意味的類似度に基づいた漫画コマ検索システムの提案

渡辺 聖陽[1,a)]　伊藤 正彦[1,b)]

概要：本研究では，漫画作品の増加に伴って好みの作品を探すことが困難になっている読者や，作画などの参考資料を求める作者を支援するため，文章入力によって特定の「漫画のコマ」を探索できるシステムを提案する．漫画データセット Manga109 から切り出したコマ画像と検索時に入力する検索文に対し，CLIP を用いて画像と検索文の特徴量ベクトルを抽出し，類似度計算を行う．計算の結果，類似度の高い TOP5 を表示し，類似度が一定以上になったものを UMAP を使用して散布図として表示する．探索を行った結果，具体的な動作や物体については意図に近い画像が提示されることが確認できたが，抽象的な表現では誤った検索結果も生じるなどの課題があることが明らかになった．

1. はじめに

近年，漫画を読むことのできる媒体は既存の雑誌形式だけでなく，デジタル上で読むことのできる媒体も登場している．漫画を読むことができる媒体が増えたことによって，作者にとって作品を発信する機会が増えている．その影響で，漫画の作品数は年々増加しており，読者は無数の選択肢の中から好みの漫画を見つけることが難しくなっている．現在の漫画検索システムでは「恋愛」や「バトル」などのジャンルのように漫画ごとに付けられたタグや，キーワードでの検索は可能である[*1]．しかし，漫画ごとの細かい特徴を使用することで検索することのできる漫画検索システムは調べた限り存在しない．

また，多種多様な漫画が増えてきている中，漫画を描く人向けの支援ツールは少ない．数少ない支援ツールとして，漫画に登場するキャラクターの設定やタイトルなどを考えるもの[*2]や，事前に用意されたパターンを使って漫画を作成できるもの[*3]などがあるが，実際にプロの漫画家が漫画を描く際に使用しているような構図や，キャラクターや特定の場面での演出の技法など，漫画の技術を学べる支援ツールは存在していない．

以上の漫画を読む「読者」側と漫画を描く「作者」側の課題を解決するために，文章入力による漫画検索システムを作成する．漫画検索システムとしては MaRU[1] という先行研究があるが，本研究では単なる検索システムの構築だけでなく，利用者にとって最適な UI の模索も行う．具体的には，「主人公とヒロインのリズムのいい会話がある」のような自身の好みや，「雨の日に傘を差して佇む女の子」のような参考にしたい場面の特徴を文章で入力して，読者は好みの漫画を，作者は参考にしたい場面の検索を行う．検索の際には入力された文章と漫画の画像の類似度を計算し，類似度の高い TOP5 と類似度に基づいた散布図を表示する．

2. 関連研究

相澤ら [2][3] は，漫画の学術研究に使える Manga109 というデータセットを作成し公開した．Manga109 は 1970 年代から 2010 年代までの間にプロの漫画家によって描かれた 109 冊の作品からなるデータセットである．データセットにはキャラクターに関するデータ，コマ，セリフのデータがあり，それぞれのデータに位置や大きさのデータも入っている．また，漫画を 1 ページごとに分けた画像データも入っている．漫画の学術研究に使えるデータセットの中でもここまで多くのデータが入っているデータセットはない．そのため，本研究では Manga109 データセットを用いることにする．

Shen[1] らは，漫画を画像処理と言語処理を用いて検索システムの MaRU を提案した．彼らは，物体検出モデルを用いてページ内のコマとセリフの吹き出しを検出し，MangaOCR[*4]を用いて検出された吹き出しからセリフを抜き出してテキスト化する．その後，SentenceBERT[4] や CLIP を組み合わせることでセリフとシーンの両方に対する自然言語での検索を実現した．

[1] 北海道情報大学
[a)] s2221102@s.do-johodai.ac.jp
[b)] imash@do-johodai.ac.jp
[*1] ブックライブ，https://booklive.jp/search/detail
[*2] コミコパ，https://lp.comic-copilot.ai/
[*3] マンガフィールド，https://lp.manga-field.net/
[*4] https://github.com/kha-white/manga-ocr

