文章编号：1003-0077(2019)07-0136-07

ISO/IEC 10646 国际编码标准下的香港电脑汉字编码及字形原则

熊 丹，陆 勤

（香港理工大学 电子计算学系，香港）

摘 要：在 ISO/IEC 10646 国际编码标准中，香港使用的汉字载于 H 列。该文介绍了如何在 ISO/IEC 10646 国际编码标准下进一步完善香港电脑汉字的扩展机制及 H 列字符字源资料的编码方案。由于目前 H 列的很多字形并未完全反映香港的实际习惯写法，因此香港制定了一套适用于香港常用写法的电脑汉字参考字形，该文介绍了此套字形的原则。

关键词：电脑汉字编码；字形；字符集

中图分类号：TP391 **文献标识码**：A

Character Encoding and Glyph Principles for Hong Kong's Chinese Computer Systems under the ISO/IEC 10646

XIONG Dan, LU Qin

(Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China)

Abstract: The Chinese characters used in Hong Kong are listed in the H column in the ISO/IEC 10646. This paper introduces the further improvements for the extension scheme for the characters in Hong Kong's Chinese computer systems and for the encoding scheme of the character resource references in the H column. Since the current glyphs for the Chinese characters in the H column do not really reflect the actual shapes of the glyphs commonly used in Hong Kong, the Reference Glyphs for Chinese Computer Systems in Hong Kong is developed, and the principles for this set of reference glyphs are presented.

Keywords: character encoding for Chinese computer systems; glyph; character set

0 引言

因香港的汉字书写系统与台湾相同，因此香港的电脑平台普遍沿用台湾工业标准“大五码”编码标准[1]。为补充《大五码字符集》的不足，香港政府建立了《香港增补字符集》。初版于 1999 年发布[2]，在 2016 年前共更新三次[3-5]。但其中只收录了 ISO/IEC 10646 国际编码标准[6]未收录、而香港需要使用的字符。即在与 ISO/IEC 10646 国际编码标准接轨的机制上，仅包括“纵向扩展”的方式。对于 ISO/IEC 10646 国际编码标准已收录、但须反映香港书写习惯的字形，则未予以增收。随着这方面的需求渐增，有必要增设相应的扩展机制，进一步完善《香港增补字符集》在 ISO/IEC 10646 国际编码标准下的编码方案。经过各方努力，这一新方案于 2016 年 5 月通过《香港增补字符集-2016》[7]正式公布。

在 ISO/IEC 10646 国际编码标准中，不同国家、地区的字形分别展示于不同列，香港使用的汉字载于 C(中国)列下的 H(香港)列。但是，目前 H 列的很多字形并不能完全反映香港的实际习惯写法。鉴于此，为方便厂商提供香港实际使用的字形，促进本地化技术的发展，有必要制定一套适用于香港的电脑汉字字形，以清晰列出 ISO/IEC 10646 国际编码标准中 H 列的字形，为香港电脑汉字提供方便且易于对照的参考字形，供业界参考。经过三年多的整理，这套“香港电脑汉字参考字形”[8]已于 2017 年 5 月随《香港增补字符集-2016》正式公布。

收稿日期：2018-07-10 **定稿日期**：2018-07-19

基金项目：香港特别行政区政府创新科技署创新及科技基金(ITS/317/12)

本文中，字形“特指构成每个方块汉字的二维图形。构成汉字字形的要素是笔画、笔数及汉字部件的位置关系等。”[9]本文描述的对象为电脑系统所使用的汉字，虽然目前通用的电脑平台均提供多种字体①，如宋体、楷体、黑体等，但本文对字形的描述并不针对或局限于某种字体。“字符”包括汉字和符号（如部首、数字、标点符号、货币符号等），本文中的汉字不仅包括中国使用的中文字，还包括其他国家、地区使用的汉字。近代语言学家把汉字称为“表意文字”，也有学者指出不妥，因为汉字包括大量表音的成分[10]。本文并不深究此名称是否妥当，而跟从 ISO/IEC 10646 国际编码标准中的名称，也使用“表意文字”指来自中国、日本、韩国及其他国家和地区的汉字。

