第一工科大学研究報告
第38号(2026)pp.31-37

深層学習 OCR を用いた在留カード文書画像からの情報抽出と構造化

伊藤 千恵　　馮 一峻

第一工科大学　工学部　情報・AI・データサイエンス学科　東京上野キャンパス（〒110-0005 東京都台東区上野 7-7-4 ）

Information Extraction and Structuring from Residence Card Document Images Using Deep Learning-based OCR

Chie Ito　　Yijun Feng

Abstract: In this study, we designed and evaluated an information processing method that automatically extracts key information from residence card document images using deep learning-based OCR and outputs it as structured data. In recent years, the increasing number of international students has created a strong demand for improving the efficiency of residence card verification tasks at universities. To achieve flexible information extraction without relying on fixed templates, we constructed a processing pipeline that combines deep learning-based OCR with semantic reasoning using a Vision-Language Model. The extracted results are further processed using regular expressions and rule-based normalization to generate structured data in CSV/JSON formats. Experiments conducted on 20 scanned images and 11 smartphone-captured images demonstrated high extraction accuracy of 90-100% for seven items, including residence card number, name, and date of birth. For the address field, detailed analysis using the Character Error Rate (CER) confirmed that a certain level of recognition performance was achieved at the character level. The proposed method achieved an average processing time of 5.39 seconds per image on a GPU environment. These results indicate that the proposed method is useful as an information extraction framework for identity documents that can flexibly adapt to institutional changes and layout variations.

Key words: Deep Learning-based OCR, Document Image Understanding, Information Extraction

1. はじめに

日本の大学では、留学生数の増加に伴い、在留資格や有効期限を管理する在留カード確認業務の重要性が高まっている。特に 2010 年代以降、外国人留学生数は 20 万～30 万人規模で推移しており(図 1)[4]、大学職員による在留管理業務の作業負荷は年々増大している。

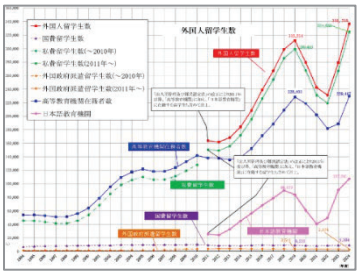

図 1 外国人留学生数の推移(人)[4]

一方、多くの教育機関では在留カード情報を目視で確認し、手入力で管理する運用が依然として行われており、作業時間の増加や入力ミスが課題となっている。このため、在留カード情報を自動的にデジタル化・構造化する仕組みが求められている。

近年は深層学習ベース OCR[7]（Optical Character Recognition）の発展により高精度な文字認識が可能となっているが、制度変更やフォーマット改訂への対応を考慮すると、固定テンプレートに依存しない柔軟な情報抽出手法が必要である。そこで本研究で

は、深層学習ベース OCR と後処理を組み合わせ、在留カード文書画像から主要項目を自動抽出し、構造化データとして管理可能な情報処理手法の確立を目的とする。

2. 関連研究と研究目的

Vision-Language Model(VLM)は、画像とテキストを統合的に扱う手法として近年注目されている。DeepSeek-VL2[1]は、Mixture-of-Experts(MoE)構造を採用することで高効率な推論を実現した VLM であり、高解像度画像に対して動的タイル分割に基づく視覚エンコーディングを行うことで、文書画像中の細かな文字領域に対しても高い認識性能を示す。さらに、文書理解や表形式データを含む多様な視覚・言語データで事前学習されており、従来は困難であった複雑な文書 OCR や文書理解タスクにも対応可能な汎用性を有している。

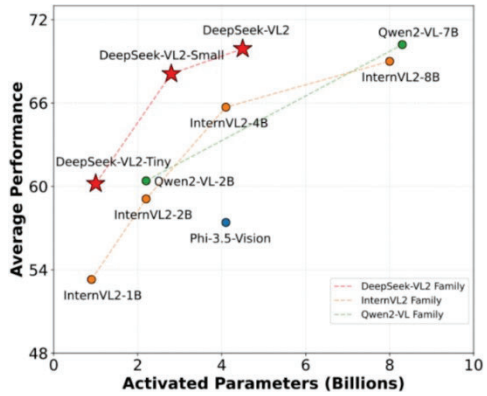

図 2 異なるオープンソースモデルにおける活性化パラメータ数と平均性能の関係[1]

