



HOKKAIDO UNIVERSITY

Title	分類体系に注目した決定木の構築に基づく走査型電子顕微鏡画像を用いた生物の分類法
Author(s)	朴, 君; 小川, 貴弘; 長谷山, 美紀
Citation	電子情報通信学会論文誌 D, 情報・システム, vol. J98-D(5), 823-834
Issue Date	2015-05-01
Doc URL	https://hdl.handle.net/2115/58867
Rights	copyright©2015 IEICE
Type	journal article
File Information	37964.pdf



分類体系に注目した決定木の構築に基づく走査型電子顕微鏡画像を用いた生物の分類法

朴 君^{†a)} 小川 貴弘[†] 長谷山美紀[†]

A Biological Classification Method Using Scanning Electron Microscope Images via Taxonomy-Based Decision Tree

Jun PIAO^{†a)}, Takahiro OGAWA[†], and Miki HASEYAMA[†]

あらまし 本論文では、走査型電子顕微鏡 (Scanning Electron Microscop, SEM) で撮像された画像を用いた生物の分類法を提案する。提案手法では、生物学者が構築した分類体系に注目し、各ノードにその下位のノードへの分類を行う分類器を割り当てることで、決定木を構築する。これにより、構築された決定木を用いて生物の分類が可能となる。このように、生物の分類に有用な分類体系の構造を導入することで、画像特徴のみに注目する生物の分類法と比較して、高精度な分類が期待できる。また、提案手法では、更なる精度向上のため、以下の二つの処理も導入する。まず、分類体系において、画像特徴が類似する異なるノードに注目した決定木の変更を行い、誤分類を抑制する。次に、同種の生物の異なる撮像倍率の画像に対し、モーフィングを施すことで学習データの充足を行う。以上によって、提案手法では、SEM で撮像された生物の高精度な分類が可能となる。本論文の最後では、提案手法の有効性を確認するための実験結果を示す。

キーワード 走査型電子顕微鏡画像、生物分類、分類体系、決定木、モーフィング

1. ま え が き

現在の分類学において、生物の形態を基に、既知生物種の分類、対象生物の分類体系^(注1)上の位置付けを行う方法は分類学者の経験則に基づいて行われており、分類の際の基準は分類学者の判断に委ねられている [1], [2]。したがって、近年、分子情報を活用する「DNA バーコーディング^(注2)」への注目が高まっている。しかしながら、DNA 情報による種判別を大量の生物に適用することは現実的でない。

一方、走査型電子顕微鏡 (Scanning Electron Microscop, SEM) の発達によって、微細構造の観察により、異なる種間の違いを明確にすることが試みられている。このような背景から、近年生物の SEM による撮像画像が増加しており、これらの SEM 画像の分

類を自動かつ高精度に行う手法が必要となっている。

これまで、SEM 画像を対象とした研究は行われていないものの、生物を一般の撮像機器で撮像した画像から分類を行う手法は種々提案されている [3]~[5]。これらの手法は、いずれも生物の画像から画像特徴量を算出し、算出された特徴量を用いて分類器を構築することで、生物の自動分類を可能としている。しかしながら、これら従来手法を先に述べた目的に用いるためには、以下に示す二つの問題が存在する。

問題 1: 生物の画像特徴のみに注目し、生物の分類に有用な分類体系の構造は用いられていない。

問題 2: SEM 画像による生物の種の分類を行う際、学習データとなる生物の画像を大量に収集することが困難である。

これらの問題に対して以下のような解決策が考えら

[†] 北海道大学大学院情報科学研究科, 札幌市
Graduate School of Information Science and Technology,
Hokkaido University, Kita 14, Nishi 9, Kita-ku, Sapporo-shi,
060-0814 Japan

a) E-mail: pakukun@lmd.ist.hokudai.ac.jp

DOI:10.14923/transinfj.2014JDP7106

(注1): 「種」をまとめて「属」、幾つかの進化的に近い「属」をまとめて「科」、以下同様に「目」「綱」「門」「界」の順番で、段々と大きなまとまり (分類階級) を作ることで構築されたもの [2]。

(注2): 短い遺伝子マーカー (ある性質をもつ生物個体に特有の DNA 配列) を利用して DNA 配列から種を特定する手法 [2]。

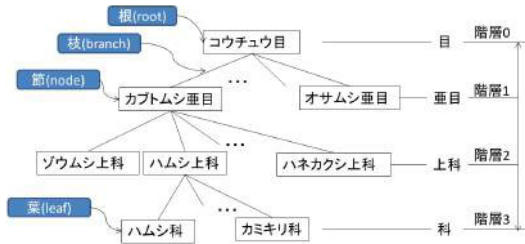


図 1 分類体系の木構造の例

Fig. 1 Example of the tree structure of the taxonomy.

れる。まず、問題 1 については、現在の生物学の分野において、生物の形態を基に既知生物種から構築されている分類体系の利用が考えられる。分類体系は、図 1 に示すように生物の種に共通する特徴によってまとめられた各科、目、綱、門、界をそれぞれノードとした階層的な木構造である。したがって、この木構造に基づいた分類を導入することで精度の向上が期待できる。ただし、SEM 画像では生物の形態特徴ではなく、生物の表面構造などに注目している場合が多く、分類体系をそのまま適用すると、誤分類が生じる可能性が存在する。また、Gorb らは生物がどのような表面構造を採用するかは、進化の系統による分類とは関係なく、環境適応によって選択されることを明らかにしている [6]。このことから、分類体系において異なる位置に存在するが、同じ環境に生息することで、画像特徴が類似する生物が存在する可能性がある。したがって、その誤分類を抑制する処理の導入により、高精度な分類が可能になると考えられる。次に、問題 2 については、最初に与えられる学習データ（以降、初期学習データ）から新たな学習データを自動的に生成することで、その充足を行う方法が考えられる。一般に、生物学者は対象生物に対して様々な撮像倍率で複数の SEM 画像を撮像しているため、その中には同じ部位を異なる倍率で撮像した画像のペアが存在する。そのため、このような画像のペアを選択し、その間の中間画像を自動的に生成すれば、学習データの充足が可能になると考えられる。

そこで本論文では、分類体系に注目した決定木の構築に基づく SEM 画像を用いた生物分類法を提案する。提案手法では、生物学者が作成した分類体系に注目して決定木を構築し、これに基づいた生物の分類を実現する。具体的には、分類体系の各ノードに、その下位のノードへの分類を行う分類器を割り当てることで実現する。また、分類体系において類似した画像特徴を

もつ異なるノード間での誤分類を抑制するため、分類体系から生成される決定木の変更を行う。以上により、画像特徴のみに注目する生物の分類手法で、分類精度に限界が存在した問題 1 が軽減される。次に、提案手法では、更なる精度の向上のため、学習データの充足を行う。具体的には、同一種の生物を撮像した SEM 画像の中から、同一部位を異なる倍率で撮像した SEM 画像のペアの間の画像を自動的に生成し、それらを初期学習データに加えることで、学習データの充足を行う。これにより、学習データを十分に収集することが困難であった問題 2 を解決する。以上により、提案手法では分類体系に注目した決定木に基づく生物の高精度な分類が可能となる。

