

Bases de Dados de Cheques Bancários Brasileiros

Cinthia O. de A. Freitas¹; Marisa Morita¹; Luiz E. Soares de Oliveira¹;
Edson Justino¹; Abdénaim El Yacoubi¹; Edouard Lethelier¹;
Flávio Bortolozzi¹; Robert Sabourin^{1,2}

¹Pontifícia Universidade Católica do Paraná - PUCPR
Programa de Pós-Graduação em Informática Aplicada - PPGIA
Laboratório de Análise e Reconhecimento de Imagens de Documentos - LARDOC
R.Imaculada Conceição, 1155 CEP:80215-901 - Curitiba - PR - Brasil
cinthia@ppgia.pucpr.br

²Ecole de Technologie Supérieure – ETS
Dep. de Genie de la Production Automatisee,
Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle -LIVIA
1100,rue Notre-Dame Ouest - Montréal(Québec) H3C 1K3 - Canada
sabourin@gpa.etsmtl.ca

Resumo

Este artigo apresenta as diversas etapas da formação das bases de dados de cheques bancários brasileiros com o objetivo de atender os diversos trabalhos de pesquisa em realização na área de processamento digital de documentos manuscritos. Descreve-se a metodologia empregada no estabelecimento das bases de dados, coleta, aquisição digital e armazenamento das imagens. Apresentam-se as principais características dos cheques bancários brasileiros e a situação atual das bases de dados, permitindo assim, auxiliar outros grupos de pesquisadores da área de documentos manuscritos.

Palavras-Chaves: bases de dados, documentos manuscritos, cheques bancários.

1. Introdução

Este trabalho descreve o processo de formação das bases de dados de imagens de cheques bancários brasileiros para as diversas pesquisas da PUC-PR, desenvolvidas no Laboratório de Análise e Reconhecimento de Imagens de Documentos – LARDOC com aplicações em Estruturação Lógica de Cheques, Segmentação, Reconhecimento do Valor Numérico, do Extenso, da Data e de Assinaturas.

Para avaliar o desempenho de qualquer processo na área de reconhecimento de documentos manuscritos, através da aplicação de processos estatísticos (por exemplo: Modelos Escondidos de Markov – HMM ou Redes Neurais) necessita-se fundamentalmente de bases de dados que permitam a validação dos métodos empregados. Assim, as bases de dados com o objetivo de atender estas aplicações necessitam de um grande número de exemplares de imagens, a fim de permitir a modelização da variabilidade das formas manuscritas sem considerar as limitações inerentes ao problema do reconhecimento.

Deste modo, independente do processo de reconhecimento escolhido será possível fornecer a melhor adequação entre os métodos e a base de dados, chamada de base de aprendizagem. Para tal, o número de exemplares coletados por tipo de dado é fundamental para uma perfeita avaliação. Outros aspectos como, resolução da imagem e níveis de cinza, também são elementos importantes, pois tratam da qualidade da imagem usada.

O presente trabalho contribui com informações no sentido de auxiliar demais pesquisadores na elaboração de bases de dados de imagens de cheques bancários manuscritos, bem como permitir que

pesquisadores interessados tenham conhecimento das bases disponíveis e possam realizar testes com novas técnicas e metodologias.

As bases de dados de imagens digitais de cheques bancários brasileiros constituem CD-ROM's e encontram-se disponíveis no LARDOC/PUC-PR. A formação destas bases contam com o suporte do Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq através do Projeto: Tratamento Automático de Documentos sob n°. 520324/96.

Este artigo está organizado em 5 seções como a seguir: a Seção 2 apresenta o estado da arte relacionando as bases de dados existentes, a Seção 3 detalha a criação das bases de dados de laboratório de cheques bancário, a Seção 4 está voltada para as bases de dados de cheques reais e a Seção 5 apresenta algumas conclusões, destacando a importância das bases de dados para os pesquisadores.

