複数の形式からなる複合データへのメタデータ付与

鈴木 釈規[1,a)]　池田 大輔[2,b)]

概要：近年では，動画，画像及び文章などの複数の形式からなるデータ（複合データ）が増加している．メタデータにより効率的な参照が可能となるが，メタデータ付与の従来手法は，単一形式のデータを利用した手法に限られる．複合データの豊富な情報活用により，メタデータ付与の正確性を向上することが期待できる．本論文では画像に付与されるメタデータの曖昧性解消と画像へのメタデータ付与の二つの研究課題に取り組む．曖昧性解消では付与されたメタデータと画像分類により取得した画像情報を組み合わせた手法を，画像へのメタデータ付与では，画像からテキスト特徴量への変換とそれに基づく分類手法を提案する．従来手法との比較評価により提案手法の有効性を示す．

キーワード：複合データ，画像，テキスト，メタデータ付与

Assignment of Metadata for Compound Data Consisting of Multiple Data Formats

Tokinori Suzuki[1,a)]　Daisuke Ikeda[2,b)]

Abstract: In recent years, the amount of multimodal data consisting of multiple data formats such as mixtures of videos, images and texts, which we call the *compound data*. Although metadata makes it possible to access such data quickly, most effort of previous studies on automatic assignmennt of metadata have been devoted to single format of data that cannot make full use of rich information of the coumpound data. In this paper, we propose a method utilzing the compound data for each of two research tasks: disambiguation of metadata assigned to images and automatic metadata annotation to images. We demonstrate their effectiveness with experiments.

Keywords: Compound Data, Image, Text, Assignment of Metadata

1. はじめに

近年，ネット上で共有されるデータが多様化している．例えばソーシャルメディアサービス（SNS）では複数の形式からなデータがコンテンツになる．画像 SNS サイト Instagram[*1]のコンテンツは，投稿画像・動画と，それに対するユーザのコメントからなるデータとみなせる．本論文では，こうした複数の形式から構成されるデータを複合データと呼称する．複合データへの効率的な参照とするメタデータ付与は重要な課題である．

メタデータ付与に関する従来研究の多くは，単一形式データの利用に限られる．例えば，文章データ中の特定の単語が指し表す実体を特定して，知識体系に対応付けすることもメタデータ付与とみることが出来る．特に，知識体系としてウィキペディア[*2]が用いられる場合，ウィキフィケーション [1, 2, 11] と呼称される．この課題では主に文章データを利用する手法に限られる．また，画像へのメタデータ付与では，参照用キーワードを割り当てる画像アノ

[1] 九州大学大学院システム情報科学府
Graduate School of Information Science and Electrical Engineering, Kyushu University
[2] 九州大学大学院システム情報科学研究院
Faculty of Information Science and Electrical Engineering, Kyushu University
[a)] suzuki.tokinori.070@s.kyushu-u.ac.jp
[b)] daisuke@inf.kyushu-u.ac.jp

[*1] https://www.instagram.com/

[*2] https://www.wikipedia.com/

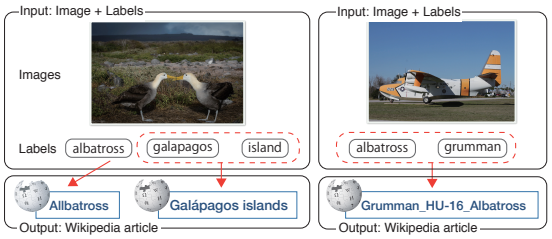

図 1 画像キーワード曖昧性解消の概要
Fig. 1 Task overview of the disambiguation of keywords assigned to images.

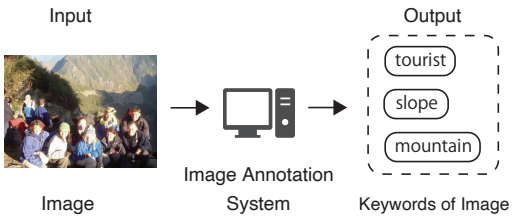

図 2 画像アノテーション課題の概要
Fig. 2 Task overview of the automatic image annotation task.

テーション課題 [7, 8, 16, 17] が研究されている．多くの付与手法が提案されているが，それらは画像情報のみの利用に限られる．

複合データは単一形式データよりも豊富な情報を持つ．豊富な情報の活用により，メタデータ付与の正確性を向上することが期待できる．しかしながら，複合データの活用についてはこれまでに研究がされてこなかった．本論文ではメタデータ付与に関する課題を具体例として，複合データを活用方針の検討を行う．研究課題として 1）画像に付与されるメタデータの曖昧性解消と 2）画像へのメタデータ付与の二つに取り組む．1）に対しては，単一形式データを用いる手法に対して，複合データの情報を加える方針，2）に対しては，複数のデータ形式を利用して，課題に適したデータを用いる方針の二つに基づく手法を提案する．

1）上述の様に，画像データは，画像の単一形式というより複合データの形式でウェブ上でやり取りされる．つまり，画像データはユーザによる説明目的のキーワード（ユーザラベルと呼ぶ）が付与されている [13]．ユーザラベルは，予め決まったキーワードから選ばれる形式ではなく，ユーザにより自由に記述される．そのために，一部のユーザラベルは曖昧性を持つ．例えば，図 1 は，画像共有サイト Flickr[*3]に投稿された写真である．図中の両方の写真に“albatross”というユーザラベルが付与されている．“albatross”は，左の写真では鳥のアホウドリを示し，右の写真では，飛行機のモデル名を示す．ユーザラベルの曖昧性により，同一検索キーワードに対する検索結果に，異なる画像が混在してしまう．