本研究では，MaRU では十分に考慮されていない「ユーザビリティ」に着目して，直観的に使用しやすい UI の設計を行う．

PromptMagician[5] は画像生成 AI を利用する際に入力するプロンプトの最適解を見つけ出すためのシステムである．このシステムでは，最初にプロンプトなどの情報を入力する．入力したプロンプトを基に画像の生成と，用意されていた画像データとの類似度を計算する．2 つの処理後に生成した画像と類似度を計算した画像を表示する．表示されている画像を選択することで詳細な情報を確認することができる．類似度の計算には，OpenAI が 2021 年に公開した CLIP というマルチモーダルモデルが使われている．本研究で作成するシステムにも CLIP を使用する点や，検索の手順とレイアウトなどに参考にするべき点が多くある．

本研究と PromptMagician の違いは使用目的にある．PromptMagician は，どのプロンプトが最適なのかを調べるためのもので，「文章」に重きを置いている．それに対して，本研究は自身の好みや参考にできる場面などを探すもので，「画像」に重きを置いている．

3. 提案システム

3.1 システム概要

本研究のシステムは，使用者が自身の好みの漫画や参考資料にしたい漫画を調べることを目的として使用するものである．本研究で作成するシステムの概要は図 1 に，システムの利用手順は図 2 に示す．最初に，利用者に文章を入力してもらう．次に，入力した文章に基づいて，用意した画像データとの類似度計算を行い，その結果を散布図として表示する．プロットされた点を選択することで，より詳しい情報を表示することができる．以降の節では，それぞれの手順と使用する技術についてより詳しく説明する．

3.2 Manga109 データセット

本研究では Manga109 データセットを使用している．使用する漫画は「よし まさこ」提供の「愛さずにはいられない」，「菅野 博之」提供の「天晴れ！カッポーレ」，「八神 健」提供の「ありさ 2」を使用する．既存の画像データでは漫画 2 ページ分の画像であり，情報量が多すぎるため類似度計算には適切ではなかった．そのため，本研究では画像データをデータセット内のコマの位置のデータを使用して 1 コマごとに切り取ったものを使用する．また，セリフの情報が特徴量に含まれている場合が考えられるため，セリフ部分を白く塗りつぶしたコマ画像を作成した．これによって情報量を削減して，検索文のイメージとより類似したコマを探し出すことができるようにした．

3.3 検索文の入力

最初に，利用者は図 1(A) に自身の好みの場面や参考にしたい場面を文章として入力する．ここで入力する文章とは，「主人公とヒロインのリズムのいい会話がある」のような好みの漫画の特徴や，「雨の日に傘を差して佇む女の子」のような参考にしたい漫画の場面などである．入力後，入力された検索文を googletrans を用いて英訳[*5]を行う．英訳を行う理由として，入力に日本語を使用すると，類似度を計算する際に類似度が正確に測れないことが探索の際に確認されたからである．

3.4 類似度の計算

入力された検索文と漫画のコマ画像との類似度の計算について説明する．事前に切り抜いた漫画のコマ画像全ての特徴量を算出する．検索文を入力後，検索文の特徴量の算出を行う．特徴量の算出後に，保存していたコマ画像の特徴量と検索文の特徴量との類似度をコサイン類似度を用いて計算する．

特徴量の算出には，OpenAI が公開している CLIP(Contrastive Language-Image Pre-training)[6] を使用する．CLIP は，インターネット上から収集した 4 億組の画像とテキストのペアを用いた大規模な学習により，画像とテキストという異なるデータ形式の類似度を計算することのできるモデルである．その結果，画像の特徴とテキストの特徴を共通の 512 次元の特徴量ベクトルに変換することができる．本研究では，事前にコマ画像の特徴量を算出して保存する際と，検索時に入力された検索文の特徴量を算出するのに使用する．

