196	1.212	1.208
dígito 9	1.182	1.190	1.237
TOTAL de dígitos por base	12.800	12.800	12.800
TOTAL de números	1.600	1.600	1.600

Tabela 3.5: Exemplos de valores gerados

R$ 0,60	R$ 846519,86	R$ 76748,10
R$ 232268,70	R$ 992924,35	R$ 0,16
R$ 37587,25	R$ 7496,62	R$ 55664,00
R$ 9201,57	R$ 10,71	R$ 700,72

Para cada valor gerado aleatoriamente, obtiveram-se os extensos respectivos. A partir desses extensos realizou-se a contagem por caracteres e por palavras do vocabulário que pretende-se obter ao término do processo de coleta dos cheques. A Tabela 3.6 apresenta alguns exemplos de extensos obtidos a partir dos valores numéricos gerados aleatoriamente. As Tabelas 3.7 e 3.8 apresentam a contagem realizada sobre as bases considerando caracteres e palavras, respectivamente. Na Tabela 3.8 indica-se como Base 1 os valores estimados e como Base 1* os valores coletados de palavras manuscritas.

Tabela 3.6: Exemplos de valores numéricos por extensos

Quinhentos mil reais e cinqüenta centavos
Cento e vinte e dois mil, setecentos e setenta e quatro reais e noventa centavos
Setenta e três mil e noventa reais
Oitocentos e trinta e dois mil, quatrocentos e noventa e um reais e um centavo

Tabela 3.7: Contagem dos caracteres

LETRAS	**BASE 1**	**BASE 2**	**BASE 3**
a	7.561	7.750	7.612
c	4.649	4.771	4.800
d	1.004	1.006	967
e	21.369	21.902	21.699
h	312	293	311
i	7.647	7.732	7.607
l	1.566	1.561	1.564
m	1.965	2.022	1.975
n	10.405	10.578	10.559
o	9.004	9.197	9.254
q	2.018	2.097	2.085
r	5.583	5.731	5.613
s	10.369	10.628	10.495
t	12.249	12.634	12.370
u	2.241	2.407	2.305
v	3.222	3.251	3.279
z	1.070	1.040	1.070
ê	421	422	412
ü	475	482	484
TOTAL	103.130	105.504	104.461

Tabela 3.8: Contagem das palavras

Palavras	BASE 1	BASE 1*	BASE 2	BASE 3
Um	399	332	464	415
Dois	435	350	417	403
Três	421	336	422	412
Quatro	404	322	425	427
Cinco	423	332	422	457
Seis	398	302	404	362
Sete	383	267	426	420
Oito	390	305	402	402
Nove	396	318	400	420
Dez	53	47	37	32
Onze	60	50	34	55
Doze	37	29	45	48
Treze	47	37	46	43
Quatorze	46	35	47	52
Quinze	50	45	38	47
Dezesseis	35	30	50	48
Dezessete	48	36	45	46
Dezoito	41	34	33	44
Dezenove	56	45	51	57
Vinte	459	382	476	460
Trinta	454	372	510	422
Quarenta	456	338	499	453
Cinqüenta	475	376	482	484
Sessenta	441	356	442	481
Setenta	475	370	465	453
Oitenta	477	363	465	457
Noventa	439	316	435	453
Cem Cento	10 302	7 269	7 302	6 315
Duzentos	299	249	328	289
Trezentos	298	236	286	309
Quatrocentos	275	218	313	311
Quinhentos	312	240	293	311
Seiscentos	282	219	304	277
Setecentos	299	248	318	299
Oitocentos	288	252	312	305
Novecentos	291	236	304	307
Mil	1.556	1.286	1.551	1.554
Real Reais	10 1.572	 1.262	10 1.573	10 1.574
Centavo Centavos	25 1.556	 1.094	18 1.567	13 1.569
e	7.537	-	7.801	7.653
TOTAL	22.710	11.948	23.269	22.955

3.2.5. Processo de Coleta dos Cheques de Laboratório

Inicialmente criou-se um cheque de laboratório, com base no Modelo Único de cheque bancário brasileiro, instituído pelo Conselho Monetário Nacional, o qual estabelece a distribuição de informações nos cheques, o modelo padrão de

preenchimento, as dimensões mínima e máxima e seus elementos componentes [BANCO CENTRAL DO BRASIL, 1983], conforme descrito no Apêndice A.