本文主要内容如下：第 1、2 节分别介绍 ISO/IEC 10646 认同规则、扩展机制，以说明国际标准对香港此项工作的要求。第 3 节阐述根据新的需求，如何补充和完善香港电脑汉字的扩展机制和编码方案。第 4 节通过实例详细说明制作香港电脑汉字参考字形的原则。第 5 节简要进行总结。

1　ISO/IEC 10646 认同规则

在 ISO/IEC 10646 国际编码标准中，来自中国、日本、韩国及其他国家和地区的表意文字会经过一套认同（Unification）规则[11]整合，然后获分配 ISO/IEC 10646 码位，继而被统称为“中日韩统一表意文字”。基于此规则，被认同的字虽然形体不同，但不会分开编码。根据认同规则的定义，抽象形状（abstract shape）的差异不被认同，在这种情况下，即使异体字[12]也会分别编码，例如，部件的数量不同（如“采”“採”）、部件的相对位置不同（如“够”“夠”）、结构不同（如“群”“羣”）等。而具体形状（actual shape）的差异则被认同，如笔画的旋转方向不同（如“敝”“　”的首两笔）、笔画是否接触（如“爪”“　”的第三、四笔）、笔画的起笔处或收笔处是否穿出（如“急”“　”的第四笔）等。

在 ISO/IEC 10646 国际编码标准字符表[6]中，不同国家、地区的字形分列展示，包括 C（中国）、J（日本）、K（韩国）、V（越南）。其中 C 列下又分为三列：中国内地（G）、香港（H）、台湾（T）。每个字形下面标示其字源资料（Source Reference），即原有国家、地区的字源标准索引，以显示其来源。图 1 展示了 ISO/IEC 10646 国际编码标准字符表中的两组不认同字：够，夠；羣，群。

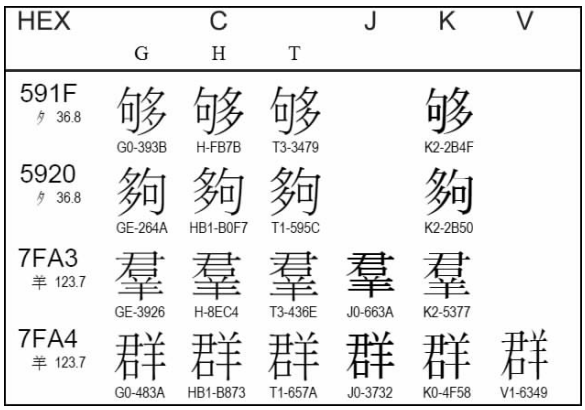

图 1　ISO/IEC 10646 国际编码标准中的不认同字例

在认同规则下，不同国家、地区的字形可以有所差异，因这些字被认同，所以获分配同一个 ISO/IEC 10646 码位。图 2 展示了 ISO/IEC 10646 国际编码标准中的认同字例，图上圈中所示的差异均被视为可认同，不会因这些差异而分开编码。

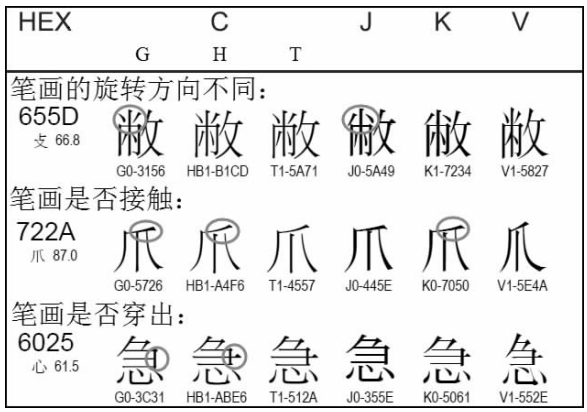

图 2　ISO/IEC 10646 国际编码标准中的认同字例

2　ISO/IEC 10646 扩展机制

2.1　概述

为了方便各个国家、地区的字符集与 ISO/IEC 10646 国际编码标准接轨，使其字符获分配 ISO/IEC 10646 码位，传统的扩展机制包括纵向扩展（Vertical Extension）及横向扩展（Horizontal Extension）。纵向扩展用于增收 ISO/IEC 10646 国际编码标准中尚未编码的新字；横向扩展用于为已编码的字符补充某个国家、地区的字形。但是，这两个