本文の構成を以下に示す。まず、2. では提案手法で用いる分類体系について説明する。次に、3. では、分類体系に注目した決定木の構築による SEM 画像を用いた生物の分類法を提案する。最後に、4. では、提案手法の有効性を確認するため、実画像を用いた実験を行い、その結果について考察を行う。

2. 分類体系

本章では、文献 [2] で紹介されている分類体系 (Taxonomy) について説明する。分類学の具体的な作業は、自然の生物世界の中から種 (Species) と呼ばれる個体集合を見付けだし、それに名前を付け、体系化することである。分類学者は種を基本的な要素として、生物をグループ、すなわちタクソン^(注3)に分割し、それを一連の段階、つまり分類体系に配置してきた。現在、用いられている分類体系は、リンネが提案したものである (表 1)。分類階層の各レベルには名前が付けられており、これが分類階級である。形態的に似た特徴を共有している種は属 (genus) にまとめられ、同様にして属は科 (family) に、科は目 (order) に、目は綱 (class) に、綱は門 (phylum) に、そして門は最も広いレベルである界 (kingdom) に、更に界はドメイン (domain) にまとめられる。また、階層的な分類体系は木構造の分類である (図 1)。このとき、最も上位の階層にあるノードを根ノード (root node) と呼び、一般的にこのノードから分類が始まる。また、子ノード (child node) をもたないノードを葉ノード (leaf node) と呼び、一般的にこのノードで分類が終わる。分類を

(注3)：タクソン (taxon, 複：タクサ, taxa) とは、生物の分類において、ある分類階級に位置付けられる生物の集合のこと。

表 1 リンネ式階層分類体系に用いられる階層の主な種類.
★印は基本的な階級 (必ず用いる階級) を示す. 文献 [2] により抜粋

Table 1 The main types of hierarchy in the Hierarchy Taxonomy of Linnaean. ★mark shows the basic rank, and these ranks are always used. Quoted from reference [2].

	階級	rank	語尾
	★ 界	Kingdom	
	★ 門	Phylum	
	亜門	Subphylum	
	上綱	Superclass	
国際動物	★ 綱	Class	
命名規約	亜綱	Subclass	
の規定を	下綱	Infraclass	
受けない	コホート (区)	Cohort	
階級	上目	Superorder	
	★ 目	Order	
	亜目	Suborder	
	下目	Infraorder	
	上科	Superfamily	-oidea
	★ 科	Family	-idae
	亜科	Subfamily	-inae
科階級群	上族	Supertribe	
	族	Tribe	-ini
	亜族	Infratribe	
属階級群	★ 属	Genus	
	亜属	Subgenus	
種階級群	★ 種	Species	
	亜種	Subspecies	

行う際には、木構造の根ノードから分類を行い、分類結果に沿って枝を選択し、選択された枝と結ばれている子ノードで、親ノード (parent node) と同様に分類を行う。この作業を葉ノードに到達するまで繰り返す。このような木構造をもつ分類体系は、生物が共有する形態特徴に基づいて生物学者により構築されたため、画像処理による生物の分類を行う際にも有用な知識になると考えられる。

3. SEM 画像を用いた生物の分類法

本章では、SEM 画像を用いた生物の分類法について説明する。提案手法ではまず、分類体系に注目した決定木を構築する。次に、構築された決定木を用いた生物の分類を行う前に、更なる精度向上のため、以下の処理を導入する。

処理 1 分類体系において同じ親ノードの下に属する類似ノード同士を順次統合することで、新たな木構造を構築する。また、分類体系において異なる親ノードに属する類似ノードを検証するノード (以降、検証ノード) を構築された決定木に加える。

処理 2 モーフィング [7] を用いて新たな学習画像を

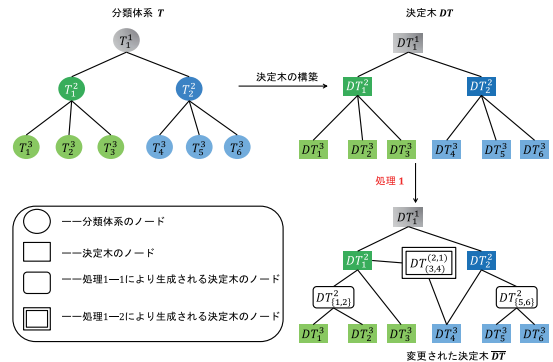


図 2 分類体系に注目した決定木構築の概念図

Fig.2 Diagram of the procedures for constructing the decision tree.

生成し、得られた新たな画像を初期学習データに追加することで、学習データを充足する。

このとき、**処理 1** は、1. において述べた**問題 1**を軽減するための処理であり、**処理 2** は**問題 2**を解決するための処理である。提案手法では、以上の**処理 1**と**処理 2**を導入することで、より高精度な生物の分類を可能とする。

以降では、提案手法における処理の詳細について説明する。まず、**3.1**では、分類体系に注目した決定木の構築法 (**処理 1**を含む) について説明する。次に、**3.2**では、モーフィングを用いた新たな学習データの充足 (**処理 2**) について説明を行う。最後に、**3.3**では、構築された決定木によって生物を分類する方法について説明する。

3.1 分類体系に注目した決定木の構築法

本節では、図 2 に示す分類体系に注目した決定木の構築法について説明する。今、分類体系を T 、その各ノードを $T_j^i (i = 1, 2, \dots, H; j = 1, 2, \dots, N_i)$ とする。このとき H は分類体系における階層の数、 N_i は階層 i に含まれるノードの数である。提案手法では、図 2 に示す分類体系 T の各ノード T_j^i に対して、その子ノードへの分類を行う分類器を与えることで、生物の分類を可能とする決定木 DT を構築する。また、より高精度に分類を行うための**処理 1**は、以下の二つの処理により実現される。

処理 1-1 決定木 DT において、同じ親ノードに属するノード同士を順次統合することで、決定木の変更を行う。

処理 1-2 決定木 DT において、異なる親ノードに属する類似ノード間に対して検証ノードを与える。

なお、図2の右下の図の例では、類似ノードが DT_1^2 と DT_4^3 であり、検証ノードは $DT_{(3,4)}^{(2,1)}$ である。ここで、検証ノード $DT_{(3,4)}^{(2,1)}$ は、ノード DT_1^2 あるいは DT_4^3 に分類されたテストデータに対し、ノード DT_1^2 と DT_4^3 との間で再度検証する役割を果たす。したがって、 $DT_{(3,4)}^{(2,1)}$ は DT_4^3 の親ノードとは異なる役割になる。以下では、上記二つの処理について、具体的な説明を行う。

