© 2012 一般社団法人映像情報メディア学会

論 文

SVDDを用いた顕微鏡画像からの新種深海底生物の検出および分類体系上の位置の推定法

Support Vector Data Dscription-Based Method for Finding New Benthic Species and Estimating Their Taxonomy Position from Microscopic Images.

長谷川 尭史†，正会員 小 川 貴 弘†，渡邉 日出海†，正会員 長谷山 美紀†

Takashi Hasegawa†, Takahiro Ogawa†, Hidemi Watanabe† and Miki Haseyama†

Abstract This paper presents a support vector data description (SVDD)-based method for finding new benthic species from microscopic images and its application to taxonomy position estimation. First, the proposed method generates hyperspheres that represent taxonomic species taxa of known species and enables automatic species classification. Furthermore, weight estimation of visual features based on multiple kernel learning (MKL) is used in this approach to realize automatic weighting of categorical traits that are traditionally determined by taxonomists. Next, based on the traditional taxonomic classification scheme, the proposed method merges the hyperspheres of similar species and generates new hyperspheres that represent ultra-species taxa in higher hierarchies. Then, from the obtained results, a new decision tree, whose nodes are hyperspheres of species taxa and ultra-species taxa, is constructed. By using this decision tree, new benthic species can be found from target samples, and their taxonomic positions can also be estimated.

キーワード：SVDD，決定木，生物多様性，新種生物探索，体系分類

1. ま え が き

地球上には，非常に多くの種類の生物が存在しており，国際自然保護連合*によると，現在の地球上に存在している生物は200万-5000万種程であると推定されている．このように，地球上の生物多様性**は充分に把握されていないことが報告されている[1]．特に，深海底は高水圧のため調査が困難であり，多くの領域が調査されていないことから，その生物多様性については，一部を除き未だ明らかにされていない．一方で，近年，高水圧に耐え2000m以上の深海底に到達することが可能なRemotely Operated Vehicle (ROV)が開発されている[2]．このROVの利用によって，これまでは不可能とされていた領域の深海底泥サンプルの取得が可能となり，さらに船上でハイビジョン映像をリアルタイムで取得・解析し制御することも可能となったことで，深海底の調査可能領域が以前と比較して大幅に拡大した．このような背景から，現在，深海底泥中に生息する生物の新種探索を目的とした生物多様性に関する研究が盛んに行われている．

現在の分類学において，新たに発見された生物個体に対する調査手順は**図1**に示すように行われる[1]．この図に示されるように，分類学者はまず，新たに発見された生物個体に対し，既知生物種との形態的な比較 (図中，点線で囲まれた部分) を行う．形態的に類似している場合には次のステップに進むが，形態的に既知の生物種群と明らかに異なる場合は新種の可能性のある生物種と判定し，生殖的隔離を確認した上で新種生物の記載を行う．このように分類学における形態的な分類は，重要な位置づけとなっている．また，既知の生物種のいずれかと形態的に類似している生物は，その後，DNAを用いた分子系統解析等の手法で調査される．しかしながら，DNAを用いた解析のためには，サンプル個体を破壊しなければならず，新種生物の登録の際に必要となるホロタイプ***として指定する個体標本を得ることができなくなる等の問題がある．深海底生物の調査においては，生物1種類あたりのサンプル個体数を多く望めないため，採集されたサンプルの中からまず新種の可能性の高いものを形態的な評価指標によって識別し，DNAを

2011年11月30日受付，2012年3月8日再受付，2012年4月24日採録
†北海道大学 大学院情報科学研究科
（〒060-0814 札幌市北区北14条西9丁目，TEL 011-706-6078）
†Graduate School of Information Science and Technology, Hokkaid University
(Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido, 060-0814)

* http://www.iicnredlist.org/static/stats

** すべての生物 (陸上生態系，海洋その他の水海生態系，これらが複合した生態系その他生息または生育の場を問わず．) の間の変位性．

*** 種を初めて記載する際に指定する唯一の標本であり，博物館や大学等の公的機関に預けられる．

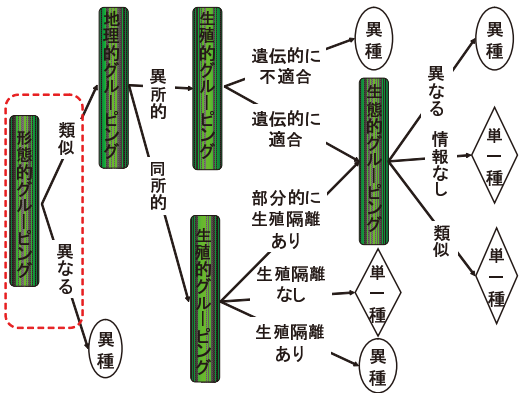

図 1　新たに発見された生物個体に対する調査手順．分類学者は新たな生物個体が発見された際に，まず形態的な分類を行う．

Fig.1 Investigation procedure of newly discovered organisms. Taxonomists firstly perform the morphological classification when new individual organisms are discovered.

用いた分子系統解析を行う対象のサンプル個体を最小限とすることが必要である．以上の理由から，特に深海底生物の調査においては形態的な調査を高精度に行うことがより重要となる．

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，新種の可能性のある生物の検出，および対象生物の分類体系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おり，判定の際の数量的判断基準や分類形質の重み付けは，分類学者個人の判断に委ねられている．そのため，例えば，同じ生物群での分類を試みる際も，取り組んだ分類学者により予想分類体系が異なるという問題が生じている [3)~5)]．特に深海底生物の調査においては，今後数多くの新種の発見が期待され，その都度分類体系が変化すると考えられるため，最適な分類の判断基準を設けることは容易ではない．

これらの問題に対する解決策として，生物が撮像された画像の処理により，上述の枠組みの自動化を行うことが有用であると考えられる．このとき，生物分類および新種の発見を行う際に，学習データとなる生物種の深海底サンプルは充分な個体数が得られないことから，汎化性を有する識別アルゴリズムを用いることが必要であると考えられる．また，これに加えて，識別の際に分類形質の重み付けの自動化を行うことで，分類の判断基準の設定を自動で行うことが可能となる．

これまでにも，文献 6)-8) 等，生物が撮像された画像の処理による生物分類システムがいくつか提案されている．しかしながら，それらは特定の生物門の中の既知の生物種同士の分類のみを行っており，また特徴の選択について考慮されていないことから，その精度が充分でないという問題があった．さらに，新たな生物が発見された際に，その生物の新種の可能性の評価を行い，分類体系上の位置を推定する手法はこれまでに提案されていない．

