Patents

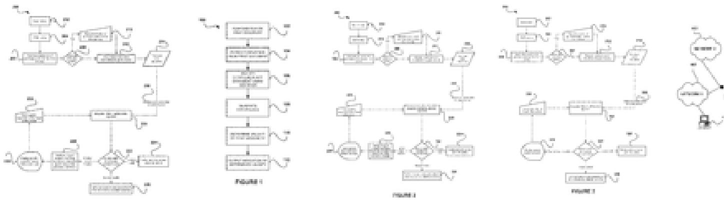
9767354

Global geographic information retrieval, validation, and normalization

Abstract

According to one embodiment, a computer-implemented method includes: capturing an image of a document using a camera of a mobile device; performing optical character recognition (OCR) on the image of the document; extracting an identifier of the document from the image based at least in part on the OCR; comparing the identifier with content from one or more reference data sources, wherein the content from the one or more reference data sources comprises global address information; and determining whether the identifier is valid based at least in part on the comparison. The method may optionally include normalizing the extracted identifier, retrieving additional geographic information, correcting OCR errors, etc. based on comparing extracted information with reference content. Corresponding systems and computer program products are also disclosed.

Images (7)



Classifications

H04N1/40062 Discrimination between different image types, e.g. two-tone, continuous tone

View 20 more classifications

US9767354B2
United States

Download PDF Find Prior Art Similar

Inventor: Stephen Michael Thompson, Jan W. Amtrup, Anthony Macciola

Current Assignee : Kofax Inc

Worldwide applications

2016 US 2017 US

Application US15/146,848 events

2016-05-04	Priority to US15/146,848
2016-05-04	Application filed by Kofax Inc
2016-11-10	Publication of US20160328610A1
2017-09-19	Publication of US9767354B2
2017-09-19	Application granted
Status	Active
2029-02-10	Anticipated expiration

Show all events

Info: Patent citations (695), Non-patent citations (161), Cited by (29), Legal events, Similar documents, Priority and Related Applications

External links: USPTO, USPTO PatentCenter, USPTO Assignment, Espacenet, Global Dossier, Discuss

Claims (18)

Hide Dependent

What is claimed is:

1. A computer-implemented method, comprising:

capturing an image of a document using a camera of a mobile device;

performing optical character recognition (OCR) on the image of the document;

extracting an identifier of the document from the image based at least in part on the OCR;

comparing the identifier with content from one or more reference data sources, wherein the content from the one or more reference data sources comprises global address information; and wherein the content from the one or more reference data sources is derived from geographic information organized in one or more of a proprietary address database and an open source address database; and wherein deriving the content from the geographic information comprises:

obtaining the geographic information from one or more of the proprietary address database and an open source address database; and

parsing the geographic information according to a set of predefined heuristic rules, wherein the set of predefined heuristic rules are configured to normalize the global address information obtained from the one or more sources according to a single convention for representing address information; and

determining whether the identifier is valid based at least in part on the comparison.

2. The method as recited in claim 1, wherein the identifier consists of characters selected from a predefined alphabet, wherein the predefined alphabet consists of one or more of numerals, alphabetic characters, and symbols.

3. The method as recited in claim 1, wherein the identifier comprises a partial or complete address.

4. The method as recited in claim 1, wherein the identifier comprises one or more of:

a street name, a street number, a block number, a unit number, a city name, a county name, a municipality name, a state name, a state abbreviation, a country name, a country abbreviation, and a ZIP code.

5. The method as recited in claim 1, the comparing comprising fuzzy matching the identifier with the content from the one or more data sources.

6. The method as recited in claim 1, wherein the identifier is validated based at least in part on determining a fuzzy match exists between the identifier and at least a portion of the global address information, wherein the fuzzy match is characterized by no more than two character mismatches between the identifier and at least the portion of the global address information.

7. The method as recited in claim 1, comprising locating the identifier within the image based on a connected components analysis.

8. The method as recited in claim 1, wherein the OCR is performed only on a portion of the image determined to depict the identifier.

9. The method as recited in claim 1, comprising determining a locality associated with the identifier; and

wherein the set of predefined heuristic rules are selected based on the locality determined to be associated with the extracted identifier.

10. The method as recited in claim 1, wherein deriving the content from the geographic information comprises populating the one or more data sources with the content, wherein the content consists of geographic information parsed using the set of predefined heuristic rules.

11. The method as recited in claim 1, wherein deriving the content from the geographic information comprises normalizing the geographic information to expand one or more abbreviations present in the geographic information; and

wherein the content excludes abbreviated geographic information.

12. The method as recited in claim 1, comprising normalizing the extracted identifier prior to comparing the identifier with content from one or more reference data sources, wherein the normalizing is performed according to one or more predefined business rules corresponding to a particular locality.

13. The method as recited in claim 1, comprising determining a locality corresponding to the extracted identifier, and