今、決定木 DT において、ある親ノード DT_j^i に属する子ノードを $DT_k^{i+1} (k \in \mathcal{L}_j^i)$ とする。ただし、 $\mathcal{L}_j^i$ は親ノード DT_j^i の下位に属する子ノードの k に関するインデックス集合とし、 $|\mathcal{L}_j^i|$ をその要素数、すなわち属する子ノードの数であるとする。また、決定木 DT における DT_j^i の子ノード $DT_k^{i+1} (k \in \mathcal{L}_j^i)$ に対し、異なる親ノード $DT_{j'}^{i+1} ((i', j') \neq (i, j))$ に属する子ノードを $DT_{k'}^{i+1} (k' \in \mathcal{L}_{j'}^{i+1})$ とする。以下では処理 1-1 及び処理 1-2 についてそれぞれ説明を行う。

処理 1-1

同じ親ノード DT_j^i に属するある子ノード $DT_{k_a}^{i+1}, DT_{k_b}^{i+1} (k_a, k_b \in \{\mathcal{L}_j^i | k_a \neq k_b\})$ の学習データである $M_{k_a}^{i+1}, M_{k_b}^{i+1}$ 枚の SEM 画像群 $\mathbb{I}_{k_a}^{i+1}, \mathbb{I}_{k_b}^{i+1}$ の特徴ベクトル $\mathbf{x}_{k_a m_a}^{i+1} (m_a = 1, 2, \dots, M_{k_a}^{i+1})$, $\mathbf{x}_{k_b m_b}^{i+1} (m_b = 1, 2, \dots, M_{k_b}^{i+1})$ を用いて、次式により簡便に定義される相違度 $\text{Dif}(DT_{k_a}^{i+1}, DT_{k_b}^{i+1})$ [8] を算出する。

$$\text{Dif}(DT_{k_a}^{i+1}, DT_{k_b}^{i+1}) = \frac{d(DT_{k_a}^{i+1}, DT_{k_b}^{i+1})}{(\sigma_{DT_{k_a}^{i+1}} + \sigma_{DT_{k_b}^{i+1}})} \quad (1)$$

ただし、 $d(DT_{k_a}^{i+1}, DT_{k_b}^{i+1})$ はノード $DT_{k_a}^{i+1}$ と $DT_{k_b}^{i+1}$ 間の距離を表し、

$$d(DT_{k_a}^{i+1}, DT_{k_b}^{i+1}) = \|\mathbf{c}_{DT_{k_a}^{i+1}} - \mathbf{c}_{DT_{k_b}^{i+1}}\| \quad (2)$$

と定義される。このとき、 $\mathbf{c}_{DT_k^{i+1}}$ はノード DT_k^{i+1} に属する学習データの中心ベクトルであり、次式により定義される。

$$\mathbf{c}_{DT_k^{i+1}} = \frac{1}{M_k^{i+1}} \sum_{m=1}^{M_k^{i+1}} \mathbf{x}_{km}^{i+1} \quad (3)$$

更に、 $\sigma_{DT_k^{i+1}}^2$ はノード DT_k^{i+1} における学習データの分散であり、次式で与えられる。

$$\sigma_{DT_k^{i+1}}^2 = \frac{1}{M_k^{i+1} - 1} \sum_{m=1}^{M_k^{i+1}} \|\mathbf{x}_{km}^{i+1} - \mathbf{c}_{DT_k^{i+1}}\|^2 \quad (4)$$

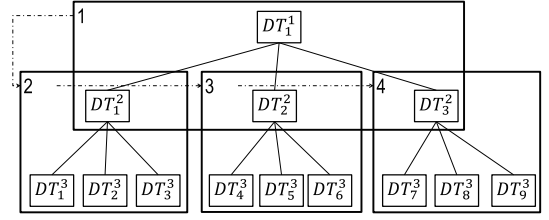


図3 2分木構築の流れ図

Fig. 3 Flow diagram of the binary tree construction.

Algorithm 1 ConstructTree(DT_j^i)

Input: ある親ノード DT_j^i に属する子ノード $DT_k^{i+1} (k \in \mathcal{L}_j^i)$.

Output: 根 (root) ノードが DT_j^i である 2 分木.

- 1: インデックスの集合 $\mathcal{U} = \{s_1, s_2, \dots, s_R\} (R = |\mathcal{L}_j^i|)$ を用意する。ただし、 $\mathcal{U} = \mathcal{L}_j^i$ である。
- 2: $s_{a,opt}, s_{b,opt} \leftarrow \arg \min_{s_a, s_b} \text{Dif}(DT_{s_a}^{i+1}, DT_{s_b}^{i+1})$. なお、 $s_a, s_b \in \{\mathcal{U} | s_a \neq s_b\}$ である。
- 3: $DT_{s_{a,opt}}^{i+1}$ と $DT_{s_{b,opt}}^{i+1}$ を統合した新たな親ノード DT_s^{i+1} を生成する。ただし、 $s \leftarrow s_{a,opt} \cup s_{b,opt}$ である。
- 4: $\mathcal{U}$ から $s_{a,opt}$ と $s_{b,opt}$ を除き、 s を加える。
- 5: $R \leftarrow R - 1$.
- 6: R が 1 になるまで 2-5 を繰り返す。

提案手法では、 $|\mathcal{L}_j^i|$ 個の子ノード $DT_k^{i+1} (k \in \mathcal{L}_j^i)$ の中から、算出された $\text{Dif}(DT_{k_a}^{i+1}, DT_{k_b}^{i+1})$ に基づいて、相違度が最も低いノード同士を順に統合する。最後に、上述の作業を親ノード DT_j^i に統合されるまで繰り返す。これにより、親ノード DT_j^i とその子ノード $DT_k^{i+1} (k \in \mathcal{L}_j^i)$ の 2 階層間で 2 分木の構造が得られる。

上で示したとおり、2 分木の構造を求める際には、子ノードの中から式 (1) を最小とする二つのノードが順に選ばれていく。したがって、得られる 2 分木の構造は、二つの子ノードの選ぶ順序によらず、唯一無二となる。本文では、以上までに示す処理を $\text{ConstructTree}(DT_j^i)$ と呼び、その具体的なアルゴリズムを Algorithm 1 に示す。提案手法では、図3の例に示すように、分類体系を表す決定木 DT の 2 分木への変更を以下の手順により行う。

Step1 決定木 DT の最上位の階層 $i = 1$ のノード DT_1^1 (root node) に属する子ノード $DT_k^2 (k \in \mathcal{L}_1^1)$ に対し、 $\text{ConstructTree}(DT_1^1)$ に基づいて二つの階層間での 2 分木を構築する。

Step2 $i = i + 1$

Step3 決定木 DT の階層 i に移動し、同じ親ノード DT_j^i に属する子ノード $DT_k^{i+1} (k \in \mathcal{L}_j^i)$ に対し、

$\text{ConstructTree}(DT_j^i)$ に基づいて 2 分木を構築する。これらの処理を $j = 1, 2, \dots, N_i$ に対して行う。

Step4 $i = H - 1$ になるまで, **Step2** と **Step3** を繰り返す。

上の処理に示すとおり提案手法では, 分類体系を表す決定木の上位の階層から下位の階層の順番で, 二つの階層間ごとに $\text{ConstructTree}(DT_j^i)$ を施し, 最終的な 2 分木を求める。ただし, この処理は 2 階層間ごとに行うため, その順序によらず, 最終的に変更された決定木は唯一となる。