3.5 可視化

類似度の算出後，類似度の値が一定を超えたコマ画像の可視化を行う．可視化の際には，図 1(B) のように，コマ画像を散布図として表示する．散布図の表示には，512 次元の特徴量を 2 次元に圧縮するために UMAP[7] を使用する．表示されている点は，特徴が似ている画像が固まって表示されている．散布図の点の色分けは，作品名による色分けと，類似度の高さによる色分けの 2 つのパターンを用意した．表示されている点を選択することで，(C) に類似度や漫画のページ数などの詳細な情報を表示する．また，(D) に類似度が高い順の TOP5 を表示する．

3.6 利用手順

最初に図 1(A) の部分で検索文の入力，検索する画像の種別，散布図の色分けを選択する．検索後に類似度が一定以上になった画像を (B) に散布図として表示し，(D) に類似度の高い順で TOP5 を表示する．上部のスライダーを操作することで，設定した類似度未満の画像を表示しないようにできる．(B) に表示されている点を選択することで，

[*5] https://pypi.org/project/googletrans/

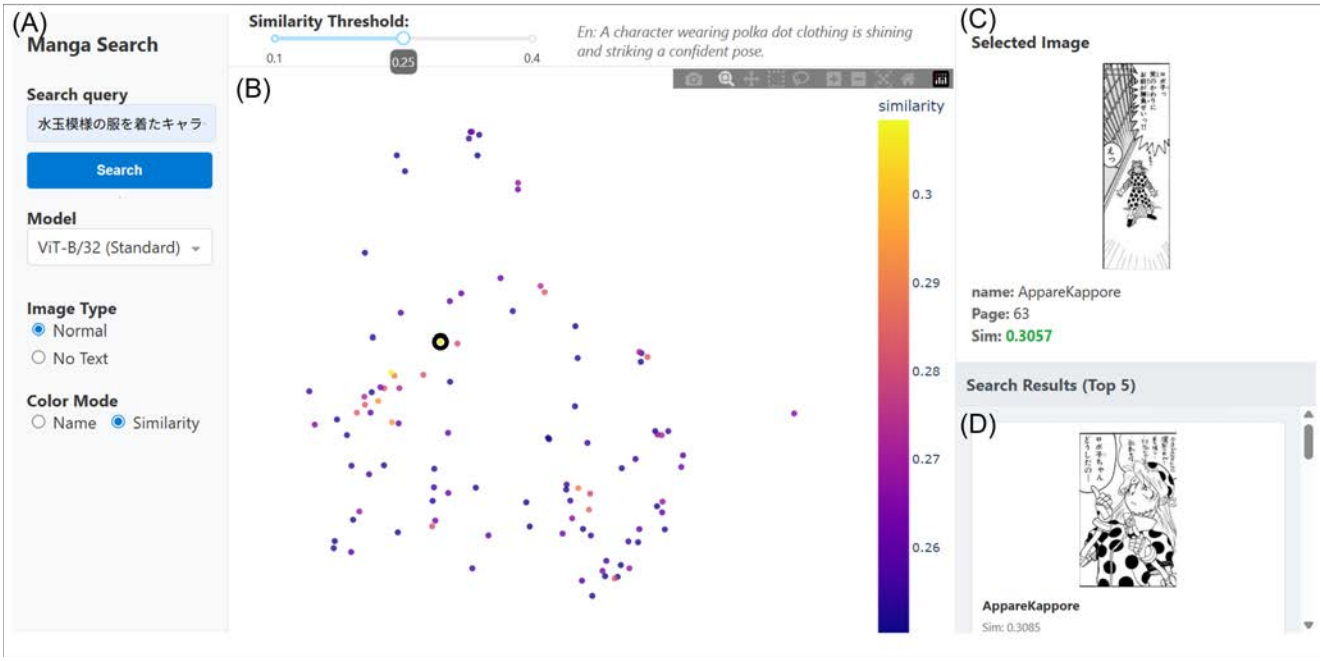

図 1　提案システムの UI 概要：(A) 検索入力，(B) 散布図，(C) 詳細表示，(D) 上位 5 件

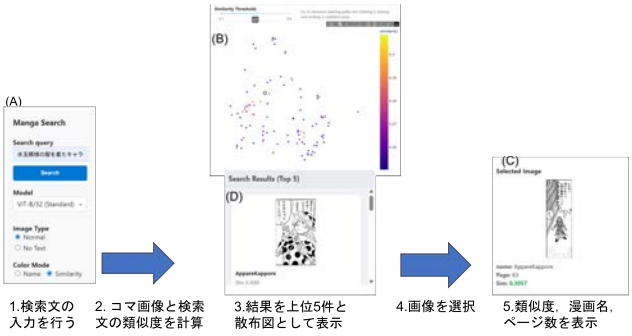

図 2　システム利用手順

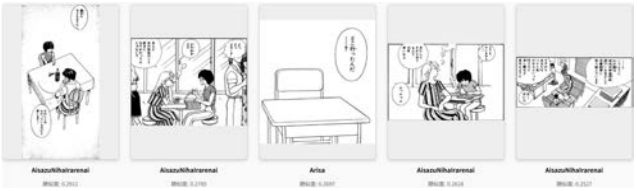

図 3　探索事例：テーブルを挟んで，二人の女性が座る（通常）

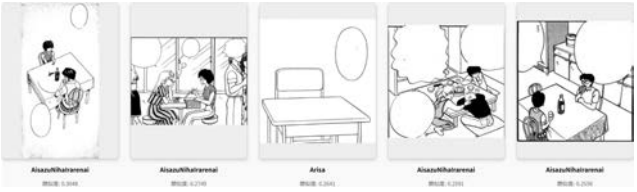

図 4　探索事例：テーブルを挟んで，二人の女性が座る（白塗り）

(C) に選択された点の画像とその詳細情報を表示する．

4. 探索事例