Procedeu-se, então, a coleta dos cheques de laboratórios preenchidos. Este processo foi realizado de forma que as pessoas preenchessem em uma folha em branco, sob uma mesa de luz, tendo como referência uma folha com quatro cheques de laboratórios seguidos. As pessoas foram orientadas a proceder da seguinte forma:

- cheque 1: somente o valor numérico;
- cheque 2: somente a 1ª linha do valor por extenso;
- cheque 3: a 2ª linha do valor por extenso;
- cheque 4: a data.

Isto foi necessário, para que fossem eliminadas quaisquer interferências entre o preenchimento das diversas linhas, das próprias linhas do cheque de laboratório e outras marcas existentes. Desta forma, pode-se iniciar o trabalho de pesquisa, considerando o cheque já segmentado, ou seja, processado e extraído somente as informações manuscritas. A Figura 3.1 exemplifica um cheque de laboratório preenchido. Ressalta-se que a coleta de dados de assinaturas em cheques foi realizado em processo e formulário distintos dos aplicados na coleta dos dados aqui em questão.

Todas as pessoas que participaram como voluntárias, forneceram autorização para uso dos dados coletados. Ao participar do processo, as pessoas recebiam o valor numérico e a data (previamente definidos) gerados aleatoriamente e estavam livres para preencher os dados da forma que costumam preencher um cheque real, conforme apresentado na Figura 3.2, para os valores por extenso.

Puderam ser utilizadas, durante o preenchimento, canetas esferográficas (preta ou azul), tinteiro ou de ponta porosa. O preenchimento a lápis não foi permitido, pois o mesmo não é permitido em cheques reais, além de gerar imagens de baixa qualidade.

Segundo as Normas do Banco Central do Brasil [BANCO CENTRAL DO BRASIL, 1983] o preenchimento de um cheque pode ser feito a mão ou a máquina, pelo próprio emitente ou por terceiros. Obviamente, a assinatura deve ser de próprio punho do emitente, admitindo-se a chancela mecânica. No que concerne ao valor por extenso as normas estabelecem que é obrigatória a grafia por extenso do valor numérico correspondente, sendo facultativa a grafia dos centavos somente por extenso.