本研究では，ユーザラベルを百科事典の項目に対応づけることにより，曖昧性を解消をすることを目指す．この曖昧性解消を，画像とユーザラベルを入力として受け付けて，ユーザラベルに対応するウィキペディアの記事を特定する課題（画像ラベル曖昧性解消）として取り組む．この課題に対して，ユーザラベルと，画像情報を画像分類ラベルとして取り出し，両情報から対応項目を特定する手法を提案する．評価実験では提案手法は逆順位の尺度で 0.6 を示し，従来の単一形式手法より高い値となった．

[*3] https://www.flickr.com

2）画像へのメタデータ付与として，画像を説明するキーワードを付与する画像アノテーション課題 [12] に取り組む．本課題の概要を図 4 に例示する．図中のハイキングをしているグループについての入力画像に，“tourist”，“slope”，“mountain”というキーワードの付与を行う．付与されたキーワードによって，画像への効率的な参照が可能となる．

多くの既存手法は，画像とキーワード群の組みを訓練データとする教師有り学習手法である．訓練データに用いられる正解キーワードは人手によって付与されるため，キーワードの揺れが起こりやすい．例えば，人が写る画像へのキーワードであっても，観光地で記念に撮影された画像では「観光客」，スポーツを楽しむ人たちについての画像であれば「プレイヤー」が付与されたり，より一般的な「女性」が付与される場合もある．これは訓練データの品質ではなく，画像を説明するキーワードを付与するという人の主観性に依存する当課題の性質に由来する．訓練データ中のキーワードの一貫性が低くなり，教師有り学習手法の学習に悪影響を及ぼす．

この問題に対して，画像をテキスト特徴量に変換する手法を提案する．提案手法では，画像とキーワードの百科事典文章の組みを教師データとするニューラルネットワークを定義し，変換器の訓練を行う．この変換器により，上述の「観光客」の画像から「女性」の画像と共通する文脈情報を持つ特徴量（例えば，「人に関係する」）を学習することが可能になる．このように変換したデータとキーワードの組みを訓練データとして，キーワード付与器を学習する．評価実験では，画像アノテーション課題で，提案手法による変換データを用いた分類では既存手法と比較して，アノテーション性能を F 値で約 0.1 向上した．

これ以降の本論文の構成は次の通りである．第 2 節では，上述の二つの研究課題の関連研究を紹介する．第 3 節では，画像ラベル曖昧性解消についての提案手法と評価実験について説明する [14, 18]．第 4 節では，画像アノテーション課題についての提案手法と評価実験について説明する [15, 19]．第 5 節では，考察及び今後の研究についてをまとめる．

2. 関連研究

本節では，画像ラベル曖昧性解消と適用可能な手法及び，画像アノテーションに対する既存手法を紹介する．

2.1 画像ラベル曖昧性解消の関連研究

画像ラベル曖昧性解消課題に対して，入力画像を画像分類することによる曖昧性解消が考えられる．近年では，畳み込みニューラルネットワークを用いた画像分類 [6] の研究が盛んにされており，高い分類性能を達成している．ただ画像分類の本課題へ適用する際には，次の様な制約がある．様々な視覚的な観点があるラベルの曖昧性解消は困難である．例えば，図 1 の「ガラパゴス諸島」では，島の砂浜や島全体の鳥瞰図などが考えられる．次に，教師有り学習による分類器の構築には，大量の訓練データが必要になる．同図で，航空機の“Grumman HU-16 Albatross”といった詳細なモデルを表す百科事典項目を網羅する，十分な量の画像データを用意することは現実的に困難である．

ユーザラベルのテキストを対象にした曖昧性解消の手法として考えられる．文章中の特定の単語が指し表す百科事典項目を特定する課題（ウィキフィケーション）に対して，様々な手法が提案されている．記事中の他ページへのリンク，アンカーテキストに基づいて，候補への類似度計算手法 [2] や，ウィキペディア全体のネットワーク構造を用いる手法 [11] が提案されている．また言語的な特徴である，照応関係と固有表現を用いる手法 [1] なども提案されている．これらの研究は，新聞記事のような，英単語で数百単語からなる文章を対象にしている．しかしながら本研究が対象にする画像ラベルは，一つの画像につき，数単語であり文脈情報が限られる．

2.2 画像アノテーションの関連研究

画像アノテーション課題は，ImageCLEF ワークショップ [16] を中心に研究が行われてきた．本節では比較的新しい手法を取り上げて紹介するが，既存手法は画像の単一データを対象にしており，複合データを対象としない．近傍画像からキーワードを特定する検索として，一般的に疎なベクトルとなる画像特徴量をクラスタリングによってグループ化する手法 [17] や，色の分布やテクスチャによる特徴量を距離に用いる手法 [8] が提案されている．最近では，深層学習を用いる半教師有り手法も提案されている [7]．この手法では畳み込みニューラルネット画像分類器と画像の低次元特徴量を用いた分類器（SVM）の二つの分類器を訓練することにより，輪郭などの高次元の特徴と低次元の特徴を組み合わせる．

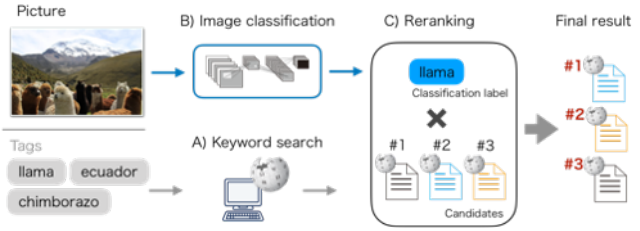

図 3 提案手法の流れ
Fig. 3 Workflow of the proposed method.

