Logical Structure Analysis by Typographic Characteristics Extraction

DUFFY Laurence, LEBOURGEOIS Frank, EMPTOZ Hubert
Laboratoire de Reconnaissances de Formes et Vision
INSA de Lyon - Bât 403 - 20 Ave. Albert Einstein
69621 Villeurbanne Cedex. Email : duffy@rfv.insa-lyon.fr

Abstract :

We propose to study the logical structure of a document through its typographic characteristics. So we present a new method of regrouping letters and words in homogeneous font families, which doesn't necessitate to explicitly recognise the font.

This analysis will allow us to extract a part of the logical information which is carried by words typographic features.

1 Context

Logical structure analysis of document is an important theme of research [NIYOGI95], [SATO95], [FISCHER95], indeed this logical structure stands between the content and the layout of the document. It is an inevitable way if we want to attain the way of « the printed document comprehension ». Generally, we link up this logical structure with a hierarchy of visual components which are observable by the reader. In a textual document, these components are titles, sections, sub-sections, lists, ...

Automatic recognition of logical structure, will only occur for document in which it is contained in graphical elements (In other terms, this logical structure will have to be extractable from the numerical image of the document). This extraction will be more easy if the publisher has created the document paste-up in order to accentuate this structure, which can help the reader to understand the text.

The publishers often use some different possibilities they have (Choice police, size, grease, and slope) in order to mark the logical structure. Figure 1 particularly well illustrates these typographical contributions. For example a bold word, with a big size, will often be representative of document significant topics, whereas document basis text will be written by means of a middle and normal police.

World Affairs	
Bosnia: So Much for the Vaunted	
'Safe' Areas	10
Russia: A Troubled Heartbeat	13
Islam: An Imam Is Murdered in the Mosque	13
Israel: Does God or the Government Call	
the Shots?	14
Burma: 'We Are All Behind Her'	15
South Korea: Miraculous Rescues	15
'Dispatches'	16
China: Hail the Iron Rice Bowl	16
Defense: 'Gangstas' in the Ranks	17

Fig. 1. Typographically rich document

Each font (police, size, grease, slope) carries a logical significance, it is fundamental to *perform an elaborate typographical analysis* of the document. We therefore suggest to distinct different fonts within documents, in order to obtain a best recognition of its logical structure.

In a first time, recognise the font name, wouldn't provide more information about the document logical structure. We are only interested in the fact that some words are written by means of the same font. So we begin to distinguish all characters patterns and to regroup them in families.

Our proposition supposes that layout and typography are coherent, so we assume that publisher respects the utilisation of the current police declination.

2 Font Analysis Using Patterns Redundancy

In a first time, we want to group together words which have the same font, but we don't want to identify their font.

2.1 Utilisation of Patterns Redundancy in OCR

Patterns redundancy is a technique used to regroup printed letters which have a similar pattern, in order to compress documents, or to decrease the number of patterns to recognise [LEBOU92] [CASEY82].

It consists in comparison of each letter with other letters, by partial template matching. All letters having a similar pattern, are joined to the same label (a family) which corresponds to the same prototype. On the other hand, a letter could produce several prototypes. Figures 2.a 2.b 2.c illustrate this method. We could note that the letter « a » produces several prototypes. Each of them differs from the other, by few pixels.

For the Olympics, hundreds of Korean-Americans have flown to Seoul, according to managers of Korean Air. Joining them were at least 50 students, who are serving as translators at the games.

Fig. 2.a Original text

Fig. 2.b. Prototype images

This method which allows to regroup all letters having a similar pattern, simplifies OCR software job ; recognition of prototypes permits to recognise entire text.

1.2.3 4.5.6 7.8.9.10.11.12.13.14
For the Olympics, I
 .15.16.17.3.18.17.14 2.19 20.2
i, hundreds of Ko-
 24.31.3 39.2.12.40.31.40.41 42.5.18.10

Fig. 2.c. Labelled text

2.2 Pattern Redundancy for the Location of Different Fonts

We want to group into families, letters which are written with the help of a same font.