先に述べたとおり, 提案手法では, 分類体系の各ノードにその子ノードへの分類を行う分類器を与えることで, 生物の分類を試みる。このとき, ある親ノード DT_j^i の子ノード $DT_k^{i+1} (k \in \mathcal{L}_j^i)$ として, $|\mathcal{L}_j^i|$ 個の分類対象が含まれ, それらが互いに類似していることにより, 高精度な多クラス分類を行うことが困難となる場合がある。したがって, 段階的に 2 クラス分類を行う問題となる 2 分木の構造へと変更することで, その問題の軽減を試みる。また, 上述の処理によって生成された新たな 2 分木の各ノードにその下位のノードへの分類を行う分類器を与えることで, 分類を可能とする。

処理 1-2

提案手法では, より高精度な生物の分類を行うため, 決定木 DT において, 異なる親ノードに属する類似ノード間に対して検証ノードを与える。具体的には, ある親ノード DT_j^i に属する子ノード $DT_k^{i+1} (k \in \mathcal{L}_j^i)$ に対し, 誤分類が生じる可能性が最も高い類似ノード $DT_{k_s}^{i_s+1}$ を, 異なる親ノード $DT_{j'}^{i'} ((i', j') \neq (i, j))$ に属する子ノード $DT_{k'}^{i'+1} (k' \in \mathcal{L}_{j'}^{i'})$ の中から選択する。具体的には, 式 (1) と同様に, 次式に示す相違度を算出し, しきい値 T_h より小さく, その値が最小となる場合の $DT_{k'}^{i'+1}$ を類似ノード $DT_{k_s}^{i_s+1}$ として選択する。

$$DT_{k_s}^{i_s+1} = \arg \min_{DT_{k'}^{i'+1}} \text{Dif}(DT_k^{i+1}, DT_{k'}^{i'+1})$$

$$\text{s.t. } \text{Dif}(DT_k^{i+1}, DT_{k'}^{i'+1}) < T_h \quad (5)$$

次に, 選択されたノード $DT_{k_s}^{i_s+1}$ とノード DT_k^{i+1} 間に検証用のノード $DT_{(i_s+1, k_s)}^{(i+1, k)}$ を追加する。最後に, 追加された検証用ノード $DT_{(i_s+1, k_s)}^{(i+1, k)}$ にノード DT_k^{i+1} とノード $DT_{k_s}^{i_s+1}$ を分類する分類器を与える。式 (5) に示すとおり, 提案手法では, 相違度 $\text{Dif}(DT_k^{i+1}, DT_{k'}^{i'+1})$

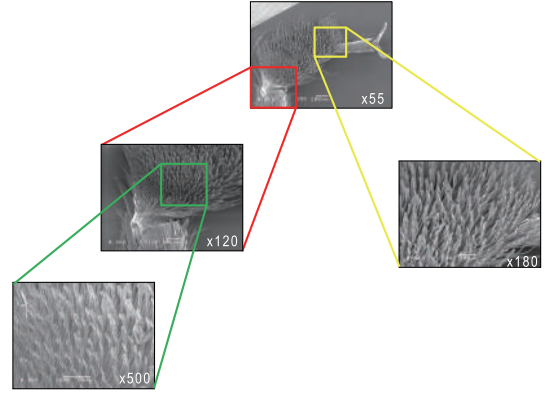


図 4 ナミクシヒゲハネカクシ種の異なる撮像倍率の SEM 画像間の関係

Fig. 4 Relationship between the different magnification of SEM images of *Velleius dilatatus*.

が最小, なおかつ, しきい値以下となるノードを類似ノードとして選択する。つまり, 相違度が最小のノードを一つ選択し, その相違度に対してしきい値処理を施すことで, 類似ノードであるか否かを決定しているため, 選択されるノードは一つとなる。このように, 提案手法では類似ノードを一つに限定し, 選択を行っているが, 実際には複数のノードが類似している場合が存在する。このような場合, 我々はそれら全てを含む親ノードが類似ノードとなり, その一つのみが類似ノードとして選択されると仮定する。

生物がどのような表面構造を採用するかは, 進化の系統による分類とは関係なく, 環境適応によって選択されることを Grob らは明らかにしている [6]。したがって, 決定木 DT の中には, 分類体系において画像特徴が類似する異なるノードが存在すると考えられる。そのため, 決定木によって分類を行う際に, 誤った子ノードへの分類が進む可能性がある。そこで提案手法では, あらかじめ互いに類似するノードを見付け出し, それらのノードへ分類が到達した際に, 上記の誤りが発生していないかを検証するノードを用い, 誤分類の抑制を試みる。

以上の処理 1-1 と処理 1-2 により, 全てのノードで分類器を有する新たな決定木 $\overline{DT}$ (図 2 右下) の構築が実現される。

3.2 モーフィングを用いた新たな学習データの生成法

本節では, モーフィングを用いた新たな学習データの生成法 (処理 2) について説明する。図 4 に示すよ

うに、同一種の SEM 画像には、同じ部位を異なる倍率で撮像した画像のペアが存在する。そこで、提案手法ではモーフィング [7] を用いて、これらの画像のペア間において中間画像を生成することで、自動的に学習データの充足を行う。モーフィングとは、ある画像から異なる画像に滑らかに変化する過程の中間画像を生成する手法である。モーフィングでは、処理対象となる二つの画像に対して対応付けを行い、それに基づいてワープ処理とクロスディゾルブ処理を行うことで、中間画像を生成する^(注4)。このとき、提案手法では、モーフィングにおける対象画像間の対応付けを Scale Invariant Feature Transform (SIFT) [9] により行う。

提案手法ではまず、最下位のノードに存在するノード $T_j^H (j = 1, 2, \dots, N_H)$ の初期学習データである全ての SEM 画像 $\tilde{I}_j^H \in \{\tilde{I}_{j1}^H, \tilde{I}_{j2}^H, \dots, \tilde{I}_{jM_j^H}^H\}$ から、同じ部位を異なる倍率で撮像した画像のペア $\tilde{I}_{jm}^H$ と $\tilde{I}_{jm'}^H$ の選択を行う。ここで、 M_j^H はノード T_j^H に属する画像の総数を表す。なお、提案手法では、SIFT により対応付けされた点の割合が $\frac{F_{m,m'}}{\min(F_m, F_{m'})} > F_h$ を満たす画像のペア $\tilde{I}_{jm}^H$ と $\tilde{I}_{jm'}^H$ を選択する。ただし、 $F_{m,m'}$ は画像 $\tilde{I}_{jm}^H$ と $\tilde{I}_{jm'}^H$ が対応付けされた点の数であり、 F_m 及び $F_{m'}$ はそれぞれ画像 $\tilde{I}_{jm}^H$ と $\tilde{I}_{jm'}^H$ の特徴点の数を表し、 F_h はしきい値である。上述のしきい値処理により、モーフィングを行うことが困難な画像のペアを事前に除去することが可能となる。次に、選択された画像のペアを対象画像とし、モーフィング [7] を行う。最後に、生成された中間画像を初期学習データ $\tilde{I}_j^H$ に追加することで、新たな学習データ $\mathbb{I}_j^H$ を構築する。