3. 画像ラベル曖昧性解消

第 2.1 節で紹介した従来手法は大きく画像のみやユーザラベルのみを利用する方策であったが，どちらの方策にも欠点があった．画像とユーザラベル両方を用いることにより，従来手法の欠点を補い有効な曖昧性解消が行える．画像情報として詳細な物体を網羅する情報ではなくても，予め定義できる情報でも少ないユーザラベルの情報を補える．例えば，図 1 のユーザラベル曖昧性解消では，左の画像では「鳥」という画像情報は，“albatross”がアホウドリへ対応する，右の画像では「飛行機」が同ユーザラベルが飛行機のモデル名“Grumman HU-16 Albatross”に対応するヒントになる．

画像情報を活用するために，画像分類器を用いて分類キーワード（画像ラベル）として取り出す．ユーザラベルと画像ラベルを組み合わせた類似度計算により，対応するウィキペディア記事を特定する．本節では，画像ラベル曖昧性解消に対して，ユーザラベルと画像の両方を活用して百科事典項目を特定する手法を提案する．評価実験では，提案手法の性能評価を行う．

3.1 ユーザラベルと画像ラベルを用いた曖昧性解消手法

提案手法は，大きく二段階の検索により，ユーザラベルに対応するウィキペディア記事の特定を行う．図 3 に提案手法の流れを示す．同図 A) で，初めのキーワード検索により候補記事の取得する．B) では，入力画像について画像分類を行い，画像ラベルの取得を行う．二回目の検索 C) では，ユーザラベルと画像ラベルを用いて類似度計算によって候補記事の順位づけを行う．以下，各過程について説明する．

A) キーワード検索では，曖昧性解消の対象ユーザラベル（クエリ）に対して，ウィキペディア記事候補を取得する．例えば，図 3 でのクエリは“llama”などである．キーワード検索には，Apache Solr[*4]検索エンジンを用いて実装し，順位づけには BM25 を用いた．この検索で得られた全ての候補記事を C) の再順位づけに用いる．

B) 入力画像を画像分類することにより，画像ラベルを

[*4] http://lucene.apache.org/solr/

取得する．取得する画像ラベルについて二つの方針を設定し，それぞれの画像分類器を構築する．B-1) 一つ目の方針は，画像から 1 件の主要な画像ラベルに分類する．画像ラベルは，対象とする画像群が既知であるとしてデータセットのカテゴリを用いる（カテゴリ分類器）．例えば，図 3 の画像では，カテゴリである“llama”ラベルを取得する．この方針では画像ラベル数が比較的少ないため，高い分類性能が期待できる．B-2) 二つ目の方針は，画像中の情報をなるべく多くの画像ラベルに分類する（多ラベル分類器）．例えば，図 3 で，カテゴリ画像ラベル“llama”に加えて“alp”（山脈）ラベルが得られれば，“chimborazo”（アンデス山脈の火山) クエリの対応記事特定に役立つ．この分類器では，一枚の入力画像に対して複数の画像ラベルを取得する．

B-1) カテゴリ分類器には，畳み込みニューラルネット Resnet [6] を用いて分類器を構築する．カテゴリの画像ラベルに分類するために，評価実験に用いるテストコレクションのカテゴリについて分類器を訓練する．大規模画像認識タスク ILSVRC [12] に定められる 1,000 クラスについて事前学習済みの Resnet モデルを用いて，転移学習を行う．具体的には，ネットワークの最終層の全結合層を，カテゴリ数をサイズとする層で置き換えて訓練を行う．二つのコレクションカテゴリの訓練データを別途収集して分類器の訓練を行う．

B-2) 多ラベル分類器には，ILSVRC タスクの 1,000 クラスについての訓練された分類器を用いる．これらのクラスは，クラスの階層関係が整理された画像データベース ImageNet[*5]の物体クラスから，あるクラスがその他のクラスの祖先とならない様に選択されている．このクラス体系には動物名やスポーツ，ガジェットなどが含まれる．これらの画像ラベルは，様々な状況を表す画像のユーザラベルに役立つと考えられる．多ラベル分類器の実装には，ImageNet の合計 120 万件の画像について訓練がされる Resnet を用いた学習済み分類器を用いた．多ラベル分類器は，画像について複数ラベルを出力するが，予測確率が高い上位のラベルを画像ラベルとして用いる．B-1,2) 両分類器の訓練詳細は第 3.2 節にまとめる．

C) キーワード検索 A) による候補ウィキペディア記事を，ユーザラベルと画像分類ラベル B) を用いた類似度計算により再度順位付けを行う．キーワード検索過程での候補記事の順位及び類似度は用いない．再順位付けでは，ユーザラベル（図 3 で，“llama”，“ecuador”など）と画像ラベル（同図で，“llama”）の単語と，候補記事の単語間の分散表現 [9] における類似度を計算する．

