CF-005

CLIP を基盤とした古典籍画像検索における 検索対象フィルタリングとファインチューニング

CLIP-Based Image Retrieval of Classical Japanese Books with Search Target Filtering and Fine-Tuning

山本 将也 [1] 佐藤 真一 [2] 大山 敬三 [2,3] 藤田 悟 [1]
Masaya Yamamoto Shin'ichi Satoh Keizo Oyama Satoru Fujita

法政大学 大学院 情報科学研究科 [1]
Graduate school of Computer and Information Sciences, Hosei University

国立情報学研究所 [2]
National Institute of Informatics

国文学研究資料館 [3]
National Institute of Japanese Literature

1. はじめに

近年、デジタル化された膨大な古典籍画像の利活用が課題となっている。これらに含まれる重要な挿絵や図表は、メタデータが未整備なため人手での探索が極めて困難である。本研究ではこの課題に対し、画像と言語を同一のベクトル空間で扱う CLIP モデルを応用し、自然言語や類似画像により、メタデータに依存しない効率的な検索基盤の構築を目指す。

本稿では物体検出による検索対象の拡張、不要領域のフィルタリング、複数画像クエリによる検索性能の改善、そしてファインチューニングによるドメイン固有語彙への対応を行い、これらの複数の手法を統合的に実装した。本システムが、人文科学と情報科学を繋ぐ学際的な研究基盤として、研究者の効率的な資料探索を支援することを期待する。

2. 先行研究

近年の AI 研究において、マルチモーダルモデルが注目を集めている。CLIP (Contrastive Language-Image Pre-training) [1] はその代表例で、画像と自然言語を共通のベクトル空間に写像することで、両者の意味的特徴を統一的に扱うことができる。テキストによる画像検索や生成など、多様なタスクに応用することが可能である。

CLIP を用いた画像検索における検索結果を、より利用者の意図に沿うよう向上させるためのアプローチも存在する。CLIP-Branches [2]は、利用者の「適合/不適合」フィードバックからその場で軽量な分類器を学習し、検索結果を再順位付けする。特定のドメイン知識の検索には、より根本的なファインチューニングが有効である。その効率的な手法として CLIP-Adapter [3]があり、巨大なモデル本体を凍結し軽量な Adapter 層のみを学習させることで、計算コストを抑えることができる。

さらに、CLIP を一枚の複雑な画像に適用する前の処理として、画像から物体領域を抽出する手法も重要な関連技術となる。オープンセット物体検出が主流となりつつあるが、このアプローチは検出対象とする物体をカテゴリカルに定義するのではなく、より汎用的な検出を目指すものである。代表例に Grounding DINO[4]がある。

3. 問題設定

国文学研究資料館が提供する約 45 万枚の古典籍画像を対象とする。これらの画像は手書きの文字が大部分を占めるが、研究対象として価値の高い挿絵や図表も多数含まれている。本研究の目的は、これらの画像群から自然言語または画像クエリを用いて、所望の挿絵や図表を効率的かつ高精度に検索するシステムを構築することにある。

4. 提案手法 (図 1)

4.1 物体検出技術による検索対象拡張

CLIP は入力画像を低解像度（224×224 ピクセル）に縮小するため、挿絵などの細部情報が失われ、検索対象から漏れる可能性がある。この問題を解決するため、まず高解像度の元画像 $N_{\text{orig}} \fallingdotseq$ 45 万枚に対し、自然言語クエリで物体領域を検出できる Grounding DINO を適用し、画像内の部分領域を網羅的に抽出する。汎用的なクエリ「object .」を用いることで、元画像に含まれる主要な視覚要素を持つ部分画像を切り出し、検索対象画像群に加えた。この処理により、検索対象画像群の総数は$N_{\text{total}} \fallingdotseq 380$ 万に拡張された。

4.2 ベクトル化とフィルタリング

次に、日本語特化の事前学習済み CLIP モデル「clip-japanese-base[5]」 の画像エンコーダを用いて、N_{total} 枚の検索対象候補画像をそれぞれ$D = 512$次元の特徴量ベクトルに変換する。このベクトルは、後段の処理のためノルムが 1 になるよう正規化する。4.3.1, 4.3.2 節で後述するとおり、この特徴量ベクトル群からなる行列 $V_{\text{total}} \in \mathbb{R}^{N_{\text{total}} \times D}$が検索インデックスとして機能することになる。