提案したシステムを使用して探索を行った事例を示す．探索事例として，探索事例 1 と探索事例 2 の 2 つの探索を行う．1 つ目に，CLIP の精度とセリフを塗りつぶした画像の有効性の検証のための探索を行った．検索文に具体的な表現を含めた例，抽象的な表現を含めた例，どちらの表現も含めた例の 3 つを，普通のコマ画像とセリフを白く塗りつぶした画像の 2 つの場合で探索を行った．ここで記述している具体的な表現とは，「花柄の服を着ている」のような具体的な物や「走っている」などの具体的な行動などを示す．また，抽象的な表現とは，「泣いている」などの感情表現や，「恋をしている男女」のような複雑な人間関係を示す．2 つ目に，散布図上の点の集まりが類似しているかを検証のための探索を行った．

4.1 探索事例 1：具体的な表現の例

最初に，具体的な表現を検索文とした場合の探索事例を紹介する．検索文は，「テーブルを挟んで，二人の女性が座る」とし，翻訳後の文章は「Two women sit across the table」である．通常の事例を図 3 に，白く塗りつぶした事例を図 4 に示す．2 つの事例を見ると，概ね検索文と一致しているコマ画像が表示されている．一致していない物に関しても，検索文内のテーブルなどの特徴は捉えられているため，具体的な表現を検索文として使用することは有効であることが示された．また，通常の事例とセリフを塗りつぶした事例の間には TOP5 の順番が入れ替わっただけのように大きな違いはなく，有効性は見られなかった．

4.2 探索事例 1：抽象的な表現の例

次に，抽象的な表現を検索文とした場合の探索事例を紹

図 5 探索事例 1：泣いている女性（通常）

図 6 探索事例 1：泣いている女性（白塗り）

図 7 探索事例 1：水玉模様の服を着たキャラクターが，輝きながら自信満々なポーズを決めている（通常）

図 8 探索事例 1：水玉模様の服を着たキャラクターが，輝きながら自信満々なポーズを決めている（白塗り）

介する．検索文は，「泣いている女性」とし，翻訳後の文章は「crying girl」である．通常の事例を図 5 に，白く塗りつぶした事例を図 6 に示す．どちらの事例も，上位 3 件に関しては検索文と概ね一致しているが，4 件目と 5 件目は一致していない．この理由として，CLIP は水のしずくのような涙滴型を涙の表現として見ていると考えられる．また，通常の事例とセリフを塗りつぶした事例には，あまり相違点がなく，有用な効果は見られなかった．