2. Bases de Dados Existentes

Existem diversas bases de dados de cheques bancários em nível internacional, porém as mesmas não se enquadram nas especificações necessárias para atender os objetivos dos diversos estudos em desenvolvimento. 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 uso 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 (*thresholding*) aplicados. Além do fato, 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 [GUI95][CÔT97];
- 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 [MOH95][KIM96];
- United State National Institute of Standards and Technology (NIST) - Maryland, USA: as bases de dados de correlatas com a área de reconhecimento de manuscritos são as de número 1, 3 e 7. A base de dados 1 contém números e textos manuscritos (cada texto de um mesmo

escritor), compondo um total de 273.000 números e 707.700 caracteres alfabéticos. A base de dados 3 contém a base de dados 1 adicionada aos caracteres segmentos dos formulários preenchidos para a base de dados 1, totalizando 313.389 caracteres isolados. A base de dados 7 contém 83.000 imagens binárias de caracteres manuscritos em caixa alta [HUL94][HUL93][MON95] ;

- 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 [GIL93][GRA99][YAC96][LET96][AVI96];
- Electrotechnical Laboratory in Japan (ETL) - Japão: a base de dados como um todo é subdividida 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 [HUL93];
- University of Essex (ESSEX) - Inglaterra: base de dados criada a partir da preenchimento à mão de envelopes padrões na Inglaterra [HUL93];
- 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 [VIA99];
- Empresa de Correios Brasileiros (ECT) - Brasil: possui 50.000 imagens de envelopes postais, 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 [KOE97].

Relativamente aos trabalhos realizados com assinaturas devem ser destacadas as bases utilizadas por Sabourin e Plamondon [SAB90] com 248 assinaturas genuínas e 224 fotocópias das genuínas, Huang e Hong [HUA96] com 504 assinaturas genuínas e 3024 falsificações, por Ammar, Yoshida e Fukumura [AMM85] com 200 assinaturas genuínas e 200 falsificações, por Brocklehurts [BRO85] com 2820 assinaturas genuínas, por Nemcek e Lin [NEM74] com 600 assinaturas genuínas e 120 falsificações, por Sabourin e Plamondon [PLA88] com 800 assinaturas genuínas e por Rigoll e Kosmala [RIG98] com 280 assinaturas genuínas e 60 falsificações. As características de cada um das bases de dados depende do número de assinantes, número de espécimes coletados por assinantes, número de pessoas utilizadas como imitadores das assinaturas, além das características peculiares de hábitos e costumes dependendo do país de origem das bases de dados.

O que concerne aos trabalhos em desenvolvimento no LARDOC em aplicações de reconhecimento de cheques bancários brasileiros é a necessidade de uma base de dados com palavras em língua portuguesa, bem como com as características dos escritores brasileiros no preenchimento das informações no cheques bancários e suas respectivas assinaturas.

Desta forma, optou-se pela criação de bases de dados de laboratório de cheques bancários brasileiros para atender as necessidades das pesquisas a cerca de reconhecimento do valor numérico, valor por extenso, data e assinaturas manuscritos. E, ainda, paralelamente a criação de uma base de dados de cheques reais para as necessidades dos estudos de estruturação lógica de cheques e segmentação das informações manuscritas em cheques bancários brasileiros.

3. Criação das Bases de Dados de Laboratório dos Cheques Bancários

A criação de bases de dados, sejam essas, de laboratório ou real, não constitui uma tarefa fácil. Pode-se dizer, que esta etapa é muito 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 das bases a serem geradas com as finalidades já apresentadas.

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

Por todas estas razões, as pesquisas 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.

3.1. Estabelecimento das Bases de Dados

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

- 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,
- existência das palavras: real, reais, centavo e centavos,
- domínio das datas: dia (01 a 31), mês (janeiro a dezembro) e ano(1997 a 2020).

O reconhecimento da parte manuscrita dos cheques bancários exige a existência de 3 bases, denominadas como : BASE 1 - Base de Aprendizagem (A), BASE 2 - Base de Validação (V) e BASE 3 - Base de Teste (T). A dimensão de cada uma das bases é de 1600 imagens de cheques de laboratórios, considerando-se para isto o número de 20 (vinte) repetições para cada uma das 8 (oito) posições possíveis (xxx.xxx,xx). Para as bases de dados de assinaturas foram coletadas 5200 imagens de assinaturas, sendo 40 amostras de assinaturas para 100 escritores diferentes. Para os 60 primeiros escritores foram coletadas 10 amostras de falsificações simples e 10 amostras de falsificações servis.