We propose to classify letters by means of the patterns redundancy method previously evoked, in order to be independent from OCR program results. So we have obtained prototypes and labels of each characters.

For us, the words notion is most important than the characters notion, we have therefore completed our treatment by regrouping together words which have some common prototypes and so the same font, in homogeneous families.

In order to do that, we assign a number family to the first word, and then *we propagate this number family to each other word, which at least have one common prototype*. A part of the result of the treatment is illustrated figure 3.

Fig. 3. Part of one of the family

2.3 Importance of Words First Letter

Adaptation of this technic to our problematic raised some difficulties.

In a first time we use all word letters, in order to classify words in families. But we have quickly notice that the use of the first letter regroups, capital and minuscule

words of a same font, in a same family because the word first letter is often a capital letter. It is very important to differentiate capital and minuscule fonts because these families could carry a different logical significance.

A solution had been to ignore the word first letter in order to process the regrouping in families.

But this raised a new problem : the rejection of all words composed of one character, and of the big majority of words composed of two or three characters. Our prototypment never have place on first character but only on following characters, so we just have one or two characters to classify words.

This is insufficient because the more a word is long, the more the probability of one common prototype with of an other word is elevated. This problem takes more importance according that our segmentation technique is turned toward the oversegmentation, so we often obtain some part of words, that we couldn't reject.

It was therefore essential to take these first letters in account when it doesn't entail risks of confusion between minuscule and capital of a same font.

In order to verify that the first letter of each word has the same type (capital, minuscule, number) than the remainder of the word, we propose to calculate the letter and word body height. If these two heights are equal, it means that the first letter of the word has the same type than the word, so it can be use for the prototypment. Otherwise, we are in the case of one minuscule word beginning by a capital letter, and we reject this first letter.

In order to calculate words body height, and then differentiate fonts, it is necessary to *find very rigorously the basis lines* (CF figure 4).

Fig. 4. Annotations regarding the word basis lines.

3 In Search of Basis Lines

3.1 Our Method

Most of basis lines detection method [LE96] are projection-based techniques and aren't sufficient to answer ours requirements. We must indeed be able to find words basis lines of words which aren't greater than three characters, and projection-based methods don't always produce good results with short words.

So, in a first time we have formalised a technique of vertical probes throw, which marks, the high (respectively the low) of the first dark pixel meet. By summing these marks, we obtain two vertical histograms (CF figure 5).

Fig. 5. Search for basis lines

This method, compared to projection analysis, provides excellent results with words which don't exceed three characters, but could generate some mistakes with short words, specially if polices have serifs. Serif of some upstroke or downstroke characters, misleads us, and the retained basis lines are erroneous (CF figure 6).

Fig. 6. Case of the serifs letters

We have therefore interface our lines basis research system with a validation technique, which consists to verify the consistency of results, and if necessary, to relaunch research with other parameters.

3.2 Technique of Validation

Our validation method is based on fact that each word character is strictly obliged to appear in word body zone. The search of word characters in the body zone, allows us to find all of them, excepted characters that we qualified of insignificant and which aren't take in account (commas, dots, ...).

In order to seek these characters, we throw, in the middle of word body, an horizontal probe to test the presence of dark pixels in a columns interval. So we know if basis lines have been judiciously chosen. Nevertheless this technique doesn't inform us about the error type (choice of lcb too high, choice of lch too low, ...)

We are therefore went deeper into the topic and we have, for each type of characters body location in a word (CF figure 7), list the possible mistakes (CF figure 8).

Case 1
lh AB 13 m a m a n lch
lb

Case 2
lh A m i c a l m a l lch
lb

Case 3
lh n a p p a g e lch
lb

Case 4
lh A m p h i lch
lb

Errors case 1
lh a p a b a b p lch
lb

Errors case 2
lh A r r a a p l U H A p t lch
lb

Errors case 3
lh H E A p p l l lch
lb

Errors case 4
lh A H lch
lb

Fig. 7. Body location in the word

Fig. 8. Possible mistakes

Then, we could define for each case, the throw height and the optimum number of probes which validates the choice of basis lines.

