日本画像学会誌　第 58 巻　第 5 号：528–537（2019）

依頼解説

深層学習を用いたコミックの内容推定

朴　　炳宣*,**，松下　光範*

（2019.8.11　受理）

Estimating Comic Content Using Deep Learning

Byeongseon PARK*,** and Mitsunori MATSUSHITA*

Extracting information from comic images is more difficult than that from natural images such as photographs. This is because they are expressed in black and white line drawings and the contents are often drawn using special techniques such as exaggeration and simplification. Due to these characteristics, sufficient extraction accuracy cannot be obtained using general functions of conventional image processing such as SIFT（Scale-Invariant Feature Transfer）and HOG（Histograms of Oriented Grandients）.

In recent years, however, techniques of applying deep learning to comics have been proposed, contributing to a significant improvement of accuracy.

In addition, this technology can be used interpreting and estimating comic stories. This paper introduces research trends in deep learning applied to comic image processing.

Keywords: Image Processing, Content Estimation, Deep Learning, Transfer Learning, Comic Computing

コミック画像を対象とした情報抽出は，対象となる画像が白黒の線画で表現されていることや，描画内容に誇張や簡略化などコミック特有の技法が用いられていることなどの理由により，写真などの自然画像に比べて難しく，SIFT や HOG など従来画像処理で用いられてきた特徴量を用いるだけでは十分な精度が得られなかった．しかし近年，コミックに対して深層学習を応用した手法が提案され，大きな精度向上が図られるようになってきている．それに伴い，コミック画像中の物体認識や抽出だけでなく，ストーリーの解釈や推測などのより内容に踏み込んだ応用が行われている．本稿では，こうしたコミックの内容推定における深層学習の研究動向について紹介する．

キーワード：画像処理，内容推定，深層学習，転移学習，コミック工学

1. はじめに

現在，年間 1 万タイトルを超える新刊コミックが出版されているが，コミックコンテンツ内の情報（以下：内容情報）を考慮したアクセス方法は十分に発展しておらず，タイトルや著者名などの書誌情報と“ファンタジー”や“恋愛”などの大まかなジャンルを単位とした大雑把なアクセス手法に留まっている．そのため，ユーザが自分の読みたい内容をもとに作品を探すことは極めて困難であり，計算機がコミックの内容を理解し，ユーザの要求に応じて柔軟にコミックを推薦したり検索したりするシステムの実現が求められている．

こうしたサービスを可能にするためにはコミックコンテンツから内容情報を獲得する必要があるが，コミックはテキストとイラストの連携によってストーリーを構成するマルチモーダルなコンテンツであり，テキスト情報であるセリフやナレーションだけでなく，それと同一のコマ内で描かれる登場キャラクターや背景，オブジェクト（e.g.，車，サッカーボール）などのイラストによって表現された情報を抽出する必要がある．しかしながら，イラストは，作者ごとに記述方法（e.g.，デフォルメ，輪郭の太さ，目の描き方）や情報量の少なさ（e.g.，デフォルメ化による省略，限定された色のバリエーション），ブレの大きさ（e.g.，作者による表現手法の違い，演出による表現の誇張）などの問題によって，従来の画像処理技術を用いても各要素の抽出及び識別が容易ではない．

* 関西大学大学院　総合情報学研究科
〒569-1095　大阪府高槻市霊仙寺町 2-1-1
* Kansai University
2-1-1 Ryozenji, Takatsuki, Osaka 569-1095, Japan
** LINE 株式会社
〒160-0022　東京都新宿区新宿四丁目 1 番 6 号　JR 新宿ミライナタワー 23 階
** LINE Corporation
JR SHINJUKU MIRAINA TOWER 23rd FL., 4-1-6 Shinjuku, Shinjuku-ku, Tokyo, 160-0022, Japan

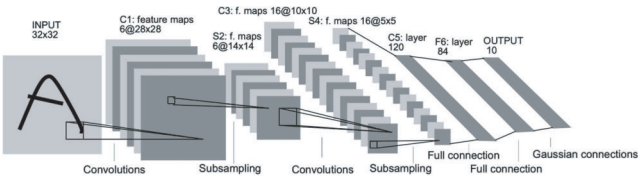

Fig. 1 The Overview of the Convolutional Neural Network (CNN)[10].

一方で，画像処理の研究分野において近年，ニューラルネットワークモデルを基盤とする深層学習アルゴリズムに関する研究が精力的に進められている（e.g.,［4，14］）．この学習アルゴリズムでは，人によって設計された特徴量を用いる従来法に比べ，コンピュータによる学習過程で特徴量を計算するため，人では設計することが困難な大規模の特徴量を得ることが可能である．このような深層学習の特徴は，上述したコミックにおける画像処理が抱えている問題点に対する分析手法として適しているといえる．こうした背景の下，本稿では深層学習を用いたコミックからの内容情報の獲得について解説する．

2. 深 層 学 習

近年ニューラルネットワークモデルを基盤とする深層学習アルゴリズムに関する研究が精力的に進められたことにより，画像処理分野における成果が飛躍的に増えつつある．深層学習が主流となる前は，画像局所特徴量と呼ばれる特徴量ベクトルを対象画像から抽出し，機械学習手法を用いて画像認識を実現する手法が主流であった．画像局所特徴量としては，局所領域でエッジ方向ごとにその強度をヒストグラム化した Histogram of Oriented Gradients（HOG）[13]や，輝度・回転・拡大縮小に不変な局所特徴量である Scale-Invariant Feature Transform（SIFT）[12]のようにヒューリスティクスに依る特徴量が広く使われていた．しかし，研究者の知見に基づいて設計に基づく特徴量は必ずしも最適であるとは限らない．これに対し，深層学習は対象とする画像データ集合から認識に有効な特徴量を自動的に抽出するアプローチであり，人間の知見や認識を遥かに凌駕する特徴量の設計を可能とした．