3.2. Vocabulário de Cheques Bancários

O vocabulário de palavras mais utilizado nos cheques bancários estudados, considerando a possibilidade de grafia dos valores numéricos, por extenso e as datas está apresentado no Quadro 01.

Quadro 01: Vocabulário de cheques bancários para o domínio do extenso e da data

ITEM	VALOR NUMÉRICO	EXTENSO CURSIVO	ITEM	VALOR NUMÉRICO	EXTENSO CURSIVO
1	1	Hum, Um	27	90	Noventa
2	2	Dois	28	100	Cem, Cento
3	3	Três, Treis	29	200	Duzentos
4	4	Quatro	30	300	Trezentos

5	5	Cinco	31	400	Quatrocentos
6	6	Seis	32	500	Quinhentos
7	7	Sete	33	600	Seiscentos
8	8	Oito	34	700	Setecentos
9	9	Nove	35	800	Oitocentos
10	10	Dez	36	900	Novencentos
11	11	Onze	37	1000	Mil
12	12	Doze	38	--	Real / Reais
13	13	Treze	39	--	Centavo /Centavos
14	14	Quatorze, Catorze	40	**	Janeiro
15	15	Quinze	41	**	Fevereiro
16	16	Dezesseis	42	**	Março
17	17	Dezessete	43	**	Abril
18	18	Dezoito	44	**	Maio
19	19	Dezenove	45	**	Junho
20	20	Vinte	46	**	Julho
21	30	Trinta	47	**	Agosto
22	40	Quarenta	48	**	Setembro
23	50	Cinqüenta, Cincoenta	49	**	Outubro
24	60	Sessenta	50	**	Novembro
25	70	Setenta	51	**	Dezembro
26	80	Oitenta			

** - A grafia do mês somente pode ser realizada por extenso, sendo proibida a representação através de valor numérico [BAN90].

Fazem parte dos estudos os seguintes itens: dez dígitos de 0 a 9 para representação dos valores numéricos, dia e ano; quatro palavras que representam a moeda corrente no Brasil: *real/reais* e *centavo/centavos*; uma palavra de ligação: *e* (sessenta e um reais); trinta e sete palavras relativas aos valores numéricos e doze palavras relativos aos meses do ano.

3.3. Geração Aleatória dos Valores dos Cheques Bancários

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

No preenchimento de cheques para estudos de reconhecimento da parte manuscrita alguns fatores são importantes: a distribuição da amostra de caracteres numéricos; a distribuição da amostra de palavras dos valores por extenso; a distribuição da amostra de dias, meses e anos da data e o número de assinaturas de uma mesma pessoa. O presente trabalho descreve a formação das bases de dados levando-se em consideração todos os fatores citados.

A geração dos valores a serem preenchidos nos cheques de laboratório foi realizada considerando um gerador de números aleatórios com distribuição uniforme entre os dígitos [0,9] e uma igual probabilidade de ocorrência. Foram controladas as vinte repetições de cada dígito, em cada posição possível, formatando-se o problema conforme o exposto pelo Quadro 02. O código de cada um dos valores apresentado no Quadro 02 representa o tipo de base de dados (*A* = aprendizagem), o tipo de dado (*N* = valor numérico) e um número sequencial de 1 a 1600 para identificação unívoca de cada um dos valores gerados aleatoriamente.

As posições foram denominadas *w3*, *w2*, *w1*, *y3*, *y2*, *y1*, *x2* e *x1* correspondendo ao valor numérico de 8 dígitos. No Quadro 02 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, de modo a complementar a geração aleatória. Visto que valores com estas características não seriam gerados aleatoriamente em quantidade suficiente.

A posição y1, na linha 6, foi fixada com o dígito 1, para que fossem gerados valores numéricos com a palavra *real*. Pois, somente, quando encontra-se o algarismo 1 nesta posição é que existe a grafia da palavra *real*, em todos os demais casos a grafia será *reais*. A posição x2, nas linhas de 1 a 9, foram fixadas com os dígitos 1,2,...,9, respectivamente, estabelecendo-se assim a possibilidade de grafia dos números de 10 a 99 centavos.