そこで本文では，深海底生物を撮像した顕微鏡画像から，Suport Vector Data Description (SVDD)[9)] を用いて新種の可能性のある生物を検出し，分類体系上の位置を推定する手法を提案する．提案手法ではまず，従来の分類学における種タクサ*を既知の生物種を正例としたSVDD による超球で表現し，これを用いて既知生物種の分類を可能とする．このとき，Multiple Kernel Learning (MKL)[13)] を導入し，各超球の算出において最適な特徴量の重み付けを行うことで，従来の分類学者が手動で行っていた生物種毎の分類形質の重み付けの自動化についても実現する．これにより，個体サンプルが少ない深海底の生物に対し，既知生物種の分類を自動で行うことが可能となる．次に，従来の分類体系の構築法に基づき，分類体系的に近いと考えられる生物種同士の SVDD 超球を順次統合することで，従来の分類学における種超タクサ**を表す SVDD 超球を算出し，得られた種タクサおよび種超タクサを表す超球をノードとした決定木を算出する．以上の手順で算出された決定木を用いて，調査対象生物の中から新種の可能性のある生物を検出し，その生物に対し分類体系上の位置を推定することが可能となる．これにより，従来の分類学における体系分類の自動化が実現され，新たな海域の生物多様性の調査を分類学者の専門的な知識を用いずに行うことが可能となる．

本文の構成を以下に示す．まず，**2** で提案手法において自動化の対象であり，現在の分類学において用いられている分類体系の構築法と，新しく発見された生物に対し分類体系上の位置を推定する方法について説明する．次に，**3** では提案手法で用いられる 1 クラス問題の識別手法である SVDD について説明する．さらに **4** では，顕微鏡画像から SVDD を用いて新種の可能性のある生物を検出し，分類体系上の位置を推定する手法について説明する．最後に **5** 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する．

2.　分類学における従来手法

本章では，文献 3)-5) で紹介されている従来の分類法について説明する．まず，**2.1** で分類学者が調査対象の生物群から分類体系を構築する方法について説明し，**2.2** で，新たに発見された種が未確定の生物に対し，分類体系上の位置を推定する方法について説明する．

2.1　分類体系の構築法

本節では，分類学者が調査対象生物のサンプルから分類体系を構築する方法について説明する．分類学者が形態的な特徴を基に生物を分類する際には，まず，手元にあるサンプルから周囲長，円形度等の形態的な特徴を表す形質の

* 生物の分類において，分類階級の中で種の段階に位置づけられる生物の集合，分類群

** 生物の分類において，分類階級の中で門，属等，種以上の段階に位置づけられる生物の集合，分類群

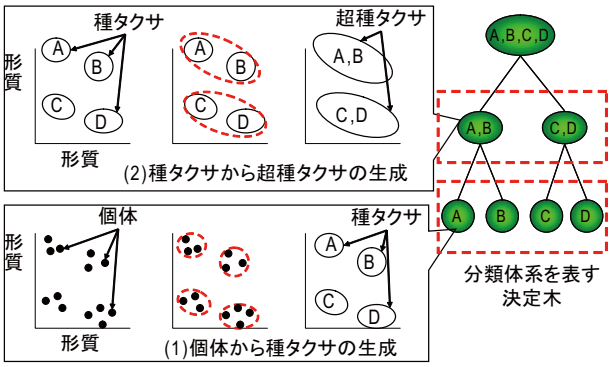

図 2 分類体系の構築手順の概念図．図中の (1),(2) でそれぞれ種タクサ，種超タクサを生成し，それらを用いた分類体系を構築する．

Fig.2 Conceptual diagram of the procedures for constructing the taxonomy. Species taxa and ultra-species taxa are respectively obtained by procedures (1) and (2), and the taxonomy is constructed from the obtained results.

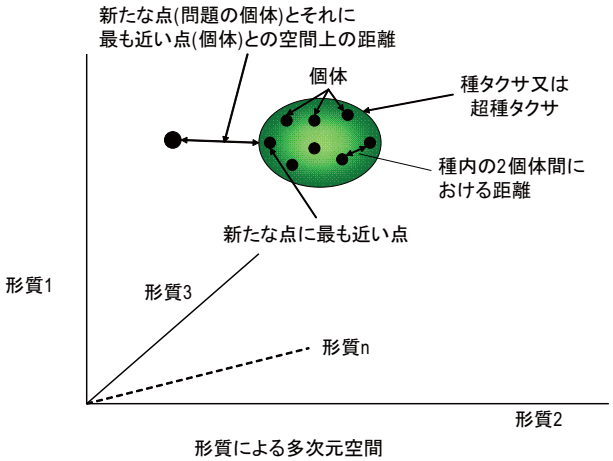

図 3 新たに発見された生物の各タクサに含まれるか否かの判定に関する概念図．個体の特徴ベクトルの集合として形成された種タクサまたは種超タクサの球に対し，新たな点とそれに最も近い点との距離と種内の 2 個体間における距離との比較により，新たな点が同種か別種かを判定する．

Fig.3 Conceptual diagram of the procedures for determining whether newly discovered organisms are included in each taxa. The determination is performed by comparing the distance between the newly discovered organism and the nearest sample in each taxa and that between the two samples within the taxa.

数量化を行い，形質を軸とした空間上にプロットする．このとき，生物種により重要視すべき分類形質は異なるため，正しい種分類を行うためには，最適な形質の選択および形質に対する重み付けを行う必要がある．次に，個体サンプルがプロットされた空間上では，複数のサンプルが固まって存在する領域があり，空間上の距離が近いサンプル同士をまとめることで種タクサを生成する (**図 2** の左下)．次に，得られた種タクサを単位とし，同様に種タクサ同士の空間上の距離を比較し，距離の近い種タクサ同士を統合することで種超タクサを生成する (図 2 の左上)．最終的に種超タクサが一つにまとめられるまで上述の作業を繰り返す．このとき，分類形質の重み付けおよびタクサを生成する際の距離評価の閾値等は，取り組んだ分類学者の判断に委ねられる．次に，得られたすべての種タクサおよび種超タクサをノードとした分類体系を表す決定木が構築される．決定木の算出は，まず種タクサを最下位ノードとし，次にタクサの統合順に上位ノードを生成する形で行われる．

2.2 分類体系上の位置を推定する方法

本節では，**2.1** で構築された分類体系を用いて，新たに発見された生物に対し，分類体系上の位置を推定する方法について説明する．新たに発見された種は既知種と同様に形質を数量化し，**2.1** で生成された各種タクサまたは種超タクサに含まれるか否かの判定を行う．このとき，いずれかの種タクサに含まれた場合は，既知の生物と同種であると判定し，それ以外は新種の可能性がある生物と判定される．具体的には，**図 3** に示されるように，種が未確定の生物の空間上の点と判定対象のタクサの中で最も近い既知生物種の個体の点との距離が，既知生物種内の個体点同士の距離より大きい場合，その生物をそのタクサの生物種とは異なる個体であると判定する．