このように自動的に学習データの充足を行うことで、学習データを大量に収集することが困難であるという問題を解決することができ、分類精度の向上が期待できる。なお、前節の $\mathbb{I}_j^H$ は、この節により生成された学習データである。

3.3 決定木を用いた生物の分類法

本節では、構築された決定木により、未知の生物の SEM 画像 (対象画像) に対して分類を行う方法について説明する。提案手法では、対象画像が与えられた際に、決定木 $\overline{DT}$ の最上位のノードにおいて分類を行い、これを最下位ノードに到達するまで繰り返す。このとき、あるノード $\overline{DT}_j^i$ に分類された際、そのノードに

検証ノード $\overline{DT}_{(i_s, j_s)}^{(i, j)}$ がつながっている場合、その検証ノードの分類器を用いて、対象画像がノード $\overline{DT}_j^i$ とその類似ノード $\overline{DT}_{j_s}^{i_s}$ のどちらに属するかを検証する。また、検証結果に基づいて、対象画像を移動し、正しいノードへの修正を行う。更に、対象画像を移動させたノードからその下位のノードへの分類を行う。上述の処理を繰り返し、対象画像が決定木 $\overline{DT}$ の最下位のノードまで到達した場合、分類を終了し、その分類結果は最終的に分類されたノードに含まれる生物であると判定する。以上のようにして提案手法では、対象画像の分類結果の修正が可能となり、分類体系に注目した決定木による高精度な生物の分類が期待できる。

4. 実 験

本章では、提案手法の有効性を確認するため、実画像を用いた評価実験を行う。実験では、昆虫を SEM で撮像した画像を用いた。なお、用いた画像は、全て昆虫の前脚画像であり、撮像倍率は 22 倍から 250 倍までとした。本実験では、分類体系において最下位のノードである 18 種類の科の画像の例を図 5 に示す。図 5 に示すとおり、各々の昆虫の前脚の画像は互いに視覚的に類似しているものが多い。また、図 6 に示す分類体系を用い、各ノードの名前と SEM 画像の画像数を表 2 に示し、用いた SEM 画像の全ての撮像倍率と各倍率とその画像数を表 3 に示す。

本実験では、提案手法における各処理の有効性を確認するため、以下の五つの手法を用いる。

比較手法: 初期学習データ $\tilde{I}_j^H (j = i, 2, \dots, N_H)$ のみを用い、分類体系に注目した決定木を導入せず、従来の分類器のみを使用した生物の分類。

手法 1: 初期学習データ $\tilde{I}_j^H (j = i, 2, \dots, N_H)$ のみを用い、生物学者により構築された分類体系に注目した決定木 DT (図 2 の右上) による生物の分類。

手法 2: 手法 1 に**処理 1** を導入した手法。つまり、修正された決定木 $\overline{DT}$ (図 2 の右下) を用いた生物の分類。

手法 3: 手法 1 に**処理 1** は導入せず、**処理 2** を導入した手法。つまり、モーフィングにより生成された新たな学習データ $\mathbb{I}_j^H (j = i, 2, \dots, N_H)$ を用いた手法 1 による生物の分類。

提案手法: 手法 2 に**処理 2** を導入した手法。すなわち、用いた学習データは手法 3 と同様。

(注4)：ワープ処理とクロスディゾルブ処理に基づくモーフィングの詳細は文献 [7] に示されている。本文では紙面の都合上、その詳細な説明は割愛する。

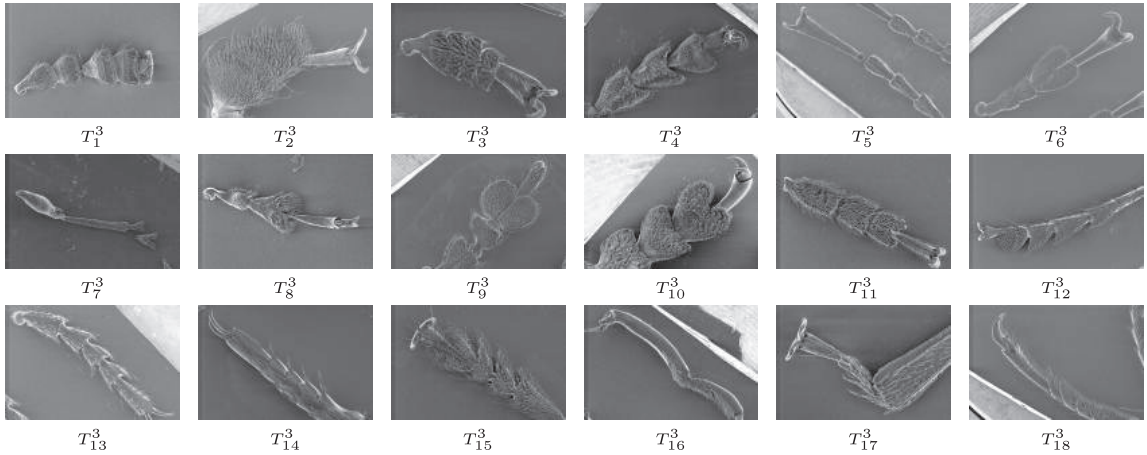


図 5 実験で用いた既知生物の例

Fig. 5 Example images of known species used in the experiment.

表 2 実験で用いた分類体系の各ノードの名前と画像数

Table 2 Name of each node of taxonomy and the number of images of each node used in the experiment.

ランク	ノード名	種名	画像数	ノード名	種名	画像数	ノード名	種名	画像数
目	T_1^1	コウチュウ目	328	-	-	-	-	-	-
上科	T_1^2	タマムシ上科	20	T_2^2	ハネカクシ上科	26	T_3^2	ホタル上科	7
	T_4^2	ゾウムシ上科	37	T_5^2	ハムシ上科	111	T_6^2	オサムシ上科	98
	T_7^2	ヒラタムシ上科	21	T_8^2	ゴミムシ上科	8	-	-	-
	T_9^2	-	-	T_{10}^2	-	-	T_{11}^2	-	-
科	T_1^3	タマムシ科	20	T_2^3	ハネカクシ科	12	T_3^3	シデムシ科	14
	T_4^3	ジョウカイボン科	7	T_5^3	オサゾウムシ科	5	T_6^3	ミツギリゾウムシ科	6
	T_7^3	オトシブミ科	5	T_8^3	ホソクチゾウムシ科	4	T_9^3	ゾウムシ科	17
	T_{10}^3	ハムシ科	50	T_{11}^3	カミキリムシ科	61	T_{12}^3	ハンミョウ科	37
	T_{13}^3	ホソクビゴミムシ科	7	T_{14}^3	ミズスマシ科	7	T_{15}^3	オサムシ科	47
	T_{16}^3	ケシキスイ科	9	T_{17}^3	オオキスイムシ科	12	T_{18}^3	ゴミムシ上科	8
	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-

表 3 実験で用いた画像の各倍率とその画像数

Table 3 The number of images in each magnification used in this experiment.