画像処理における深層学習の中でも代表的なアルゴリズムとして，LeCun らの畳み込みニューラルネットワーク（Convolutional Neural Network, CNN）[10]が挙げられる（**Fig. 1** 参照）．

CNN の特徴は，従来のニューラルネットワークモデルに基づく学習モデルの構造に加え，畳み込み層（Fig. 1 における C_1，C_3，C_5）およびプーリング層（Fig. 1 における S_2，S_4）と呼ぶ特殊な層を交互に接続した構造を持つことである．この特徴を除き，CNN の基本構造や学習過程は従来のニューラルネットワークモデルと同様である．CNN におけるある層の j 番目のユニットに，その直前の層の入力 $y_i(i=1,\cdots,m)$ の重み付き和にバイアスが加算された

$$x_j=b_j+\sum_{i=1}^{m}y_iw_{ij} \qquad (1)$$

が入力される．この x_j を，活性化関数と呼ばれる非線形関数に入力した時の応答 $y_j=f(x_j)$ がこのユニットの出力となり，次の層の各ユニットに入力される．活性化関数 f は，古典的なシグモイド関数 $f(x_j)=1/(1+e^{-x_j})$ のほか，近年は $f(x)=\max(x_j,0)$ という関数を持つユニット（rectified linear unit, ReLU[11]）がよく使われるようになっており，収束性や学習速度の向上に貢献している．通常，CNN の出力層付近には隣接層のユニット間を全て結合した層を 1 層以上配置する（Fig. 1 における F_6）．クラス分類が目的の場合，最後の出力には，目的のクラス数と同数のユニット n 個を配置し，これらのユニットへの入力 $x_j(j=i,\cdots,n)$ を，多項ロジスティック関数（もしくはソフトマックス関数）

$$p_j=\frac{e^xj}{\sum_{k=1}^{n}e^xk} \qquad (2)$$

を用い（$\sum_{j=1}^{n}p_j=1$ になる），出力とする．

Fig. 2 The Example Results of the Liu's Method[8].

上記の構造から構成されたネットワークによる学習は，ラベル付き学習サンプル集合を対象に，各サンプルの分類誤差を最小化することで行う．なお，出力層にあるユニットは，ソフトマックス関数による正規化（式（2）参照）により，対応するクラスに対する確率 $p_1,\cdots,p_n$ を出力する．分類誤差は，入力サンプルに対する理想的な出力 $d_1,\cdots,d_n$ と，実際の出力 $p_1,\cdots,p_n$ の乖離を交差エントロピー

$$C=-\sum_{j=1}^{n}d_j\log p_j \qquad (3)$$

によって測る．この C が小さくなるように，各層のフィルタの係数 h_{ijk} および同各ユニットのバイアス b_k，さらに出力層に設置した全結合層の重みとバイアスを調整する．

これらの過程で得られるパラメータによって得られるパラメータこそ，人によって設計された特徴量を用いる従来法との大きな違いであり，これによって人では設計することが困難な大規模の特徴量を得ることが可能である．これまでのことから深

Fig. 3　The Example Results of the Yagcioglu's Method[9].

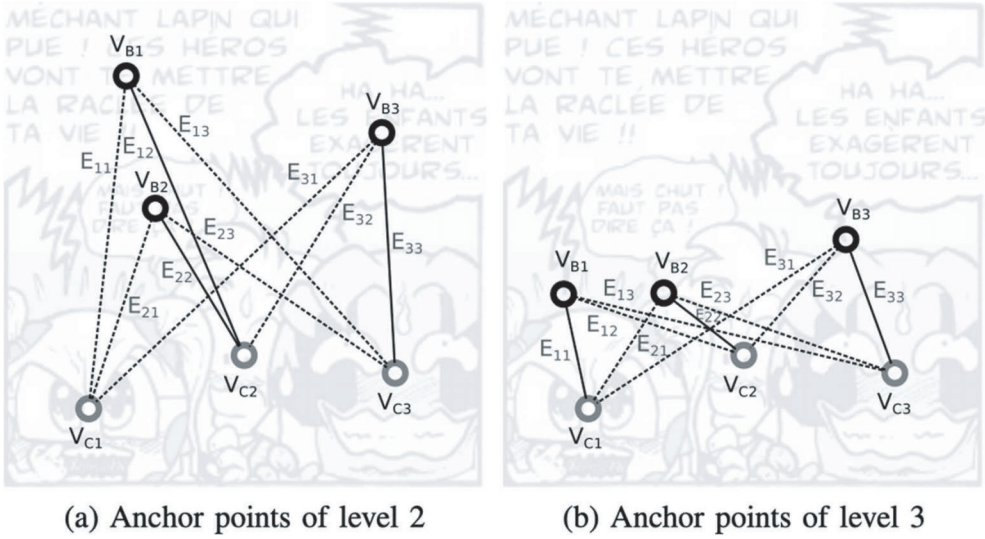

Fig. 4　The Example Results of the Rigaud's Method[27].

層学習は様々な分野での活用が進んでおり，画像処理においても目覚ましく成果を挙げている[14-16]．その中でも，画像の意味推定に密接なタスクとして挙げられる物体検出や説明文の生成に関する成果として，Liu らや Yagcioglu らの成果が挙げられる．Liu ら[8]は CNN を用いた高性能の物体検出モデルとして，Single shot multibox detector（SSD）を提案した．SSD は，入力画像から領域スキャンのアプローチを使わずに CNN で直接物体の位置を検出する One-Stage（Shot）と呼ばれるアプローチの 1 種であり，既存の手法では Bounding Box の出力を出力層だけで行っていたのに対し，SSD では CNN の複数の層から物体の Bounding Box を出力する（**Fig. 2** 参照）．また，Yagcioglu ら[9]は，入力画像に視覚的に類似した画像のキャプションから抽出された文ベクトルの平均ランクによって生成される分布意味ベースの形式に変換することで，精度と汎用性の高い説明文生成モデルを作成した（**Fig. 3** 参照）．

