様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 ２５ 日現在

機関番号：１４２０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2023
課題番号：２０Ｋ０６８２４

研究課題名（和文）画像解析と群集メタバーコーディングによる統合的生物多様性モニタリング法の開発

研究課題名（英文）An integrative biodiversity monitoring with computer vision and community DNA metabarcoding

研究代表者
藤澤　知親（FUJISAWA, Tomochika）

滋賀大学・データサイエンス学系・助教

研究者番号：１０７９２５２５

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：本研究はDNAメタバーコーディングと画像解析技術を組み合わせた昆虫群集のモニタリング手法をを開発することを目的として実施した。深層学習による画像解析では土壌コアサンプルの昆虫画像の分類をおこなった。画像の品質や撮影方法・トレーニングデータの不均一さによる分類性能の低下を防ぐ方法の実装と性能検証を行った。またDNA配列データから種レベルの同定を行うための深層学習モデルを実装し、既知の種の分類と未知の種を検知する手法の性能を評価した。その結果深層学習モデルの配列データの同定は極めて正確であることがわかったが、一方で未知の種の発見は既知種の同定より困難な状況があることなどがわかった。

研究成果の学術的意義や社会的意義
昆虫群集は生態系の健全な機能を維持するために重要な働きを担っていると考えられている。しかし昆虫群集のモニタリング調査は専門知識を持った人材の不足などにより大規模に行うことが難しい。機械学習による画像解析やDNA分類はモニタリング調査を簡便に行うための有用な手法と考えられている。本研究では深層学習モデルによる画像データ・DNAデータの分類の精度評価に加え、機械学習モデルの既知の問題点に対する対処方法を探った。特にデータベースの不均一さや不完全さといった分類モデルの性能低下につながる問題に対する方法の実用性を検証した点が主要な学術的な意義である。

研究成果の概要（英文）：We developed monitoring methods for insect communities using computer vision and DNA metabarcoding in this project. We classified insect images taken from soil core samples with a deep learning model and evaluated its performance. We also implemented and tested methods to alleviate performance reductions due to heterogeneity of training databases. In addition to the image analyses, we developed a deep learning model for identification of insects with DNA barcoding fragments, and evaluated its performance for identification of the known species as well as detection of the unknown species. The model correctly identified the known species, but the detection of the unknowns was more difficult in some conditions typical of metabarcoding studies.

研究分野：生物情報学

キーワード：機械学習　メタバーコーディング　画像解析

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

生物群集の詳細な構成とその時間変化を知ることによって、我々は生態系を理解しその健全な機能維持を行うことができる。しかし、生物群集の大規模かつ詳細なモニタリングは専門知識をもつ人員の不足などによりこれまでは容易ではなかった。とりわけ昆虫群集のモニタリングは膨大な多様性によりその傾向が顕著だった。

DNA 配列決定技術の効率化と DNA を用いた生物分類の統計手法の発展にともなって、この状況は変わりつつある。群集メタバーコーディングや環境 DNA 技術が実用化されるのにともない、昆虫群集の種構成や遺伝的多様性を DNA により大規模かつ詳細ににモニタリングする可能性が開けてきた。

一方、DNA による生物のモニタリング法のもっとも大きな問題点はサンプルの形態や個体数の情報が失われることである。この問題点を解決する 1 つのアプローチが画像情報をもちいて DNA 配列の情報を補うことである。画像解析技術を用いた生物の同定は近年多くの研究が行われており、特に深層学習のパッケージが広く普及したことにより、いくつかの研究はすでに成功を収めている(Valen et al. 2019)。またハイスループット撮影技術と組み合わされた画像分類モデル(例えば Wührl et al. 2021)は配列ベースのモニタリングと容易に組み合わせられ、その欠点を補完することができる。

画像解析を広く生物のモニタリングに利用するためには解決が必要な点が複数存在する。例えば多くの画像分類モデルは訓練を行った画像と異なる条件下で撮影された画像を正確に分類できないことが知られている(例えば博物館標本のデータで訓練し野外で撮影された写真を分類する状況 Knyshov et al. 2021)。また多くの場合訓練用データベースは不完全であり、対象の分類群を含んでいない状況や含んでいても下位分類群の構成が異なる場合などが存在する。機械学習モデルはこのような不完全かつ不均一な訓練データをもちいた状況下でも正確な生物同定ができる必要があるが、その性能は詳しく調べられていない。

２．研究の目的

機械学習の分類モデルをもちいて画像と DNA 配列による生物同定の手法を開発し、実データで性能を評価する。特に生物多様性モニタリングへの応用に必要な既存モデルの問題点（特に訓練データの不均一性・不完全さ）の影響を調べ、それらを解消する方法を検討する。