分散表現では，単語の周辺の文脈に基づき近い文脈を持つ単語ほど類似するベクトルとなるように学習がされる．単語ベクトルによる類似度計算によって，第 3.1 節 B) で述べた限られた画像ラベルの欠点を補うことが出来る．例えば，図 3 の画像について，B-2) の画像ラベル“alp”が得られていても，単語マッチ検索では火山“chimborazo”記事との間で語彙が異なり類似度計算が出来ないが，単語ベクトルを用いることにより，この問題を解決できる．

曖昧性解消の対象であるクエリラベルに 1 件について候補ウィキペディア記事の集合 T とユーザラベルの単語集合 Q と画像ラベルの単語集合 L から次の類似度を計算することにより，再順位付けを行う．

$$\mathrm{sim}(\mathrm{T},\mathrm{Q}) = \alpha \sum_{w_q \in Q} \mathrm{sim}(\mathrm{T},\mathrm{w_q}) + (1-\alpha) \sum_{w_l \in L} \mathrm{sim}(\mathrm{T},\mathrm{w_l}), \quad (1)$$

ここで，$\mathrm{w_q}$ と $\mathrm{w_l}$ はそれぞれユーザラベルの単語と画像ラベルの単語を表し，α は，ユーザラベル部と画像ラベル部の類似度の重みづけパラメータである．ウィキペディア記事の集合 T は，事前実験から記事タイトル部の単語集合を用いた．各項の計算を次に説明する．

$$\mathrm{sim}(\mathrm{T},\mathrm{w_l}) = \frac{1}{|T|} \sum_{w_l \in T} \mathrm{sim}(\mathrm{w_t},\mathrm{W_l}) \times \mathrm{score}(\mathrm{w_l}), \quad (2)$$

$\mathrm{w_t}$ はウィキペディア記事の単語を表し，$\mathrm{score}(\mathrm{w_l})$ は画像分類器のソフトマックス関数によって出力される画像ラベルの確率である．これは分類により得られる画像ラベルが必ずしも正しいわけでないので，確率により類似度の重み付けを行う．式 (1) の左辺のユーザラベルの類似度も同様に計算するが，式 (2) の重み付けはない．単語間の類似度は，単語ベクトルのコサインによって計算する．

$$\mathrm{sim}(\mathrm{w_1},\mathrm{w_2}) = \cos(\mathbb{w}_1,\mathbb{w}_2) = \frac{\mathbb{w}_1 \mathbb{w}_2}{|\mathbb{w}_1||\mathbb{w}_2|},$$

2 つの単語 w のベクトル表現 $\mathbb{w}$ の内積によって計算する．

3.2 評価実験

評価実験では，既存手法及び提案手法の性能を比較評価する．評価には ImageCLEF コレクション [16] と動物名コレクション（詳細は [14, 17] ）を用いた．表 1 にテストコレクションの基本統計量を示す．ImageCLEF では 1,197 枚の画像に対する 10,573 件のユーザラベル，動物名コレクションでは 450 枚の画像に対する 2,280 件のユーザラベルを曖昧性解消の対象とする．同表中の対応記事数はユーザラベルに対応するウィキペディア記事の異なり数を表す．

比較手法には提案手法を含めた六つの手法を評価した．キーワード検索は，第 3.1 節 A) に対応する手法である．ウィキフィケーション手法は，テキストを対象にしたウィキフィケーション [2] をユーザラベルに適用した手法である．`word2vec` 手法は，第 3.1 節 C) でユーザラベルだけ

[*5] https://www.image-net.org/

表 1　テストコレクションの統計量
Table 1　Statistics of two test collections used in the experiments.

テストコレクション	画像数	ユーザラベル数	対応記事数
ImageCLEF	1,197	10,573	904
動物名	450	2,280	207

を用いて再順位付けを行う手法である．つまり式 (1) 中の右辺第二項の類似度を用いない．以上がユーザラベルのみをウィキペディアへの対応付けに用いる手法である．画像検索手法は，対象の入力画像とウィキペディア記事の見出し画像の類似度を計算することにより順位付けを行う．第 3.1 節 B-2) 多ラベル分類器の最終層を画像の特徴ベクトルとして取り出し [10]，入力画像と記事の画像のベクトルのコサイン類似度を計算する．

提案手法 1，2 はそれぞれ，第 3.1 節 B-1) と B-2) に対応する．手法 1 のカテゴリ分類器の訓練には，上述の 2 つのコレクションの各カテゴリに対して，ImageNet から収集した 300〜600 件の画像を用いた．分類器の訓練には，収集データの 7 割を訓練用，3 割を検証用に分割して，訓練を行った．ネットワークの訓練設定は，ミニバッチサイズを 32，最適化関数には確率的勾配降下法を用いた．初期学習率，荷重減衰，モーメンタムは，それぞれ 10^{-4}，10^{-6}，0.9 とした．手法 2 の出力ラベル数は，それぞれのコレクションの平均ユーザラベル数に近い，動物名コレクションで 5 件，ImageCLEF で 8 件を用いた．

