deepcopy OCR algorithm design

Balázs Nyírő - diogenesz@pergamen.hu

Updated: Saturday 4th January, 2020

https://github.com/BalazsNyiro/deepcopy

0.1 OCR algorithm based on char analysing

0.1.1 What does 'char analysing' mean?

There are really good methods to recognise rasterised text with statistical methods, for example. Character analysing means that the algorithm study character attributes and sort the detected pixel blocks (marks) into classes based on their main attributes.

The algorithm first of all developed on latin based abc systems - I hope later it can be used to analyse other writing systems, too.

First of all the method is tested with plain text without frames, images and multiple columns - the more complex questions will be discussed later.

0.1.2 Keywords

- **background color** - a color (range) that doesn't contain information
- **text color** - pixels in a color range that means something, we want to detect them. During mark detection we can reach color informations, too - so in the future the program will be able to restore colorised text.
- **frame** - every text is in a frame. If you have an empty paper with one word in the middle of it, the page borders are the natural frame of it.
  - invisible frame - there are a natural distance between the visible characters on a page which is small enough to explicit define the position of the next character in a word. If the distance is bigger, the reader can separate words but he is able to switch from word to word in same line. If the distance is too big between words on a page then it's not well defined that after a word which one is the next: the next one in right or the next from the line below the current one. So, invisible frame means a text block where the characters and words are near enough to form a standalone unit - and it has enough distance from other texts to be separated.
  - visible frame - the other case of text separations are visible frames/borders. It's an obvious separation solution but in a children book or in a magazine a design background element can be a line that is detected as a border - and for a human reader it isn't mean a real frame. The detection of frames can be complex.
- **relative distance** - A rasterised image can contain more or less pixels about the same source page - so the number of pixels doesn't mean exact distance information. With the text analysing the algorithm define basic units based on characters: x, n, m width and height, baseline positions, etc. Because

- **glyph** - A glyph is a modifier. For example letter **e** is different from **é** - in this situation the acute changes the meaning of letter **e**. In this case the accent is a glyph (meaning modifier). But the accent in English **i** isn't a meaning modifier because letter **ı** (dotless i) doesn't exist in English alphabet so the dot on i isn't a modifier, isn't a glyph. Because later the algorithm will be used to detect different languages where the same sign can be a modifier or not a modifier, we use a new definition: mark instead of glyphs.

- **mark** - a separated information block. Character **i** has two marks, because the dot is separated from the body of the letter. One or more marks form a sign: a char, or a question mark, or anything that has independent meaning - this is the smaller unit of the text. In European alphabets a sign typically means a character but ligatures can be represented as one sign (ff ligature for example) but they mean more than one character. In Asian alphabets a sign can mean words for example - but in that case the algorithm try to define the original sign.

0.1.3 The attributes of characters

0.1.4 Character classification