



- (51) International Patent Classification:
G06K 9/00 (2006.01) *G06K 9/72* (2006.01)
- (21) International Application Number:
PCT/US2019/057509
- (22) International Filing Date:
23 October 2019 (23.10.2019)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
16/185,737 09 November 2018 (09.11.2018) US
- (71) Applicant: MICROSOFT TECHNOLOGY LICENSING, LLC [US/US]; One Microsoft Way, Redmond, Washington 98052-6399 (US).
- (72) Inventors: FLORENCIO, Dinei Afonso Ferreira; Microsoft Technology Licensing, LLC, One Microsoft Way,

Redmond, Washington 98052-6399 (US). **ZHANG, Cha**; Microsoft Technology Licensing, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). **NAHMIAS, Gil Moshe**; Microsoft Technology Licensing, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). **DAI, Yu-Yun**; Microsoft Technology Licensing, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US).

(74) Agent: MINHAS, Sandip S. et al.; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME,

(54) Title: SUPERVISED OCR TRAINING FOR CUSTOM FORMS

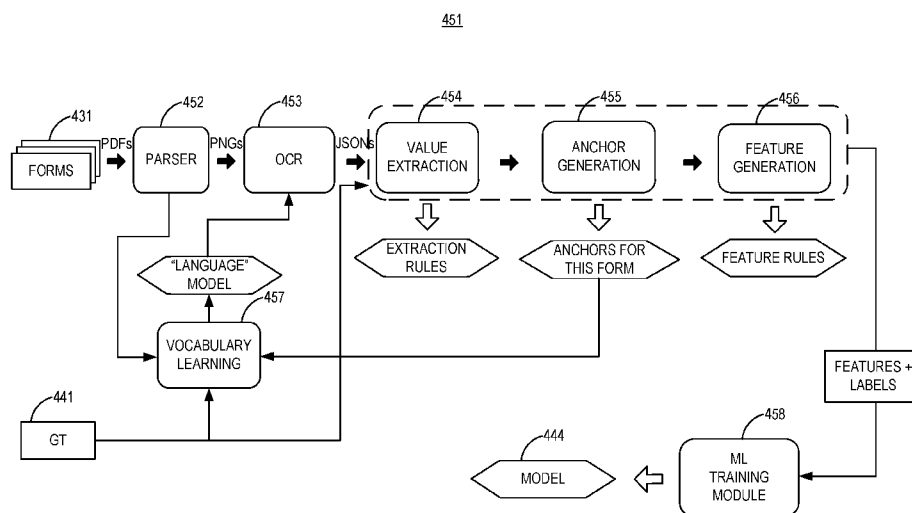


FIG. 4

(57) Abstract: The disclosed technology is generally directed to optical character recognition for forms. In one example of the technology, optical character recognition is performed on a plurality of forms. The forms of the plurality of forms include at least one type of form. Anchors are determined for the forms, including corresponding anchors for each type of form of the plurality of forms. Feature rules are determined, including corresponding feature rules for each type of form of the plurality of forms. Features and labels are determined for each form of the plurality of forms. A training model is generated based on a ground truth that includes a plurality of key-value pairs corresponding to the plurality of forms, and further based on the determined features and labels for the plurality of forms.



MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ,
OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA,
SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN,
TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

(88) Date of publication of the international search report:

14 January 2021 (14.01.2021)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2019/057509

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06K9/00 G06K9/72
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, COMPENDEX, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2018/033147 A1 (BECKER RICHARD J [US] ET AL) 1 February 2018 (2018-02-01)	1-12,14
Y	0047, 0010, 0018- 0025, 0054-0055, 0061, 0014, 0037.; figures 2,3,5,6,8	13,15
A	JP H07 200726 A (NIPPON ELECTRIC ENG) 4 August 1995 (1995-08-04)	1
Y	US 2015/347861 A1 (DOEPKE FRANK [US] ET AL) 3 December 2015 (2015-12-03) paragraph [0063]	13
	----- -/-	



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

1 December 2020

Date of mailing of the international search report

22/12/2020

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Versluis, Anton

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2019/057509

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	PERE CONSTANS: "A Simple Extraction Procedure for Bibliographical Author Field", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 4 February 2009 (2009-02-04), XP080309334, abstract, V -----	15

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2019/057509

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☒ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☒ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-12, 14

Subject-matter for OCR-ing a form and in particular, determining which texts corresponds to which field and determine the type of form for a plurality of forms using a training model based on that plurality.

2. claim: 13

The subject-matter of claim 1, with generation of a language model based on the anchors.

3. claim: 15

The subject-matter of claim 1, with modification of the plurality of forms based on line breaks and generating extraction rules.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2019/057509

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2018033147	A1	01-02-2018	
		AU 2017302248 A1	30-08-2018
		AU 2020200251 A1	06-02-2020
		CA 3017646 A1	01-02-2018
		EP 3491578 A1	05-06-2019
		US 10621727 B1	14-04-2020
		US 2018033147 A1	01-02-2018
		WO 2018022160 A1	01-02-2018

JP H07200726	A	04-08-1995	NONE

US 2015347861	A1	03-12-2015	NONE