3. Support Vector Data Description

本章では，提案手法で用いられる SVDD[9] についての説明を行う．SVDD は単一の対象クラスに属するサンプルを用いて学習を行い，入力したサンプルが対象クラスに属するか否かの識別を可能とする 1 クラス問題の識別手法である．SVDD では，識別器として特徴空間上において，対象クラスに属するサンプルを包含する最小の識別超球を算出する．この超球による識別は，入力したサンプルが内部に存在するとき対象クラスに属するサンプル，外部に存在するとき対象クラスに属さないサンプルとして判定が行われる．また，通常の SVDD では単一の対象クラスに属するサンプル (正例) のみを用いて超球を算出するが，対象クラスに属さないサンプル (負例) も同時に学習に用いることで，より高精度な識別超球の算出が可能であることが知られている [9]．このとき，得られる超球は，特徴空間上において正例を内部に，負例を外部に配置するものとなる．

正例 $\mathbf{x}_i^{pos}(i = 1, 2, ..., N^{pos})$ および負例 $\mathbf{x}_j^{neg}(j = 1, 2, ..., N^{neg})$ が与えられたとき，最も単純な超球は，**図 4** に示されるように，正例をすべて内包し，負例を含まない最小の超球として定義される．したがって，次式に示される 2 次計画問題を解くことにより超球の算出を行う．

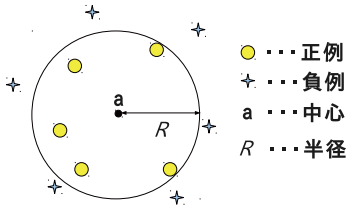

図 4　SVDD により得られる超球の概念図
Fig.4 Conceptual diagram of the hypersphere obtained by SVDD.

$$\min_{R,\mathbf{a},\xi^{pos},\xi^{neg}} R^2 + C^{pos}\sum_{i=1}^{N^{pos}}\xi_i^{pos} + C^{neg}\sum_{j=1}^{N^{neg}}\xi_j^{neg}$$
$$\text{s.t. } ||\phi(\mathbf{x}_i^{pos}) - \mathbf{a}||^2 \leqq R^2 + \xi_i^{pos},$$
$$||\phi(\mathbf{x}_j^{neg}) - \mathbf{a}||^2 \geqq R^2 - \xi_j^{neg},$$
$$\xi_i^{pos} \geqq 0, \quad \xi_j^{neg} \geqq 0, \quad \forall i,j \tag{1}$$

ここで，$\mathbf{a}$ は超球の中心ベクトル，R は半径である．また，C^{pos} および C^{neg} は体積と誤りのトレードオフを制御するパラメータであり，ξ_i^{pos} および ξ_j^{neg} はスラック変数である．また，$\phi(\cdot)$ は高次元特徴空間 [10] への非線形写像を表す．このとき，高次元空間へ非線形写像したベクトルの内積を，低次元な原空間上の演算によって求めることを可能とするため，カーネル関数が用いられる．カーネル関数は以下の式で表される．

$$k(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x})'\phi(\mathbf{y}) \tag{2}$$

カーネル関数は現在様々なものが提案されているが，使用するカーネル関数の種類やその際に設定するパラメータは学習の精度に大きく影響するため，学習の際には適切なカーネル関数およびパラメータを選択する必要がある．

以上の学習により求められる超球の中心 $\mathbf{a}$ および半径 R を用いて，識別対象として入力されたサンプル $\mathbf{x}$ が対象クラス $\boldsymbol{P}$ に属するか否かを次式により決定する．

$$\begin{cases} \mathbf{x} \in \boldsymbol{P} & \text{if } f_{svdd}(\mathbf{x}) \geqq 0 \\ \mathbf{x} \notin \boldsymbol{P} & \text{otherwise} \end{cases}$$
$$f_{svdd}(\mathbf{x}) = R^2 - \| \phi(\mathbf{x}) - \mathbf{a} \|^2 . \tag{3}$$

上式において，出力値 $f_{svdd}(\mathbf{x})$ は $\phi(\mathbf{x})$ と超球の中心 $\mathbf{a}$ との距離 $|| \phi(\mathbf{x}) - \mathbf{a} ||^2$ に対して単調に減少する．したがって，$f_{svdd}(\mathbf{x})$ が大きな値をとるとき，特徴空間上における $\phi(\mathbf{x})$ と $\mathbf{a}$ との距離は小さくなる．超球の中心 $\mathbf{a}$ は，特徴空間上における対象クラスの分布の中心であるため，識別対象から算出される特徴ベクトルが特徴空間上において中心に近くなるほど，識別対象は対象クラスと類似した特徴を持つこととなる．

4. SVDD を用いた新種の可能性のある深海底生物の検出および分類体系上の位置の推定法

本章では，深海底の生物を対象に，顕微鏡画像から SVDD を用いて新種の可能性のある生物を検出し，さらに新種の可能性があると判定された生物に対して，分類体系上の位置を推定する手法について説明する．提案手法の流れを**図 5** に示す．提案手法ではまず，顕微鏡画像から抽出される生物部位に対して生物の形態に着目した特徴量を求め，算出された特徴ベクトルを用いた SVDD による学習により，既知生物種毎の超球を算出する．このとき，学習においては，既知の生物種毎に Multiple Kernel Learning[13] を用いて，特徴量の重み付けを反映させた最適なカーネル関数を用いる．次に，算出された既知の生物種毎の超球を用いて分類体系を表す決定木を算出する．最後に，算出された決定木を用いて種が未確定の生物の中から新種の可能性のある生物を検出し，同時に分類体系上の位置の推定を行う．

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，**4.1** で顕微鏡画像から生物領域を抽出し，さらに抽出された生物領域から算出する特徴量について説明する．次に，**4.2** では，SVDD を用いて種タクサを表す超球を算出する方法について説明する．さらに，**4.3** では，従来の分類体系の構築法に基づき，SVDD を用いて分類体系を表す決定木を構築する方法について説明する．最後に，**4.4** では，**4.3** で算出された決定木を用いて，調査対象生物の中から新種の可能性のある生物を検出し，さらに同時にその生物の分類体系上の位置を推定する手法について説明する．

4.1 顕微鏡画像からの生物領域の抽出および特徴量の算出法

本節ではまず，顕微鏡画像から生物領域を抽出する方法を説明し，次に提案手法においてその領域から算出する特徴量について説明する．まず，生物が 1 個体ずつ撮像されている顕微鏡画像から，生物部分と背景を区別するための 2 値画像を生成する．一般に生物を対象とした顕微鏡画像では，上から顕微鏡の照明を照射するため，生物領域と非生物領域で輝度値に差が生じる．そのため，提案手法では，背景と生物部分の輝度値の差を利用し，大津の手法 [11] による 2 値画像生成を行う．次に，生成された 2 値画像から生物部位の輪郭画素の位置を求める．深海生物を対象とした顕微鏡画像例，およびその画像から生成された 2 値画像，算出した輪郭を朱線で描写した画像を**図 6** に示す．

次に，生成された 2 値画像と輪郭から特徴量を算出する．生物の種類を判定する特徴としては，幾何学的な特徴や色特徴等が考えられるが，深海底で得られたサンプルは通常，エタノールによる固定により脱色され本来の色を失う等，保存方法により変化が生ずる場合がある．そこで提案手法では，幾何学的な特徴のみに着目し，回転不変かつスケール不変な以下の 7 種類の特徴量を算出する (**表 1**)．以下で，各々の特徴量について説明する．

特徴量 1　生物領域と輪郭領域の面積比

まず，形態を表す特徴として，得られた 2 値画像および輪郭から，生物領域の画素数と輪郭画素数の比が考

表 1　提案手法において算出される特徴量の概要
Table 1 Overview of the features calculated by the proposed method.