一方、テンプレートベース OCR[2]は、運転免許証や身分証明書など固定レイアウト文書を対象とした従来手法であり、高精度な認識が可能である反面、文書ごとにテンプレート作成が必要であり、レイアウト変更への柔軟性に乏しい。そのため、将来的な制度変更やフォーマット改訂が想定される在留カードのような文書への適用には課題が指摘されている。

これらの関連研究を踏まえ、本研究ではテンプレートに依存せず、DeepSeek-VL2 の高解像度文書処理能力および意味理解能力に着目し、在留カード画像からの情報抽出へ応用する。

本研究の目的は、文字認識に加えて意味理解を伴う深層学習ベース OCR を用い、在留カード画像から主要項目(在留カード番号、氏名、在留資格、在留期間等の 9 項目)を自動抽出し、CSV 形式などの構造化データとして管理可能な情報処理パイプラインを設計・評価することである。OCR 結果に対して正規表現やルールベースの整形処理を適用することで、表記揺れの吸収および項目単位での正規化を行い、実務利用を想定した抽出精度の実現を目指す。評価には、認識精度(項目別正解率および文字誤り率(CER))ならびに処理時間を指標として用い、提案手法の有効性を検証する。本研究により、大学における留学生在籍管理業務への適用を想定し、テンプレート作成に依存しない情報抽出手法として、一定のレイアウト変動に対しても運用コストの低減が期待される。

3. 研究方法

3.1 提案手法

本研究では、文字認識に加えて意味推論を含む深層学習ベース OCR として DeepSeek-VL2-Tiny[1]（DeepSeek-VL2 系列）を用い、在留カード画像から在留カード番号、氏名（漢字・ローマ字）、生年月日、性別、国籍、住所、在留資格、在留期間、許可年月日の 9 項目を自動抽出し、実務利用を想定した構造化データへ変換する提案手法を設計・実装した。提案手法の全体的な処理フローを図 3 に示す。

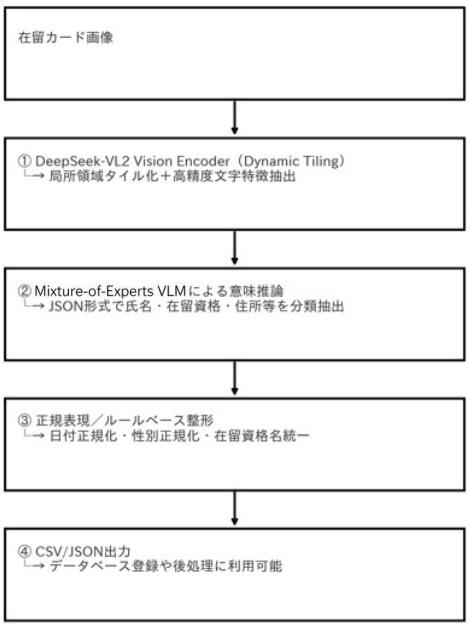

図 3 DeepSeek-VL2-Tiny を用いた提案手法の処理フロー

本提案手法では、まず与えられた在留カード画像（図 4）に対して DeepSeek-VL2-Tiny [1]の Vision Encoder（Dynamic Tiling）を適用し、動的タイル分割に基づく高精度な文字特徴抽出を行う。続いて、Mixture-of-Experts （MoE） 構造を有する Vision-Language Model（VLM）による意味推論処理を行い、在留カードに含まれる主要な 9 項目を JSON 形式で抽出する。次に、抽出結果に対して正規表現およびルールベース整形処理を適用し、日付形式や性別表記の標準化、在留資格名称の統一を行う。最後に、整形済みデータを CSV／JSON 形式の構造化データとして保存する。

図 4　在留カードサンプル画像[6]