Quadro 02: Geração aleatória dos valores numéricos para a BASE 1 (Aprendizagem)

CÓDIGO	POSIÇÃO								
Fixada em:	w3	w2	w1	y3	y2	y1	,	X2	x1
AN001	5	6	1	3	1	0	,	1	0
AN002	3	7	2	8	0	0	,	2	0
AN003	0	8	4	0	0	0	,	3	0
AN004	0	2	0	0	0	0	,	4	0
AN005	1	0	0	0	0	0	,	5	0
AN006	0	0	0	0	0	1	,	6	0
AN007	2	3	6	2	6	8	,	7	0
AN008	1	5	8	0	4	9	,	8	0
AN009	9	6	2	5	8	5	,	9	0
AN010	3	9	5	0	5	4	,	7	0
AN011	4	1	4	2	1	9	,	0	0
AN012	9	1	6	0	7	0	,	3	0
AN013	7	0	8	6	8	9	,	6	0
AN014	7	9	9	1	1	4	,	4	0
AN015	9	6	6	1	1	7	,	6	0
AN016	0	0	7	9	2	3	,	7	0
AN017	2	1	5	7	6	9	,	3	0
AN018	7	5	0	0	6	7	,	8	0
AN019	8	5	4	7	6	4	,	8	0
AN020	6	3	0	1	3	9	,	9	0

No caso específico das datas, elas foram geradas utilizando uma distribuição uniforme e igual probabilidade de ocorrência do dia, mês e ano em cada sub-base. Para cada mês, os dias 1 até 31 foram fixados, levando em consideração os meses que se limitam ao dia 30, 28 ou 29 (no caso do mês de fevereiro quando o ano é bissexto), e o ano foi gerado aleatoriamente no intervalo [1997-2020]. No caso do ano foi solicitado para as pessoas escreverem somente os dois últimos dígitos do mesmo durante o procedimento da coleta. Entretanto, nem todas as pessoas respeitaram essa solicitação, ocasionando a presença de quatro dígitos no campo do ano em nossa base de dados. Devido a geração do mês e do dia estarem fixas, em cada sub-base vai existir a mesma quantidade de ocorrências dessas informações. Entretanto, como o ano foi gerado aleatoriamente isso não vai ocorrer.

3.4. Estatística 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], conforme mostrado no Quadro 03. Para os dígitos foram estabelecidas onze classes: dígitos de 0 a 9, caracter # utilizado antes e após a grafia do valor numérico e os sinais de pontuação ponto e vírgula que são utilizados para marcação da milhar e parte decimal respectivamente (por exemplo: R\$ 1.200,45). O Quadro 03 apresenta também o total de dígitos já coletados através da denominação Base Atual, sendo que a composição dessa bases é: 89% de dígitos isolados, 9% de dígitos conectados em dupla e 2% de dígitos conectados em tripla. Além disto, verificou-se que 26% dos dígitos conectados em dupla e 55% dos dígitos conectados em tripla são formados com dígitos zero.

Quadro 03: Total de dígitos aleatórios gerados

TOTAL de DÍGITOS	BASE 1 (Aprendizagem)	BASE 2 (Validação)	BASE 3 (Teste)	BASE ATUAL
Dígito 0	1926	1662	1778	1841
Dígito 1	1244	1233	1263	1349
Dígito 2	1230	1266	1200	1516
Dígito 3	1220	1264	1186	1484
Dígito 4	1181	1284	1243	1458
Dígito 5	1260	1235	1299	1450
Dígito 6	1156	1200	1168	1383
Dígito 7	1205	1254	1218	1451
Dígito 8	1196	1212	1208	1297
Dígito 9	1182	1190	1237	1404
#	---	---	---	991
”	---	---	---	1362
TOTAL	12.800	12.800	12.800	16.986

Para cada valor gerado aleatoriamente, obtiveram-se, por processo automático, os extensos respectivos. A partir desses extensos realizou-se a contagem das palavras do vocabulário. Para os valores correspondentes as datas realizou-se o mesmo procedimento.