撮像倍率	x22	x23	x25	x27	x30	x33	x35	x37	x40	x43	x45	50	x55	x60	x65	x70	x75	x80
画像数	8	2	6	5	1	3	4	5	3	4	6	2	13	10	7	7	10	9
撮像倍率	x85	x90	x95	x100	x110	120	x130	x140	x150	x160	x170	x180	x190	x200	x220	x230	x250	
画像数	1	5	13	14	15	14	12	14	13	10	19	11	13	17	11	18	20	

なお、手法 3 と提案手法では、各画像のペアから生成される中間画像の枚数を実験的に 1 枚と設定した。その枚数は、各画像のペアから生成される画像の枚数と分類精度間の関係を表す図 7 に示すように、分類精度 (後述の式 (9)) が最大となっていることから、そのように設定を行っている。

決定木の各ノードに与える分類器として、従来よりベンチマーク手法として用いられる Support Vector Machine (SVM) [10] と Support Vector Data Description (SVDD) [11] を採用し、それぞれの分類精度を算出する。ここで、SVM, SVDD のカーネル関数

はいずれもガウシアンカーネルであり、パラメータは grid search により決定する。また、本実験では各 SEM 画像から SIFT [9] 特徴量を抽出し、Bag-of-Features (BoF) [12] アプローチによりベクトル量子化した 100 次元の特徴量を用いる。このとき、提案手法では、BoF のアプローチを利用する際、基準特徴は学習データである SEM 画像から SIFT 特徴量を抽出し、それらに対して k-means クラスタリングを施すことで得られたものを利用している。更に、 $T_h = 0.25, F_h = 0.1$ とする。

また、分類精度の定量評価として Leave-one-out

表 4 recall, precision, F 値及び正解率の比較. 太字は上科及び科においてそれぞれ最大となる値を示している

Table 4 Comparison of recall, precision, F-measure and accuracy. The value shown as bold fonts represents the maximum value for Superfamily and Family, respectively.

分類器	精度	比較手法		手法 1		手法 2		手法 3		提案手法	
		上科	科	上科	科	上科	科	上科	科	上科	科
SVM	Recall	0.42	0.25	0.42	0.28	0.43	0.29	0.52	0.34	0.51	0.36
	Precision	0.61	0.34	0.61	0.41	0.60	0.40	0.56	0.36	0.60	0.36
	F-measure	0.46	0.25	0.46	0.30	0.47	0.31	0.51	0.33	0.52	0.35
	正解率	0.65	0.47	0.65	0.52	0.66	0.54	0.67	0.53	0.70	0.58
SVDD	Recall	0.32	0.17	0.32	0.17	0.33	0.18	0.34	0.19	0.34	0.20
	Precision	0.43	0.25	0.43	0.20	0.44	0.26	0.42	0.25	0.43	0.30
	F-measure	0.35	0.18	0.35	0.16	0.36	0.18	0.36	0.19	0.37	0.21
	正解率	0.51	0.30	0.51	0.31	0.52	0.33	0.52	0.34	0.53	0.36

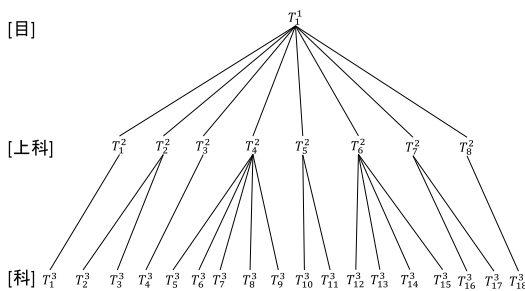


図 6 実験で用いた分類体系 T
Fig. 6 Taxonomy T used in the experiment.

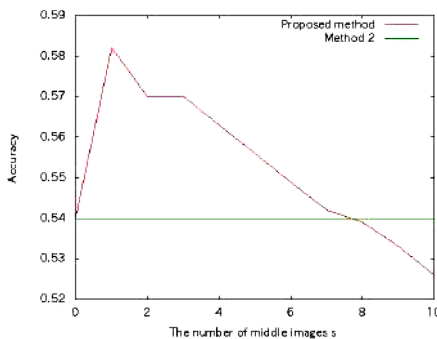


図 7 モーフィングによる各画像のペアから生成された画像の枚数 s と提案手法の正解率間の関係

Fig. 7 The relationship between accuracy of the proposed method and the number of middle images s , where the middle images are generated from each image pair by image morphing.

法 [13] を採用する. 更に, 定量評価指標として, 次式に示す再現率 (Recall), 適合率 (Precision), F 値 (F-measure) 及び正解率を用いる.

$$\text{Recall} = \frac{\text{分類された中で正解である数}}{\text{全正解数}} \quad (6)$$

$$\text{Precision} = \frac{\text{分類された中で正解である数}}{\text{分類された数}} \quad (7)$$

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (8)$$

$$\text{正解率} = \frac{\text{正しく分類されたテストデータの数}}{\text{全てのテストデータの数}} \quad (9)$$

実験の結果を表 4 に示す. 表 4 の比較手法における SVM は, 多クラス分類を行う one-vs-one [14] であり, SVDD は文献 [11] と同様に行う. なお, 表 4 には最終結果として求める科の分類結果だけでなく, その上の層の上科の分類結果も合わせて示している. なお, 表 4 に示す Recall, Precision, F-measure の値は各上科また各科から算出された値の平均値である.

分類体系に注目した決定木の有効性に関する考察

表 4 において, 手法 1 の SVM による科分類に関する正解率及び F-measure の値が比較手法を上回っていることから, 分類体系を表す決定木による生物の分類の有効性の確認ができた^(注5). つまり, 分類学者が分類体系を構築する際, 生物の形状特徴に基づいたクラスタリングを行うことから, 分類体系において同じノードに属する子ノードの間では, 特徴がある程度類似し, 我々はそれを表す決定木を用いることで, 生物分類の高精度化を可能とした.

なお, 上科における SVDD による分類では, 正解率は高くなっているものの, F-measure が低下している. Recall, Precision 及び F-measure は各上科で求められる値を平均することで得られており, ある上科において, 低いものが存在する場合, 平均の値が低下する. これは, 異なる上科の間での学習データ数の偏

(注5): 上科については, 比較手法と手法 1 は同一の手法となるため精度の差は発生しない.

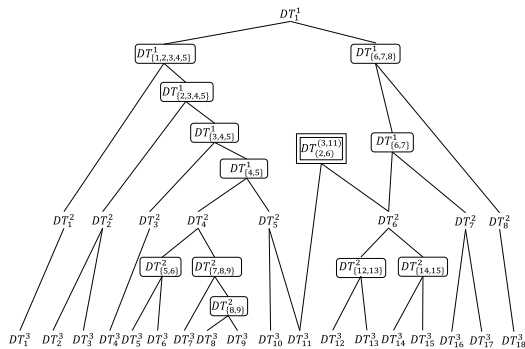


図 8 変更された分類体系を表す決定木 $\overline{DT}$
Fig. 8 Modified taxonomy-based decision tree $\overline{DT}$.

