Universal Multiple-Octet Coded Character Set UCS

ISO/IEC JTC1/SC2/WG2 IRG N2496
Date: 2021-09-09

Source:	China
Meeting:	IRG#57 (online)
Title:	Activity Report
Status:	Member's submission
Actions required:	FYI
Distribution:	IRG
Medium:	Electronic
Pages:	3

1. Progress of the revision of GB 18030 *Information technology -- Chinese coded character set*

GB 18030 *Information technology -- Chinese coded character set* is one of the most important mandatory national standards of PRC. It has been revised once and the latest edition is GB 18030-2005. The third edition is now under review by the related authority.

In consideration of the great difficulty and cost of supporting such an enormous number of encoded characters, the third edition of GB 18030 sets up three implementation levels for different types of products to choose:

Implement Level 1--including 1-byte/2-byte characters and CJK Extension A.

Implement Level 2--including all characters of Implement Level 1 and the *The General Purpose Normalized Hanzi List*（*通用规范汉字表*）

Implement Level 3--including all characters of the coded character set

GB 18030 has so far been adopted by many mainstream operating system companies, font companies, and typesetting software companies.

For instance, Microsoft (China) Co.Ltd. claimed that all of their software products could at least reach implement level 1 of GB 18030, while Windows, Office and database products could reach implement level 2 (actually those software products are able to hit implement level 3 as long as fonts including all characters of the coded character set are installed).

HuaguangFont and FounderType, two Chinese font companies as well as typesetting companies, said that both of them have font products meeting the requirement of GB 18030 implement level 3 and including all encoded CJK unified ideographs in ISO/IEC 10646:2020. And all typesetting software products of those two companies could at least reach implement level 1.

2. Progress of the Project “Chinese Characters Repertoire” (中華字庫)

2.1 Fonts

The Project “Chinese Characters Repertoire” has finished the process of glyph collection. There are 790 thousand of 1 million fonts made by the end of 2021, which have been included in the Font named "The Intermediate Font". The Intermediate Font is the interim results of the project.With a long term of research on this basis, the project will select 300 thousand fonts from it to make the final product "The Achievement Font". One third has been selected so far.

2.2 Future Submission

Most of the glyphs of The Intermediate Font are variants. According to the new PnP rules, the variants need adequate research before submission. On the other hand, the project also found 57 thousand unencoded independent characters, which need careful textual research on the shape and meaning. In the subsequent work, the project will give priority to submitting independent characters.

3. Research of CJK ideograph font synthesis method and online handwriting style transfer model from Peking University

3.1 CJK ideograph font synthesis method (FontRL)

Peking University proposed FontRL, a novel method for CJK ideograph font synthesis by using deep reinforcement learning. Specifically, they first train a deep reinforcement learning model to obtain the Thin-Plate Spline (TPS) transformation that is able to modify the reference stroke skeleton in an average font style into the skeleton of a required style for each stroke of every unseen CJK ideograph character. Afterwards, they utilize a CNN model to predict the location and scale information of these strokes, and then assemble them to get the skeleton of the corresponding character.

Finally, they convert each synthesized character skeleton into the glyph image via an image-to-image translation model. Both quantitative and qualitative experimental results demonstrate the superiority of the proposed FontRL compared to the state of the art.

3.2 Online handwriting style transfer model

Peking University proposed a novel Sequence-to-Sequence model based on metric-based meta learning for the arbitrary style transfer of online CJK ideograph handwritings. Unlike most existing methods that treat CJK ideograph handwritings as images and are unable to reflect the human writing process, the proposed model directly handles sequential online CJK ideograph handwritings. Generally, our model consists of three sub-models: a content encoder, a style encoder and a decoder, which are all Recurrent Neural Networks. In order to adaptively obtain the style information, they introduce an attention-based adaptive style block which has been experimentally proven to bring considerable improvement to our model. In addition, to disentangle the latent style information from characters written by any writers effectively, they adopt metric-based meta learning and pre-train the style encoder using a carefully-designed discriminative loss function. Then, our entire model is trained in an end-to-end manner and the decoder adaptively receives the style information from the style encoder and the content information from the content encoder to synthesize the target output. Finally, by feeding the trained model with a content character and several characters written by a given user, our model can write that CJK ideograph character in the user's handwriting style by drawing strokes one by one like humans. That is to say, as long as you write several CJK ideograph character samples, our model can imitate your handwriting style when writing. In addition, after fine-tuning the model with a few samples, it can generate more realistic handwritings that are difficult to be distinguished from the real ones. Both qualitative and quantitative experiments demonstrate the effectiveness
and superiority of our method.

END.