このような深層学習は，写真などの自然画像よりも色や線のバリエーションが顕著に少なく，作者ごとに記述方法（e.g., デフォルメ，輪郭の太さ，目の描き方）が異なるコミックに対する分析に適している．

3.　深層学習を用いたコミックの画像処理

自然画像と同様にコミック画像においても，深層学習が登場する前は SIFT や HOG といた人間によって設計された特徴量を用いた画像処理が主流となっていた．コミックの画像処理における主な研究タスクとしては，コミックのコマや吹き出しなどのレイアウトに関する要素の認識[26-28]や，テキストの認識[22]が挙げられる．その中でも，Rigaud ら[27]はコミックのレイアウト要素の認識に関する取り組みとして，各コマの話者を推定するための特徴量として，各コマとコマのアンカー，キャラクターの顔の位置の距離を用いている（**Fig. 4** 参照）．また，Praneesh ら[22]は，コミックのテキストの認識に関する取り組みとして，分類対象の集合を内的結合と外的分離が達成されるような部分集合の分割による分類手法であるクラスタリングの一種である Fuzzy c-Means[3]を活用し，コミックに含まれた色

Fig. 5 The Example Results of the Praneesh's Method[22].

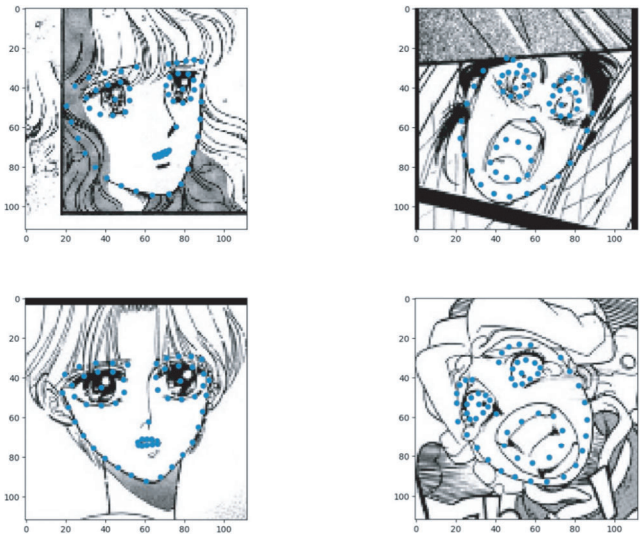

Fig. 6 The Example results of the Striker's method[38] (The images from the left, top to bottom are taken from : Ningyoushi by Omi Ayuko, PlatinumJungle by Shinohara Masami. The images from the right, top to bottom are taken from : RisingGirl by Hikochi Sakuya, ParaisoRoad by Kanno Hiroshi).

ごとの領域のセグメンテーションによって文字認識の精度を向上させた（**Fig. 5** 参照）．しかしこれらの研究では，コマとキャラクターの位置が一律に決まる関係にある状況や，色領域が明確である状況に高い精度を示すことは可能だったものの，位置関係が複雑になった状況（e.g.，コマに描かれていないキャラクターの発話，キャラクターが複数いる時）や色や形状が複雑に入れ混じっている状況では精度が低減することが確認された．上述したタスク以外にもキャラクターの識別[24, 25]のようにコミックの内容の中心となる要素の認識に関する試みもあったが，認識の精度や汎用性が低いという限界が存在していた．これらのことによって，人間による設計では，拡張し続けるコミックの表現に対応できる柔軟な特徴量を生み出すことは困難であると考えられる．

一方で，近年では深層学習を取り入れることによって，このような問題点を解決することが可能となったという報告が多数挙げられている．コマやテキストの認識において既存の手法よりも性能の向上[33, 36, 41]が見られた上に，キャラクターの識別においても著しい性能の向上が確認された[34, 35]．さらに，深層学習によるアプローチも用いることでより高度な認識について精力的に研究が行われている．Strikter ら[38]は，自然画像に写った人間の顔ランドマークを推定する Deep alignment network[6]を応用することで，コミックに描かれたキャラクターの顔ランドマークを検出できるモデルを提案し，東京大学大学院相澤清晴・山崎俊彦研究室によりとりまとめられた日本のコミックの画像データセット Manga109 データセット*1 [37]において 80% の検出精度を実現した（**Fig. 6** 参照）．Strikter らのモデルは，自然画像に比べ色のバリエーションが制限的であり，作者によって表現方法が異なるという難点を持つコミックでも高い精度を示しており，Strikter らのモデルの検出結果を応用することで，キャラクターの感情の推定やキャクターの唇と目にアニメーションを与えるといった発展が期待できる．また，Ogawa ら[40]は，既存の CNN をベースとしたモデルでは，大規模なデータベースを必要とする問題点を解決するために SSD[8]を応用することでコミックのコマやテキスト，キャラクターを限定的なデータ量でも複合的に検出できる SSD300-fork を提案した（**Fig. 7** 参照）．Ogawa らのモデルは，コミックの要素の複合的検出の先行研究である Rigaud ら[28]の手法と比べ，コマの検出性能は劣るものの，キャラクターの認識性能は優位であった上に，複数の作者の作品においても高精度を示すなど，高い汎用性を示した．

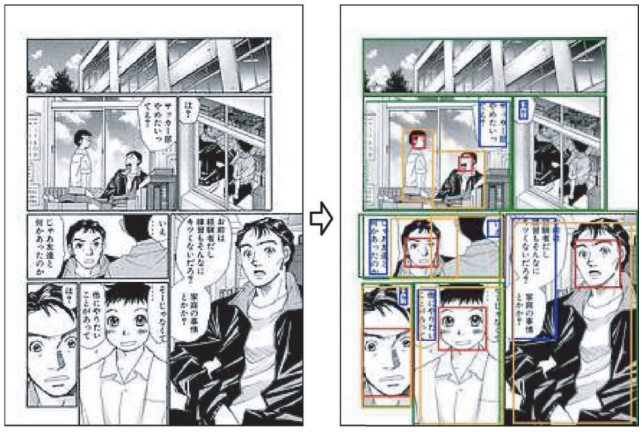

Fig. 7 The Example results of the Ogawa's method[40] (The images from YamatoNoHane by Saki Kaori).

