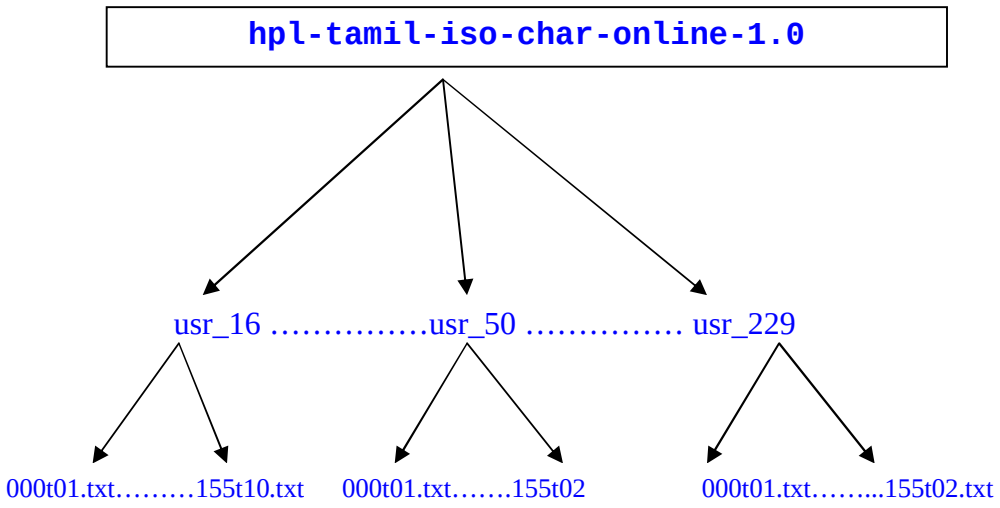
Dataset *hpl-tamil-iso-char-online-1.0 & hpl-tamil-iso-char-offline-1.0*

Updated Jun 8, 2006

The dataset hpl-tamil-iso-char-online-1.0 contains samples of the 156 character classes collected from different writers using a TabletPC application. Most writers contributed two samples per class ("trials"), a few contributed as many as 10. 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-tamil-iso-char-offline-1.0. The offline data contains bilevel TIFF images derived from the online data. Details of the directory structure, file contents and statistics regarding the number of samples per class are provided in the following sections

1. Directory Structure



- *hpl-tamil-iso-char-online-1.0* directory is the root directory for the online data.
- *hpl-tamil-iso-char-offline-1.0* directory is the root directory for the offline data.
- Ink data is organized by writer into subdirectories of the form usr_<*user-id*>, e.g., usr_16 corresponds to the 16th writer.
    - User-ids are in the range 16 … 229, but not contiguous
- 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 …155
    - Trial-id is in the range 01 … 02 for most users, 01 … 10 for some. However they are not guaranteed to be contiguous since bad samples may have been removed.

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.
- Since the ink was captured using Microsoft Ink Picture Controls on a TabletPC, spatial

resolution and sampling rate correspond to the interpolated ink returned by the control, and are higher than what is supported by the hardware.

```
.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

Id No	No of samples
0	527
1	521
2	522
3	509
4	518
5	525
6	513
7	522
8	512
9	501
10	497
11	522
12	517
13	497
14	508
15	508
16	524
17	523
18	513
19	492
20	518
21	521
22	519

23	508
24	516
25	515
26	503
27	503
28	512
29	498
30	521
31	477
32	271
33	509
34	469
35	508
36	490
37	515
38	512
39	512
40	511
41	516
42	506
43	520
44	511
45	518
46	519
47	519
48	518
49	515
50	518
51	521
52	524
53	511
54	500
55	271
56	511
57	475
58	515
59	486
60	513
61	524
62	510
63	520
64	504
65	502
66	518
67	508

68	514
69	509
70	512
71	514
72	500
73	507
74	502
75	509
76	513
77	511
78	270
79	512
80	509
81	519
82	502
83	502
84	506
85	509
86	516
87	512
88	514
89	508
90	516
91	515
92	521
93	510
94	520
95	494
96	502
97	507
98	519
99	488
100	460
101	501
102	495
103	512
104	505
105	525
106	522
107	522
108	523
109	521
110	522
111	509
112	519

113	418
114	495
115	507
116	511
117	507
118	528
119	504
120	512
121	488
122	443
123	426
124	272
125	470
126	457
127	455
128	438
129	266
130	440
131	439
132	516
133	487
134	507
135	516
136	524
137	523
138	524
139	506
140	513
141	520
142	523
143	508
144	525
145	516
146	508
147	510
148	518
149	520
150	511
151	484
152	271
153	508
154	472
155	516