INTELLIGENT NEWS DETECTION AND EXTRACTION

Mahmut Kurşun[1]

INTRODUCTION

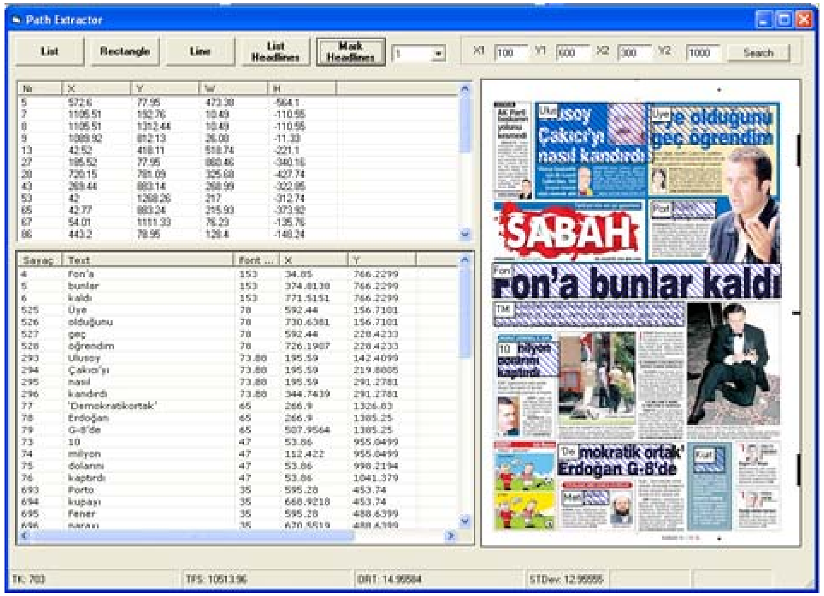
Newspapers in Turkey and in general have several editions, depending on which region they will be transported. As an example Turkey's one of the most sold newspaper Sabah is printed on four main print facilities in Turkey, namely in Istanbul, Ankara, Izmir and Adana. And depending on the distance from the print facility there are two main editions, long reaching first edition and second edition (city edition or latest edition). And there will be several changes between the first and second edition. The overall process is highly complicated, and since all the data is transferred to the print facilities page by page, the whole process is prone to some errors. All the matching pages of each edition should be carefully analyzed for the changes, therefore first the matching pages are to be controlled and listed. And there is a huge amount of work for web in the extraction of the news from the newspaper PDF. All available commercial PDF to Text extraction software unfortunately could not handle this specific case of SABAH, where the layout and design of the pages and text are highly sophisticated. So with this motivation the object of this report is to automate the underlying process, so that all the news from the PDF data will be automatically identified and extracted.

METHODOLOGY

The problem attacking strategy can be outlined in two main operations, the first is to work on the page-view images of the PDF, and the second is to work on the PDF itself. With the help of image processing, comparison of the page-view images of the first and second editions can be performed. It will be also possible to find out the matching pages of the first and second editions. Actually the main aim of this project is to automatically extract the headlines from any Sabah page. With the help of the 3 Heights PDF Export API to many and noisy valuable information is available to be used. But there is a need of visualization of the gathered data so that the data can be interpreted for further use. Therefore a new software has been developed which will help in the visualization of the gathered data on the page-view image of the relevant PDF page. And also statistical values like the average font size, and standard deviation of the font sizes are calculated. So every possible heuristic can be easily constructed, and every possible heuristic can be instantly tried for its effectiveness. So news extraction algorithm has been developed.

RESULTS

The differences in the pages of the editions could be found first graphically than textually. Automatic page matching algorithm is developed which works with almost 98% accuracy. The News Extraction Algorithm works with an accuracy of 96.2% on the non-sport news pages, and with 71.4% on sport only pages. The overall accuracy of the algorithm is 88.5%. These are quite satisfactory numbers for the industry, when compared to the benefits arising from the use of the software. Actually there is a huge performance degradation of the algorithm when used on the sport-only pages. The main reason for this is the special layout of the sport-only pages. Mainly the headlines of sport pages are converted to images, so that many visual effects are applied on them. So the when extracting no textual information could be extracted as text, only image extraction is possible. For future work OCR (Optical Character Recognition) techniques can be implemented to solve this issue, and maybe to make this software a more generic software which can solve the headlines and news regions even from page-view images only.


Path Extractor
List
Rectangle
Line
List Headlines
Mark Headlines
1
X1 100
Y1 600
X2 300
Y2 1000
Search
Ulusoy Çakıcı'yı nasıl kandırdı
Üye olduğunu geç öğrendim
SABAH
Fon'a bunlar kaldı
10 milyon dolarını kaptırdı
'Demokratik ortak' Erdoğan G-8'de

Nr	X	Y	W	H
5	572.6	77.95	473.38	-564.1
7	1105.51	192.76	10.49	-110.55
8	1105.51	1312.44	10.49	-110.55
9	1088.92	812.13	26.00	-11.33
13	42.52	418.11	518.74	-221.1
27	185.52	77.95	860.46	-340.16
28	720.15	781.09	325.68	-427.74
43	269.44	883.14	268.99	-322.85
53	42	1268.26	217	-312.74
65	42.77	883.24	215.93	-373.92
67	54.01	1111.33	76.23	-135.76
86	443.2	78.95	128.4	-148.24

Sayaç	Text	Font ...	X	Y
4	Fon'a	153	34.85	766.2299
5	bunlar	153	374.8138	766.2299
6	kaldı	153	771.5151	766.2299
525	Üye	78	592.44	156.7101
526	olduğunu	78	730.6381	156.7101
527	geç	78	592.44	228.4233
528	öğrendim	78	726.1907	228.4233
293	Ulusoy	73.88	195.59	142.4099
294	Çakıcı'yı	73.88	195.59	219.8805
295	nasıl	73.88	195.59	291.2781
296	kandırdı	73.88	344.7439	291.2781
77	'Demokratikortak'	65	266.9	1326.83
78	Erdoğan	65	266.9	1385.25
79	G-8'de	65	507.9564	1385.25
73	10	47	53.86	955.0499
74	milyon	47	112.422	955.0499
75	dolarını	47	53.86	998.2194
76	kaptırdı	47	53.86	1041.379
693	Porto	35	595.28	453.74
694	kupayı	35	668.9218	453.74
695	Fener	35	595.28	488.6399
696	narayı	35	670.5519	488.6399

TK: 703 TFS: 10513.96 ORT: 14.95584 STDev: 12.95555

Figure 1. Algorithm Development Helper

CONCLUSION

The differences in the pages of the editions could be found graphically and textually. The developed automatic page matching algorithm works with almost 98% accuracy. And finally developed Automatic Headline Detection and News extraction algorithm performed with 96.2% accuracy on non-sport news pages. The algorithm performed with 88.5% accuracy on the total of the news pages.

[1] Grad. Student., Boğaziçi Univ., Ind. Eng., Bebek 80815 Istanbul, Turkey kursun@sabah.com.tr