① 指“同一汉字由于各种原因（历史演变、书写、印刷等）而形成的各种不同体式。”[9]

机制并不允许为可认同的异体字分别编码。一种新的技术——注册“表意文字异体字序列”(Ideographic Variation Sequence,简称IVS),补充了这两个机制的不足,为电脑系统提供了一套符合统一码标准的方案,使ISO/IEC 10646国际编码标准中原本不允许分开编码的异体字得以编码。第2.2～2.4节详细解释这三个扩展机制。

2.2 纵向扩展

随着时代的推进和社会的发展,汉字也在不断演变,尤其是在互联网发达的当今,新的汉字层出不穷。当一个新字在某个国家或地区广泛流通时,无论是文字工作者,还是普通民众,都会强烈希望这个新字能被电脑系统所支持,从而能在日常生活、工作中得以正常使用。纵向扩展的机制便使此类新字被纳入ISO/IEC 10646国际编码标准,获得ISO/IEC 10646码位。电脑平台、软件、字体、输入法等开发商便有规范可遵从,从而使得这些新字能在电脑系统中得以支持。图3列举了一个纵向扩展的样例,汉字“⿰石示”在中国用作地名,而之前未被纳入ISO/IEC 10646国际编码标准,因此电脑系统未能支持,网络上出现很多因无法在电脑上输入此字而求助的案例。2015年,由中国提出申请,希望能将此字作为“迫切需要编码的字符”(urgently needed characters,UNC)纳入ISO/IEC 10646国际编码标准。2017年发布的ISO/IEC 10646:2017[6]中,此字获分配码位U+2E014①,被纳入中日韩统一表意文字扩展区F(CJK Unified Ideographs Extension F)。

图3 纵向扩展的中日韩统一表意文字字例

2.3 横向扩展

事实上,各个国家、地区使用的绝大多数字符在ISO/IEC 10646国际编码标准中已编码,目前的最新版本ISO/IEC 10646:2017[6]已收录超过130 000个字符(包括汉字和符号)。然而,各个国家、地区的书写习惯有所不同,因此常用的字形存在差异,而这些差异在认同规则下不允许被分开编码,因此需要通过横向扩展的机制来补充其习惯使用的字形,以反映该国家、地区实际的写法。例如图4所示的“豏”,如需补充越南使用的字形,则通过横向扩展的方式将其纳入V列。

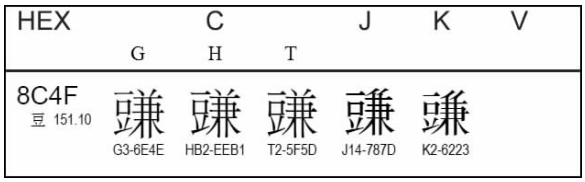

图4 横向扩展的中日韩统一表意文字字例

2.4 IVS技术

自古以来,汉字一直存在大量异体字,即“汉字通常写法之外的一种音同、义同,只是字形笔画或结构不同的字。”[12-13]现今,有些异体字仍会被使用。但是,受限于认同规则,有些异体字并不允许与其对应的正体字分开编码。如果ISO/IEC 10646国际编码标准中已收录某个国家、地区所使用的正体字字形,则该异体字便无横向扩展的空间,因此无法被电脑系统所支持。为了弥补这一不足,统一码联盟(The Unicode Consortium)建立了一套汉字字形定义技术,通过注册IVS,ISO/IEC 10646国际编码标准中原本不允许分开编码的异体字得以编码,从而能被电脑系统所支持。该方案的原理是,在一个“基本字”(base character)②后加上一个异体选择符(variation selector)[14],使其组合成IVS,用于定义与已编码汉字认同的异体字,并通过注册添加到表意文字异体字数据库(ideographic variation database,IVD)中,使该异体字得以在ISO/IEC 10646体系下编码,并保留异体字的字形。目前,异体选择符共有240个,编码为U+E0100至U+E01EF。以常用字“割”为例,ISO/IEC 10646国际编码标准字符表[6]中所列的各国家、地区字形如下:

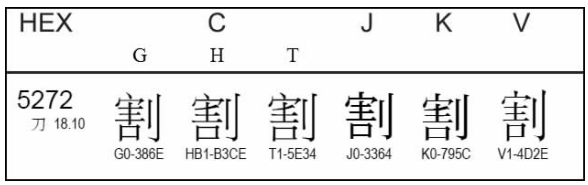

图5(a) 注册IVS的基本字字例

如图5(a)所示,J列已列出了日本的代表字形,但在日本还会使用其他字形。由于ISO/IEC 10646

① 通常加一个前置的“U+”来表示。

② 是一个ISO/IEC 10646中已编码的中日韩统一表意文字。

国际编码标准中已无横向扩展的空间，因此通过注册 IVS 来对这些异体字编码，以便电脑系统能支持。

在名为“Adobe－Japan1 collection”[15]的 IVD 集合中，日本以“割”(U＋5272)为基本字，共注册了 3 个异体字。图 5(b)展示了 IVD 中这些异体字的字形及相关注册信息。比较这些异体字与图 5(a)中 J 列所示的基本字的字形，中间的异体字字形和基本字差异较细微(左部部件“害”的三横笔相对长度不同)，而左、右两个异体字字形差异较明显(左部部件“害”的横笔变异为撇笔，且竖笔的收笔处向下穿出)。在 IVD 中，左边会以较大字号列出基本字的 ISO/IEC 10646 码位，即“割”的码位“5272”，异体字的下部则显示其异体选择符(E0100 至 E0102)及注册主体根据一定规则[16]提供的标识符(identifier)。

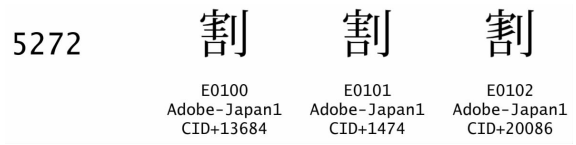

图 5(b)　注册 IVS 的异体字字例

3　香港电脑汉字编码方案

3.1　扩展机制

《香港增补字符集》的前四个版本[2-5]均是使用纵向扩展的方式增收字符，尚未有横向扩展的需求。例如，图 6 所示的“禧”，之前并未收录在 ISO/IEC 10646 国际编码标准中。因在香港用作人名，且已见于香港身份证上，属于已流通的用字，因此《香港增补字符集-2008》[5]增收此字，继而通过纵向扩展的方式纳入 ISO/IEC 10646:2011[17]中。

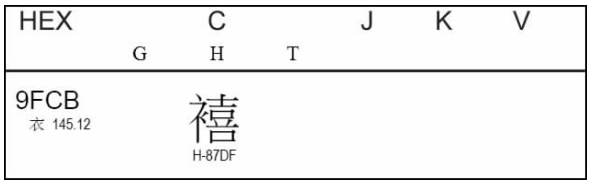

图 6　香港电脑汉字纵向扩展字例

而在整理《香港增补字符集-2016》[7]的过程中，因各种原因出现不同类型的横向扩展需求，主要分为三类。

(1) 因“原字集分别编码原则”(Source Separation Rule)[11]而需横向扩展的汉字：有些可认同的汉字在 ISO/IEC 10646 国际编码标准中因原字集分别编码原则获分配两个不同的码位，若修改其中一个码位的字形，会出现两个不同的码位对应同一个字形的情况，从而破坏 ISO/IEC 10646 的编码原则。以“兌”(U＋514C)和“兑”(U＋5151)为例，在 ISO/IEC 10646 国际编码标准中，这两个字因原字集分别编码原则而分别编码。香港的常用写法是“兑”，但反映香港字形的 H 列是“兌”(U＋514C)，而没有“兑”(U＋5151)(因为 U＋5151 不在《大五码字符集》和《香港增补字符集》内)。但是，不能直接把 U＋514C 这个码位的字形修改为“兑”，因为这样会令 U＋514C 和 U＋5151 两个码位对应相同的字形。因此，《香港增补字符集-2016》增收“兑”(U＋5151)后，再以横向扩展的方式纳入 ISO/IEC 10646 国际编码标准。《香港增补字符集-2016》共增收了 22 个此类汉字。图 7 展示了“兌”(U＋514C)、“兑”(U＋5151)在 ISO/IEC 10646 国际编码标准中的字形及香港横向扩展的字形。