Os Quadros 04 a 07 apresentam a contagem prévia realizada sobre as bases considerando as palavras do extenso e as datas (dia, mês e ano), respectivamente. Os Quadros 04 e 05 apresentam também o total de palavras já coletadas através da denominação Base Atual. Ressalta-se que a grafia por extenso dos valores numéricos pode apresentar para uma mesma palavra, grafia diferenciada, tais como: *hum* e *um*, *três* e *treis*, *quatorze* e *catorze*, *cinquenta* e *cincoenta*, etc.. Além, é claro, da possibilidade dos erros de ortografia cometidos pela pessoa que realiza o preenchimento do cheque.

Quadro 04: Estatística das palavras

PALAVRA	BASE 1	BASE 2	BASE 3	BASE ATUAL	PALAVRA	BASE 1	BASE 2	BASE 3	BASE ATUAL
Um	399	464	415	333	Trinta	454	510	422	372
Dois	435	417	403	351	Quarenta	456	499	453	339
Três	421	422	412	338	Cinquenta	475	482	484	372
Quatro	404	425	427	322	Sessenta	441	442	481	352
Cinco	423	422	457	335	Setenta	475	465	453	370
Seis	398	404	362	301	Oitenta	477	465	457	363
Sete	383	426	420	297	Noventa	439	435	453	316
Oito	390	402	402	305	Cem / Cento	312	309	321	276
Nove	396	400	420	320	Duzentos	299	328	289	249
Dez	53	37	32	47	Trezentos	298	286	309	235
Onze	60	34	55	50	Quatrocentos	275	313	311	208
Doze	37	45	48	29	Quinhentos	312	293	311	238
Treze	47	46	43	37	Seiscentos	282	304	277	221
Quatorze	46	47	52	35	Setecentos	299	318	299	238
Quinze	50	38	47	45	Oitocentos	288	312	305	249
Dezesseis	35	50	48	30	Novencentos	291	304	307	235
Dezessete	48	45	46	35	Mil	1556	1551	1554	1286
Dezoito	41	33	44	35	Real / Reais	1582	1583	1584	1262
Dezenove	56	51	57	43	Centavo / Centavos	1581	1585	1582	1079
Vinte	459	476	460	381	e	7537	7801	7653	--
					TOTAL	22.710	23.269	22.955	11.936

Quadro 05: Estatística dos meses

MÊS	BASE 1	BASE 2	BASE 3	BASE ATUAL
Janeiro	134	134	134	198
Fevereiro	133	133	133	163
Março	133	133	133	178
Abril	134	134	134	194

Maio	133	133	133	188
Junho	133	133	133	147
Julho	133	133	133	164
Agosto	133	133	133	140
Setembro	133	133	133	159
Outubro	133	133	133	148
Novembro	134	134	134	169
Dezembro	134	134	134	168
TOTAL	1.600	1.600	1.600	2.016

Quadro 06: Estatística prévia dos dias

DIA	BASE 1	BASE 2	BASE 3	DIA	BASE 1	BASE 2	BASE 3
1	52	52	52	17	52	52	52
2	52	52	52	18	51	51	51
3	52	52	52	19	51	51	51
4	52	52	52	20	51	51	51
5	52	52	52	21	51	51	51
6	52	52	52	22	51	51	51
7	52	52	52	23	51	51	51
8	52	52	52	24	51	51	51
9	52	52	52	25	51	51	51
10	52	52	52	26	51	51	51
11	52	52	52	27	51	51	51
12	52	52	52	28	51	51	51
13	52	52	52	29	56	56	56
14	51	51	51	30	53	53	53
15	51	51	51	31	49	49	49
16	51	51	51	TOTAL	1.600	1.600	1.600

Quadro 07: Estatística prévia dos anos

ANO	BASE 1	BASE 2	BASE 3	ANO	BASE 1	BASE 2	BASE 3
1997	61	66	74	2010	63	80	66
1998	65	77	72	2011	73	70	58
1999	62	54	73	2012	68	57	65
2000	76	68	65	2013	56	76	87
2001	62	61	65	2014	68	64	59
2002	64	69	63	2015	67	68	77
2003	70	53	76	2016	63	68	77
2004	74	68	55	2017	66	68	52
2005	73	74	67	2018	61	82	78
2006	65	65	69	2019	65	72	72
2007	67	58	60	2020	72	69	62
2008	66	54	54	TOTAL	1.600	1.600	1.600
2009	73	59	54				

Observou-se nas bases de dados que 72% dos exemplares de palavras manuscritas do extenso dos cheques são da categoria cursiva, 13% constituem a categoria caracteres disjuntos (caixa alta), 7% encontram-se na categoria caracteres disjuntos com espaçamento regular e 8% apresentam escrita mixada, ou seja, misturando os demais tipos de escrita, e ainda incluindo a categoria de palavras com caracteres disjuntos com a presença de eventuais vínculos entre caracteres.