しかし、検索対象の元画像には挿絵のない文書が含まれることに加え、検索対象に追加した部分画像群にも文字や紙の汚れなどが多く含まれるため、挿絵検索の目的においてノイズとなる。そこで、検索品質向上のため、これらの不要な画像を予め除去するフィルタリング処理を行う。まず、約 1000 枚の画像に対して「採用（鮮明な絵など）」「除外（文書、文字、不鮮明な画像など）」のラベルを手動で付与した。このデータセットを用いて、画像特徴量ベクトルからラベルを予測する 2 値分類器（例：決定木）を学習させた。この分類器を用いて全画像を分類し、「採用」と判断された$N_{\text{filtered}} \fallingdotseq 114$ 万件の画像に対応する要素のみを保持した$V_{\text{filtered}} \in \mathbb{R}^{N_{\text{filtered}} \times D}$を構築した。これを$V_{\text{total}}$の代わりに用いることで検索対象画像群が最適化され、精度と効率の向上が期待できる。

4.3 検索

4.3.1 単クエリ検索

テキストまたは画像クエリから、対応する CLIP エンコーダを用いてクエリベクトル$q \in \mathbb{R}^D$を生成する。qとV_{filtered}との間で内積計算（正規化済みのためコサイン類似

度に相当）を行い、類似度の高い上位 k 件（例：$k = 100$）の画像を検索結果として提示する。

4.3.2 複数画像による検索

「蔀」のような古典籍特有の語彙は、事前学習済みモデルが十分に理解できず、単一のテキストクエリでは良好な結果が得られない。この課題に対し、本研究では、検索システムの利用者が別途自前で用意した、参考資料となる複数の画像を入力とし、それらの視覚的特徴の共通点を抽出したクエリベクトルを構成することで、所望の検索結果を得る手法を提案する。

まず、入力とするL枚の画像から得られる特徴量ベクトル群 $\{p_1, \ldots, p_L\}$を用意する。次に、これらのベクトル群の各次元j $(j = 1, \ldots, D)$ において、平均μ_jと標準偏差σ_jから変動係数$CV_j = \sigma_j / |\mu_j|$を算出する。ただし、分母の絶対値は、CLIP 特徴ベクトルの成分が負の値を取りうることを考慮した暫定的措置である。CV_jが低い上位d_{selected}個の次元を選択し、そのインデックス集合をSとする。

これは、変動係数が小さい次元は入力画像間で値のばらつきが少なく、対象に共通する本質的な特徴を示唆するという仮説に基づいている。このインデックス集合 S が定義する部分空間上で検索を行うことで、ノイズとなりうる無関係な次元の影響を排除する。具体的には、まず入力画像の平均特徴ベクトル q_{avg} を計算し、そこからインデックス集合 S に対応する次元のみを抽出したのち再正規化を施し、部分空間上のクエリベクトル $\tilde{q}_{\text{avg}} \in \mathbb{R}^{d_{\text{selected}}}$ を生成する。同様に、検索対象となる特徴量行列 V_{filtered}にも次元抽出と再正規化を適用し、検索時専用の縮小特徴 $\tilde{V}_{\text{filtered}} \in \mathbb{R}^{N_{\text{filtered}} \times d_{\text{selected}}}$ を作成する。最後に、$\tilde{q}_{\text{avg}}$ と $\tilde{V}_{\text{filtered}}$ との間で内積計算を行い、単クエリ検索時同様、類似度に基づき検索結果を提示する。この一連の処理により、入力画像群の共通点を考慮した頑健な検索を実現する。

4.4 対話的な検索改善

4.4.1 検索結果の補正

本研究では、検索結果を対話的に改善する補正機能も実装した。CLIP-Branches の手法を参考に、利用者が上位の検索結果に与えた少数の「適合/不適合」フィードバックと、N_{filtered}枚の画像からランダムに追加した「不適合」の例から、その場限りの軽量な決定木分類器を学習させる。この分類器で全検索対象をフィルタリングし、適合すると予測された画像のみを検索対象とし新たな結果を提示する。検索を繰り返し、フィードバックを追加するほど、検索精度の向上が期待できる。本稿では手法の紹介に留め、定量的な評価は今後の課題とする。

