

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property  
Organization  
International Bureau

(43) International Publication Date  
02 January 2025 (02.01.2025)



(10) International Publication Number  
**WO 2025/007016 A1**

(51) International Patent Classification:

G06V 30/148 (2022.01) G06F 40/103 (2020.01)  
G06V 30/414 (2022.01) G06V 30/18 (2022.01)  
G06V 30/416 (2022.01) G06V 30/19 (2022.01)

(21) International Application Number:

PCT/US2024/036198

(22) International Filing Date:

28 June 2024 (28.06.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/511,553 30 June 2023 (30.06.2023) US

(71) Applicant: **GREENHOUSE SOFTWARE, INC.**  
[US/US]; 18 W 18th St, New York, New York 10011 (US).

(72) Inventor: **XYLOURIS, Triantafyllos**; c/o Greenhouse  
Software, Inc., 18 W 18th St, New York, New York 10011  
(US).

(74) Agent: **HUTTER, Christopher R.** et al.; Cooley LLP,  
1299 Pennsylvania Avenue NW, Suite 700, Washington,  
District of Columbia 20004 (US).

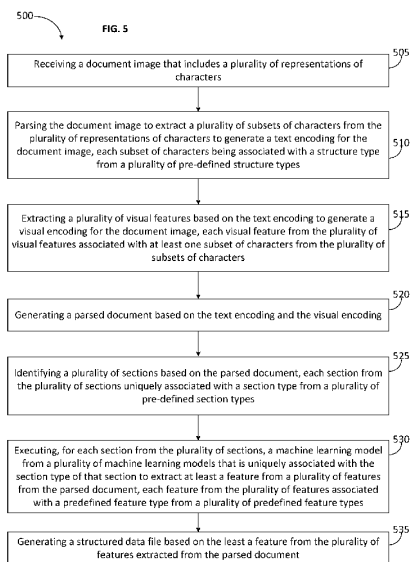
(81) Designated States (unless otherwise indicated, for every

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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every

kind of regional protection available): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, 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).

(54) Title: METHODS AND APPARATUS FOR EXTRACTING DATA FROM A DOCUMENT BY ENCODING IT WITH TEXTUAL AND VISUAL FEATURES AND USING MACHINE LEARNING



(57) Abstract: An apparatus including a processor caused to receive document images, each including representations of characters. The processor is caused to parse each document image to extract, based on structure type, subsets of characters, to generate a text encoding for that document image. For each document, the processor is caused to extract visual features to generate a visual encoding for that document image, each visual feature associated with a subset of characters. The processor is caused to generate parsed documents, each parsed document uniquely associated with a document image and based on the text and visual encoding for that document image. For each parsed document, the processor is caused to identify sections uniquely associated with section type. The processor is caused to train machine learning models, each machine learning model associated with one section type and trained using a portion of each parsed document associated with that section type.

WO 2025/007016 A1

[Continued on next page]

**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))*