4.3 探索事例 1：どちらの表現も含めた例

最後に，どちらの表現も含めたものを検索文とした場合の探索事例を紹介する．検索文は「水玉模様の服を着たキャラクターが，輝きながら自信満々なポーズを決めている」とし，翻訳後の文章は「A character wearing polka dot clothing is shining and striking a confident pose.」である．通常の事例を図 7 に，白く塗りつぶした事例を図 8 に示す．2 つの事例を見ると，どちらも「水玉模様の服を着たキャラクター」という部分は一致している．しかし，「輝きながら自信満々なポーズを決めている」という文は，

図 9 アテンションマップ（成功例）

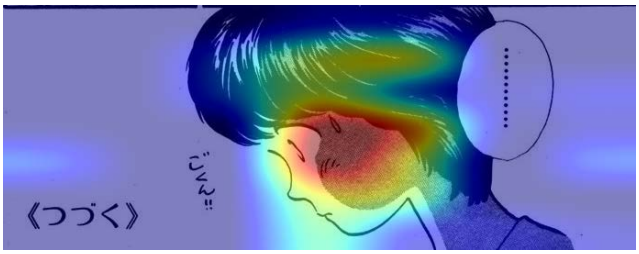

図 10 アテンションマップ（失敗例）

図 11 アテンションマップ（塗りつぶし画像）

図 7 の 5 つ目以外は該当していない．また，通常の事例と塗りつぶした事例には，あまり相違点がなく，有効な効果は見られなかった．

以上の探索事例から，検索文として具体的な表現を使用する場合は，検索文と一致した結果が出るが，抽象的な表現を使用した場合は，検索文と一致していない結果が出る場合があることが明らかになった．また，情報量の削減やより検索文と一致している漫画を探すためにセリフの塗りつぶしを行ったが，有効な効果は見られなかった．