特徴量	説明
生物領域と輪郭領域の面積比	2 値画像から生物領域の画素数を求め，輪郭の画素数との比を算出
外接矩形の縦横比	2 値画像から生物を内包する最小の矩形を算出し，矩形の長辺と短辺との比を算出
外接矩形と生物の面積比	算出された外接矩形内の画素数と生物領域画素数の比を算出
外接凸包と生物の面積比	2 値画像から生物領域を内包する最小の凸包を算出し，凸包内の画素数と生物領域の画素数の比を算出
頭部と尾部の面積比	外接矩形の重心座標を求め，重心座標を通り外接矩形の長辺と直行する直線で生物を 2 分割し，それらの面積の比を算出
輪郭のモーメント	算出された 2 値画像のモーメントとして，回転，スケール不変性を有する Hu 不変モーメント [12] を算出
生物の実サイズ	顕微鏡画像取得ソフトを用いて撮像する際に記録される倍率情報を用いて，画像 1 画素あたりの面積を算出し，それを用いて生物の実際のサイズを算出

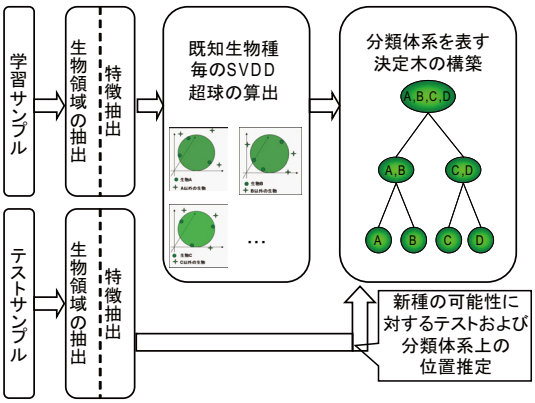

図 5　提案手法の流れ．提案手法ではまず，学習サンプルとテストサンプルから同様の方法で生物領域を抽出し，特徴量を求める．次に，学習サンプルの特徴ベクトルから種タクサを表す SVDD 超球を算出し，それを用いて分類体系を表す決定木を算出する．最後に，算出された決定木を用いてテストサンプルに対し新種の可能性に対する判定および分類体系上の位置の推定を行う．

Fig.5 Brief procedures of the proposed method. First, the proposed method extracts regions including organisms from training and test samples in the same manner, and their features are calculated. Futhermore, the hyperspheres of SVDD representing taxa are calculated from feature vectors of the training samples, and the decision tree representing the taxonomy is constructed from them. Finally, from the obtained decision tree, the detection of new species and their taxonomy position estimation are performed.

えられる．この特徴量は，細長い生物では値が小さく，逆の場合は値が大きくなると考えられるため，生物の大まかな太さを表現できると考えられる．また，足等がなく形態が単純なものと，複雑なものとの間でもその値に差が出ると考えられる．具体的には，生物領域の画素数を S_1，輪郭の画素数を S_2 としたとき，生物領域と輪郭領域の面積比 f_1 は以下の式で算出される．

$$f_1 = \frac{S_1}{S_2} \tag{4}$$

特徴量 2　外接矩形の縦横比

(a)　(b)　(c)

図 6　(a) 深海底生物を対象とした顕微鏡画像例．(b)2 値画像．(c) 算出した輪郭を朱線で描写した画像

Fig.6 (a)Example of a microscopic image including a benthic organism, (b)Binary image, (c)Image whose extracted contors are shown in red.

生物の形態は多種多様であるが，大まかな形態の特徴として，先にも述べた通り生物の体の細長さは重要な特徴と考えられる．縦横比のみでは線形動物門等，体が湾曲している生物を完全に特徴づけることは困難であるが，円形度の高い生物等との差を表現する際には有効であると考えられる．そこで，提案手法では，算出された輪郭画素を内包する最小の矩形 (外接矩形) を算出し，その矩形の長辺と短辺の比を算出することで生物の縦横比を求める．外接矩形の長辺と短辺の長さをそれぞれ L_l，L_s とすると，外接矩形の縦横比 f_2 は以下の式で表される．

$$f_2 = \frac{L_l}{L_s} \tag{5}$$

特徴量 3　外接矩形と生物の面積比

上記で特徴量 2 として外接矩形の縦横比を算出したが，縦横比が近い値を持つ生物種同士でも，その中に占める生物領域の割合は様々であり，外接矩形と生物の面積比を算出することで，それらの差を表現することが可能であると考えられる．具体的には，外接矩形と生物の面積比 f_3 を以下の式で算出する．

$$f_3 = \frac{L_l \times L_s}{S_1} \tag{6}$$

特徴量 4　外接凸包と生物の面積比

算出された輪郭情報を用いて，生物を内包する凸包 (外接凸包) を求めることが可能である．外接凸包は上述で求めた外接矩形と比較し，より生物の形態に沿った形となる．この外接凸包と生物領域の面積の比を求めることで，生物の中でも足等の形により輪郭が複雑な

ものと，凹凸の少ないものとの差が表現可能であると考えられる．凸包内の画素数を S_3 としたとき，外接凸包と生物の面積比 f_4 は以下の式で表される．

$$f_4 = \frac{S_1}{S_3} \tag{7}$$

特徴量 5　頭部と尾部の面積比

生物の中には，線形動物門等のように頭部から尾部まで全体的に体幅が変わらないものや，逆にクマ目のように極端に頭部が大きく尾部が細長い生物が存在する．したがって，生物領域を中心で 2 分し，頭部と尾部の面積の比を算出すれば，これらの生物を有効に分けることが可能であると考えられる．ここで，提案手法では既に外接矩形を求めているため，生物領域をその外接矩形の長辺で 2 分し，それらの比を求めることで，頭部と尾部の面積比を算出することが可能であると考えられる．以上のようにして求めた頭部の面積画素数を S_4，尾部の面積画素数を S_5 としたとき，頭部と尾部の面積比 f_5 は以下の式で算出される．

$$f_5 = \frac{S_4}{S_5} \tag{8}$$

特徴量 6　輪郭のモーメント

生物の形態の特徴を表す指標として，モーメントの算出が有効であると考えられる．モーメントを表す特徴量は現在様々なものが提案されているが，本手法ではその中でも回転，スケール，平行移動に対し頑健なものを用いる必要がある．したがって，以上の三つに対して頑健である Hu モーメント不変量 [12] を輪郭のモーメント f_6 として用いる．