このように，深層学習の登場によって，コミックにおける画像処理は既存研究に比べて性能が向上すると共に，研究領域の拡張にも貢献した．ただし，現状ではコミックに描かれた要素の同定が注目されており，コミックの連続したコマとページによって生まれるストーリーを把握するといった内容理解に関する取り組みはまだ活発的ではない．最近の研究成果では，Daiku ら[42]によるユーザが読みたい内容のページが検索できる検索システムの実現を目的とした CNN に基づくジャンル推定モ

*1 http://www.manga109.org/ja/（2019 年 8 月 10 日存在確認）

デルの提案が挙げられる．Daiku らは，コミックの内容は単一のジャンルに限らず，複合的なテーマが混在する点に着目し，各ページに描かれた内容をジャンルとして推定するモデルを提案した．Daiku らの研究が実現することでユーザの趣向に合わせていた作品の推薦が可能になることが期待できる．しかし，現状では限られた作品のみ安定した精度を示している上に，コミックの単行本のページをすべて利用して学習を行うため，実用化に向けたコストが膨大であることが課題となっている．

4.　表紙からの内容情報の獲得

上述したように，コミックにおける画像処理は，自然画像と同様に深層学習によって大きな進展が生まれた．しかし，これらの手法は，キャラクターの登場位置やセリフ情報を抽出することはできるが，これらの情報のみで，コミックのコンテンツとしての中核となる要素であるストーリーを把握することは容易ではない．コミックのストーリーはキャラクターを中心として展開され，その展開はキャラクターの状態（e.g.，身体的・心理的状態，位置，衣服，持ち物）や行動（e.g.，姿勢，動作，発言）の変化によって表現される．つまり，コミックの内容理解を実現するためには，キャラクターの出現位置や頻度以外のストーリーを形成する要素の抽出が必要となる．

4.1　コミックにおける表紙

そこで，著者らはよりストーリーに密接なコミックの内容情報の獲得手法として各コミックの表紙に着目した[2]．コミックの表紙は，ユーザは読んだことのない本を購入する際，重要な役割を持つ．例えば，あるユーザがコミックを選ぶために **Fig. 8** にある読んだことのない３つの作品の表紙を見ているとする．ユーザは図 Fig. 8 の（a）と図 Fig. 8 の（b）のコミックの内容を知らなくとも，表紙の中のキャラクターの服装や持ち物（e.g.，Fig. 8 の（a）の場合は“ローブ”と“剣”），Fig. 8 の（b）の場合は“制服”と“スーツ”）を見ることで，二つのコミックが題材やコンセプトを扱っていることが明確にわかる．一方で，Fig. 8 の（b）の表紙は，Fig. 8 の（c）と題材は異なるものの，似ているコンセプトに基づいていることが分かる．このように，ユーザーは表紙に描かれた情報を読み取ることで，本文の内容を推測し自分の趣向に合うコミックの選別を行う情報として用いることが可能である．よって，本稿では深層学習に基づく表紙に描かれたオブジェクトの識別を行うことで，コミックの内容を把握するための情報源として活用する．

4.2　表紙に含まれる内情情報

読者がコミックの表紙から読み取ることができる情報は下記のように大別できる．

- 絵柄

  絵柄情報は，作者がどのような記法を用いて作品を表現しているかに関する情報を指す．コミックは絵画のように作者の描写によって自由に表現されるコンテンツである．よってコミックの描写には線の強弱や配色といった，作者固有の習慣が現れる．さらに，作者は作品に対して，人間のデフォルメの度合いや，オブジェクトの描写の細かさにおいて，意図的に本来の習慣とは異なる描写記法を選ぶことがある．これは，コミックの絵柄は作品の雰囲気を大きく左右することから，より作品のテーマに合った演出を施すための工夫と言える．例えば，Fig. 8 の（c）の作品は，他の作品よりも使われている色の種類が少ない上にデフォルメ化された人物が描かれていることから，穏やかな雰囲気が伺える．一方，図 Fig. 8 の（a）は図 Fig. 8 の（c）に比べ人物やオブジェクト描写が細かく，特に人物の表情がより詳細に描かれているため，キャラクターの緊迫した様子が表現されている．以上のことから，絵柄情報は読者がまだ読んだことのないコミックへの第１印象を決定づける上で最も直感的な領域に関わる役割を持つと言える．

- ストーリー

  ストーリー情報は，表紙に表現された本編の内容に関す

(a) Seishinki Vulnus (by Yuzuru Shimazaki)

(b) Aris ² (by Ken Yagami)

(c) Everyday Osakana-chan (by Uka Kuniki)

Fig. 8　The Samples of the Comic's Cover.

Fig. 9 The Example results of the Saito's method[43].