4.4 アテンションマップによる可視化

4.2 節にて，TOP5 の結果の中で検索文と一致しない画像が表示されたことについて検証を行った．検証を行

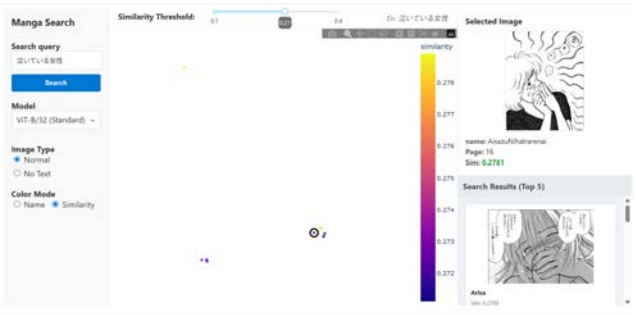

図 12 探索事例 2：類似画像の近接事例

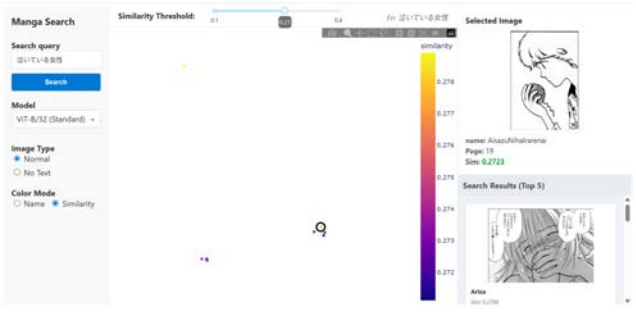

図 13 探索事例 2：類似画像の近接事例 2

うため，Grad-CAM (Gradient-weighted Class Activation Mapping)[8] を用いて CLIP の判断をアテンションマップとして可視化した．対象にしたのは，図 5 の 1 件目と 5 件目の画像である．なお，これらのアテンションマップにおいて，赤色はモデルが強く注目している領域を，青色は注目していない領域を表している．検索文としては 4.2 と同じく「crying girl」である．1 件目の画像を図 9 に，5 件目の画像を図 10 に示す．この 2 つの事例を見ると，どちらの画像も涙や汗を表現するための涙滴形に注目していることが分かる．これは，CLIP が涙と汗の表現として使われている涙滴形の判別ができていないということであり，涙滴形のように 2 つ以上の意味を持つ表現を CLIP が判別することは困難であることが示唆された．

また，塗りつぶした画像の有効性を確認するために図 9 のセリフを塗りつぶしたものとの比較を行った．塗りつぶしたコマ画像を図 11 に示す．事例の比較を行うと，どちらの事例も涙滴形に大きく注目しており，セリフの部分には注目していないことが分かる．この結果から，通常の場合と塗りつぶした場合には相違点はなく，有効性は限定的であると考えられる．

以上の結果を踏まえて，今後の研究方針として，より検索文と一致した結果を表示するために，CLIP を漫画に特化させたモデルにするための再学習を行うことや，抽象的な表現も使用できるような検索文の模索が考えられる．

4.5 探索事例 2：散布図の探索

この探索は，散布図上の点の集まりがどれだけ類似しているかを検証するための探索である．図 12 と図 13 の 2 つの探索を行った．選択している点は黒丸で囲われた点である．選択した結果を見ると，どちらの場面も人の横顔が書かれている場面であり，散布図上の集まりは類似していると判断できる．

5. おわりに

本研究では，ユーザーが文章を入力し，Manga109 データセットの画像データとの類似度の計算を行うことで，類似した漫画の画像を表示するシステムを提案した．類似度の計算は，CLIP を用いて特徴量を算出し，コサイン類似度を用いて類似度を計算した．

4 章で実際に探索を行った結果，期待した画像が表示された一方，類似していない画像が表示されるなどの課題も存在した．また，情報量の削減や類似した漫画を探すためにセリフの塗りつぶしを行ったが，あまり有効な効果は見られなかった．

今後は，より検索文と一致した漫画を検索するために，CLIP の再学習を行うことや，抽象的な表現も使用できるような検索文の模索を行う．またその他にも，システムの UI を利用者が使いやすいように改善することや，新たに画像を入力として使用し，「画像×画像」のような元々保存されているコマ画像と入力されたコマ画像との類似度を計算して検索を行うことができるようなシステムを追加する．最終的には，実際に作成したシステムを利用してもらい，システムの使いやすさや，システムが類似した画像として選んだ画像と利用者が選んだ画像がどれだけ一致しているかなどの評価実験も行っていく．

参考文献

[1] Shen, T. et al.: MaRU: A Manga Retrieval and Understanding system connecting vision and language, *arXiv [cs.IR]* (2023).
[2] Matsui, Y. et al.: Sketch-based Manga Retrieval using Manga109 Dataset, *Multimedia Tools and Applications*, Vol. 76, No. 20, pp. 21811–21838 (2017).
[3] Aizawa, K. et al.: Building a Manga Dataset “Manga109” with Annotations for Multimedia Applications, *IEEE MultiMedia*, Vol. 27, No. 2, pp. 8–18 (2020).
[4] Reimers, N. and Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, *EMNLP-IJCNLP 2019*, pp. 3982–3992 (2019).
[5] Feng, Y. et al.: PromptMagician: Interactive Prompt Engineering for Text-to-Image Creation, *IEEE VIS 2023* (2023).
[6] Radford, A. et al.: Learning Transferable Visual Models From Natural Language Supervision, *ICML 2021*, pp. 8748–8763 (2021).
[7] McInnes, L. et al.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, *arXiv preprint arXiv:1802.03426* (2018).
[8] Selvaraju, R. et al.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, *ICCV 2017*, pp. 618–626 (2017).