本研究で提案する手法の特徴は、固定レイアウトに依存せず、深層学習ベース OCR と意味推論を組み合わせることで、在留カードに含まれる主要項目を柔軟に抽出できる点にある。Mixture-of-Experts 構造を有する VLM による意味理解に基づき、氏名、住所、在留資格などの代表的な項目を文脈情報から分類可能である。さらに、正規表現およびルールベースによる整形処理を組み合わせることで、実務利用を想定したデータ品質を確保している。抽出結果は CSV／JSON 形式で出力され、大学における在留管理業務への適用を想定している。

3.2 実験設定

本研究の処理パイプラインは、Windows 11 搭載の GPU マシン上（RTX 3070）において Python（バージョン 3.11）により実装した。照明条件や解像度などが異なる画像を評価データとして用いることで、提案手法の実運用環境における汎用性を検証することを目的とする。主要な実行環境を表 1 に示す。

表 1 主要な実行環境

項目	設定内容
OS	Windows 11
GPU	NVIDIA GeForce RTX 3070
CPU	Intel Core i9-12900
メモリ	16 GB
Python バージョン	Python 3.11
使用ライブラリ	Transformers / PyTorch 2.0.1 + cu118 / Pillow
DeepSeekモデル	DeepSeek-VL2-Tiny
入力画像解像度	スキャン画像：300 dpi
評価データ形式	スキャン画像 20 枚 / スマートフォン撮影画像 11 枚

本研究では、計算資源と処理時間の制約を考慮し、DeepSeek-VL2 系列の中から軽量モデルである DeepSeek-VL2-Tiny[1]を採用した。その上で、提案手法の有効性を評価するため、在留カード画像を対象とした実験を行った。使用したデータは、撮影条件の異なる 2 種類の在留カード画像から構成される。

1 つ目は、大学に過去在籍していた留学生の在留カードをスキャナで読み取った画像 20 枚である。2 つ目は、現在在籍している留学生の在留カードをスマートフォンで撮影した画像 11 枚である。個々の在留カードに含まれる顔写真、在留カード番号、住所、生年月日について匿名化処理を施した（図 5）。この処理により、個人情報は保護されつつも、評価に必要な文字列の視認性およびレイアウト構造は保持されている。

図 5　匿名化処理を施した在留カード画像

各画像に対して 3.1 の提案手法を適用し、在留カード画像から 9 項目を自動抽出した。抽出結果は JSON 形式で出力し、正規表現およびルールベースの後処理によって項目ごとに整形した後、CSV 形式の構造化データとして保存した。評価にあたっては、次の 3.3 の項目別正解率、文字誤り率 CER（Character Error Rate）および処理時間を算出した。

3.3 評価指標

本研究では、評価指標として項目別正解率、文字誤り率（CER：Character Error Rate）、および処理時

間を用いた。これらにより、提案手法が大学等における実務運用に求められる精度および効率性を満たし得るかを評価する。

項目別正解率は、各項目について抽出結果が正解データと完全一致した件数を評価対象データ総数で除した割合として定義する。正解判定は完全一致に基づき、氏名項目については漢字表記およびローマ字表記の双方が一致した場合に正解とした。9 項目それぞれについて項目別正解率を算出し、項目単位での抽出精度を評価した。

なお、住所および国籍項目は改行や表記揺れ、部分一致が生じやすく、完全一致評価のみでは認識性能を適切に反映できない可能性がある。そのため、補助指標として文字レベルの誤り率である CER（Character Error Rate）を併用した。CER は式（1）で定義される[5] 。

$$CER = (S + D + I) / N \qquad (1)$$

式（1）　CER（Character Error Rate）の定義[5]

式（1）において、S は置換文字数、D は削除文字数、I は挿入の誤り文字数、N は元の正解文の総文字数を表す。CER が小さいほど認識性能が高く、0 の場合は完全一致を示す。一方、挿入誤りが多い場合には CER が 1 を超えることがある。

近年の Vision-Language Model（VLM）は高い認識性能を示す一方で、推論計算量が大きく、実運用においては処理時間の評価が重要な課題とされている[1]。この背景を踏まえ、実運用を想定した性能評価として、図 3 に示した一連の処理フローに基づき、在留カード画像を入力してから、深層学習ベース OCR 処理、意味推論および項目構造化処理を経て CSV 形式の出力が完了するまでの処理時間を測定した。在留カード画像 31 枚を対象に処理を行い、全画像の処理完了までに要した総時間を測定した上で、1 枚あたりの平均処理時間を算出した。この平均処理時間を用いて、処理効率の観点から提案手法の有用性を評価した。

