April 19, 2019

METHOD FOR AUTOMATIC FILE NAMING OF SCANNED DOCUMENTS

HP INC

Method for automatic file naming of scanned documents

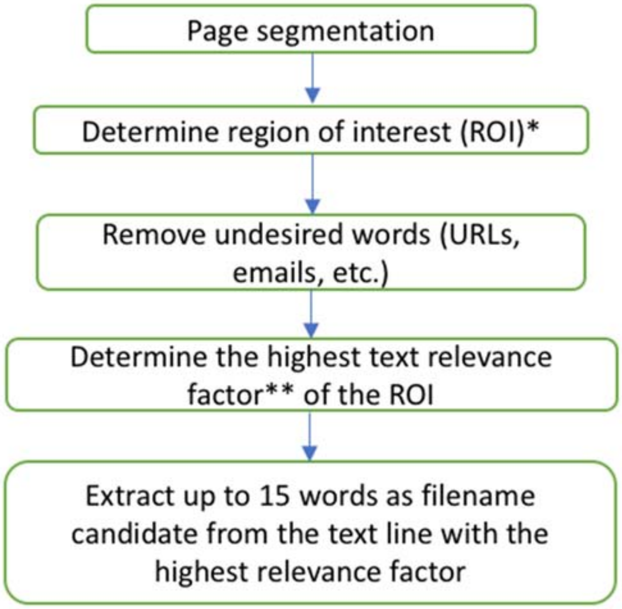
Abstract

When scanning a document, the goal in most cases is to obtain another printed copy. However, there are many situations in which the user would like to keep a digital copy of the document (e. g., in a cloud storage service or in the user's own local device). To this end, the user usually needs to choose and type a meaningful file name, which makes it easier to find that document again in the future. The task of choosing and typing a filename is usually tedious and, depending on the number of documents that are to be scanned, impractical. The proposed invention is a simple method to suggest a file name that relies on Optical Character Recognition (OCR) and the layout of the text boxes in the page extracted by the OCR engine.

Description

Scanning documents is part of an everyday workflow. Scanning receipts, invoices, contracts and many other forms of document are central to business, and often takes considerable time from the user. Performing it in mobile devices, although simplifying the task, still requires the user to properly name the scanned file - which can be time consuming and error prone if done manually. To solve this problem, we describe a process to automatically extract the title form a file.

Our solution is based on the extraction of the top ranked element from a list of possible candidates for a meaningful file name, based on a well-defined heuristic. The full pipeline is shown in figure 1. which considers the text block with the highest text relevance factor as the best option for a meaningful file name. Figure 1 illustrates a step-by-step workflow of the solution and more details of the approach follow below.


Page segmentation
Determine region of interest (ROI)*
Remove undesired words (URLs, emails, etc.)
Determine the highest text relevance factor** of the ROI
Extract up to 15 words as filename candidate from the text line with the highest relevance factor

* Region of interest: the area between the top and the bottom word on the page segmentation;

** Text relevance factor:
(line height * line width) / word count

Figure 1: File name suggestion workflow

The first steps of the pipeline consists in extracting geometric information from the document layout, such as the position and size of textual elements in the page. Then, a Region of Interest (RoI) is determined from the layout information. The text from the RoI is then analyzed and known patterns of texts such as emails and URLs are discarded. Most of this process is performed by off-the-shelf Optical Character Recognition software libraries.

Next, text relevance factor (TRF) metric is computed for every line of text in the RoI. As described in figure 1, TRF metric is (line height * line width) / word count, and assigns larger values to lines which are considered more relevant according to the size of the text block and the number of words. The blocks of text are then ranked according to this metric. Finally, a small number (e. g., 15) of words from the top ranked block are selected and suggested to the user as a meaningful filename for the scanned document. Table 1 illustrates how the text relevance factor is calculated based on the geometric information extracted from an example of a document.

	Rank	**Suggested filename**	**Width**	**Height**	**Words**	**Top**	**TRF**

	1	THE REINVENTOR AWARDS	1577	116	4	1652	44081
	2	The quest to imagine the future and make it happen	1438	71	10	313	9897
	3	Opens worldwide 16 July	650	68	4	2392	8658
	4	THE WAY	189	52	3	2146	1130

Table 1: Table describing the file name suggestions ranked by TRF

In some cases, the top ranked suggestion might not be considered the most relevant name by a user for a document or the process described above may fail to find any suggestions at all. Therefore, an extra step to be considered on the described method may present not just one, but several options of possible file names to the user. If none of the options are considered relevant, users may then manually provide a file name of their choosing as a fallback.

Disclosed by João Melo, Ricardo Piccoli, Vinicius Lafourcade, Ricardo Ribani and Rafael Borges, HP Inc.