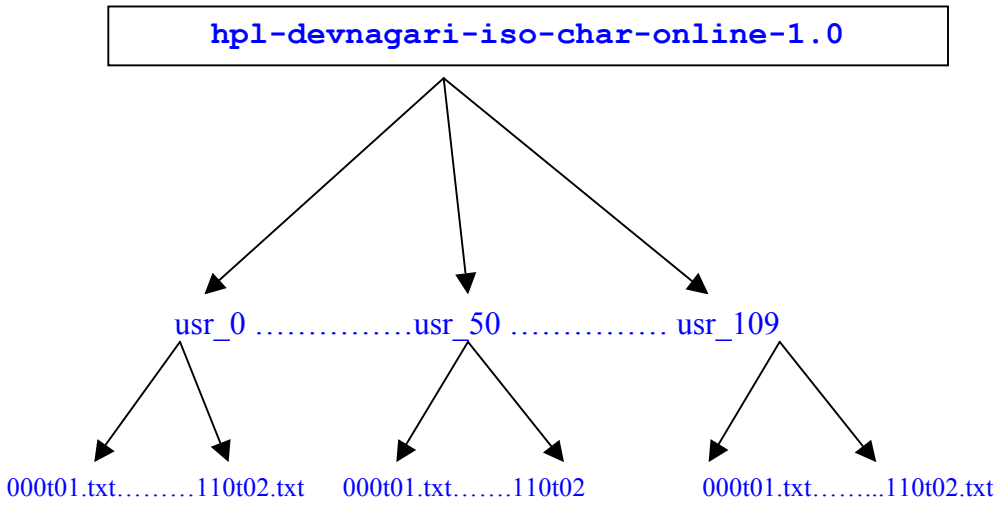
Dataset *hpl-devnagari-iso-char-onine-1.0* and *hpl-devnagari-iso-char-offline-1.0*

Updated Jun 9, 2009

The dataset hpl-devnagari-iso-char-online-1.0 contains samples of the 111 character classes collected from different writers on ACECAD DigiMemo (A4 sized) using a AcecadDigimemoDCT application. The data collection forms were designed using the Adobe Designer, and were printed on A4 sheets. The writers were asked to write on the printed A4 sheets using ACECAD Digimemo. Most writers contributed two samples per class ("trials"). Details of the directory structure, file contents and statistics regarding the number of samples per class are provided in the following sections.

The directory structure is same for both the online and offline datasets except that the name of the root directory in the offline case would be hpl-devnagari-iso-char-offline-1.0. The offline data contains bilevel TIFF images derived from the online data.

1. Directory Structure



• *hpl-devnagari-iso-char-online-1.0* directory is the root directory for the online data.

• Ink data is organized by writer into subdirectories of the form usr_<*user-id*>, e.g., usr_16 corresponds to the 17th writer.

User-ids are in the range 0…109, but not contiguous. Samples of four writers (usr_9, usr_58, usr_72, and usr_95) were found to be very noisy and were discarded.

• Ink data is stored in files of the form <*3-digit class-id*>t<*trial-id*>, e.g. 008t03.txt implies the 3rd trial of the character with class-id 008.

• Class-id is in the range 000 …110

• Trial-id is in the range 01 … 02. However, some samples written by a user might have been found noisy and hence discarded.

2. Ink File Contents

• In each file, ink data is represented in UNIPEN v1.0 format, as shown below.
• The channels reported for each ink point are X,Yand T. Files corresponding to some users have valid T (time) values for the first and last points of each stroke, with intermediate values set to 0. For other users, the time channel is set to 0 for all points.

```
.VERSION 1.0
.HIERARCHY CHARACTER
.COORD X Y T
.SEGMENT CHARACTER
.X_POINTS_PER_INCH 2500
.Y_POINTS_PER_INCH 2500
.POINTS_PER_SECOND 1200
.PEN_DOWN
935 523 0
935 523 0
935 523 0
935 520 0
935 517 0
935 514 0
935 511 0
.PEN_UP
```

For more information on UNIPEN format please refer to http://unipen.nici.ru.nl/unipen.def

3. Samples per class

Character Id.	Number of Samples
0	201
1	205
2	206
3	202
4	205
5	203
6	204
7	199
8	204
9	204
10	204
11	202
12	204
13	203
14	206
15	202
16	203
17	207
18	199
19	201
20	205

21	202
22	206
23	204
24	203
25	205
26	205
27	203
28	199
29	202
30	197
31	203
32	205
33	202
34	203
35	204
36	208
37	206
38	208
39	208
40	205
41	202
42	206
43	205
44	205
45	201
46	200
47	204
48	199
49	205
50	203
51	205
52	206
53	206
54	207
55	206
56	206
57	207
58	196
59	202
60	203
61	203
62	203
63	199
64	203
65	193
66	200
67	199
68	199

69	198
70	202
71	198
72	194
73	200
74	196
75	176
76	197
77	207
78	199
79	197
80	186
81	203
82	197
83	202
84	202
85	205
86	205
87	202
88	196
89	199
90	190
91	206
92	205
93	193
94	199
95	198
96	184
97	205
98	193
99	205
100	197
101	204
102	206
103	185
104	201
105	193
106	189
107	196
108	190
109	179
110	190