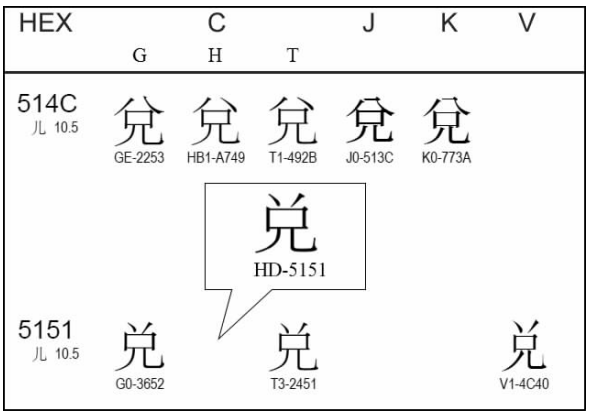

图 7　香港电脑汉字横向扩展字例

(2) 有流通需要而横向扩展的汉字：“鮫鰊”作为一种食用鱼，是香港的常用词汇。“鮫”(U＋9B9F)早已收录于《香港增补字符集-1999》[2]；因既非大五码字符也未被《香港增补字符集》收录，“鰊”(U＋9C47)未被视作香港使用的汉字列入 ISO/IEC 10646 国际编码标准的 H 列。因此，《香港增补字符集-2016》增收此字，并将以横向扩展的方式纳入 ISO/IEC 10646 国际编码标准。

(3) 有流通需要而横向扩展的符号：由于大五码编码标准产生于欧元区成立之前，因此《大五码字符集》中未包括欧元货币符号“€”(U＋20AC，名为“EURO SIGN”)，该符号已被各国家、地区的电脑系统所支持且广泛使用，因此，《香港增补字符集-2016》增收此符号。

目前，香港没有注册IVS的案例，但已具备可行的机制。例如，“丽”(U+4E3D)，图8中H列所示的为香港使用的字形，如果今后香港需要如G列所示的字形(上部的横为一笔)，因无横向扩展的空间，可通过注册IVS的机制增收并获得编码。

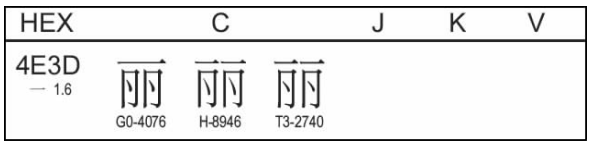

图8 香港可注册IVS的字例

3.2 编码方案

香港使用的汉字载于ISO/IEC 10646国际编码标准的H列，它不仅包括《香港增补字符集》的汉字，还包括香港使用的大五码字符(虽然香港对该字符集的字符无所有权)。《香港增补字符集》的前四个版本仅定义了此字符集的字源资料编码格式为H-XXXX(“XXXX”是该字符的大五码编码)。随着扩展机制的补充，H列字符字源资料的编码格式也相应进行了扩充，并在《香港增补字符集-2016》中进行了定义，具体如下：

(1) H-XXXX：表示《香港增补字符集-2008》的字符，其中，“XXXX”是该字符的大五码编码。(《香港增补字符集-2008》是《香港增补字符集》大五码编码部分的最后版本，之后，香港不再为《香港增补字符集》新增收的字符提供大五码编码。)

(2) HB-XXXX：表示大五码字符，具体格式为“HB0-XXXX”、“HB1-XXXX”和“HB2-XXXX”，分别代表大五码符号区、常用字集和次常用字集，其中“XXXX”是该字符的大五码编码。

(3) HC-XXXX：表示后续以纵向扩展的形式加入ISO/IEC 10646国际编码标准的新字符，其中，“XXXX”是“0001”至“9999”的顺序编号。《香港增补字符集-2016》尚无以“HC”定义的字符。