る情報を指す．コミックは映画やアニメーションと同様，絵や動画などの視覚的演出のみならず，それらの組み合わせによって構成される「ストーリー」こそがコンテンツの核心であると言える．そのため，「この作品にはどのようなストーリーが含まれているのか」というストーリーに関する情報は，ユーザの嗜好に大きく影響を与える．よって，作者は作品の表紙に対して読者が作品のストーリーを推測できるように情報を与えることが多い．さらに，表紙は本の外側となる1枚の紙の分だけの容量しか持たず，そこに描くことができる情報は限りがあるため，作者は作品のストーリーにおいて最も重要だと思われる情報を選定し，イラストとして表現する．表紙に表現されるストーリー情報は，主に「シナリオ」と「コンセプト」に分けられる．まず，表紙に含まれるシナリオでは，表紙に描かれたキャラクターの表情や行動，位置関係によってキャラクターの立ち位置や関係性，本編のシーンが断片的に表現される．例えば，図 Fig. 8 の(b)と Fig. 8 の(c)の作品のキャラクターは，温厚な表情を浮かべている上に両隣に並んでいることから，友好的な関係が伺える．一方で図 Fig. 8 の(a)のキャラクターは剣を持って緊張している様子から，女性を守っているよう見える．また，表紙に含まれるコンセプトでは，キャラクターの衣服や持ち物，背景や周囲のオブジェクトなどの描画要素によって，作品のストーリーの題材やモチーフ，時代的・文化的背景が表現される．例えば，図 Fig. 8 の(a)の作品のキャラクターはローブをつけていたり，剣を持っていたりするなど，古代の西洋の文化に類似したコンセプトが伺える．一方で，図 Fig. 8 の(b)と Fig. 8 の(c)の作品のキャラクターは現代の制服やスーツ，T シャツなど，現代の文化に近いコンセプトが伺える．このように，読者はこれらの描画要素によって，図 Fig. 8 の(b)と Fig. 8 の(c)が類似したコンセプトを持ち，図 Fig. 8 の(a)が異なるコンセプトを持つことを認識することが可能となる．

このような表紙に含まれた情報を全て活用することで，読者のコミック検索に用いる内容情報をより容易に抽出することが可能となる．そこで著者らは，表紙情報に基づくコミックの情報アクセス手法の発端として，表紙に含まれた要素の中でも作品の「コンセプト」に関する情報を抽出し，コミックを関連づける手法について考察する．

コンセプト情報は他の要素よりも「知らない作品を探す」という状況において，情報量の個人差が少ない要素であると考えられるため，検索システムへの有効な活用が期待できる．

5. コミックの表紙からの情報抽出

5.1 モデル設計

コミックのように人によって描かれた画像に対する情報抽出の成果として，Saito らによる Illustration2Vec[43]が挙げられる．Saito らは画像の意味をベクトル化して解釈し，意味に基づいた画像検索を実現するために，Network In Network（NIN）モデル[45]と VGG[46]モデルを組み合わせた CNN モデルを提案した．Saito らの方法では，入力されたイラストを 4,096 次元のベクトルに写像し，イラストを表す特徴空間を構築するために，イラストからバイナリ属性（タグ）を予測するための CNN モデルを訓練する．**Fig. 9** は，Saito らの提案モデルによって得られる出力結果の例を示す．Saito らの手法を応用することによって，表紙に描かれた描画要素を推定することが可能になると考えられる．しかし，Saito らが学習対象とした画像はインターネット上のイラスト投稿サイトに投稿された 100 万枚以上のタグ付きイラストであるため，大規模な学習サンプルとして活用することが可能であった．一方で，本稿の対象となるコミックの表紙は，本手法に適した形で整理されたサンプルが存在しない．よって，サンプルの数は限定的となり，学習を繰り返すことによってパラメータを調整するため大量のデータが必要となる CNN の特性上，Saito らの手法を倣ったとしても倣ったとしてもいい成果を得られるとは限らない．

そこで，我々の提案手法では学習済ネットワーク（Pre-trained network）を用いた Fine-tuning を行った．Fine-tuning は，Pre-trained network のパラメータを固定し出力層のみを置き換えることで，Pre-trained network を純粋な特徴抽出器として用いる方法である転移学習（Transfer Learning）の一種である．転移学習を活用することで，大量のデータによって得られた特徴量を手軽に利用することが可能となる．一方，Fine-tuning は，転移学習のように Pre-trained network の出力層を置き換える以外にも，出力層に隣接する一部の層のパラメータに初期値にリセットし，誤差逆伝搬法による学習を進める方法である．よって Fine-tuning は，データが少なくとも Pre-trained network の豊富な特徴量を利用しつつ，Pre-trained network のパラメータをより新しいタスクに適したものに調整（Tuning）できることから，高い精度を得られるという長点を持つ．

このような特性を持つ Fine-tuning では，初期値となる Pre-trained network 自体の性能が最終的なモデルの精度に大きな影響を与える．そのため，一般的に ImageNet[44]の 1,000 クラスのデータを用いたコンペティション型ワークショップである ImageNet Large-scale Visual Recognition Challenge（ILSVRC）で優秀な成績を獲得したモデルが用いられること

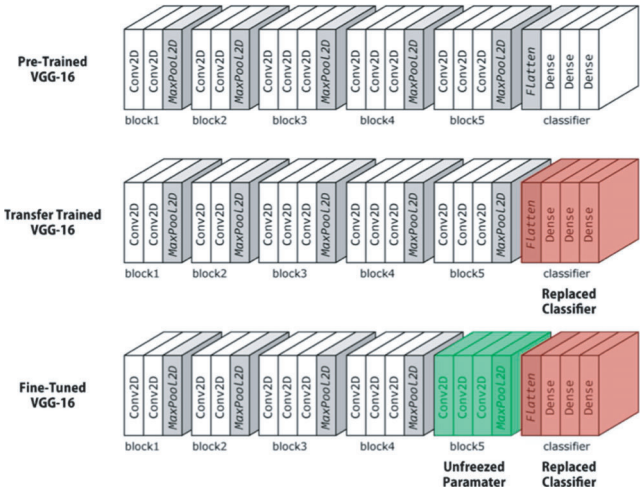

Fig. 10 The structure of the VGG-16 and the samples that applied the transfer learning and fine tuning.