4.4.2 ファインチューニング

古典籍特有の語彙には、「蔀」のように事前学習済みモデルの知識だけではテキスト検索が困難なものが存在する。本研究では、このような語彙を検索可能にするため、本画像検索システムをアノテーションに活用した、効率的なファインチューニングの枠組みを導入する。具体的には、4.3.2 節の複数画像検索などを用いて特定の概念に合致する画像群を発見し、テキスト注釈を付与することで、高品質な（画像, テキスト）ペアのデータセットを効率的に構築できる。

このデータセットを用いて CLIP モデルをファインチューニングする際、計算コストが課題となる。もしモデル全体を再学習すると、$N_{\text{filtered}} \fallingdotseq 114$ 万件の画像に対し、元画像からの特徴量抽出（[元画像] → [512 次元ベクトル]）をすべて計算しなおす必要があり、非効率的である。

そこで、本研究では CLIP-Adapter の手法を採用する。この手法では、CLIP 画像エンコーダの重みを固定したまま、その後段に軽量なニューラルネットワーク（Adapter）を接続し、この Adapter のパラメータのみを学習する。このアプローチの最大の利点は、Adapter が「既存の特徴量」を入力とし、「新しい特徴量」を出力する写像として機能する点にある。すなわち、ベースモデルV_{filtered}に対し、学習済み Adapter による変換関数A_θを適用すると、ファインチューニング後の新しい特徴量行列 V'_{filtered}は、以下の式で高速に算出できる。

$$V'_{\text{filtered}} = A_\theta(V_{\text{filtered}})$$

これにより、索引全体の更新コストを削減しつつ、モデルを古典籍ドメインに効率的に適応させることが可能とな

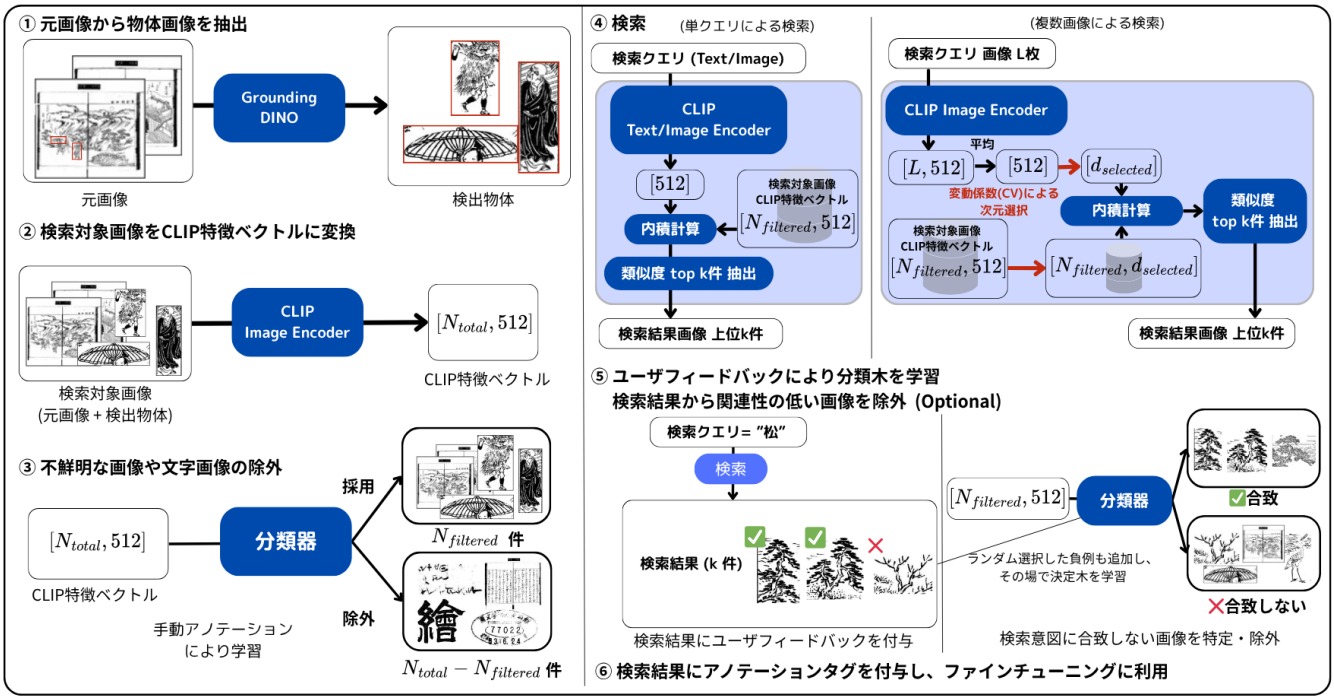

図 1 システムの全体像 (画像出典(1)・図中の画像は実際の動作ではなく、手法を説明するためのものです)