３．研究の方法

1)昆虫画像の分類モデル

Valan et al.(2019)が提唱した深層転移学習モデルを用いた昆虫画像の分類をおこなった。共同研究者から提供された土壌コアサンプルの昆虫群集の顕微鏡写真を用いて分類モデルの性能評価を行った。また、学習用データの不均一さの分類性能への影響を調べるため、モデル訓練用画像データに複数のデータベース由来の画像を用いて分類を行った。既存のデジタル一眼レフカメラによって撮影された同一分類群の画像・オンラインで公開されている他地域の昆虫画像などを訓練データとして訓練をおこない、画像の撮影方法や地理的な由来の違いが分類に与える影響を調べた。加えて異なるソースのデータを同時に用いて訓練を行うドメイン適応のアルゴリズムを利用し分類の精度を向上を試みた。また分類の予測確率を用いて訓練データに存在しないグループの画像が判定できるかを確認した。

2)DNA メタバーコーディングデータの分類モデル

畳み込みニューラルネットワークを用いて DNA 配列断片を分類するモデルを実装し、その性能評価を行った。訓練およびテストデータに Barcoding of life database (BOLD)に登録されている主要な昆虫の分類群の配列データを用いて、種レベルの分類を行った。加えて、データベースが不完全な状況を考慮し、深層学習による分布外検知モデルを実装した。分布外検知モデルは対象が訓練データ内に存在するクラスのサンプルかどうかを判定するモデルであり、分類モデルはまず対象が訓練データ内の既知のグループの配列かどうかを判定した後に種レベルの分類を行う。訓練に使われなかったグループの配列を用いてこの分布外検知モデルの性能評価を行い、BLAST などの配列類似度用いた方法と比較した。

４．研究成果

1)昆虫画像の分類モデルにおいては、従来の報告と同様に深層転移学習モデルは科レベルの分類を 98%程度の正答率で行うことができた。しかし、訓練用画像のデータに分類対象画像と異なる方法で撮影されたものを用いたときには分類の性能が大きく低下した。一方で、別地域で同じ撮影方法を用いて得られた画像を使って訓練を行った場合性能低下が少なかった。異なるソースのデータを同時に用いて訓練するドメイン適応のアルゴリズム(DANN, Domain adversarial neural network, Ganin et al. 2015)を用いると、撮影方法の違いによる分類の性能の低下を軽減できた。このことから画像解析を用いて生物多様性評価を行うときには訓練用データベースの構築・訓練アルゴリズムの選定に注意が必要であることが示唆された。

2)深層学習による DNA 配列分類モデルは極めて正確に種レベルの同定が行えることがわかった(種レベルの正答率 97%)。また既存の方法（配列類似度を用いた方法）と比較して配列断片が短い状況でも正答率の低下が少なかった。一方で未知の配列の検知性能は深層学習モデルも含め比較した全ての手法で高くなく、特に配列が短いときには全ての手法が 15～20%程度の未知配列を既知のものと判定した。この結果は配列データベースが不完全なときには配列分類モデルは真の多様性を過小評価する可能性があることを示唆している。

参考文献

Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, March M, Lempitsky V. 2016. Domain-Adversarial Training of Neural Networks. J Mach Learn Res. 17(59):1–35.

Knyshov A, Hoang S, Weirauch C. 2021. Pretrained Convolutional Neural Networks Perform Well in a Challenging Test Case: Identification of Plant Bugs (Hemiptera: Miridae) Using a Small Number of Training Images. Jockusch E, editor. Insect Syst Divers. 5(2).

Valan M, Makonyi K, Maki A, Vondráček D, Ronquist F. 2019. Automated Taxonomic Identification of Insects with Expert-Level Accuracy Using Effective Feature Transfer from Convolutional Networks. Syst Biol. 68(6):876–895.

Wührl L, Pylatiuk C, Giersch M, Lapp F, von Rintelen T, Balke M, Schmidt S, Cerretti P, Meier R. 2022. DiversityScanner: Robotic handling of small invertebrates with machine learning methods. Mol Ecol Resour. 22(4):1626–1638. doi:10.1111/1755-0998.13567.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　1件／うちオープンアクセス　1件）

１．著者名 Tomochika Fujisawa, Victor Noguerales, Emmanouil Meramveliotakis, Anna Papadopoulou, Alfried P. Vogler	４．巻 -
２．論文標題 Image-based taxonomic classification of bulk insect biodiversity samples using deep learning and domain adaptation	５．発行年 2023年
３．雑誌名 Systematic Entomology	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1111/syen.12583	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	山本　哲史 (YAMAMOTO Satoshi) (10643257)	国立研究開発法人農業・食品産業技術総合研究機構・農業環境研究部門・研究員 (82111)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
キプロス	University of Cyprus			
英国	Imperial College London			