(4) HD-XXXX[X]：表示以横向扩展的形式纳入的汉字，其中，“XXXX[X]”是该汉字的ISO/IEC 10646码位。如字符在ISO/IEC 10646的基本平面①，编码为4位；如在其他平面，则为5位。《香港增补字符集-2016》共有23个以“HD”定义的字符。

(5) HE-XXXX：表示以横向扩展的形式纳入的符号，其中，“XXXX”是该符号的ISO/IEC 10646码位。《香港增补字符集-2016》共有1个以“HE”定义的符号。

4 制作香港电脑汉字参考字形的原则

4.1 总体原则

虽然香港对《大五码字符集》的字符无所有权，但因香港会使用，因此ISO/IEC 10646国际编码标准的H列也列出了大五码字符，字源资料的编码代号为“HB”(如3.2节所述，分为“HB0”、“HB1”及“HB2”)。目前ISO/IEC 10646国际编码标准中，“HB”所示的字形并非香港提交，因此很多字形并不能完全反映香港的实际习惯写法。所以，有必要制作一套完整的香港电脑汉字字形，涵盖《香港增补字符集》和《大五码字符集》的所有汉字，以在ISO/IEC 10646国际编码标准中更清晰、明确地显示H列的所有字形，供业界参考。经过三年多的整理，香港特区政府于2017年5月随《香港增补字符集-2016》公布了正式文件《香港电脑汉字参考字形》[8]。

从标准化的角度考虑，兼顾整齐、美观等因素，这套参考字形的基本原则是在部件层面具备一致性。《常用字字形表(二零零零年修订本)》[18]以手写楷书列出了香港常用字的习惯写法，为了反映香港的书写习惯，香港电脑汉字参考字形以此为主要参考资料。沿用香港政府此前制定《香港电脑汉字楷体字形参考指引》、《香港电脑汉字宋体(印刷体)字形参考指引》[19]的原则，这套参考字形楷体和宋体采用相同的字形规律。

从美学的角度考虑，不同的字体在设计上或有差异，这套香港电脑汉字参考字形仅用于清晰显示ISO/IEC 10646国际编码标准中H列的字形，并不限制字体开发商所采用的字体风格，也不用于规定香港社会的日常用字。

4.2 具体原则

4.2.1 部件层面的一致性

香港电脑汉字参考字形的基本原则是在部件层面具备一致性，在审视字形时，主要遵从《常用字字形表(二零零零年修订本)》，或根据此字形表中常用字部件的写法进行类推。当《常用字字形表(二零零零年修订本)》中无字形可参考或无法类推时，则参考《康熙字典》[20]，或从字源角度审视。当参考资料不一致、或因其他原因有待讨论时，则提交香港中文

① ISO/IEC 10646国际编码标准架构内的基本多文种编码平面，简称基本平面，也称“平面0”，码位由0000至FFFF。

界面咨询委员会工作小组会议讨论决定。

在部件层面具备一致性要求同一部件的写法一致。本文中，部件指“由笔画组成的具有组配汉字功能的构字单位”[21]，提及“部件”时为泛指，不区分成字部件和非成字部件、基础部件和合成部件。例如，表 1 列出了“如”在中国内地①、香港、台湾②的常用字形，香港常用字形的左偏旁③写法明显不同于另外两个字形（香港字形的第三笔为提，且收笔处不穿出撇笔）。根据一致性原则，含左偏旁“女”的字，无论是作为整个汉字的左偏旁（如“好”、“姑”、“妙”等），或汉字合成部件的左偏旁（如“呶”、“菇”、“努”等），香港的常用写法为表 1 的第二列所示。

表 1　左偏旁“女”的不同字形样例

如	如	如
（中国内地）	（香港）	（台湾）

部件层面具备一致性并非强求所有部件的写法完全相同，还需考虑部件所在的部位和字义。例如，当“女”并非整个汉字或汉字合成部件的左偏旁时，香港的习惯写法为“女”（第三笔为横，且横笔、撇笔互相穿出），例如“女”、“　”、“”。这是“女”的原形，称为“主形部件”，而左部部件“女”是其变体部件，虽含义相同，但因部位不同而写法有所变异（笔势上横变提）。据研究，这种笔形④变异的原因主要包括为求书写连贯、结构紧凑、构形美观[23]。