特徴量 7　生物の実サイズ

実際の生物の分類において，その生物の大きさは非常に重要な特徴となる．例えば，形態が似た生物種同士でも，大きさが異なる場合，同じ生物種である可能性は低い．また，分類体系においても通常個体の大きさが近いもの同士が分類体系上で近縁となっている場合が多い．以上の理由から，生物の実サイズは非常に有用な特徴量であると考えられる．ここで，顕微鏡画像取得ソフトを用いて撮像する際に倍率情報が記録されるため，その倍率情報を用いて 1 画素あたりの面積 S_{pixel} の算出が可能である．この S_{pixel} と生物領域の画素数 S_1 を用いて，以下の式により生物の実サイズ f_7 を求める．

$$f_7 = S_{pixel} \times S_1 \tag{9}$$

4.2　SVDD を用いた種タクサを表現する超球の算出法

本節では，SVDD を用いて従来の分類学における種タクサを表現する超球を算出する手法について説明する．提案手法では，**4.1** で算出された特徴量を用いて特徴ベクトルを生成し，既知生物種の中の一種類の特徴ベクトルを正例，それ以外の生物種を負例とした SVDD 超球を既知生物種毎に算出する．このとき，各超球の算出において最適な特徴量の重み付けを行うため，MKL による特徴統合を導入する．MKL は近年用いられている特徴統合法であり，複数の特徴量のカーネルを線形結合することにより統合カーネルを生成する．最適なカーネルのサブカーネルに対する重み β_j を学習する問題は，以下の式のように示される．

$$K_{combined}(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{K} \beta_j k_j(\mathbf{x},\mathbf{y})$$
$$with \qquad \beta_j \geqq 0, \sum_{j=1}^{K} \beta_j = 1. \tag{10}$$

ここで，K はカーネル数 (特徴量数) を表す．提案手法では，各特徴量におけるカーネルの重み係数を決定する方法として，文献 14) で提案されている MKL を SVM のフレームワークで解く手法を SVDD に対し適用する方法を導入する．具体的な手順を以下に示す．

Step1　β_j を均等重みとする．

Step2　β_j を固定して統合カーネルを単一のカーネルとみなし，通常の SVDD 学習を行いサポートベクターの重み係数 $\alpha_i (i = 1, .., N)$ を求める．ただし，α_i は文献 9) で用いられているものと同じものである．

Step3　求めた α_i を固定して，全学習データの識別境界面までの距離の合計値が増加するように β_j を変化させる．

Step4　終了条件に達するまで Step2, Step3 を繰り返す．

ここで，N はサポートベクター数である．各超球の算出の際には，このように統合されたカーネルを用いる．これにより，従来の分類学者が手動で行っていた生物種毎の重み付けの自動化が実現される．また，従来の分類学における種タクサとして，1 クラス識別器の SVDD における超球を用いることで，従来手法の概念に基づいたタクサの表現が可能となる．さらに，SVDD は学習の際に正例だけでなく，他の生物種を負例として学習することが可能なため，1 クラス識別器の中で最も高精度な分類が可能となる．

4.3　分類体系を表す決定木の算出法

本節では，従来の分類体系の構築法に基づき，分類体系を表す決定木を算出する手法について説明する．提案手法では，まず，**4.2** で算出された種タクサを表す各既知生物種の SVDD 超球を用いて，従来の分類体系の構築法における種超タクサを表す SVDD 超球を算出する．ここで，従来の手法では，各種タクサ同士の距離が近いものから順に統合し種超タクサを生成するため，これに対する自動化の方法としては，各既知生物種を正例とした SVDD 超球の中心ベクトルからの距離を比較する方法等が考えられる．しかしながら，提案手法では，各超球を特徴選択により異なる特徴空間上で算出しているため，各超球間の距離を直接評価することはできない．したがって，提案手法では，複数

の生物種を正例とした超球に対する分類精度を算出し，精度の最も高い超球で正例となっている生物種を分類体系上の距離が近い生物とみなし超球の統合を行うことで，種超タクサを表現する超球を生成する．具体的な手順を以下に示す．

Step1　種タクサまたは種超タクサを表す SVDD 超球の中から二つを選択し，各超球の正例のサンプルの両方を正例，その他を負例とした SVDD 超球を算出し，交差検定により分類精度を求める．

Step2　**Step1** の作業を存在する超球ですべての組合せに対して行い，最も分類精度の良い超球を決定する．

Step3　**Step2** において決定された超球を種超タクサを表現する超球とし，**Step1** に戻る．

Step4　すべての超球が統合されるまで **Step1-Step3** を繰り返す．

これにより，様々な決定木のパターンの中で精度の高い決定木を算出することが可能となる．

次に，上述により算出された種タクサおよび種超タクサを表す SVDD 超球を用いて，分類体系を表す決定木を算出する．算出される決定木では，まず種タクサを表す超球を最下位ノードとし，超球の統合順に上位ノードとする．最上位ノードは，すべての既知の生物種を正例とした超球となる．これにより，従来すべて人手で行われていた分類体系の構築の自動化が実現される．また，超球の統合の際に最も精度の良い組合せで統合を行っているため，考えられる 2 分木の中で最も高精度な分類が可能となる．以上により，個体サンプル数が少なく，分類体系の見直しのサイクルが短い深海底生物調査において，統一した評価による分類体系の構築が実現される．

4.4　新種の可能性のある生物の検出および分類体系上の位置の推定法

本節では，**4.3** により算出された決定木を用いて，種が未確定の生物 (調査対象生物) の中から新種の可能性のある生物を検出し，同時にその生物の分類体系上の位置を推定する手法について説明する．具体的な手順の概念図を**図 7** に示す．まず，学習サンプルと同様に算出した調査対象生物の特徴ベクトルに対し，決定木の最上位ノードを構成する超球に含まれるか否かの判定を行う．超球に含まれた場合は，次のノードに進み次のノードを構成する二つの超球に対する判定を同様に行う．含まれない場合は，新種の可能性のある生物と判定する．同様に最下位ノードまで上述の作業を繰り返し，最終的に最下位ノードに含まれた場合は，その超球を算出する際に正例とした生物と同種であるとみなし，決定木のいずれかのノードで超球に含まれなかった場合は，新種の可能性のある生物種と判定される．これにより，新種の可能性があるか否かだけでなく，決定木の中のどのノードで外れたかを確認することにより，その生物の分類体系上の位置を推定することが可能となる．

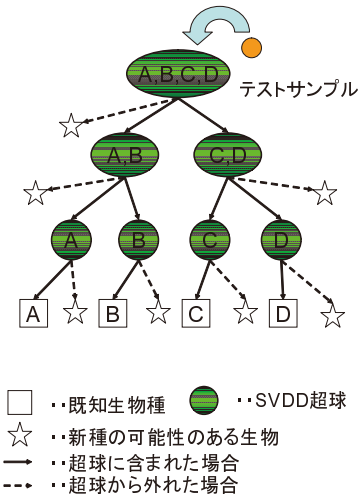

図 7　分類体系を表す決定木によりテストサンプルの新種可能性を評価する手順の概念図．調査対象生物の特徴ベクトルに対し，決定木の最上位ノードから順に超球に対するテストを行い，新種の可能性があるか否かの判定，および分類体系上の位置の推定を行う．

Fig.7 Conceptual diagram of the procedure for finding new species from test samples based on the decision tree representing the taxonomy. By testing the provided samples based on the calculated hypersperes from the top node in the decision tree, the new species detection and the taxonomy position estimation are performed.