る。一方、テキストエンコーダは古典籍特有の語彙そのものを学習する必要があるため、全体をファインチューニングの対象とする。

5. 実験

5.1 検索対象分類モデルの選定

4.2 節で述べた検索対象の品質フィルタリングに用いる 2 値分類器（採用/除外）の選定のため、性能比較実験を行った。比較対象として、ランダムフォレスト、LightGBM、XGBoost、SVM、MLP の 5 つの代表的な機械学習モデルを評価する。

手動でアノテーションした約 1000 件の画像データから抽出した CLIP 特徴量を説明変数として使用し、そのうち 80%を学習、20%を検証データとして分割した。各モデルの主要なハイパーパラメータは、5 分割交差検証を用いたグリッドサーチにより最適化を行った。モデルの性能は、クラス不均衡を考慮できる加重平均 F1 スコア（weighted F1-score）を主指標として評価した。

5.2 複数画像検索

4.3.2 節で提案した複数画像検索手法、特に変動係数（CV）を用いた次元選択の有効性を検証するため、性能比較実験を行った。ある特定のテーマ(「冠」、「蔀」)に合わせてインターネット上で収集した参考画像(それぞれ 6, 8 枚)から代表クエリベクトルを作成し、提案手法である変動係数（CV）に基づく次元選択、比較対象としてのランダムな次元選択、そして次元選択を行わない全次元（512 次元）使用の 3 つの方法で検索精度を比較した。

次元選択を行う CV ベースの手法とランダム選択の手法については、選択する次元数を 64, 128, 192, 256 の 4 パターンで評価した。各手法で得られた検索結果について、上位 100 件における適合率（Precision@100）を算出し、性能を比較した。ランダム選択については、結果の頑健性を示すため 10 種類の異なる乱数シードでの試行を集計し、その統計値を算出する。

5.3 ファインチューニング (Visual Adapter, テキストエンコーダの学習)

N_{filtered}枚に含まれる画像の一部に対し、前述までの検索システムを活用して手動でテキストキャプションの付与を行い、数百件の（画像, テキスト）ペアからなるデータセットを作成した。付与したキャプションには「几帳」など古典籍特有語を含む。これを用いて、CLIP-Adapter の手法に基づきモデルのファインチューニングを行った。

ベースモデルには clip-japanese-base を使用した。その画像エンコーダの重みは完全に凍結し、後段にボトルネック構造（次元数を 1/4 に圧縮後、ReLU を介して復元）とスキップ接続を持つ Visual Adapter 層を接続した。学習では、この Adapter 層とテキストエンコーダ全体のパラメータのみを更新対象とした。

学習は、バッチサイズ 64、学習率 5e-5、3 エポックの設定で、AdamW オプティマイザを用いて対照学習損失（InfoNCE Loss）を最小化するように行った。学習完了後、学習済みの Visual Adapter を用いて、索引対象の全画像特徴量を高速に更新した。

6. 結果

6.1 検索対象画像除外用分類モデルの選定実験

各モデルを未学習の検証データで評価した結果を表 1 に示す。ランダムフォレスト[1]が、加重平均 F1 スコア 0.99 と他のモデルを大きく引き離して最も高い性能を示した。クラス 0（除外）、クラス 1（採用）ともに適合率・再現率が非常に高く、実用に十分な精度を達成している。次点で SVM と XGBoost が F1 スコア 0.92 と良好な性能を示したが、ランダムフォレストには及ばなかった。MLP と LightGBM は F1 スコア 0.91 であった。