評価尺度には，順位付きリストの評価尺度である平均逆順位 (MRR) を用いる．正解ウィキペディア記事が上位にあるほど，1 に近い高い値となる．また順位付きリストの上位 1 件または 10 件に正解があれば 1 を取る尺度，上位 1，10 件の再現率（R@1，R@10）も計測に用いる．

実験結果を表 2，3 に示す．表中の各数値はマクロ平均であり，$*$ 印はキーワード検索手法と出力を比較して，1% 水準の有意差があったことを示す．MRR では，動物名コレクションで提案手法 1 が 0.609，ImageCLEF コレクションで提案手法 2 が 0.719 を示している．提案手法が各コレクションで一番良い性能を示している．またこれらの値は，キーワード検索手法よりも高い値である．`word2vec` 手法は，二つのコレクションで二番目に高い性能を示している．特に，表 3 の MRR で 0.714 を示し，手法 2 の 0.719 に近い値である．ユーザラベルのみを用いる三手法（キーワード検索，ウィキフィケーション，`word2vec`）は，同一のクエリに対して特定の記事を高い順位で出力していた．例えば，“jaguar”がクエリの場合では動物の「ヒョウ」の記事を 1 位に，自動車メーカー「ジャガー」の記事を 50 位前後に出力していた．

表 2　動物名コレクションの実験結果
Table 2　Results of the ILW methods on the Animal Name collection.

手法	MRR	R@1	R@10
キーワード検索	0.509	0.471	0.577
ウィキフィケーション	0.523	0.481	0.601
`Word2vec`	0.526	0.495	0.581
画像検索手法	0.075	0.037	0.158
提案手法 1	*0.609	0.558	0.684
提案手法 2	*0.583	0.518	0.679

表 3　ImageCLEF コレクションの実験結果
Table 3　Results of the ILW methods on the ImageCLEF collection.

手法	MRR	R@1	R@10
キーワード検索	0.627	0.595	0.706
ウィキフィケーション	0.628	0.595	0.708
`Word2vec`	0.714	0.711	0.717
画像検索手法	0.209	0.130	0.384
提案手法 1	*0.715	0.711	0.718
提案手法 2	*0.719	0.711	0.730

4. 画像アノテーション課題

本節では，画像アノテーション課題に対する画像データからテキストデータへのデータ変換，及び変換データを利用した分類手法を提案する．図 4 はデータ変換による画像アノテーションの概要図である．第 2.2 で紹介した画像とキーワードを組みを訓練データに用いる教師有り学習手法は，訓練データ中のキーワードの一貫性の低さがアノテーションの課題であった．例えば，図 4 中の右の画像では，上の画像には“tourist”のキーワードが付与され，下の画像には“woman”が付与されている．上の画像の観光客の団体に女性が含まれているが，“woman”キーワードはない．この様なキーワード揺れにより類似するキーワードに関する学習が難しくなる．

訓練データの一貫性が低い問題に対して，提案手法では，図 4 下段に示す様に，画像をキーワードに対応するウィキペディアのテキストデータに変換を行ことにより，類似するキーワードの特徴量を共有を出来る様にする．図では，“tourist”を持つ画像 1 と“woman”を持つ画像 2 の変換データが“people”という特徴量を共有する．

4.1 データ変換による画像アノテーション手法

図 5 に提案手法の流れを示す．初めに，データ変換器の画像と変換テキストの組みからなる訓練データを用意する．同図 A) で，畳み込みニューラルネットワーク (CNN) を用いた画像分類器から画像の特徴量を取り出す．同図 B) で，

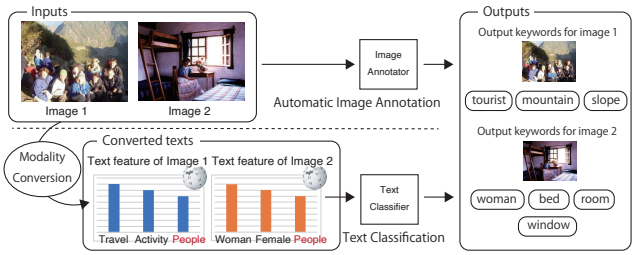

図 4 データ変換による画像アノテーションの概要
Fig. 4 Overview of the automatic image annotation by data converting.

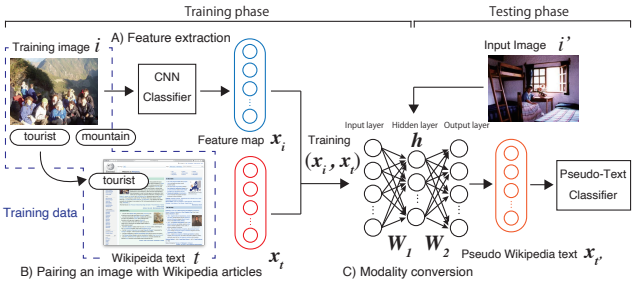

図 5 データ変換手法の流れ
Fig. 5 Workflow of the data converting method.