5.　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 GCOE 新種探索プロジェクト [2] で採集されたベーリング海深海底泥中に含まれる生物を対象とした．なお，用いた顕微鏡は LEICA M205C であり，撮像は LEICA DEC290 を用いて行った．

取得画像で既知生物種において，30 枚以上の画像が得られたものについて，形態的な特徴を基に分類し，既知生物種に含まれないものを未知生物種とした．結果，A,B,C,D,E,F の 6 種類の既知生物種における画像数はそれぞれ，A70 枚，B93 枚，C123 枚，D75 枚，E58 枚，F35 枚，未知の生物種 20 枚となり合計 474 枚となった．各既知生物種の画像例および未知生物種の画像例をそれぞれ**図 8**，および**図 9** に示す．

本実験では，すべての顕微鏡画像に対して正解データ (Ground Truth) を与えるため，専門知識を有する者が A〜F の 6 種類および未知生物種の分類を行った．なお，A〜F の生物種は，具体的に，下記のように定義される．

生物 A：Nematoda
生物 B：Harpacticoida
生物 C：Cumacea
生物 D：Gammaridea

生物 E：Ampeliscidae
生物 F：Crustacea
未知生物種：上記生物 A～F に属さない生物
以上のことから，未知生物種は，生物 A～F に属さない生物であり，真に新種であるとは限らない．本文の実験において，実際に新種を Ground Truth として与えることは困難であったため，我々は未知生物種を上記のように設定した．したがって，未知生物種は，生物 A～F 以外の既知の生物種である可能性がある．なお，提案手法において，実際に新種の発見を行う際の作業は，本実験での作業から変化しないため，上記設定による妥当性の低下はないものと考えられる．

次に，本実験における Ground Truth の作成について述べる．分類学における従来法では，分類形質の重み付けおよびタクサを生成する際の距離評価の閾値等を決めていくことで，生物の分類を可能としている．したがって，実際に Ground Truth を生物の画像に対して付与していく際には，これらの手続きを行う必要がある．しかしながら，本実験で用いた生物 A～F は先にも述べたとおり，その生物種名が既にわかっていることから，直接専門的知識を有する者が判断し，分類を行った．つまり，上記手続きを省略した形で Ground Truth の作成を行った．

前章までに示したように，提案手法では種タクサを SVDD の超球で表現し，さらにこれに基づいた決定木の構築により新種の可能性のある生物の検出と分類体系上の位置の推定を可能とする．したがって，次の二つの実験により，手法の有効性を確認する．**実験 1** において種タクサを SVDD の超球で表現することの有効性を検証する (**5.1**)．具体的に，既知生物種 A～F 毎に SVDD の超球を算出し，これを用いた生物分類の精度および新種の可能性のある生物の検出精度を評価する．**実験 2** では，提案手法における決定木の構築に関する有効性の検証を行う (**5.2**)．具体的に，SVDD の超球に加えて構築される決定木を用いて，**実験 1:**と同様に，生物分類精度と新種の可能性のある生物の検出精度を評価し，決定木構築の有効性について検証する．なお，新種の生物を実際に取得し，真の分類体系上の位置を与えることは困難であることから，本実験において生物の分類体系上の位置の推定の精度検証は，決定木を用いた際の種タクサおよび種超タクサを表現する各超球における生物の分類の精度で評価を行うこととする．また，本実験において提案手法で使用するカーネル関数は，RBF カーネルとした．

5.1 実験 1:SVDD 超球による種タクサの表現に関する評価

本節では，提案手法において種タクサを SVDD の超球で表現することの有効性を検証する．本実験ではまず，既知生物種 A～F に対し，OCSVM[15] と SVDD を用いた場合の分類精度をそれぞれ確認し，次に，それらのテスト画像と未知生物種の画像を用いて，新種の可能性のある生物の検出精度を算出する．ここで，OCSVM は，1 クラス識別器に関するベンチマーク手法であることから，提案手法の比較として有効であると考えられる．なお，本実験の目的は，1 クラス問題の識別手法のベンチマークである OCSVM と比較した場合においても同様の精度が出力されるという点で，SVDD の利用の妥当性を示すものである．したがって，SVDD が SVM の考え方をベースとする識別器の中で最も高い精度を保証することを示すものではない．

表 2　10 分割交差検定による学習画像数とテスト画像数
Table 2 The number of training and test images used in the 10-fold cross-validation.

	学習画像数	テスト画像数
A	63	7
B	84	9
C	111	12
D	68	7
E	53	5
F	32	3

本実験において，既知生物種の分類は，各生物種毎に識別器の構築を行い，対象の生物に対してすべての識別器により識別を行うことで，どの生物であるかの判定を行った．なお，複数の生物種に含まれると判定された場合には，識別器の出力が最も大きなクラスに分類することとした．また，新種の可能性のある生物の検出については，対象の生物がすべての生物種の識別器を用いた際に，どのクラスにも属さないと判定された場合に，新種であると判定することとした．先にも述べた通り，本実験において，真の新種をテストデータとして与えることは困難であることから，未知生物種は，生物 A～F 以外の生物種を与えている．ただし，それらは確実に A～F 以外の生物であることが専門の知識を有する者によって確認されている．したがって，本文において，我々は A～F 以外の生物を「新種の可能性のある生物種」とみなし，実験を行った．なお，実際に提案手法によって新種生物を発見する際には，本実験と同様の手続きが行われることとなる．したがって，検証，評価の方法として本実験は妥当性を欠くものにはならないと考え，以上の評価方法を採択した．

本実験では，分類精度および識別精度の定量評価として 10 分割交差検定を採用し，評価のための学習サンプルとテストサンプルは異なる画像を利用して実験を行った．具体的な学習サンプルとテストサンプルの数を，**表 2** に示す．10 分割交差検定による再現率 (recall)，適合率 (precision)，F 値 (F-measure) を次式で算出した．

$$\text{recall} = \frac{\text{検出されたうち正解である数}}{\text{全正解数}} \tag{11}$$

$$\text{precision} = \frac{\text{検出されたうち正解である数}}{\text{全検出数}} \tag{12}$$

$$\text{F-measure} = \frac{2.0 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \tag{13}$$