6.2 変動係数（CV）に基づく次元選択の検証実験

変動係数（CV）を用いた次元選択手法の有効性を検証するため、テーマ「冠」「蔀」の 2 種類で性能評価実験を行った。その結果を図 2 に示す。

提案手法である CV を用いた次元選択は、両方のテーマにおいて、次元選択を行わないベースライン（全次元使用）の精度を概ね上回る結果を示した。テーマ「冠」では 192 次元で最高精度 0.64 を達成し、テーマ「蔀」でも全次元使用（0.38）を上回る最高精度 0.44（256 次元）を記録した。

比較対象であるランダムな次元選択は、10 回の試行で性能が大きく変動し、その最大値と最小値の差が大きい。提案手法は、すべてのケースでランダム選択の平均値を上回った。

6.3 ファインチューニング

本ファインチューニングの有効性を検証するため、学習時の損失（Loss）の推移と、実際の検索結果の変化を評価した。

図 3 に学習時の損失の推移を示す。学習ステップの進行に伴い、損失は順調に低下し、モデルが与えられたデータに適応していったことが確認できる。

次に、ファインチューニングに用いたキーワード「几帳」によるテキスト検索の定性的な結果を図 4 に示す。学習前は、無関係な画像が表示されるのみで、目的の画像を検索することはできなかった。一方、学習後は、所望の画像群が正しく検索結果の上位に表示された。なお、図 4 の右側で提示した画像は、ファインチューニングに直接使用していないが、少量のアノテーション画像から学習した「几帳」

表 1 各分類モデルの性能比較（検証データ）

	Precision	Recall	F1	Accuracy
Random Forest	0.99	0.99	0.99	0.99
XGBoost	0.92	0.92	0.92	0.92
LightGBM	0.92	0.92	0.92	0.92
SVM	0.91	0.91	0.91	0.91
MLP	0.91	0.91	0.91	0.91

表 2 Random Forest モデルのクラス別性能 (検証データ)

	Precision	Recall	F1	Support
0 (not keep)	1.00	0.99	0.99	161
1 (keep)	0.97	1.00	0.99	68

1 グリッドサーチによって決定したランダムフォレストのハイパーパラメータは以下に示す：
n_estimators=500, max_depth=50, min_samples_leaf=5, min_samples_split=2, max_features='sqrt'

の特徴に類似していると判断され、検索結果上位に露出したものである。

7. 考察

7.1 検索対象画像除外フィルタリングの効果

実験結果から、ランダムフォレストが本タスクにおいて最適な分類器であると結論付けられる。0.99 という高い F1 スコアは、CLIP の特徴量ベクトルが、挿絵や図表といった検索対象として価値のある画像と、本文の文字や汚れといったノイズとを、高精度で分離できる情報を含んでいることを示唆している。

このフィルタリングにより、検索索引のサイズを約 30%に縮小し（380 万件→114 万件）、計算効率を向上させると同時に、検索結果のノイズを大幅に低減させることが可能となり、検索の利便性が向上する。

7.2 変動係数（CV）に基づく次元選択の効果

実験結果は、提案手法（CV による次元削減）が、性能のばらつきが大きいランダム選択に比べ、安定して高い精度を出す頑健なアプローチであることを示している。本手法の精度は全次元使用時と比較して同等かそれ以上で、さらに次元削減により計算効率も高まるという二重の利点を持つ。これにより、利用者は試行錯誤に頼ることなく、一度の高速な検索で信頼性の高い結果を得られる。

一方で、ランダム選択が本手法を上回るケースがあったことは、変動係数（CV）が有効なヒューリスティクスではあるものの、常に最適解を与えるわけではない可能性も示唆している。より高度な特徴選択アルゴリズムを探求することは、今後の重要な課題である。

7.3 ファインチューニングの効果

学習ロスの順調な低下と、それに伴う検索性能の向上は、ファインチューニングが有効に機能したことを示している。特に、学習データに含まれていない類似画像までが検索結果に浮上した事実は重要である。これは、本モデルが単に（画像, テキスト）のペアを丸暗記したのではなく、より汎用的な「几帳らしさ」を視覚的特徴とテキストラベルの両面から捉え直し、その対応関係を学習したことを示唆している。

この汎化能力の獲得は、CLIP-Adapter アーキテクチャが既存の強力な視覚特徴を維持しつつ、テキストエンコーダ側で新たな語彙の意味を適切に構築した結果と考えられる。前述のとおり再計算にかかる時間も短く抑えられ、実用的なドメイン適応手法であると結論付けられる。