りによるものと考えられ、これを考慮した解決法については、今後の課題とする。

決定木の変更の有効性に関する考察

表 4 において、手法 2 の科の分類に関する正解率及び F-measure の値が手法 1 を上回っていることから、同じ親ノードに属するノードの順次統合に基づいた新たな木構造の構築及び異なる親ノードに属する類似ノード間における検証ノードの追加が有効であることが確認できる。また、この処理により修正された決定木 $\overline{DT}$ を図 8 に示す。なお、提案手法では、式 (5) に示すしきい値の値を分類精度 (式 (9)) が最大になるように設定し、検証用ノードを決めている。生物の微細構造は生物が生息している環境に関連するため、分類体系において類似する画像特徴をもつ異なるノードが存在する可能性がある。我々はその誤分類の抑制を可能にすることで、生物の分類精度の向上を実現した。

モーフィングにより生成された中間画像の有効性に関する考察

表 4 において、手法 3 の正解率及び F-measure の値が手法 1 を上回っていることから、モーフィングによる学習データの自動的な充足の有効性の確認ができた。また、提案手法におけるモーフィングにより生成された中間画像の例を図 9 に示す。更に、どの撮像倍率間で中間画像が何枚作成されたかを示す図 10 から、生成された中間画像の総数は 104 枚であることが確認できる。一般的に、学習データとなる生物の SEM 画像を大量に収集することが困難であり、我々はモーフィングを施すことで、同じ部位を異なる倍率で撮像した同種の生物の画像ペアの中間画像を自動的に生成し、生物分類の精度を向上させた。

また、生物の表面は、ナノからマイクロに至る領域

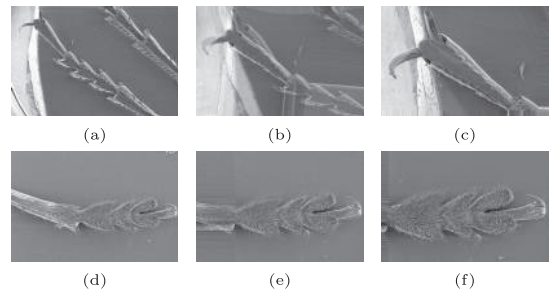


図 9 モーフィングにより生成された中間画像の例。画像 (a)(x23) と (c)(x60) はオサムシ科の初期学習データの中から選択された画像ペアであり、画像 (d)(x55) と (f)(x95) はハムシ科の初期学習データの中から選択された画像ペアである。画像 (b) はオサムシ科の画像ペアから生成された中間画像であり、画像 (e) はハムシ科の画像ペアから生成された中間画像である

Fig. 9 Example middle images generated by morphing. The images (a), (c) and (d), (f) are image pairs selected from initial training data of Carabidae and Chrysomelidae, and these magnifications are x23, x60, x55, x95, respectively. The middle images (b) and (e) are generated from image pairs of Carabidae and Chrysomelidae, respectively.

において階層的な構造を有し、その表面構造が大きく異なることが明らかにされている [6]。したがって、SEM で撮像した生物の画像では、同じ種の同一の撮像部位であっても、異なる特徴をもつ場合が存在する。そのため、初期学習画像に存在しない撮像倍率の画像がテストデータとして与えられた際、分類精度が低下する可能性がある。提案手法において、モーフィングを用いることは、単に学習データの充足だけではなく、異なる撮像倍率の画像のペアの間の画像を生成することで、初期学習画像に存在しない撮像倍率の画像を生成し、上記の問題を軽減する効果があると考えられる。

ただし、本実験では提案手法の精度が最大になるように、各画像のペアから生成された画像の枚数を実験的に決めているが、学習に十分な画像の枚数については、対象データによっても異なり、適応的な設定が必要になると考えられる。したがって、その適応的な設定については今後の課題として検討する必要がある。

提案手法の有効性に関する考察

表 4 より、モーフィングによる学習データの充足 (処理 2) を導入した手法 3 に対し、提案手法の正解率及び F-measure の値が全て上回っていることから、決定木の変更 (処理 1) とともに処理 2 を組み合わせて利用することの有効性が推測される。その理由について

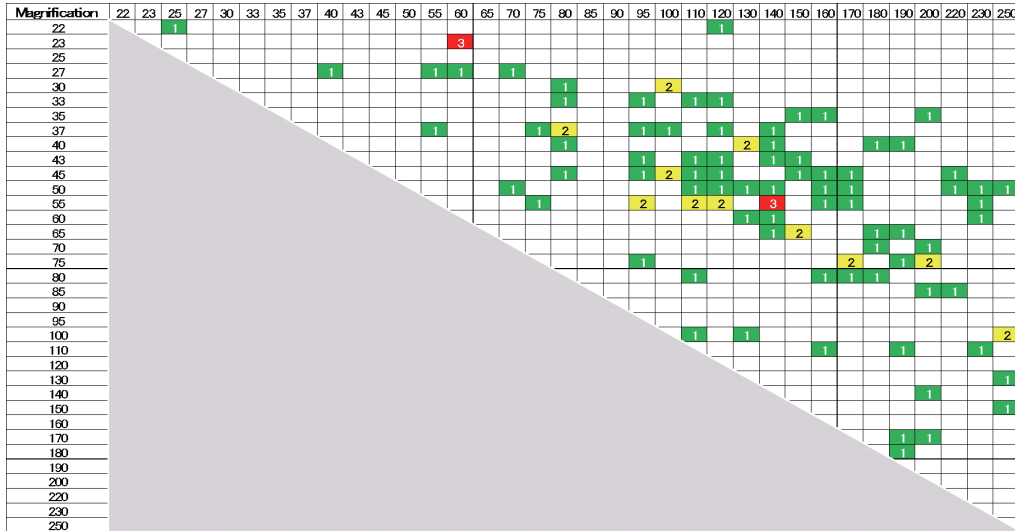


図 10 処理 2 においてモーフィングを行う際に用いる画像ペアの撮像倍率の関係
Fig. 10 The relationship of magnification of image pairs to be used in the morphing of procedure 2.

表 5 検証用ノードによる検証の前後におけるオサムシ上科 DT_6^2 とカミキリムシ科 DT_{11}^3 に分類精度の変化に関する比較

Table 5 Comparison of the classification accuracy of Caraboidea DT_6^2 and Cerambycidae DT_{11}^3 before and after varification by the verification node.

提案手法			
検証前	検証後	正	誤
		125	3
正			
誤		11	1

手法 2			
検証前	検証後	正	誤
		120	2
正			
誤		4	9

て以下に示す.

提案手法では, ある親ノード DT_j^i に, $|\mathcal{L}_j^i|$ 個の分類対象である子ノード $DT_{k+1}^i (k \in \mathcal{L}_j^i)$ が含まれ, それらが類似していることにより, 高精度な多クラス分類を行うことが困難であるため, 2 分木の構築 (処理 1-1) を導入した. ここで, モーフィングによる学習データの充足 (処理 2) を行った場合, 異なる倍率の間の画像を補間するため, 分類対象である子ノードの互いの相違度が初期学習データを用いた場合よりも低くなる. 実際に, 本実験における八つの上科に注目した場合, 28 個の組み合わせの中で, 処理 2 を導入することにより, 18 個の組み合わせで式 (1) の相違度が低くなったことが確認された. したがって, 処理 1-1 を導入するきっかけとなった問題が顕著となるため, 処理 2 を行う際, 処理 1-1 と組み合わせることが必要になると考えられる.