4. 実験結果

本章では、3.1 節で提案した手法を在留カード画像に適用した際の実験結果について述べる。評価には、スキャナで取得した在留カード画像 20 枚およびスマートフォンで撮影した在留カード画像 11 枚の 2 種類のデータを用いた。実験の結果、在留カード番号、氏名、生年月日、性別、在留資格、在留期間、許可年月日といった項目では高い認識精度が得られた。一方、国籍および住所の項目では、他の項目と比較して認識精度が相対的に低い傾向が確認された。

表 2 は、スキャナで取得した在留カード画像 20 枚を対象として、在留カード番号、氏名（漢字・ローマ字)、生年月日、性別、国籍、住所、在留資格、在留期間、許可年月日の 9 項目について、OCR による抽出結果を評価した結果を示す。その結果、在留カード番号、氏名、生年月日、性別、在留資格、在留期間、許可年月日の 7 項目では高い正解率を示し、90～100%の抽出精度が確認された。一方、国籍および住所の項目では誤認識が発生し、正解率はそれぞれ 65%、25%と低い結果となった。

表 2　在留カード（スキャン 20 枚）の項目別正解率

項目	正解数	誤り数	全体（20）	正解率（%）
在留カード番号	19	1	20	95%
氏名	18	2	20	90%
生年月日	20	0	20	100%
性別	20	0	20	100%
国籍	13	7	20	65%
住所	5	15	20	25%
在留資格	20	0	20	100%
在留期間	19	1	20	95%
許可年月日	19	1	20	95%

表 3 は、スマートフォンで撮影した在留カード画像 11 枚を対象とした評価結果を示す。評価に用いたサンプル在留カード画像には匿名化処理を施しており、顔写真はマスキング処理を行い、在留カード番号、住所、生年月日についても仮の番号・住所・生年月日に置き換えている。ただし、文字列およびレイアウト構造は保持されているため、スキャナ画像と同様に 9 項目すべてを評価対象とした。その結果、在留カード番号、氏名（漢字・ローマ字)、生年月日、性別、在留資格、在留期間、許可年月日の 7 項目において 90%～100%の高い正解率が確認された。これは表 2 に示したスキャナ画像 20 枚の評

価結果と同程度の精度である。一方、国籍および住所の正解率はそれぞれ 54.5%、45.4%と低く、表 2 のスキャナ画像と同様の傾向が確認された。

表 3　在留カード（スマートフォン撮影 11 枚）の項目別正解率

項目	正解数	誤り数	全体（11）	正解率（%）
在留カード番号	11	0	11	100%
氏名	11	0	11	100%
生年月日	11	0	11	100%
性別	11	0	11	100%
国籍	6	5	11	54.50%
住所	5	6	11	45.40%
在留資格	11	0	11	100%
在留期間	11	0	11	100%
許可年月日	10	1	11	90.90%

続いて、特に正解率の低かった住所項目における認識性能をより詳細に分析するため、式（1）で定義した Character Error Rate（CER）を算出した。表 4 に、住所項目についてスキャン画像 20 枚とスマートフォン撮影画像 11 枚の平均 CER を比較した結果を示す。

表 4　住所項目の CER 平均比較
（スキャン vs スマートフォン）

データ種別	枚数	CER平均
スキャン画像	20	0.36
スマホ撮影画像	11	0.27

表 4 より、住所項目の CER はスキャン画像で平均 0.36、スマートフォン撮影画像では 0.27 であり、スマートフォン撮影画像の方が相対的に CER が低い結果となった。CER は値が小さいほど文字認識精度が高いことを示す指標であるため、この結果からスマートフォン撮影画像の方が住所文字列の認識誤りが少ない傾向が確認された。