が多い．そこで我々は，Saito らのレイヤ構造のベースとしても活用された VGG-16[46] の Pre-trained network を採用した．VGG-16 は，2015 年の ILSVRC で発表された Oxford Visual Geometry Group の 16 層 CNN（VGG-16，**Fig. 10** 参照）であり，畳み込み層と全結合層を連結したシンプルに層を増やしたネットワーク構造を持つ．構図のシンプルさから汎用性が高く，VGG-16 の構造をベースにした研究が良く見られている．

5.2 データ収集

我々の提案手法では，Fine-tuning 学習を行うための学習データとして，コミックの表紙の画像と，それらの画像に付与されたタグを用いる．コミックの表紙における既存の情報はコミックのタイトルや著者名といった書籍情報しか存在しないため，画像に描かれコンセプトに関連する情報（e.g.，キャラクターの衣服や持ち物，ストーリーの舞台）については新たにタグを付与しなければいけない．そこで，Web 書籍販売サイトである Amazon[*2] で販売されるコミックの中から，任意に選定したコミック 100 作品の 536 枚の表紙画像を集取した．次に，収集した各画像に対して，下記の要素から構成されたタグを作成した．

「名称」要素とは，表紙に描かれたオブジェクトに対する名称を指す．コミックのコンセプトを表現するための情報を作成するにあたり，各オブジェクトの名称の統一性が失われないように，「名称」要素には，固有表現（e.g.，地名，商品名）や修飾語（e.g.，色，大きさ，形状）を全て排除した名称のみを記入した．例えば，デニムパンツのようなオブジェクトがあった場合，「パンツ」のみ名称として記入している．「時代・文化カテゴリ」とは，同じ名称を持つオブジェクトを区別するための小区分を指す．例えば，表紙に「城」が描かれている場合，「城」は国や時代によって様式が異なる場合が存在する（e.g.，ノイシュヴァンシュタイン城と大阪城）．特徴が大きく異なる

[*2] https://www.amazon.co.jp/（2019 年 8 月 10 日存在確認）

Table 1 Number of output layers in our model.

Layer	Output
Flatten	25088
Batch Normalization	25088
Dense	512
Dropout	512
Dense	100

オブジェクトに対して同じタグが付与される場合，モデルの学習過程においてノイズとなる恐れがある．このようなケースであってもオブジェクトを区別できるように，（1）現代，（2）西洋，（3）東洋，（4）近未来，という 4 つのクラスを「時代・文化カテゴリ」として設けた．最後に，「補足情報」要素とは，オブジェクトに対する補足情報を指し，「名称」要素で省略された情報を記録するための要素である．「補足情報」要素を設けることによって，「名称」要素で省略された情報を任意に拡張することが可能となる．これらの規則に基づき作成したタグの例として，「デニムパンツ」の場合「パンツ：1：デニム」，「大阪城」の場合，「城：3：」がある．

データ作成の結果，上記の規則に基づき作成したタグは合計 560 種類であった．学習に用いるためのデータとして，あまりに希少なタグはモデルの精度に影響を及ぼす可能性があると考えられる．そこで，本稿では全ての表紙における出現頻度が 5 以上（5 作品以上が共通的に出現するタグ）となる上位 100 位までのタグを使用した．

5.3 学習

作成したデータから VGG-16 モデルを用いた Fine-tuning を行った．まず，Fine-Tuning のために，学習に用いるレイヤーとそのパラメータを設定した．VGG-16 モデルの Pre-trained network の出力層（Fig. 10 の Classifier）は **Table 1** のような構造を持つものに置き換えた．また，出力層から最も近いブロック（図 Fig. 10 の Block 5）にある 3 つの畳み込み層と 1 つのプーリング層のパラメータを初期値にリセットした．この過程によって，既存の VGG-16 の Pre-trained network のレイヤーまでの特徴量によって得られ入力から，新しいレイヤーのパラメータを学習させることが可能となる．最後に，学習過程におけるパラメータでは，学習率（各学習時のパラメータの更新の大きさ）は 10^4，慣性（これまであった傾向と類似した入力の受け入れやすさ）は 0.9，エポック数（学習回数）は 150 として設定した．

上記の設定に基づき学習を行う際，表紙の画像データ（合計 536 枚）の 20%（107 枚）を評価データとして用いた．CNN においてデータの数は精度と比例するため，本来用いるデータの規模は大きいことが望ましい．よって，少量のデータによって Fine-Tuning を行う場合，既存の入力データと正解データの組を活用し，入力データにのみ変化を行った新しい組を作成することで，データ拡張（Data Augmentation）を行う．本稿では，学習の際に学習データのみ（1）上下反転，（2）左右反転，（3）90° 回転，（4）270° 回転，（5）Enhanced Edge フィルタを適用，（6）強度な Enhanced Edge フィルタを適用，（7）

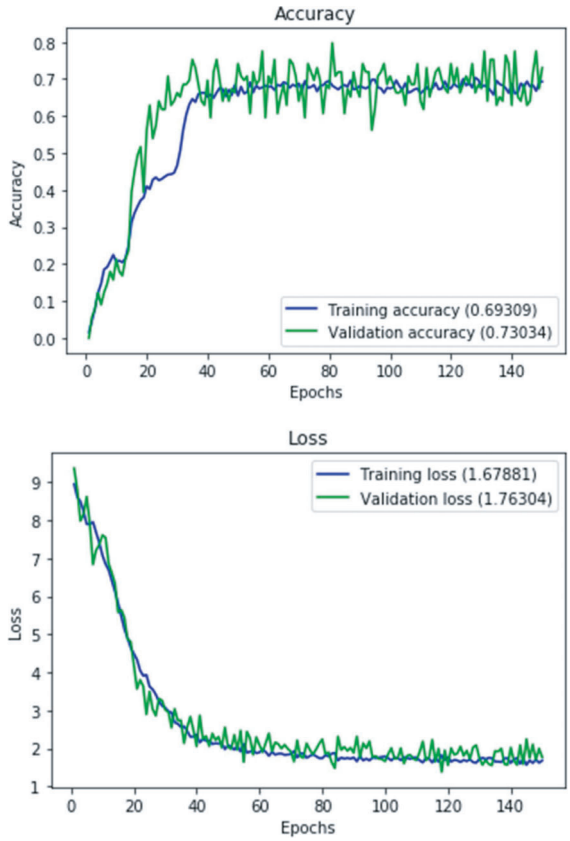

Fig. 11 The result of learning by fine-tuned VGG-16.