当构字部件的来源、含义不同时，部件的字形可能有所不同，不能强求其写法的一致。例如，来源为“月”的部件“月”（中间为两横）及来源为“肉”的部件“月”（中间为点和提）：中国内地所使用的简体字统一为“月”，但在香港、台湾则有明显的区分；而来源为“肉”的部件“月”位于右部和下部时，香港的习惯写法为“月”（首笔为竖，中间为两横），台湾则保留内部为点、提的写法。表 2 列出了这几个不同来源的相关部件在中国内地、香港、台湾常用的字形字例。如上文所述，部件所处的部位也会影响其字形。

4.2.2　捺笔变形为顿点的法则

为求轮廓整齐、结构平衡、笔画匀称，汉字书写中捺笔收敛为顿点的情况颇为常见，而这种笔形变异“不会影响汉字的构形和构意”[23]。在此之前，ISO/IEC 10646 国际编码标准中 H 列的香港代表字形无一定规律可循。因此，在整理香港电脑汉字

表 2　“月”相关部件在不同地区的常用字形字例

地区	来源：月	来源：肉⑤
中国内地	1. 左、右偏旁⑥：月，如：朦、明 2. 下部：月，如：有	1. 左、右偏旁：月，如：肚、胡 2. 下部：月，如：胃、肩、胤
香港	1. 左、右偏旁：月，如：朦、明 2. 下部：月，如：	1. 左偏旁：月，如：肚、腿 2. 右偏旁：月，如：胡 3. 下部：月，如：背、肩、胤
台湾	1. 左、右偏旁：月，如：朦、明 2. 下部：月，如：有	1. 左、右偏旁：月，如：肚、腿、胡 2. 下部：月，如：背、肩、胤

参考字形的过程中，基于部件的部位，归纳了大多数字的香港习惯写法，从而形成了《香港电脑字形原捺笔变形为顿点法则》[8]。这套法则列出了香港习惯使用的字形捺笔收敛为顿点的几条规则，且按优先级排序。此法则明确了一些总体原则，例如，捺笔右边尚有其他部件，该捺笔一律收敛为顿点（如“私”、“瓣”）；一字不两捺（如“奏”、“食”），但“入”、“水”等部件除外（如“汆”、“　”）等等。对于一些涉及字数较多且较复杂的情况，则酌情考虑。例如，当汉字的上部为左右结构，右部有捺笔时，该捺笔是否保留或收敛为顿点，则取决于上部与下部的相对宽度：下比上宽时捺笔收敛为顿点，如“警”、“熬”；下比上窄，或下部上窄下宽时，则保留捺笔，如“繁”、“堅”。此套法则从笔形变异的角度更增强了香港电脑汉字参考字形的一致性。

4.2.3　可容许的细微风格差异

虽然香港电脑汉字参考字形采用楷体和宋体字形规律一致的原则，但同时也尊重两者固有的差异，如“花”的上部：宋体“花”上部的两笔短竖略向内倾斜，显得对称；而据楷体的书写习惯，楷体“花”的第四笔为短撇。从视觉审美的角度考虑，即使同为宋

① 本文中中国内地的字例均采用中国内地常用的宋体字体（SimSun）。

② 本文中台湾的字例均采用台湾教育部标准宋体[22]。

③ “偏旁”是“合体字的构字单位的传统称呼。旧称合体字左为偏，右为旁，现在统称偏旁。”[9]

④ 指“笔画的具体形状”[21]。

⑤ “然”的左上部也含“肉”之意，但因字形差异较大，因此不在此表中列出比较。

⑥ 包括作整個字左、右偏旁及构字合成部件左、右偏旁的情况，如“明”作为字例的情况也适用于“盟”。

体，不同开发商的字体也会风格各异，如香港常用的“江”和中国内地常用的“江”，第三笔风格迥然不同。香港电脑汉字参考字形的制定和标准化并不限制此类风格差异，但其自身则会保持一致。