表 5　国籍項目の CER 平均比較
（スキャン vs スマートフォン）

データ種別	枚数	CER平均
スキャン画像	20	0.27
スマホ撮影画像	11	0.29

表 5 より、国籍項目における CER はスキャン画像では 0.27、スマートフォン撮影画像では 0.29 であり、その差は 0.02 と小さい。

実行性能評価として、1 枚あたりの処理時間を測定した結果を表 6 に示す。GPU 環境において在留カード画像 31 枚を対象に評価した結果、平均処理時間は 5.39 秒/枚、最短 2.49 秒、最長 11.89 秒であった。

表 6　GPU 環境における処理時間評価結果

指標	値
対象枚数	31枚
平均処理時間	5.39秒／枚
最短処理時間	2.49秒
最長処理時間	11.89秒

5. 考察

本研究では、深層学習ベース OCR として DeepSeek-VL2-Tiny[1]を用い、在留カード画像から主要項目を自動抽出し、その抽出精度および実行性能について評価を行った。その結果、在留カード番号、氏名（漢字・ローマ字）、生年月日、性別、在留資格、在留期間、許可年月日において 90～100%の高い正解率が得られ、提案手法が在留カード管理に必要な情報を一定の精度で抽出可能であることが確認された（表 2、表 3）。

これらの項目で高精度が得られた要因として、DeepSeek-VL2-Tiny[1]が Mixture-of-Experts 構造と動的タイル分割を用いることで、文書画像中の文字特徴を高精度に捉え、意味理解に基づいて項目を分類できる点が挙げられる。第 2 章で述べたように、DeepSeek-VL2-Tiny[1]は従来困難であった複雑な文書 OCR にも対応可能な汎用性を有しており、その特性が在留カードの主要項目抽出において有効に機能したと考えられる。

一方で、国籍および住所項目は、スキャン画像・スマートフォン撮影画像の両条件において他項目と比較して正解率が低かった。国籍項目では、カード上で性別項目と近接して配置されているため、行分割や項目対応付けの段階で両項目が混在し、誤った値が抽出されるケースが確認された。さらに、国籍項目の CER はスキャン画像 0.27、スマートフォン撮影画像 0.29 であり差は 0.02 と小さく（表 5）、撮影条件による影響は限定的で、誤りは主としてレイアウト近接に起因する混在が一因である可能性が示唆される。

住所項目は項目別正解率がスキャン画像 25%、スマートフォン撮影画像 45.4%と低かった一方で（表 2、表 3）、文字単位の認識性能を CER で評価した結果、平均 CER はスキャン画像 0.36、スマートフォン撮影画像 0.27 となり、文字レベルではスマートフォン撮影画像の方が良好であった（表 4）。このことから、住所項目では文字認識自体は比較的良好であるものの、住所が二行に分割される場合や、アパート・マンション名を含む複雑な表記構造を有する場合に、行結合や表記正規化などの後処理が不十分となり、項目単位の完全一致に至らないケースが多いと考えられる。スマートフォン撮影画像では反射や傾きなどのノイズが存在するが、文字の視認性が確保される条件では CER が改善する可能性がある。これに対し、スキャン画像では文字の潰れ等が生じた場合に文字誤りが増加し、CER 上昇につながった可能性がある。以上より、本提案手法はテンプレートベース OCR[2]に依存しない点で柔軟性を有する一方、住所のように構造的に複雑な項目では、行結合処理および表記正規化を含む後処理の強化が今後の課題である。

実行性能については、GPU 環境下において 1 枚あたり平均 5.39 秒の処理時間（表 6）で在留カード画像の解析が完了した。本処理には OCR に加えて意味推論および項目構造化処理が含まれており、既存の単純な OCR 処理と比較すると計算負荷は大きいが、文書理解を伴う VLM を用いた手法としては実運用においても適用可能な処理時間であると考えられる。これは、DeepSeek-VL2-Tiny[1]が有する高効率な推論特性が、実務利用における有効性を示唆する結果である。一方、最大処理時間が 11.89 秒となったケースでは、反射や傾きが大きい画像など、入力画像品質の影響により推論処理が複雑化した可能性が考えられる。今後は、傾き補正や反射低減といった前処理を導入することで、処理時間のばらつきを抑え、さらなる安定化が期待される。