3.5. Processo de Coleta dos Cheques

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 [BAN83].

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; e cheque 4: a data.

Para a assinatura o processo foi diferenciado devido as características inerentes ao problema de reconhecimento de assinaturas. Cada conjunto de 40 assinaturas é formado por dez folhas A4 com quatro assinaturas genuínas do mesmo escritor, totalmente isenta de qualquer elemento de fundo (decoração, linhas de base, marcações, etc.).

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 01 exemplifica o formulário do cheque de laboratório preenchido com os dados de valor numérico, extenso e data, além de apresentar 4 espécimes de assinaturas de um mesmo assinante.

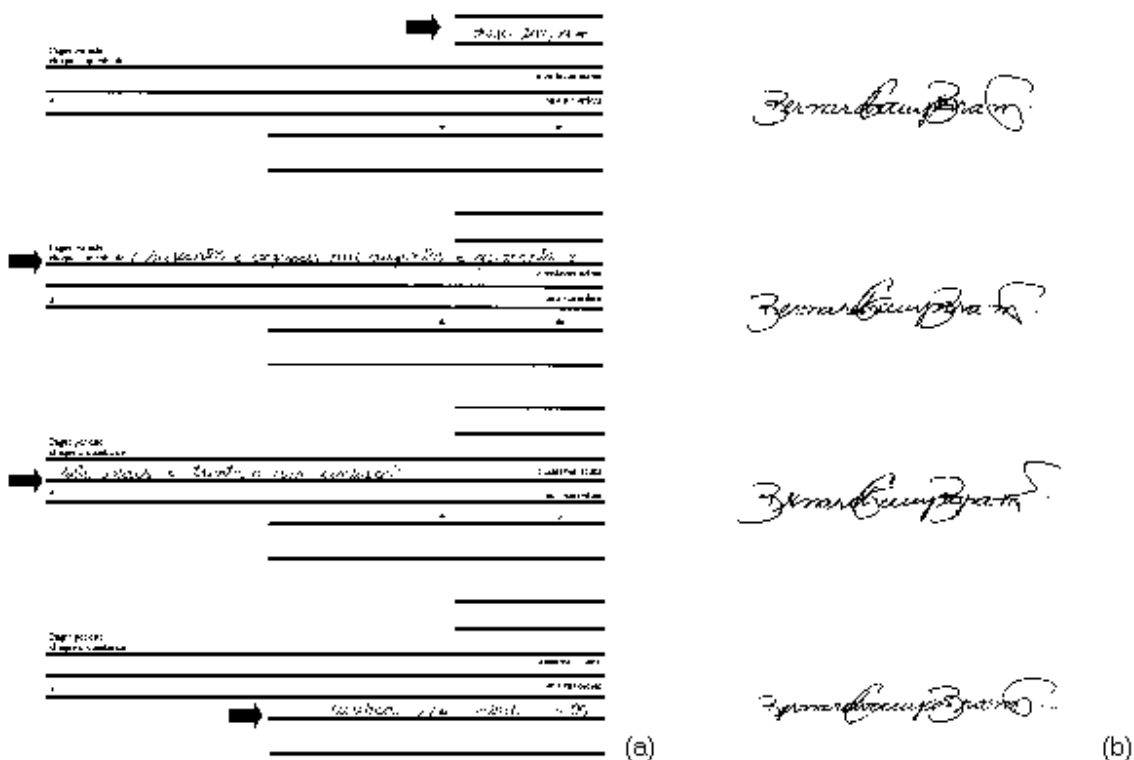


Figura 01: Cheque de laboratório com dados de: a) valor numérico, extenso, data e b) espécimes de assinaturas de um mesmo assinante

Todas as pessoas que participaram como voluntárias, forneceram autorização para uso dos dados coletados. Ao participar do processo, as pessoas receberam 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. Cada voluntário realizou o preenchimento de somente um cheque de laboratório no que se refere aos dados do valor numérico, extenso e data. Para a assinatura os voluntários seguiram as especificações apresentadas anteriormente.