画像と変換対象となるウィキペディア記事との対応づけを行う．次に，同図 C) ではこれらを訓練データとするデータ変換器を構成する．最後に，変換データ（擬似テキストと呼ぶ）を用いた分類器により，キーワード付与を行う．

4.1.1 画像からの特徴抽出

図 5 A) では，CNN を用いた画像分類器から特徴量を取り出す．画像 i が分類器に入力として与えられた時，ニューラルネットワークの最終層の特徴マップ $\boldsymbol{x_i} \in \mathbb{R}^d$ を画像の特徴量とする [10]．

本手法での分類器には，CNN の ResNet [6] を用いる．ネットワークは，224 × 224 ピクセルの画像の入力層として，畳み込み層，活性化，バッチ正規化層から構成される複数の層をブロックとして，ブロックを複数回繰り返して，プーリング層と全結合層というネットワーク構成をとる．各ブロックは，その前の層の出力と共に，前のブロックの出力（ショートカット）の 2 つを入力にとる．前ブロックの入力を明示的に入力とするショートカット構造は，各ブロックに対応する潜在的な写像を学習するように設計される．

ResNet の実装には，深層学習フレームワーク Keras の実装[*6]を用いた．ILSVRC 画像認識課題の 1,000 クラス [12] について 100 万枚以上の画像から学習されたモデルをネットワークの重みの初期値として用いた．ResNet の最後の層に，評価用データセットのキーワード数（第 4.2 節参照）のサイズの全結合層を付け加えてた．最適化には，ミニバッチサイズ 32 で，確率的勾配降下法を用いた．初期学習率を 0.1，重み減衰値を 10^{-4}，モーメンタムを 0.9 とした．

[*6] https://keras.io/application/#resnet

4.1.2 データ変換器

図 5 B) 及び C) についての画像とウィキペディアテキストの組みの訓練データとデータ変換器の構成を説明する．変換の対象にするテキストデータとして，キーワードの特徴を反映させるために，内容が充実する百科事典サイト，ウィキペディア [3] 記事のテキストを用いた．

ウィキペディア記事を変換対象として用いるには，キーワードに対応する記事が存在することが必要である．図 5 B) の例では，“tourist” キーワードを見出しとするウィキペディア記事である．第 4.2 節の評価実験で用いるテストコレクションについては，全キーワードについての見出し語があることを確認した．キーワードに対して複数の見出し語がある（ウィキペディア中で，曖昧なページ）場合は，人手で対応関係を判断した．これらの自動化には，対応関係特定の手法 [14] の適用が考えられる．キーワードと対応するウィキペディア記事の統計は，表 4 にまとめる．

ここまでに，画像とウィキペディア記事の組みからなる訓練データが用意できた．次に，画像からテキスト特徴量（擬似テキスト）への変換を学習するデータ変換器を説明する．次元削減などの教師無し課題に用いられる自己符号化器 [4] に着想を得て，本手法のデータ変換器を構成する．類似する構成により画像中の特徴をテキストに対応させることを目指す．データ変換器の構成を図 5 C) に示す．入力層，隠れ層，出力層からなる 3 層のニューラルネットワークを用いる．変換器では，画像特徴ベクトル $\boldsymbol{x_i}$ を擬似テキストベクトル $\boldsymbol{x'_t} \in \mathbb{R}^l$ への変換を学習する．

各層のニューロン数は，入力層のサイズは画像特徴の次元数 n，隠れ層のサイズ m，出力層のサイズ l とする．入力層と隠れ層かなる入力部では，次の関数 f によって，x_i を低次元の潜在表現ベクトル $\boldsymbol{h} \in \mathbb{R}^l$ に変換する．

$$\boldsymbol{h} = \mathrm{f}(\boldsymbol{i}) = \mathrm{ReLU}(\boldsymbol{W_1} \cdot \boldsymbol{x_i} + \boldsymbol{b_1}),$$

ここで，$\mathrm{ReLU}()$ はランプ関数を表し，$\mathrm{ReLU}(x) = \max(0, x)$ となる．$\boldsymbol{W_1}$，$\boldsymbol{b_1}$ は入力部のパラメータであり，それぞれ，重み行列 $\boldsymbol{W_1}$，バイアス項 $\boldsymbol{b_1} \in \mathbb{R}^m$ である．出力部では，潜在表現ベクトル $\boldsymbol{h}$ を関数 g によりテキストベクトル $\boldsymbol{x'_t}$ に写像する．

$$\boldsymbol{x'_t} = \mathrm{g}(\boldsymbol{h}) = \mathrm{sigmoid}(\boldsymbol{W_2} \cdot \boldsymbol{h} + \boldsymbol{b_2}),$$

ここで，sigmoid はシグモイド関数を表し，$\boldsymbol{W_2} \in \mathbb{R}^{(m \times n)}$ は重み行列，$\boldsymbol{b_2} \in \mathbb{R}^n$ はバイアス項であり，出力部のパラメータとなる．データ変換器は，入力 $\boldsymbol{x_i}$ と出力 $\boldsymbol{x'_t}$ の変換誤差を最小化するように学習がされる．変換誤差は，二乗誤差を変換誤差として，損失関数 $\mathcal{L}$ を次式とする．

$$\mathcal{L} = ||\boldsymbol{x_i} - \boldsymbol{x'_t}||,$$

損失関数は確率的勾配降下法を用いて最適化を行う．

4.1.3 分類手法

第 4.1.2 節で変換した擬似テキストを特徴量とするキーワード分類器を訓練する．入力画像擬 i' から変換された擬似テキスト t' が画像像説明文中のキーワード $k \in K$ が割り当てられる確率 $P(k|t')$ としてモデル化する．一つの入力に対して，複数のキーワードが付与されるため，多ラベル分類問題として扱う．出力層に，softmax 関数を用いる三層のニューラルネットワークを用いた．

$$P(k|t') = \mathrm{softmax}(\boldsymbol{W_k} \cdot \boldsymbol{h_{t'}} + \boldsymbol{b_k}),$$
$$\boldsymbol{h_{t'}} = \mathrm{sigmoid}(H_{x_{t'}} + \boldsymbol{b_m}),$$