5 结语

本文介绍了如何在 ISO/IEC 10646 国际编码标准下，根据认同规则的要求进一步完善香港电脑汉字的扩展机制。在与 ISO/IEC 10646 国际编码标准接轨的机制上，不仅包括“纵向扩展”的方式，还包括“横向扩展”，并具备了注册 IVS 的可行机制。根据这些扩展机制，完善了 ISO/IEC 10646 国际编码标准下香港字符的字源资料编码方案。在 ISO/IEC 10646 国际编码标准中，香港使用的汉字载于 H 列，但目前 H 列的很多字形并未完全反映香港的实际习惯写法，因此制定了一套适用于香港习惯写法的电脑汉字参考字形，本文介绍了此套字形所遵从的总体和具体原则。

参考文献

[1] 电脑用中文字型与字码对照表(台湾工业标准“大五码”的正式文献)[S]. 台北：财团法人资讯工业策进会，1984.

[2] 香港增补字符集-1999[S]. 香港：香港特别行政区政府，1999.

[3] 香港增补字符集-2001[S]. 香港：香港特别行政区政府，2001.

[4] 香港增补字符集-2004[S]. 香港：香港特别行政区政府，2005.

[5] 香港增补字符集-2008[S]. 香港：香港特别行政区政府，2009.

[6] ISO/IEC 10646：2017 Information technology - Universal Coded Character Set (UCS) [S]. Switzerland：ISO，2017.

[7] 香港增补字符集-2016[S]. 香港：香港特别行政区政府，2017.

[8] 香港特别行政区政府 政府资讯科技总监办公室与公务员事务局法定语文事务部. 香港电脑汉字参考字形[M]. 香港：香港特别行政区政府，2017.

[9] GB/T 12200.2-94，中华人民共和国国家标准 汉语信息处理词汇 02 部分：汉语和汉字[S]. 北京：中国标准出版社，1994.

[10] 裘锡圭. 文字学概要[M]. 北京：商务印书馆，2016：9-20.

[11] ISO/IEC 10646：2017，Annex S (informative) Procedure for the unification and arrangement of CJK Ideographs [S]. Switzerland：ISO，2017.

[12] 国发[2013]23 号，通用规范汉字表[S]. 北京：中华人民共和国教育部、国家语言文字工作委员会，2013.

[13] GB/T 12200.1-90，中华人民共和国国家标准 汉语信息处理词汇 01 部分：基本术语[S]. 北京：中国标准出版社，1990.

[14] Ken Lunde. CJKV Information Processing (2nd Edition) [M]. USA：O'Reilly Media，2008：171.

[15] The Unicode Consortium. Ideographic Variation Database [DB/OL]. [11 Sep 2016]. http://www.unicode.org/ivd/.

[16] The Unicode Consortium. Unicode Technical Standard # 37，Unicode Ideographic Variation Database [DB/OL]. [31 Jan 2017]. http://www.unicode.org/reports/tr37/.

[17] ISO/IEC 10646：2011 Information technology - Universal Coded Character Set (UCS) [S]. Switzerland：ISO，2011.

[18] 李学铭. 常用字字形表(二零零零年修订本)[M]. 香港：香港教育学院，2000.

[19] 香港特别行政区政府 资讯科技署与法定语文事务署. 香港电脑汉字楷体字形参考指引，香港电脑汉字宋体(印刷体)字形参考指引[M]. 香港：香港特别行政区政府，2002.

[20] 康熙字典[M]. 北京：中华书局，1997.

[21] GF 3001-1997，信息处理用 GB 13000.1 字符集汉字部件规范[S]. 北京：语文出版社，1997.

[22] 台湾教育部国语推行委员会. 国字标准字体宋体母稿<教育部字序>[M]. 台北：台湾教育部，1998.

[23] 王宁. 汉字构形学导论[M]. 北京：商务印书馆，2015：79-81.

熊丹(1980—)，硕士，研究助理，主要研究领域为词汇语义学、汉字编码及标准化、自然语言处理。E-mail：csdxiong@comp.polyu.edu.hk

陆勤(1960—)，通信作者，博士，教授，主要研究领域为计算语言学、词汇语义学、中文信息处理、基于自然语言处理技术的信息抽取和知识发现。E-mail：csluqin@comp.polyu.edu.hk