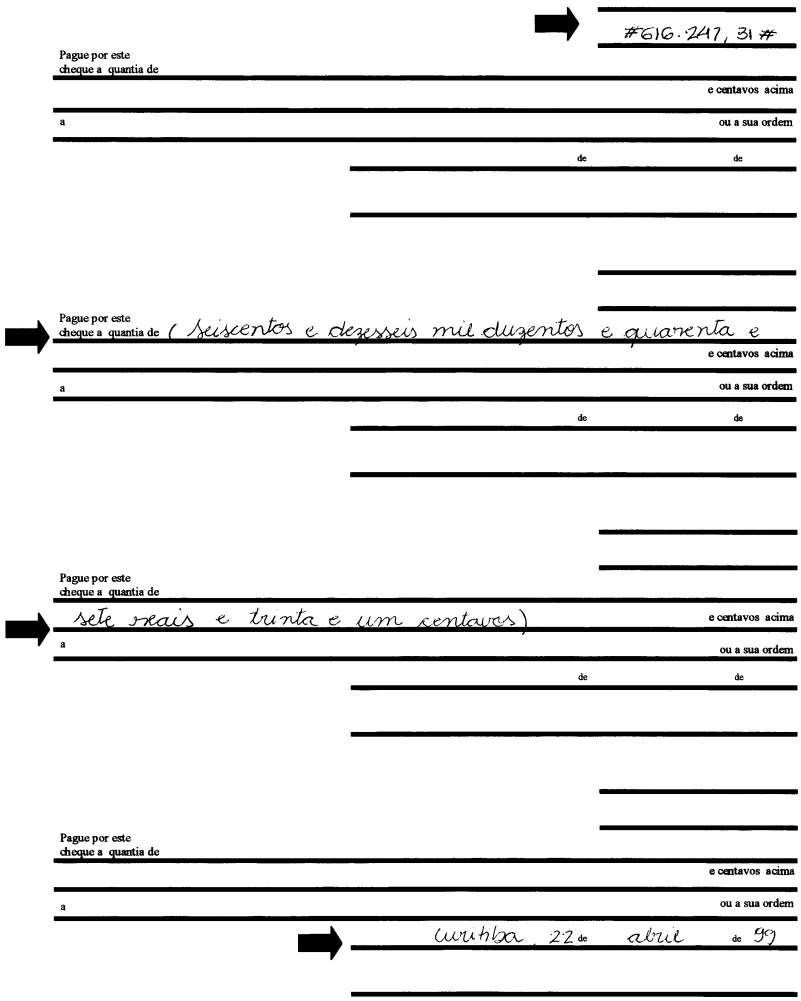

Figura 3.1: Exemplo de cheque de laboratório preenchido

Todos estes cuidados foram tomados, uma vez que, o processo de coleta dos cheques de laboratório demanda uma parcela de tempo razoável e, portanto, nenhum exemplar de coleta deve ser eliminado ou descartado.

3.2.6. Armazenamento e Recuperação das Informações

As imagens dos cheques de laboratórios coletados foram adquiridas através de *scanner* FUJITSU M30096GX, em 300 dpi, 256 níveis de cinza, 100% da escala, 80 de brilho e 128 de contraste, formatos BMP e TIFF.

O formato BMP não oferece uma boa compressão das imagens porém facilita o processamento posterior em programas para plataforma Windows. O formato TIFF é mais utilizado pelos programas de plataforma UNIX. Para o armazenamento das

imagens em CD-ROM foi utilizado o formato JPEG, o qual permite um alto grau de compressão das imagens com pouca perda de informação.