また, 図 8 からオサムシ上科 DT_6^2 とカミキリムシ

科 DT_{11}^3 の間に検証用ノードが追加 (処理 1-2) されていることが確認できる. ここで, 検証用ノードによる検証の前後におけるオサムシ上科 DT_6^2 とカミキリムシ科 DT_{11}^3 の分類精度の変化を表 5 に示す. 表 5 から, 提案手法及び手法 2 において検証用ノードによりオサムシ上科 DT_6^2 とカミキリムシ科 DT_{11}^3 の間で, 誤って分類されたテストデータが正しく修正された枚数は, 本来, 正しく分類されたテストデータが誤って分類された枚数より多いことが確認できる. また, 手法 2 より提案手法の方が, 分類精度の向上が顕著であることも確認できる. したがって, 処理 1-2 と処理 2 を組み合わせることの有効性が確認できる.

以上により, 提案手法では分類体系を表す決定木を用いて生物の分類を行う際, 決定木の変更 (処理 1) とともにモーフィングによる学習データの充足 (処理 2) を組み合わせることで, 分類精度の向上が実現された.

最後に, 提案手法に対して, SIFT 特徴量ではなく,

Histogram of Oriented Gradient (HOG) [15] 特徴量を用い、表 4 に示す比較手法と提案手法に対して実験を行った。ただし、用いた分類器は SVM である。その結果から、提案手法の精度が比較手法を上回っていることが確認された。具体的に、上科の正解率が 0.48 から 0.50 となり、科の正解率は 0.26 から 0.31 になった。したがって、提案手法は SIFT 特徴量だけでなく、HOG 特徴量を用いた際にも有効であることが確認でき、様々な特徴量を用いた場合においても、その精度の向上が期待できる。

5. む す び

本文では、分類体系に注目した決定木を用いた生物の分類法を提案した。提案手法ではまず、分類体系に注目し、分類体系の各ノードにその下位のノードへの分類を行う分類器を割り当てることで、構築された決定木を用いた生物の高精度な分類を可能とした。次に、更なる高精度化を実現するため、決定木の木構造の変更及び学習データの充足を行う処理を導入した。これらの処理を導入することで、誤分類が存在する場合、または、少ない初期学習データが与えられた場合においても、高精度な生物の分類を可能とした。本文の最後では、提案手法の有効性を確認するため、実際に SEM で撮像された昆虫の画像を対象とした実験を行った。その結果、従来の手法と比較し、提案手法はより高精度な生物の分類が可能であることが確認された。

謝辞 本研究の一部は、文部科学省科学研究費補助金新学術領域研究 (課題番号: 24120002) により行われた。

文 献

- [1] ジュディス・E・ウィンストン, 馬渡俊輔, 種を記載する-生物学者のための実際的な分類手順, 新井書店, 2005.
- [2] 馬渡俊輔, 動物分類学 30 講, 朝倉書店, 2006.
- [3] M.B. Blaschko, G. Holness, M.A. Mattar, Dimitri Lisin, P.E. Utgoff, A.R. Hanson, H. Schultz, E.M. Riseman, M.E. Sieracki, W.M. Balch, and B. Tupper, "Automatic in situ identification of plankton," Proc. 7th IEEE Workshops on WACV/MOTIONS, vol.1, pp.79-86, 2005.
- [4] C. Wen, D. Guyer, and W. Li, "Automated insect classification with combined global and local features for orchard management," ASABE Annual Meeting, 2009 Paper No.095865, 2009.
- [5] H. Zhang, Q. Huo, and W. Ding, "The application of adaBoost-neural network in storedproduct insect classification," IEEE International Symposium on IT in Medicine and Education, ITME, pp.973-976, 2008.
- [6] S.N. Gorb, Attachment devices of insect cuticle, Kluwer Academic Publishers, pp.1-305, Dordrecht, Boston, London, 2001.
- [7] T. Beier and S. Neely, "Feature-based image metamorphosis," SIGGRAPH 92 Conference Proceedings, vol.26, no.2, pp.35-42, 1992.
- [8] X. Wang and Z. Shi, "An improved algorithm for decision-tree-based SVM," 6th World Congress on Intelligent Control and Automation WCICA, vol.1, pp.4234-4238, 2006.
- [9] L. Lowe, "Distinctive image features from scale invariant key-points," International Journal of Computer Vision (IJCV), vol.60, no.2, pp.91-110, 2004.
- [10] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery 2, pp.121-167, 1998.
- [11] D. Tax and R. Duin, "Support vector data description," Machine Learning, vol.54, pp.45-66, 2004.
- [12] G. Csúrká, C.R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of key-points," ECCV Workshop on Statistical Learning in Computer Vision, pp.1-22, 2004.
- [13] C.M. Bishop, Neural networks for pattern recognition, Oxford University Press, 1995.
- [14] C.W. Hsu and C.J. Lin, "A Comparison of methods for multi-class support vector machines," IEEE Trans. Neural Netw., vol.13, pp.415-425, 2002.
- [15] N. Nalal and B. Triggs, "Histograms of oriented gradients for human detection," IEEE Conf. Computer Vision and Pattern Recognition, vol.1, pp.886-893, 2005.

(平成 26 年 7 月 28 日受付, 12 月 8 日再受付,
27 年 2 月 4 日早期公開)



朴 君 (学生員)

2013 北大大学院情報科学研究科修士課程了。現在、同大学大学院情報科学研究科博士後期課程在学中。画像処理に関する研究に従事。電子情報通信学会学生員。



小川 貴弘 (正員)

2003 北大・工・情報工学卒, 2005 同大大学院情報科学研究科修士課程了, 2007 同大学院情報科学研究科博士後期過程了, 2008 同大学情報科学研究科助教, 現在に至る. 2008 電子情報通信学会論文賞, ITE Transactions on Media Technology and Applications, Associate Editor 等を歴任. 画像復元及び映像処理に関する研究に従事. 電子情報通信学会, 映像情報メディア学会, EURASIP, IEEE 各会員. 博士 (情報科学).



長谷山美紀 (正員)

1988 北海道大学大学院工学研究科修士課程了, 同大学助手, 助教授 (1994-1995 アメリカワシントン大学客員准教授) を経て 2006 同大学院情報科学研究科教授. 博士 (工学). 画像・映像及び音響信号などマルチメディア信号処理の研究に従事. 2008 電子情報通信学会論文賞, 映像メディア学会副会長, ITE Transactions on Media Technology and Applications, Editor in Chief, 電子情報通信学会調査理事等を歴任. 電子情報通信学会, 情報処理学会, 映像情報メディア学会, IEEE, 信号処理学会各会員.