No preenchimento dos cheques, puderam ser utilizadas canetas esferográficas (preta ou azul), tinteiro ou de ponta porosa, e no preenchimento das assinaturas não foi feita nenhuma restrição quanto a cor da caneta a ser utilizada. O preenchimento a lápis não foi permitido em qualquer dos procedimentos, pois o mesmo não é permitido em cheques reais, além de gerar imagens de baixa qualidade.

Segundo as Normas do Banco Central do Brasil [BAN83] o preenchimento de um cheque pode ser feito a mão ou a máquina, pelo próprio emitente ou por terceiros. O caso de preenchimento a máquina não se aplica ao presente estudo. No que concerne ao valor por extenso as normas estabelecem a obrigatoriedade da grafia por extenso do valor numérico correspondente, sendo facultativa a grafia do extenso dos centavos [BAN90].

3.6. 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 TIF. O formato BMP não oferece uma boa compressão das imagens porém facilita o processamento posterior em sistemas de plataforma Windows. O formato TIF é mais utilizado pelos sistemas 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.

O processo de aquisição das imagens de assinaturas seguiu o mesmo procedimento descrito anteriormente, porém as imagens foram recortadas e gravadas em um arquivo no formato BMP, num retângulo de tamanho de 3x10 cm (1000x400 pixels), correspondendo à área destinada à assinatura no cheque.

Após o processo de aquisição das imagens, essas foram cadastradas segundo uma codificação, descrita na 1ª coluna do Quadro 02, 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.

4. Criação das Bases de Dados de Cheques Reais

Paralelamente ao processo de coleta dos dados de laboratório realizou-se a coleta e criação da base de dados de cheques reais. Estes cheques tem o objetivo de atender as necessidades dos trabalhos relativos ao processamento de imagens coloridas, segmentação e extração de informações específicas. Esta base de dados contém 700 cheques, entre cheques em branco e cheques preenchidos.

5. Conclusões

A criação de bases de dados de laboratórios de cheques bancários brasileiros ainda encontra-se em processo de coleta. Atualmente, foram coletados 2.016 cheques de laboratório, perfazendo um total de 16.986 dígitos e 13.952 palavras isoladas (valor numérico por extenso e meses do ano). Ao final do processo, espera-se obter uma base de dados de laboratório com 4.800 imagens de cheques.

O processo de criação das bases de dados reveste-se de grande importância para todos os pesquisadores, uma vez que as imagens contidas nas bases constituem a matéria prima para a realização dos seus trabalhos. A maior preocupação deste trabalho foi a de descrever o processo de criação das bases de dados de laboratório de cheques, uma vez que as bases de dados internacionalmente conhecidas não podem ser aplicadas no que se refere aos estudos de reconhecimento de palavras manuscritas em cheques bancários brasileiros.

As bases que estão sendo formadas: aprendizagem, validação e teste; são necessárias para o desenvolvimento de sistemas de reconhecimento de documentos manuscritos, independentemente da técnica ou método adotados. As bases de dados descritas estão sendo utilizadas nos processos de aprendizagem, validação e teste de sistemas de reconhecimento da parte manuscrita de cheques bancários brasileiros.