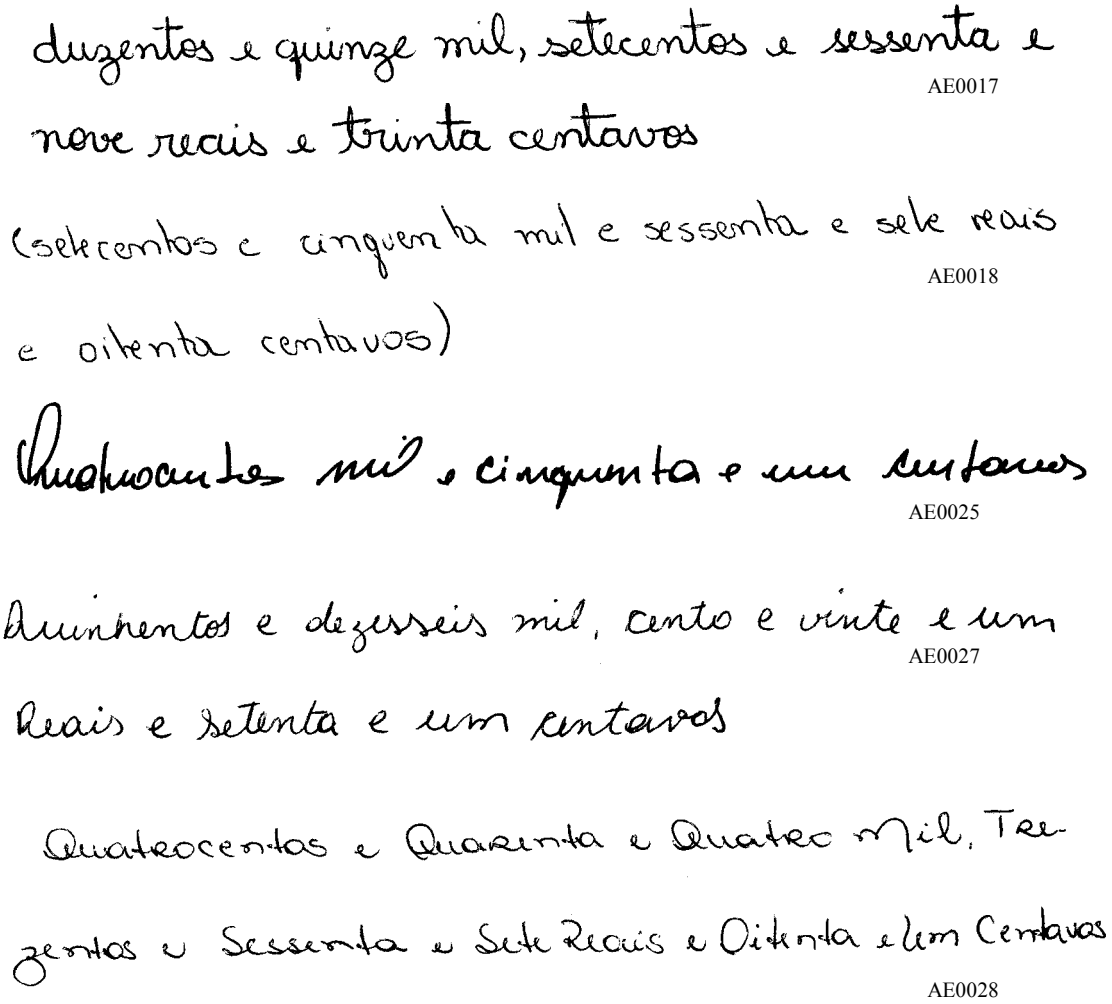

Figura 3.2: Exemplares de estilos de valores por extenso

Após o processo de aquisição das imagens, essas foram cadastradas segundo uma codificação que permite armazenar e recuperar os cheques, bem como informações específicas encontradas nos cheques.

A codificação utilizada permite que os itens que compõem o cheque bancário possam ser armazenados individualmente e ao mesmo tempo serem identificados como um único elemento, ou seja, o cheque. Os itens que compõem o cheque são:

- Valor numérico;
- Extenso;
- Data .

A estrutura lógica do código para identificação dos itens dos cheques de laboratório é composta por:

- 2 (duas) letras: L1 e L2 e
- 4 (quatro) dígitos : D1 a D4, conforme Figura 3.3.

L1	L2	D1	D2	D3	D4

Figura 3.3: Esquema lógico da codificação dos cheques de laboratório

Por sua vez, cada campo possui a seguinte descrição:

- **Campo L1:** Identifica o tipo de base, podendo ser:
  - ⇒ A = treinamento;
  - ⇒ V = validação e
  - ⇒ T = teste.
- **Campo L2:** Identifica o item do cheque contido na base, podendo ser:
  - ⇒ N = número;
  - ⇒ E = extenso;
  - ⇒ D = data.
- **Campo D1 a D4:** Número de 1 a 1600, identificando o valor numérico e a data do cheque (gerados aleatoriamente). A Tabela 3.9 apresenta e explica alguns exemplos de códigos.

Tabela 3.9: Exemplos da codificação dos cheques de laboratório

L1	L2	D1	D2	D3	D4	Descrição
A	N	0	0	0	1	Cheque da base de treinamento, com valor numérico referente ao primeiro valor numérico gerado.
A	E	1	6	0	0	Cheque da base de treinamento, com valor por extenso referente ao último valor numérico gerado.
T	D	1	2	4	5	Cheque da base de testes, com data referente ao valor numérico gerado na posição 1245.
V	E	0	0	3	3	Cheque da base de validação, com extenso referente ao valor numérico gerado na posição 33.