$\boldsymbol{x'_t} \in \mathbb{R}^d$ は擬似テキスト t' の特徴ベクトル，行列 $H_{x'_t} \in \mathbb{R}^{d \times h}$，$\boldsymbol{w_k} \in \mathbb{R}^d$，バイアス項 $\boldsymbol{b_n} \in \mathbb{R}^h$，$b_k \in \mathbb{R}$ はモデルパラメータである．隠れ層のサイズは $h = 1000$ とした．

4.2 評価実験

評価実験では提案手法と既存手法について，画像アノテーション性能を比較評価する．

4.2.1 実験設定

本実験では，画像アノテーション課題評価用の IAPR TC-12 [5] と ImageCLEF [16] の二つのテストコレクションを用いた．二つのコレクションともに，図 4 に示す様な人，動物，風景，街，スポーツなど様々な場面の画像を含む．IAPR TC-12 は，20,000 画像からなり，各画像は，英独西の三言語による画像説明文が付与される．英語の説明文に品詞解析を行い，低頻度語などを除外した名詞をキーワードとした．291 のキーワードとそれらが付与された 19,008 画像を評価に用いた．ImageCLEF では，7,291 画像のうち，人手でキーワードが付与された 3,124 画像を評価に用いた．このコレクションにおけるキーワード数は 207 である．表 4 には，両コレクションのキーワードに対応するウィキペディア記事の統計量を示す．

評価実験では，次の三つの手法を評価した．

画像分類手法 第 4.1.1 節で説明した CNN の画像分類器 [6] を用いる手法．

テキスト分類手法 画像のキーワードに対応するウィキペディア記事のテキストデータ（第 4.1.2 節に対応）の特徴量を分類に用いる．特徴量はウィキペディア記事中の単語の頻度とする．特徴ベクトルの次元数は表 4 の語彙数とする．本研究では，画像からテキストへ置き換えることを目指すため，この手法を上限手法として位置付ける．

擬似テキスト分類手法 提案手法であるデータ変換器による擬似的なテキストデータを特徴量に用いる手法である．特徴ベクトルの次元数は表 4 の語彙数とする．

画像アノテーションの評価には，各画像につき手法の出力する分類スコアの高い上位 5 件のキーワードを出力とした精度，再現率とそれらの調和平均による F 値を用いる [8]．また少なくとも一枚の画像を正しくキーワード付与できたキーワードの数も評価に用いる．精度，再現率の計算は次の通りである．

$$\text{精度} = \frac{\text{分類が正解した画像数}}{\text{手法が出力した画像数}},$$
$$\text{再現率} = \frac{\text{分類が正解した画像数}}{\text{データセット中の画像数}}.$$

4.2.2 実験結果

表 5 と表 6 に各コレクションでの実験結果を示す．表中の $*$ 印は画像分類手法との比較で，出力に 1%水準の有意差が確認出来たことを表す．二つのコレクションの結果で，テキスト分類手法が一番高い性能示した．IAPR TC-12 と ImageCLEF の F 値で 0.816 である．提案手法は，付与性能では二番目の結果となった．F 値で，IAPR TC-12 で 0.447，ImageCLEF で 0.623 を示した．提案手法では画像分類手法と比較すると高い性能である．ただテキスト分類手法と比較すると，付与性能に開きがある．

次に，キーワードをグループに分けることにより実験結果の分析を行う．キーワードを 5 つのグループ（物と動物，場所，人，建造物，その他）に分けて，それぞれのグループにおける性能を調査する．図 6 に IAPR TC-12 のグループ毎の F 値を示す．提案手法と画像分類の比較において，最も F 値の向上が高かったグループは，人と建造物であり，向上幅はそれぞれ，0.12 と 0.14 であった．例えば，

表 4 キーワードと対応するウィキペディア記事の統計
Table 4 Statistics of Wikipedia articles with titles of annotation keywords.

テストコレクション	キーワード	平均キーワード	語彙
IAPR TC-12	291	2,682	37,230
ImageCLEF	207	8,403	41,170

表 5 IAPR TC-12 コレクションの実験結果
Table 5 Results of the AIA methods on the IAPR TC-12 collection.

手法	精度	再現率	F1	キーワード数
画像分類手法	0.514	0.325	0.398	207
テキスト分類手法	0.774	0.861	*0.816	275
擬似テキスト分類手法	0.537	0.383	*0.447	260

表 6 ImageCLEF コレクションの実験結果
Table 6 Results of the AIA methods on the ImageCLEF collection.