既知生物種の分類精度，新種の可能性のある生物に対する識別精度の結果をそれぞれ**表 3**，**表 4** に示す．結果より，提

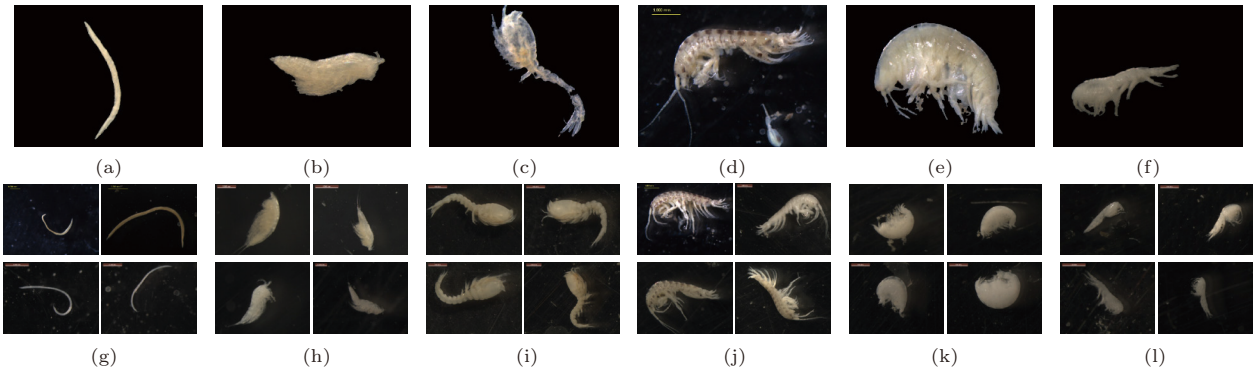

図 8 実験で用いた既知生物種の画像例．上段 (a)～(f):A～F それぞれの画像例，下段 (g)～(l):A～F 生物種毎の他の四つのサンプル画像例．
Fig.8 Example images of known species used in the experiment. Top (a)～(f): Example images of A～F, Bottom (g)～(l): Other four examples of each species.

図 9 実験で用いた未知生物種の画像例
Fig.9 Examples of unknown species used in the experiment.

表 3 実験 1 における OCSVM と SVDD(提案手法) の分類精度
Table 3 Classification accuracy of OCSVM and SVDD(proposed method) in Experiment 1.

	精度 \ 生物種	A	B	C	D	E	F	平均
OCSVM	recall	0.96	0.87	0.84	0.61	0.95	0.77	0.83
	precision	1.00	1.00	0.73	0.96	0.98	1.00	0.95
	F-measure	0.98	0.93	0.78	0.75	0.96	0.87	0.88
SVDD	recall	0.96	0.86	0.87	0.76	0.95	0.86	0.88
	precision	0.99	0.92	0.91	0.79	0.98	0.91	0.92
	F-measure	0.97	0.89	0.89	0.77	0.97	0.88	0.90

表 4 実験 1 における OCSVM と SVDD(提案手法) の新種の可能性のある生物の検出精度
Table 4 Detection accuracy of new species by OCSVM and SVDD (proposed method) in Experiment 1.

	recall	precision	F-measure
OCSVM	0.84	1.00	0.91
SVDD	0.87	0.98	0.92

表 5 OCSVM および SVDD の学習の際の MKL のパラメータ
Table 5 Parameters of MKL obtained by OCSVM and SVDD.

		1	2	3	4	5	6	7
A	OCSVM	0.16	0.15	0.15	0.11	0.16	0.09	0.17
	SVDD	0.12	0.10	0.08	0.13	0.20	0.12	0.26
B	OCSVM	0.14	0.08	0.08	0.10	0.40	0.10	0.10
	SVDD	0.12	0.13	0.07	0.11	0.40	0.08	0.08
C	OCSVM	0.13	0.07	0.08	0.10	0.21	0.14	0.26
	SVDD	0.12	0.10	0.08	0.13	0.20	0.12	0.26
D	OCSVM	0.16	0.09	0.09	0.09	0.30	0.17	0.10
	SVDD	0.17	0.13	0.05	0.10	0.33	0.11	0.10
E	OCSVM	0.09	0.21	0.19	0.08	0.18	0.06	0.20
	SVDD	0.10	0.23	0.19	0.11	0.15	0.02	0.19
F	OCSVM	0.09	0.09	0.07	0.21	0.32	0.03	0.20
	SVDD	0.13	0.11	0.04	0.24	0.25	0.02	0.22

案手法で用いられる SVDD は，高精度に既知生物種の分類および新種の可能性のある生物の識別が可能であることが確認できる． ここで，提案手法で推定されるパラメータとして，分類形質の重み付けのために与えられる SVDD による学習の際の MKL のパラメータが存在する．そこで，学習の際の MKL のパラメータを OCSVM の結果と共に**表 5** に示す．なお，これらの値は学習により自動で求まった値となる．また，タクサを生成する際の距離評価の閾値については，SVDD における超球の半径が該当すると考えられ，これらの値についても学習により自動で決定されることとなる．

ところで，表 3 より，SVDD では，生物 C が他の生物と比較して精度が大きく向上していることがわかる．特に，recall の変化に対し、precision の精度が大きく上がっていることが確認できる．この原因として，OCSVM は基本的に正例のみで学習を行っているのに対して，SVDD は SVDD-neg を利用し，負例も有効に学習に用いていることから，それにより超球の高精度な推定が可能であったと考えられる．したがって，precision が向上し，それに伴い F-measure の値も向上したと考えらる．また一方で，生物 D の精度が OCSVM，SVDD でともに低くなっている．生物 D に関する分類結果を調査した結果，生物 C,E,F との誤りが多く存在した．**図 8** を参照しても，それらの形態的特徴が比較的類似することから，高精度な分類が困難であったと考えられる．その中でも C については，特徴量 1 および特徴量 2 が C と D の間で比較的近い値となっていたこと等が原因

表 6　実験 2 の各生物における提案手法の分類精度
Table 6 Classification accuracy of each species by the proposed method in Experiment 2.