Atualmente as bases de dados de cheques de laboratório referentes ao extenso manuscrito encontram-se na seguinte situação:

- Base de Treinamento: Coleta realizada, aquisição digital (*scanner*), tamanho de 1,3Gbytes, 256 níveis de cinza. Esta base de dados encontra-se em dois formatos: base dos valores por extenso e base de palavras isoladas,
- Base de Teste: em processo de coleta, com 416 exemplares já coletados,
- Base de Validação: não iniciado processo de coleta.

Neste trabalho utiliza-se 2016 cheques sendo que após o processo de segmentação do extenso em palavras obteve-se um conjunto de 11.948 palavras. Portanto, divide-se este conjunto de palavras isoladas em 3 sub-bases, a saber:

- 60% para treinamento = 7.146 palavras,
- 20% para validação = 2.427 palavras e
- 20% para testes = 2.375 palavras.

3.2.7. Características da Base de Dados de Treinamento

A base de dados de treinamento possui relevante importância no processo de desenvolvimento de um método de reconhecimento de palavras manuscritas. Isto devido ao fato de que é sobre esta base que se realiza o aprendizado dos modelos de reconhecimento, especialmente no caso das aplicações que envolvem modelos estatísticos, como por exemplo: MEM e Redes Neurais. Para que tal tarefa tenha êxito torna-se importante o conhecimento e análise da base de dados de treinamento sob dois aspectos: o tipo de escrita e os símbolos utilizados como iniciais e terminais do extenso manuscrito em cheques bancários brasileiros.

3.2.7.1. Estilo de Escrita

A base de dados de treinamento foi avaliada quanto aos estilos de escrita cursiva encontrados. Para [TAPPERT et al., 1990] existem 5 categorias distintas de escrita cursiva, conforme já apresentado na Figura 2.11. A análise visual dos cheques de laboratórios preenchidos encontrou as seguintes categorias:

- palavras em caracteres disjuntos (caixa alta);
- palavras em caracteres disjuntos com espaçamento regular;
- palavras em escrita cursiva pura, ou seja, todos os caracteres de uma palavra são conectados dois a dois;
- palavras em escrita mixada, ou seja, misturando os demais tipos de escrita, e ainda incluindo a classe de palavras com caracteres disjuntos com a presença

de vínculos entre eventuais caracteres. Observou-se que quando o escritor gera palavras formadas por caracteres disjuntos com a presença de vínculos entre eventuais caracteres, o escritor utiliza aleatoriamente as outras 3 categorias de tipo de escrita cursiva.

A Tabela 3.10 apresenta a distribuição dos tipos de escrita na base de dados de treinamento, mostrando que a escrita cursiva é o tipo mais encontrado nesta base de dados.

Tabela 3.10: Distribuição dos tipos de escrita na base de treinamento

Tipo de Escrita	Exemplo	Porcentagem
Cursiva Pura	reais	72%
Caixa Alta	REAIS	13%
Caracteres Disjuntos	reais	7%
Outros	reais	8%

3.2.7.2. Símbolos Iniciais e Terminais do Extenso dos Cheques

Encontrou-se uma grande variedade de símbolos utilizados como iniciais e terminais no valor numérico por extenso. Porém, inicialmente deve-se explicar o que estão sendo considerados como símbolos iniciais e terminais.

Símbolo inicial e terminal é qualquer tipo de simbologia utilizada no início e/ou no final do valor por extenso em um cheque bancário brasileiro. É possível que as pessoas ao preencheram um cheque iniciem por "(" ou "#" e terminem, após a grafia do valor numérico por extenso, com ".", ")", ou simplesmente um traço longo. A Figura 3.4 exemplifica cheques da base de dados de treinamento com símbolos inicial e terminal.