Referências Bibliográficas

- [AMM85] AMMAR, M.; YOSHIDA, Y.; FUKUMURA, T.. *Automatic off-line verification of signature based on pressure feature*. Proc. Tech. Group meeting of Pattern Recognition Learn, of IECEJ, Tokyo, 1985;
- [AVI96] ÄVILA, M. *Optimisation de modeles markoviens pour la reconnaissance de L'ecrit*. Thèse de doctorat, Université de Rouen, France, 1994;
- [BAN83] Banco Central do Brasil. *Manual e normas de instruções*. Brasília, DF. mai,1983;
- [BAN90] Banco Central do Brasil. *Decreto nº 99179*. Brasília, DF, out, 1990;
- [BRO85] Brocklehurst, Er. *Computer methods of signature verification*. J. Forence Science Society - vol. 25, 1985;
- [CÔT97] CÔTÉ, M. *Utilisation d'un modèle d'accès lexical et de concepts perceptifs pour la reconnaissance d'images de mots cursifs*. Thèse de doctorat, École Nationale Supérieure des Télécommunications, France,1997;
- [GIL93] GILLOUX, M.; LEROUX, M.; BERTILLE, J.M. *Strategies for handwritten words recognition using hidden markov models*. Proc. Of the Second International Conf. On Document Analysis and Recognition. ICDAR'93. 299-304p.;
- [GUI95] GUILLEVIC D. *Unconstrained handwriting recognition applied to the processing of bank cheques*. Thesis of Doctor's Degree in the Department of Computer Science at Concordia University. Canada.1995;
- [GRA99] GRANDIDIER, F.; SABOURIN, R.; EL YACOUBI, ^a ; GILLOUX, M.; SUEN, C.Y. *Influence of word length on handwriting recognition*. International Conference on Document Analysis and Recognition. ICDAR'99. 777-780p.;
- [HUA96] HUANG, K.; YAN, H. *Off-line signature verification based on geometric feature extraction and neural network classification*. Pattern Recognition, Vol. 30, No. 1, 1997. 9-17p.;
- [HUL93] HULL, J.J. & FENRICH, R.K. *Large database organization for document images*. In: Fundamentals in Handwriting Recognition, Edited by Sabastiano Impedovo. Nato ASI Series, Vol. 124, 1993. 397-414p.;
- [HUL94] HULL, J.J. *A database for handwritten text recognition research*. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol.16, No. 5, may, 1994. 550-554p.;
- [KIM96] KIM, G. *Recognition of offline handwritten words and its extension to phrase recognition*. PhD Thesis, University of New York at Buffalo, USA, mar., 1996;
- [KOE97] KOERICH, A . L. *Uma proposta para o processamento de imagens de cheques bancários: extração e armazenamento de informações*. Dissertação de Mestrado na Universidade Estadual de Campinas. Campinas, SP, fev., 1997.190p.;
- [LET96] LETHELIER, E. *Combinaison des concepts de segmentation et de reconnaissance pour l'écriture manuscrite hors-ligne: application au traitement des montants numériques des chèques*. Thèse de doctorat, L'Université de Rouen. Rouen, France, 1994;
- [MOH95] MOHAMED, M.A. *Handwritten word recognition using generalized hidden markov models*. PhD Thesis, University of Missouri-Columbia, USA, 1995;
- [MON95] MONTOLIU, L. *Architecture multi-agents et réseaux connexionnistes. Application à la lecture de chèques manuscrits*. Thèse de doctorat, École polytechnique, France, 1995;

- [NEM74] NEMCEK, W. F.; LIN, W. C. Experimental investigation of automatic signature verification. IEEE Trans. Syst. Man. Cybernetics 4, 1974. 121-126p.;
- [PLA88] PLAMONDON, R.; LORETTE, G. Automatic signature verification and writer identification - the state of the art. Pattern Recognition vol. 22, 1988. 107-130p.;
- [RIG98] RIGOLL, G.; KOSMALA, A. *A Systematic comparison between on-line and off-line methods for signature verification with Hidden Markov Models*. 14th International Conference on Pattern Recognition – vol II, 1998. 1755-1757p.;
- [SAB90] SABOURIN, R.; PLAMONDON, R. Progress in the field of automatic handwritten signature verification systems using gray-level images. Inter. Workshop on Frontiers in Handwriting Recognition, April 2-3, Montreal, Canada, 1990;
- [VIA99] VIARD-GAUDIN, C.; LALLICAN, P.M.; KNERR, S.; BINTER, P. *The IRESTE On/OFF (IRONOFF) dual handwriting database*. International Conference on Document Analysis and Recognition. ICDAR'99. 455-458p.;
- [YAC96] EL YACOUBI, A. *Modélisation Markovienne de l'écriture manuscrite. Application à la reconnaissance des adresses postales*. Thèse de doctorat, Université de Rennes, France, 1996.