精度 \ 生物種	A	B	C	D	E	F	平均
recall	0.96	0.93	0.95	0.95	0.95	0.94	0.95
precision	0.99	0.96	0.97	0.95	0.98	0.96	0.97
F-measure	0.97	0.95	0.96	0.95	0.96	0.95	0.96

表 7　実験 2 の提案手法における新種の可能性のある生物の検出精度
Table 7 Detection accuracy of new species by the proposed method in Experiment 2.

	recall	precision	F-measure
提案手法	0.89	0.99	0.94

であると考えられる．また，矩形の形状が C,D,F で比較的近いものとなることや，E とは特徴量 5 が類似すること等が原因として考えられる．

提案手法では，従来の分類学で用いられている種タクサの性質に基づいていて，SVDD の超球を種タクサを表現する識別器として利用している．つまり，従来より分類学では，SVDD に類似した超球の考え方，つまりタクサという球を用いた分類が行われていることから，我々はそれらに従って，SVDD の超球を用いた分類が妥当であると考えた．したがって，既知の生物種に対してそれぞれ SVDD の超球が算出され，従来の分類学における生物分類および新種の可能性のある生物の識別を自動で行うことが可能となる．以上のことから，我々は本文で SVDD の超球を用いた分類について検討を行い，実験において，その妥当性について検証した．

5.2　実験 2:分類体系を表す決定木の構築に関する評価

本節では，提案手法において算出される分類体系を表す決定木の有効性を検証する．具体的には，提案手法により算出された決定木を用いて，**4.4** で説明した方法で，既知の生物種に対する分類精度および調査対象生物に対する新種の可能性のある生物の検出精度を算出する．既知の生物種分類および新種の可能性のある生物種の検出に関する recall, precision，F-measure をそれぞれ**表 6** および**表 7** に示す．

これらの実験結果から，既知生物種同士の分類精度および新種の可能性のある生物の検出の精度が高いことが確認できる．これは，分類体系を表す決定木の上位ノードにおいて，既知の生物種に含まれない生物を効率よく分類できることが要因であると考えられる．このとき，決定木は分類精度が最大となるように，既知の生物種の超球を統合していくことで構築されおり，従来の分類学における種超タクサに対応する超球を逐次算出していくことで，生物分類および新種の可能性のある生物種の検出のための分類体系を自動で求めることが可能となっている．以上により，扱う既知の生物種が多くなった場合でも，高精度な分類を行うことが可能となる．また，決定木で超球外となった位置の確認を行うことで，その生物の分類体系上の位置推定が期待できる．

6. むすび

本文では，SVDD を用いて新種の可能性のある生物を検出し，分類体系上の位置を推定する手法を提案した．提案手法ではまず，従来の分類学における種タクサを表す球を既知の生物種を正例とした SVDD 超球で表現し，これを用いて既知生物種同士の高精度な分類を可能とした．次に，種タクサおよび種超タクサを表す SVDD 超球をノードとした分類体系を表す決定木を算出し，これを用いることで，更に高精度な既知生物種同士の分類，および調査対象生物の中から新種の可能正のある生物の検出し，その分類体系上の位置の推定を行うことを可能とした．本文の最後では，提案手法の有効性を確認するため，実際に採集された深海底泥中に含まれた生物の顕微鏡画像を対象とした実験を行った．結果，従来の方法と比較し，より高精度な既知生物種同士の高精度な分類を実現し，また，新種の可能性のある生物に対し高精度な検出を実現した．

最後に，本研究は，北海道大学大学院情報科学研究科 GCOE プロジェクトの一環として行われたものである．また本研究の一部は，日本学術振興会科学研究費補助金基盤研究 (B)(課題番号:21300030) 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”, 新井書院 (2005).
2）渡邉 日出海, 金子 俊一, 長谷山 美紀, “生命異分野共同研究プロジェクト,” 信学誌, **92**, 10, pp.822-827 (2009).
3）馬渡 俊輔，1994,“動物分類学の論理-多様性を認識する方法-”, 東京大学出版会，東京，233pp.
4）岩槻 邦男，馬渡 俊輔，“生物の種多様性 -バイオディバーシティ・シリーズ 1”, 裳華房，(1996).
5）馬渡 俊輔，“動物分類学 30 講”, 朝倉書店 (2006).
6）David A, Lytle,“Automated processing and identification of benthic invertebrate samples”, *Journal of the North American Benthological Society 29(3),* pp. 867-874 (2010).
7）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”, *Proceedings of the 7th IEEE Workshops on WACV/MOTIONS* (2005).
8）C. Wen, D. Guyer and W. Li,“Automated insect classification with combined global and local features for orchard management”, *ASABE Annial Meeting,* 2009 Paper No.095865.
9）D.Tax, R.Duin,“Support vector data description”, *Machine Learning,* **54**, pp. 45-66 (2004).
10）J.Mercer,“Functions of positive and negative type and their connection with the theory of integral equations”, *Trance. Lond. Phil. Soc. (A),* **209**, pp. 415-446 (1909).
11）S. Chen, D. Li,“Image binarization focusing on objects”, *Neurocomputing,* **69**, pp 2411-2415 (2006).
12）M.Hu,“Visual pattern recognition by moment invariants”, *IEEE Transactions on Information Theory, 8:179-187,1962*, **20**, 11, pp. 1254-1259 (1998).
13）Lanckriet, G. R. G., Cristianini, N., Bartlett, P., Ghaoui, L. E. and Jordan, M. I.,“Learning the Kernel Matrix with Semidefinite Programming”, *Journal of Machine Learning Research,* **5**, pp.27-72 (2004).
14）Sonnenburg, S., Ratsch, G., Schafer, C. and Scholkopf, B,“Large Scale Multiple Kernel Learning”, *The Journal of Machine Learning Research,* **7**, pp. 1531-1565 (2006).
15）Bernhard Schokopf, John C. Platt, John Shawe-Taylor, Alex J.

Smola and Robert C. Williamson, “Estimating the Support of a High-Dimensional Distribution”, *Neural Computation* (2001).

長谷川尭史（はせがわたかし） 2010 年，北海道大学工学部情報エレクトロニクス学科卒業．同大学院情報科学研究科修士過程在学中．画像処理および生物分類学の研究に従事．

小川 貴弘（おがわ たかひろ） 2003 年，北海道大学工学部情報工学科卒業．2005 年，同大大学院情報科学研究科修士課程了．2007 年，同大学院情報科学研究科博士後期過程了．日本学術振興会特別研究員．同大学博士研究員を経て，現在同大学助教．画像復元および映像処理に関する研究に従事．博士 (情報科学)．正会員．

渡邉日出海（わたなべひでみ） 1988 年，東京理科大学理工学部応用生物学科卒業．1993 年，同大学院博士後期課程退学．博士 (理学)．北海道大学大学院情報科学研究科教授．国立遺伝学研究所，米国国立衛生研究所，理化学研究所，奈良先端大学を経て，2004 年度より現職．専門はゲノム生物科学．ヒトゲノムプロジェクト等に参加．現在深海底探索等を行っている．

長谷山美紀（はせやまみき） 1986 年，北海道大学工学部電子工学科卒業．1988 年，同大大学院工学研究科修士課程了．1989 年，同大学電子科学研究科助手．1994 年，同大学工学研究科助教授．2006 年，同大学情報科学研究科教授．ディジタル信号処理および画像処理の研究に従事．博士 (工学)．正会員．