Observa-se que os escritores não utilizam na sua grande maioria (94%) um símbolo inicial, porém o caracter "(" é o mais utilizado (75%) para iniciar o preenchimento do extenso quando o fazem. Como símbolo terminal, a maioria (63%) não utiliza qualquer símbolo, mas quando o fazem (55%), utilizam o caracter "." para terminar o preenchimento.

Estas duas características, tipo de escrita e símbolos iniciais e terminais, permitem conhecer melhor a base de dados de treinamento e avaliar as possíveis dificuldades a serem encontradas, por exemplo, no processo de segmentação do extenso em palavras.

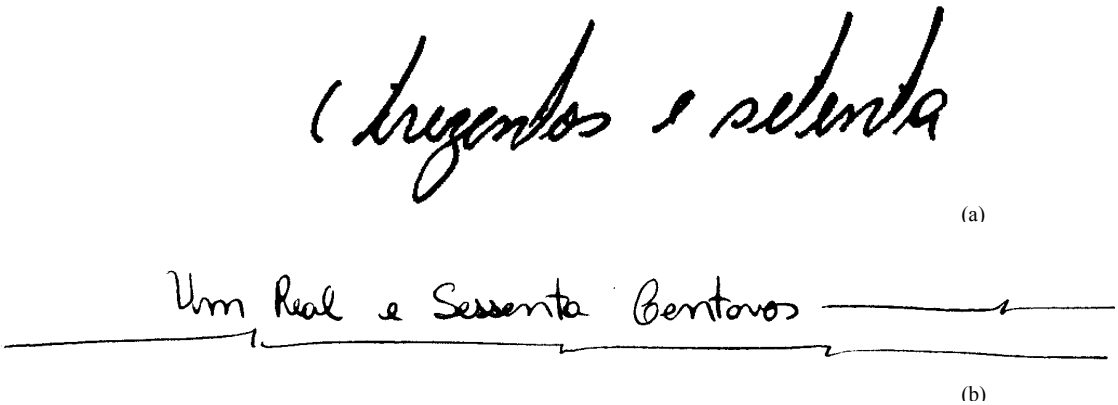

Figura 3.4: Exemplos de símbolos: a) inicial e b) terminal

3.2.8. Considerações Finais

Neste Capítulo descreve-se a criação da base de dados de laboratório, a qual se reveste de grande importância para todos os pesquisadores, visto que estas imagens constituem a matéria prima para a realização dos seus trabalhos.

A criação da base de dados de laboratórios de cheques bancários brasileiros ainda encontra-se em processo de coleta, como mencionado anteriormente. Porém, foram coletados 2.016 cheques de laboratório, perfazendo um total de 15.139 dígitos isolados e 15.636 palavras isoladas, sendo 11.948 palavras do valor numérico por extenso e 3.688 palavras relativas aos meses da data. Ao final do processo, espera-se obter uma base de dados com 4.800 imagens de cheques [FREITAS et al., 2000a]. Estas imagens constituem CD-ROM's e encontram-se disponíveis no Laboratório de Análise e Reconhecimento de Imagens de Documentos - LARDOC da PUC-PR.

Utiliza-se a base de dados descrita nos processos de treinamento, validação e teste de métodos de reconhecimento da parte manuscrita de cheques bancários brasileiros através dos Modelos Escondidos de Markov.

No que concerne ao usuário final, será a partir de bases de dados que os sistemas de automação do processamento de documentos com informações manuscritas poderão ser desenvolvidos, validados e testados. Então, esses sistemas poderão realmente reduzir o tempo de processos cotidianos e melhorar a eficiência global de sistemas bancários.

O Capítulo 4 apresenta a metodologia do trabalho através da descrição e detalhamento das estratégias, métodos e técnicas utilizados nas diversas etapas do reconhecimento das palavras isoladas do extenso manuscrito em cheques bancários brasileiros.