手法	精度	再現率	F1	キーワード数
画像分類手法	0.674	0.385	0.490	81
テキスト分類手法	0.723	0.936	*0.816	101
擬似テキスト分類手法	0.678	0.577	*0.623	101

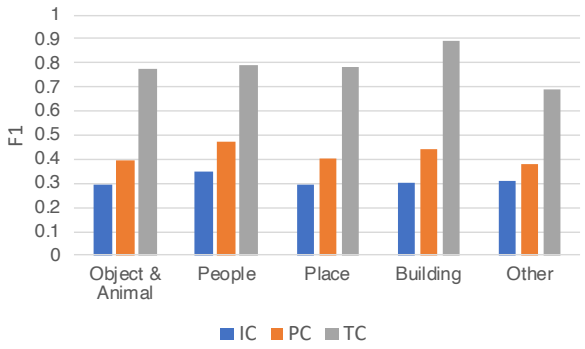

図 6 IAPR TC-12 キーワードのグループ毎の F 値.
IC：画像分類，PC：擬似テキスト分類，
TC：テキスト分類.

Fig. 6 F1 on each group of annotation keywords on IAPRTC-12.

人グループのキーワードは，画像のシーンによって変わりやすい傾向がある．例に用いてきた“torist”，“woman”や“cyclist”などである．提案手法が揺れが起こりやすいキーワードに対して，キーワード付与が成功していた．反対に，失敗事例としては，画像分類と比較して性能向上が比較的小さかった，物と動物グループが挙げられる．このグループのキーワードは，例えば“bed”や“lion”など，キーワードの物体が写っている訓練データ中の画像に対して一貫性しているものが多い．

5. おわりに

本論文では複合データへのメタデータ付与の具体例として，画像メタデータの曖昧性解消と画像へのメタデータ付与の二つの研究課題に取り組んだ．曖昧性解消では付与されたメタデータと画像分類により取得した画像情報を組み合わせた手法を，画像へのメタデータ付与では，画像からテキスト特徴量への変換手法を提案した．各課題の評価実験では，従来の単一形式データを用いる従来手法と比較して，提案手法の有効性を確認した．

本研究では，複合データの活用方針として，データを加える方針（テキストと画像の利用）と複数様式を利用したデータを変更する方針（画像からテキスト）を検討した．今後の課題の一つとして複合データの利用方針や手法の検討が挙げられる．

参考文献

[1] Cheng, X. and Roth, D.: Relational inference for wikification, *Proceedings of the 2013 Conference on Emprical Methods in Natural Language Processing*, pp. 1787–1796, (2013).
[2] Cucerzan, S.: Large-scale named entity disambiguation based on wikipedia data, *Proceedings of the 2007 Joint Conference on Emprical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 708–716, (2007).
[3] Giles, J.: Internet encyclopedias go head to head, *Nature*, 438, 900–911, (2005).
[4] Goodfellow, I., Bengio, Y. and Courville, A.: *Deep Learning*, Cambridge MA, USA (2016).
[5] Grubinger, M. *et al.*: The iaprtc-12 benchmark: a new evaluation resource for visual information systems, *Proceedings of the International Workshop OntoImage 2006 Language Resources for Content-Based Image Retrieval*, pp. 13–23, (2006).
[6] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, (2016).
[7] Li, Z. *et al.*: Collaborating cnn and svm for automatic image annotation, *Proceedings of the 2019 ACM International Conference on Multimedia Retrieval*, pp. 63–67, (2019).
[8] Makadia, A., Pavlovic, V. and Kumar, S.: Baselines for image annotation, *Int. J. Comput. Vis.*, 90(1), 88–105, (2010).
[9] Mikolov, T. *et al.*: Distributed representations of words and phrases and their compositionality, *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pp. 3111–3119, (2013).
[10] Ng, J. Y.-H., Yang, F. and Davis, L. S.: Exploiting local features from deep networks for image retrieval, *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 53–61, (2015).
[11] Ratinov, L., Roth, D., Downey, D. and Anderson, M.: Local and global algorithms for disambiguation to wikipedia, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1375–1384, (2011).
[12] Russakovsky, O. *et al.*: ImageNet large scale visual recognition challenge, *Int. J. Comput. Vis.*, 115(3), 211–252, (2015).
[13] Sawant, N., Li, J. and Wang, J. Z.: Automatic image semantic interpretation using social action and tagging data, *Multimedia Tools and Applications*, 51(1), 213–246, (2011).
[14] Suzuki, T., Ikeda, D., Galuščáková, P. and Oard, D.: Towards automatic cataloging of image and textual collections with wikipedia, *Proceedings of the 21st International Conference on Asia-Pacific Digital Libraries*, pp. 167–180, (2019).
[15] Suzuki, T. and Ikeda, D.: A modality converting approach for image annotation to overcome the inconsistent labels in training data, *Proceedins of International Workshop on Content-Based Image Retrieval in conjunction with the 25th International Conference on Pattern Recognition*, 8 pages, (2021).
[16] Villegas, M. and Paredes, R.: Overview of the imageclef 2014 scalable concept image annotation task, *Working Notes of CLEF 2014 - Coference and Labs of the Evaluation Forum*, pp. 308–328, (2014).
[17] Zhang, S. *et al.*: Automatic image annotation using group sparsity, *Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3312–3319, (2010).
[18] 鈴木釈規，池田大輔：画像認識を用いた画像ラベルの知識体系への対応付手法の構築とその展望，情報処理学会九州支部 2018 年度若手の会セミナー論文集，6 頁，(2018).
[19] 鈴木釈規，池田大輔：画像アノテーション課題からテキスト分類課題へ〜深層学習を用いたモダリティ変換の試み，火の国情報シンポジウム 2020 論文集，6 頁，(2020).