Enhanced Edge フィルタによる輪郭のみを使用，といった差分データを合計 7 種類の作成し，学習に用いた．また，特定の画像やタグに偏って学習が行われることを防ぐために，Epoch ごとに学習データと評価データを任意に再選別した．

上記の手順により学習を行った結果を **Fig. 11** に示す．150 エポック時点での学習データにおける精度は 69.3%，損失は 1.679，評価データにおける精度は 70.3%，損失は 1.763 であった．また，作成したモデルから全データに対する予測を行い，その出力（各タグに対する確信度）に対して，すべての表紙が一つ以上のタグを持つように閾値（0.1）を設定した際の正解データへの再現度は 91.8% であった．

6. おわりに

本稿では，コミックにおける画像処理に関する最新の動向を紹介しつつ，深層学習を用いたコミックからの内容情報を獲得する手法について述べた．

近年のコミックにおける画像処理では，人間によって設計された特徴量よりも膨大かつ複雑な特徴量を作成できる深層学習の導入によってレイアウト要素の識別，テキスト認識などの既存のタスクの性能が大きく向上している．しかし，現状ではストーリーといったコミックに含まれた複数の要素の組み合わせによって表現された，より高度な要素に関する取り組みは活発的に行われていない．そこで，著者らはストーリーを表現する要素として，コミックの表紙に描かれた服装や持ち物を活用し，作品のコンセプトを推定する手法を提案した．著者らは表紙からコンセプトに関連する情報を抽出するために，コンセプトを表現する情報（e.g.，衣服，持ち物）をタグとして付与したコミックの表紙の画像を用いて深層学習の一種である Fine-Tuning を活用し，約 92% の再現度を示すモデルを作成することができた．今後は，データベースの構築や検索の実現に向けたより高度な内容情報抽出（e.g.，ストーリーの理解・要約，感情推定）が可能となる手法の実現に向けた深層学習の活用と活性化を期待する．

参 考 文 献

1) B. Park, K. Okamoto, R. Yamashita, and M. Matsushita, “Designing a comic exploration system using a hierarchical topic classification of reviews,” Information Engineering Express International Institute of Applied Informatics, **3 (2)**, pp. 45–57 (2017).
2) B. Park and M. Matsushita, “Estimating Comic Content from The Book Cover Information Using Fine-Tuned VGG Model,” In Proceedings of MultiMedia Modeling (2019), **2**, pp. 650–661.
3) J.C. Bezdek. “Pattern Recognition with Fuzzy Objective Function Algorithms,” Plenum Press (1981).
4) G.E. Hinton, S. Osindero, and Y.W. The, “A fast learning algorithm for deep belief nets,” Neural Computation, **18**, pp. 1527–1544 (2006).
5) A. Krizhevsky, I. Sutskever, and G.E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” In Proceedings of NIPS (2012), **1**, pp. 1097–1105.
6) M. Kowalski, J. Naruniec, and T. Trzcinski, “Deep alignment network : A convolutional neural network for robust face alignment,” In Proceedings of the International Conference on Computer Vision & Pattern Recognition Faces-in-the-wild Workshop Challenge (2017), **3**, pp. 88–97.
7) R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” In Proceeding of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580–587.
8) W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A.C. Berg, “SSD : Single shot multibox detector,” In Proceedings of European Conference on Computer Vision (2016), pp. 21–37.
9) S. Yagcioglu, E. Erdem, A. Erdem, and R. Cakici, “A Distributed Representation Based Query Expansion Approach for Image Captioning,” In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (2015), **2**, pp. 106–111.
10) Y. Le Cun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” In Proceedings of the IEEE (1998), **86** pp. 2278–2324.
11) V. Nair and G.E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” In Proceedings of CVPR (2008).
12) D.G. Lowe, “Object recognition from local scale-invariant features,” In Proceedings of IEEE ICCV (1999), **2**, pp. 1150–1157.
13) N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), **1**, pp. 886–893.