Following the location, we therefore throw some probes to various heights and we sum characters which are crossed by the probe. The utilisation of these totals, associated to rules describing possible mistakes, assure us the validity of results, or indicate us the type of the most likely mistake (CF figures 9). Thus, when an error occurs, we seek new basis lines with new parameters.

Errors case 1

Top probe
Bottom probe

Err1 Err2 Err3

```

if nb_car_rencon par sondh =
nbcар_mot then
  if nb_car_rencon par sondb =
    nbcар_mot
  then coherent result
  else Err1
else
  if nb_car_rencon par sondb =
    nbcар_mot
  then Err2
  else Err3

```

Fig. 9. Processing of the mistake 1

Association of our method of basis lines research, with our validation technique provides very good results, even on two or three characters words (CF figures 10.a and 10.b).

Fig. 10.a. Research without validation

Fig. 10.b. Research with validation

Our basis lines research permits to validate the words first letter, and to not classify two or three characters words in reject family.

4 Results

The result of regrouping same font words in family is illustrated figure 11, a colour represents a family.

World Affairs	
Bosnia: So Much for the Vaunted	
'Safe' Areas	10
Russia: A Troubled Heartbeat	13
Islam: An Imam Is Murdered in the Mosque	13
Israel: Does God or the Government Call	
the Shots?	14
Burma: 'We Are All Behind Her'	15
South Korea: Miraculous Rescues	15
'Dispatches'	16
China: Hail the Iron Rice Bowl	16
Defense: 'Gangstas' in the Ranks	17

Fig. 11. A labelled chapter

Now we are searching families typographic characteristics. This part of the processing will be simplified because we won't search font with just one word, but with several words.

5 Conclusion

We have just presented a new method, of regrouping characters and words in homogeneous font families. This analysis, achieved with the application of one pattern redundancy technique, will allow us to extract the logical information which is carried by words typographic features.

We must have bring a new solution to problems of lines basis research, because they are very useful in order to differentiate words font, but also to develop an outstanding system of characters prototypment.

The significance of the method, we have developed, is that each process only depend on the image ; it isn't depend on the document type or on fonts data basis. So this method can be applied to every document type, specially complex and typographically rich documents.

6 Bibliography

- [CASEY82] **R.G. CASEY, G. NAGY** Recursive segmentation and classification of composite character. *6th ICPR, Intenational Conference on Pattern Recognition*, Paris, France, 1982, vol.2, p.1023-1025
- [FISCHER95] **FISCHER S., AMIN A. and DRIVAS D.** Segmentation of the Yellow Page. *Third ICDAR, International Conference on Document Analysis and Recognition*, Montréal, Canada, 1995, p. 605-609
- [LE96] **LE D.X., THOMA G.R. et WECHSLER.** Automated Borders Detection and Adaptative Segmentation for binary Document Images. *13th ICPR, Intenational Conference on Pattern Recognition*, Vienne, Austria, 1996, p.737-741
- [LEBOU92] **LEBOURGEOIS F., BUBLINSKI Z. et EMPTOZ H.** A fast and efficient method for extracting text paragraphs and graphics from unconstrained documents. *11th ICPR, Intenational Conference on Pattern Recognition*, La Hague, Hollande, 1992, p.272-276
- [NIYOGI95] **NIYOGI D. and SRIHARI S.N.** Knowledge-Based Derivation of Document Logical Structure. *Third ICDAR, International Conference on Document Analysis and Recognition*, Montréal, Canada, 1995, p. 472-475
- [SATO95] **SATO S., TAKASU A. and KATSURA E.** An Automated Generation of Electronic Library based on Document Image Understanding. *Third ICDAR, International Conference on Document Analysis and Recognition*, Montréal, Canada, 1995, p. 163-166