May 24, 2019

# A METHOD FOR IDENTIFYING AND COMPARING VERSIONS OF SINGLE OR MULTI-PAGES SCANNED DOCUMENTS

HP INC



**A Method for identifying and comparing versions of single or multi-pages scanned documents**

On many occasions users have to manipulate both printed and digital formats of a document. As they collaborate on editing and creating different versions of digital documents, the regular versioning systems are not enough to capture the dynamics involved in this process. This disclosure describes a method for identifying the version of a document and comparing the differences between one or more scanned and digital pages.
The idea takes into consideration that the system has previously calculated and stored all text diffs between document versions, similar to a usual version control system. However, differently from a usual version control system, the proposed system requires the differences to be calculated by document page. This fact is used to facilitate the search, described further in the text. Considering a case in which a user needs to find which is the version of a printed document, these are the steps to accomplish the task:

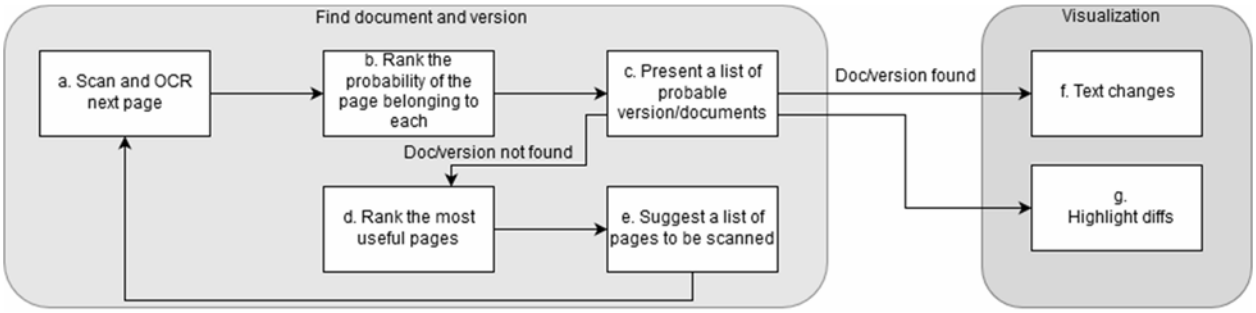


Figure 1 - Versioning method flow chart

- Scan and OCR next page: the user captures a picture from one or more pages of the target document, by using a scanner or a smartphone camera, for example. Then the application extracts the text, images and formatting style from the picture by using OCR/DNN techniques (Figure 1-a).
- (optional) Filter pages by page format: probable page format (page size, margins, fonts), detect, if possible, the document's characteristics.
- Perform a search in the versions tree: the system performs a search in the document tree, by performing diff operations between the scanned document and all pages of the first versions of all documents in the repository. Then, for each new version of each document, a merge operation is performed for all pages, so the system recovers the text for that page/version, and then a new diff operation is performed.
    - If there is one page found with page diff score equal to zero, then the system has found the document and page version (Figure 1-f and 1-g).
    - If there are more pages found with page diff score equal to zero, then the system returns a possible list of documents and versions to choose from (Figure 1-b and 1-c). The system will then calculate and suggest new pages to scan in order to narrow the search (Figure 1-d and 1-e).
    - If no page with page diff score equal to zero is returned, then the user is given a ranked list of page diff scores and is suggested a list of pages to scan (Figure 1-d and 1-e).

    - In the case of no matches, this could mean that the scanned page is a new page (meaning it is a new version or a new document); or that there was an error in the scanning process.
- Text Changes: With the selected version of the target document, the application extracts the text and layout differences between the scanned and the last document version and returns it to the application. The differences from step 5 may be highlighted in the device display (Figure 1-f and 1-g).

**OCR Next Page**

If after the searching step the version wasn't found, it can mean that the list of probable versions is still large or no there are no matches. In order to improve the search accuracy, the system can find which pages among the ones not scanned would potentially distinguish a version among the current list of probable versions. The application can then suggest new pages to be scanned, based on pages importance so expedite the process of scanning the document with a mobile device. The page importance of the version of the document could be calculated using a counter of versions that a page was modified:

$$page\ importance\ = \frac{number\ of\ versions\ that\ the\ page\ is\ changed}{total\ number\ of\ version\ of\ the\ document}$$

According to this calculation each version has a number of pages changed into them. The system will count the number of times each page appears in a new version, and use the median of this count to suggest the page with changes number closer to the median to be scanned. The user can select among the ranked list of potential versions at any time, in case the user knows something else the application didn't have access to.

During the visualization step (Figure 1-g), the application can use an Augmented Reality overlay to highlight parts of the text that are different on the target document (Figure 2).

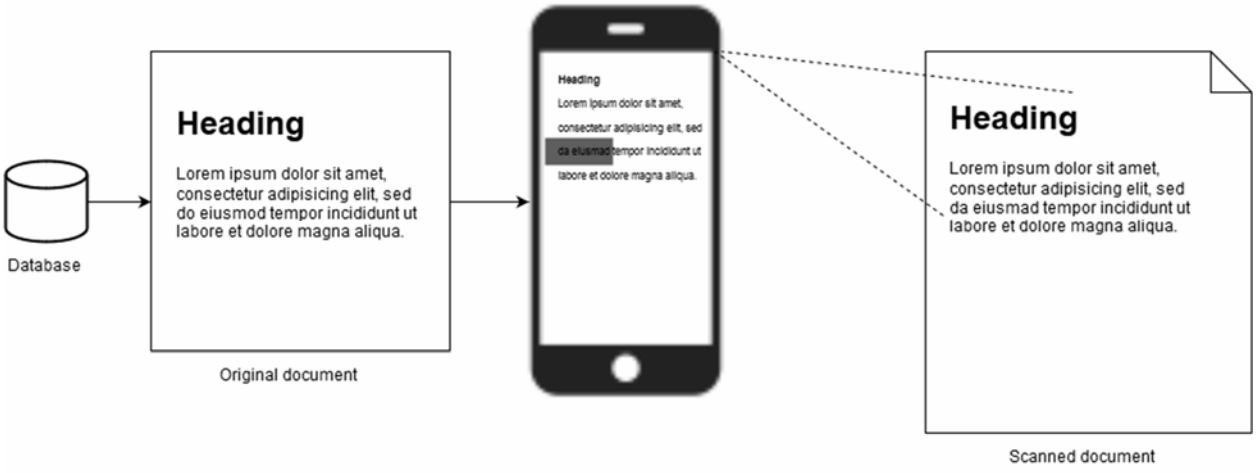


Figure 2 - Result of a text comparison using augmented reality.

It is possible that the particular version of the document cannot be found, in that case, it can offer the probability of it being an untracked version. Perhaps also providing the most likely branch from which it was edited.