以上より、本研究で提案した手法は、テンプレートに依存しない深層学習ベース OCR と意味理解を統合することで、在留カード情報抽出に対する有効性が確認された。一方で、住所項目に代表される構造的に複雑な情報に対しては、行分割処理および辞書整合を含む後処理機構の高度化が今後の重要な課題である。本研究の結果は、留学生管理業務における在留カード情報処理の自動化に向けた基礎的知見を提供する。

6. まとめ

本研究では、深層学習ベース OCR（DeepSeek-VL2-Tiny[1]）を用いて在留カード画像から主要項目を自動抽出し、構造化データとして保存可能な情報処理パイプラインを設計・評価した。その結果、在留カード番号、氏名（漢字・ローマ字）、生年月日、性別、在留資格、在留期間、許可年月日について 90〜100%の高い抽出精度が得られ、大学等における在留管理業務の効率化に有用である可能性が示された。また、GPU 環境下では平均 5.39 秒／枚で処理が完了し、実務利用への適用可能性を示す処理性能が確認された。

一方で、国籍および住所の抽出精度には課題が残り、国籍項目では性別項目とのレイアウト上の近接により、行分割や項目対応付けの段階で誤った値が抽出される傾向が確認された。住所項目について CER（Character Error Rate）による評価を行った結果、スキャン画像では 0.36、スマートフォン撮影画像では 0.27 となり、項目単位では完全一致に至らない場合でも、文字レベルでは一定の認識性能が得られていることが示された。特に住所項目では、行分割や番地位置の揺らぎ、撮影条件の影響によって誤認識が発生する傾向が確認された。

本研究は、在留カード文書画像を対象とした深層学習ベース OCR による情報抽出手法の有効性を示し、大学における在留管理業務の自動化に向けた基礎的枠組みを提示した。本研究では 31 枚の在留カード画像を対象として評価を行ったが、DeepSeek-VL2-Tiny[1]に代表される MoE 構造を有する VLM による意味推論と後処理を組み合わせることで、在留カード情報を高精度に構造化できる可能性を示した点に特徴がある。今後は、住所特有の表記揺れに対応する抽出ロジックの強化や、傾き補正・反射抑制などの前処理の導入、さらに評価データの拡充や従

来 OCR 手法との定量比較を通じて、本手法の汎用性および実用性の検証を進める予定である。

謝辞

本研究の実施にあたり、研究データをご提供いただいた国際交流センターの宇野先生に深く感謝申し上げる。また、本研究で使用した写真データをご提供いただいた研究室の皆様に心より感謝申し上げる。

参考文献

[1] Z. Wu et al., “DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding,” arXiv:2412.10302, 2024.

[2] C. Irimia, F. Harbuzariu, I. Hazi, and A. Iftene, “Official Document Identification and Data Extraction using Templates and OCR,” Procedia Computer Science, vol. 207, pp. 1571–1580, 2022.

[3] R. Iskandar and M. E. Kesuma, “Designing a Real-Time-Based Optical Character Recognition to Detect ID Cards,” International Journal of Electronics and Communications System, vol. 2, no. 1, pp. 23–29, 2022.

[4] 独立行政法人 日本学生支援機構, “外国人留学生在籍状況調査（2024 年度）”, https://www.studyinjapan.go.jp/ja/_mt/2025/04/data2024z.pdf （閲覧日：2025 年 9 月 10 日）.

[5] C. Neudecker et al., “A survey of OCR evaluation tools and metrics,” Proceedings of the 6th International Workshop on Historical Document Imaging and Processing (HIP ’21), Lausanne, Switzerland, Sep. 5-6, 2021, ACM, 2021, DOI: 10.1145/3476887.3476888.

[6] 出入国在留管理庁, “在留カード等読取アプリケーション,” https://www.moj.go.jp/isa/applications/procedures/rcc-support.html （閲覧日：2025 年 10 月 23 日）.

[7] H. Nakata et al., “OCR Text Analysis for Unstructured Document Images with Tuned Large Language Model,” Proceedings of the 38th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI), 2024.