14) A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems (2012), pp. 1097–1105.
15) C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), **1**, pp. 1–9.
16) S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, and P.H. Torr, "Conditional Random Fields as Recurrent Neural Networks." In Proceedings of the 2015 IEEE Conference on Computer Vision (ICCV 2015) International, pp. 1529–1537.
17) T. Tanaka, K. Shoji, F. Toyama, and J. Miyamichi, "Layout analysis of Tree-Structured scene frames in comic images," In Proceedings of the 20th International Joint Conference on Artificial Intelligence (2007), pp. 2885–2890.
18) K. Arai and H. Tolle, "Method for automatic E-Comic scene frame extraction for reading comic on mobile devices," In Proceedings of the 7th International Conference on Information Technology (2010), pp. 370–375.
19) H. Tolle and K. Arai, "Method for Real Time Text Extraction of Digital Manga Comic," International Journal of Image Processing, **4**, pp. 669–676 (2011).
20) C. Rigaud, N. Tsopze, J.C. Burie, and J.M. Ogier, "Robust Frame and Text Extraction from Comic Books," Graphics Recognition New Trends and Challenges : Proceedings of the 9th International Workshop on Graphics Recognition (GREC 2011), pp. 129–138.
21) S. Nonaka, T. Swano, and N. Haneda, "Development of "GT-Scan" the Technology for Automatic Detection of Frames in Scanned Comic," Fujifilim Reserch & Development, **57**, pp. 46–49 (2012) [In Japanse).
22) M. Praneesh and J.R. Kumar, "Novel Approach for Color based Comic Image Segmentation for Extraction of Text using Modify Fuzzy Possiblistic C-Means Clustering Algorithm," International Journal of Computer Applications, 1, pp. 16–18 (2012).
23) C. Guerin, C. Rigaud, A. Mercier, F.A. Boudjelal, K. Bertet, A. Bouju, J.C. Burie, G. Louis, J.M. Ogier, and A. Revel, "eBDtheque : a representative database of comics," In Proceedings of International Conference on Document Analysis and Recognition (2013), pp. 1145–1149.
24) D. Ishii and H. Watanabe, "A Study of Automatic Human Detection for Comic Image," IPSJ SIG Technical Report, **2012 AVM-76**, pp. 1–5, (2012).
25) W. Sun, J.C. Burie, J.M. Ogier, and K. Kise, "Specific Comic Character Detection Using Local Feature Matching," In Proceedings of International Conference on Document Analysis and Recognition (2013), pp. 275–279.
26) C. Rigaud, J.C. Burie, and J.M. Ogier, "Text-Independent Speech Balloon Segmentation for Comics and Manga," Graphics Recognition New Trends and Challenges : Proceedings of the 11th International Workshop on Graphic Recognition (GREC 2015), pp. 133–147.
27) C. Rigaud, N.L. Thanh, J.C. Burie, J.M. Ogier, M. Iwata, E. Imazu, and K. Kise, "Speech balloon and speaker association for comics and manga understanding," In Proceedings of International Conference on Document Analysis and Recognition (2015), pp. 351–355.
28) C. Rigaud, C. Guérin, D. Karatzas, J.C. Burie, and J.M. Ogier, "Knowledge-driven understanding of images in comic books," International Journal on Document Analysis and Recognition, **18**, pp. 199–221 (2015).
29) R. Gaikwad, and N.G. Pardeshi, "Text Extraction and Recognition Using Median Filter," International Research Journal of Engineering and Technology, **3**, pp. 717–721 (2016).
30) J.M.C. Correia and A.J.P. Gomes, "Balloon extraction from complex comic books using edge detection and histogram scoring," Multimedia Tools and Applications, **75**, pp. 11367–11390 (2016).
31) C. Rigaud, S. Pal, J.C. Burie, and J.M. Ogier, "Toward speech text recognition for comic books," In Proceedings of the 1st International Workshop on coMic ANalysis, Processing and Understanding (MANPU) (2016), **8**, pp. 1–6.
32) S. Hiroe and S. Hotta, "Histogram of Exclamation Marks and Its Application for Comics Analysis," In Proceedings of International Conference on Document Analysis and Recognition (2017), pp. 66–71.
33) Y. Aramaki, Y. Matsui, T. Yamasaki, and K. Aizawa, "Text detection in manga by combining connected-component-based and region-based classifications," In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), pp. 2901–2905.
34) W.T. Chu and W.W. Li, "Manga FaceNet : Face detection in manga based on deep neural network," In Proceedings of the ACM on International Conference on Multimedia Retrieval (2017), pp. 412–415.
35) N.V. Nguyen, C. Rigaud, and J.C. Burie, "Comic characters detection using deep learning," In Proceedings of International Conference on Document Analysis and Recognition (2017), pp. 41–46.
36) H. Yanagisawa and H. Watanabe, "Recognition of panel structure in comic images using Faster R-CNN," In Proceedings of the 5th IIEEJ International Workshop on Image Electronics and Visual Computing (IEVC 2017), **4C-2**, pp. 1–5.
37) Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using manga109 dataset," Multimedia Tools and Applications, **76**, pp. 21811–21838 (2017).
38) M. Stricker, O. Augereau, K. Kise, and M. Iwata, "Facial Landmark Detection for Manga Images," arXiv : 1811.03214 (2018).
39) H. Yanagisawa, T. Yamashita, and H. Watanabe, "A Study on Object Detection Method from Manga Images using CNN," In Proceedings of the 2018 International Workshop on Advanced Image Technology (IWAIT), pp. 1–4.
40) T. Ogawa, A. Otsubo, R. Narita, Y. Matsui, T. Yamasaki, and K. Aizawa, "Object Detection for Comics using Manga109 Annotations," arXiv : 1803.08670 (2018).
41) D. Dubray and J. Laubrock, "Deep CNN-based Speech Balloon Detection and Segmentation for Comic Books," arXiv : 1902.08137 (2019).
42) Y. Daiku, O. Augereau, M. Iwata, and K. Kise, "Comic story analysis based on genre classification," In Proceedings of International Conference on Document Analysis and Recognition (2017), pp. 60–65.
43) M. Saito and Y. Matsui, "Illustration2Vec : a semantic vector representation of illustrations," In Proceedings of SIGGRAPH ASIA 2015 Technical Briefs (SA 2015), **5**, pp. 1–4.
44) J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, "ImageNet : A large-scale hierarchical image database," In Proceedings of CVPR (2009), pp. 248–255.

45) M. Lin, Q. Chen, and S. Yan, “Network in network,” In Proceedings of International Conference on Learning Representations (ICLR 2014).
46) K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” In Proceedings of International Conference on Learning Representations (ICLR 2015).

朴　　炳宣

2019年関西大学大学大学院総合情報学研究科知識情報学専攻博士課程前期課程を修了．同年LINE株式会社に入社．音声合成技術の自然言語処理に関する開発を担当．同年関西大学院大学同学科博士課程後期課程に入学．現在，コミック工学と機械学習に関する研究に従事．情報処理学会，人工知能学会各会員．

松下　光範

1995年大阪大学大学院基礎工学研究科物理系専攻制御工学分野博士前期課程修了．同年，日本電信電話株式会社入社．2008年関西大学総合情報学部准教授．2010年同教授．自然言語理解，インタラクションデザインに関する研究に従事．博士（工学）．2003年情報処理学会論文賞，2013年LavalVirtual Award，2017年芸術科学会論文賞ほか各賞受賞．電子情報通信学会，情報処理学会，芸術科学会，ACM各会員．