8. おわりに

本稿では、CLIP を基盤とする古典籍画像検索システムに対し、索引の品質向上（物体検出・フィルタリング）、検索の頑健性と効率の改善（CV 次元選択）、ドメイン適応（Adapter ファインチューニング）といった複数の手法を統合し、その有効性を実験で示した。本研究で提案した手法は、古典籍研究にとどまらず、他分野においても専門的な画像群に対する実用的な探索基盤を提供できる可能性がある。

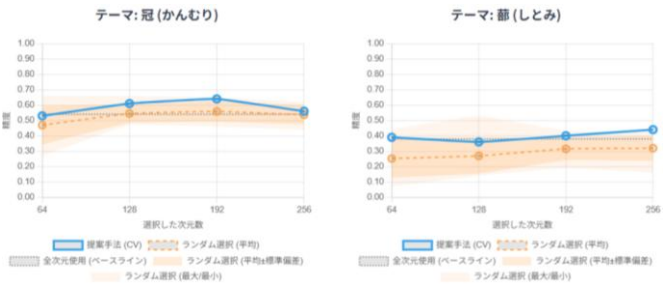

図 2 テーマ別画像群による複数画像検索時の変動係数 CV による次元選択の効果

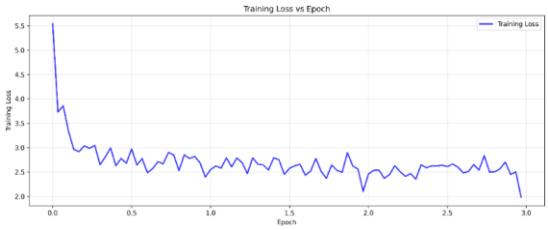

図 3 ファインチューニング 学習損失の推移

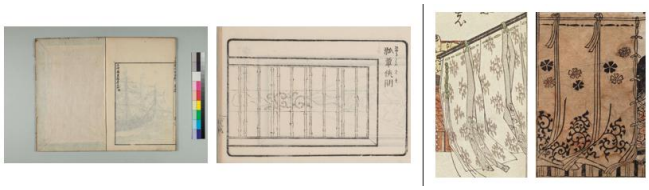

図 4 「几帳」検索上位画像からの抜粋
左 2 枚: ファインチューニング前　右 2 枚: 後・画像出典 (2)~(5)

謝辞

本研究は国文学研究資料館の「国文研プロジェクト型共同研究(萌芽研究): AI 技術を用いた大規模古典籍画像に対する新たな検索手法の研究」として実施されたものです.

参考文献

[1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021, February 26). Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020. https://doi.org/10.48550/arXiv.2103.00020

[2] Lülf, C., Martins, D. M. L., Vaz Salles, M. A., Zhou, Y., & Gieseke, F. (2024, June 19). CLIP-Branches: Interactive Fine Tuning for Text-Image Retrieval. arXiv preprint arXiv:2406.13322. https://doi.org/10.48550/arXiv.2406.13322

[3] Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., & Qiao, Y. (2021, October 9). CLIP-Adapter: Better Vision Language Models with Feature Adapters. arXiv preprint arXiv:2110.04544. https://doi.org/10.48550/arXiv.2110.04544

[4] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., & Zhang, L. (2023, March 9). Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv preprint arXiv:2303.05499. https://doi.org/10.48550/arXiv.2303.05499

[5] Yokoo, S., Okada, S., Zhu, P., Nishimura, S., & Takayama, N. (2024). clip-japanese-base. LY Corporation. https://huggingface.co/line-corporation/clip-japanese-base

画像出典

(1)『[書名なし]』(国文学研究資料館所蔵), 国書データベース, https://doi.org/10.20730/200006743
(2)『欄間雛形』(筑波大学附属図書館所蔵), 国書データベース, https://doi.org/10.20730/100272041
(3)『長崎聞見録』(奈良女子大学学術情報センター所蔵), 国書データベース, https://doi.org/10.20730/100258847
(4)『［女房三十六歌仙］』(東京藝術大学附属図書館所蔵), 国書データベース, https://doi.org/10.20730/100288299
(5)『歌仙絵抄』(東京藝術大学附属図書館所蔵), 国書データベース, https://doi.org/10.20730/100266422