屏風画像のスパース性緩和のための類似画像を用いた近傍学習

A Neighbor Space Learning by Similar Image for the Easing of Sparsity

濱上知樹

HAMAGAMI Tomoki

［論文要旨］

近年，高度な画像分類を用いた様々なアプリケーションが開発され，その学習精度の向上が期待されている。分類精度向上のために画像特徴の抽出法や分類器の学習等の改善［6］がなされてきた。特に画像特徴の抽出と分類において高い精度が得られている手法として深層学習［1］［2］［3］が注目されている。

画像に対する深層学習においては畳み込みニューラルネットワークが広く用いられ，分類カテゴリのラベルを有する教師あり画像を大量に用いて学習するのが一般的である。しかし，教師あり画像のサンプル数が十分でない場合，画像空間全体の学習が不完全となり，未知の画像の分類に失敗する。画像に対して分類カテゴリのラベルを付与することは非常に労力を要するため，教師あり画像の数は教師なし画像と比べて非常に少ない。さらに，ラベル付与に専門家の知識が必要な医療や人文系のデータでは，十分な数の教師ありデータが揃わず，大規模な学習は困難である。そのため，少数の教師あり画像から画像分類を高い精度で行える手法が求められている。

そこで本研究では，少数の教師あり画像の近傍空間を用いた学習方法を提案する。提案手法では，分類カテゴリに属する代表画像をもとに，一般画像空間中における近傍サンプルを収集し，これを用いて近傍空間の学習を行う。この方法を用いることにより，少数の教師あり画像から学習に必要なクラスタを構成し，クラスタの特徴を抽出することで分類を可能にする。提案手法の有効性を確認するため，ラベルが付与された画像が少ない事例として小袖屏風画像におけるモチーフの分類を行い，従来手法では困難であった分類が可能となったことを示す。

【キーワード】CNN，類似画像検索，画像分類，SOM，スパース空間

❶……序論

1.1　研究背景と目的

一般に教師あり画像群を用いた画像分類では，画像をある写像関数によって画像特徴空間に写像して識別面を学習する。従来では既存の画像特徴量を抽出する手法を用いることで画像特徴空間へ写像し，教師あり学習を用いて教師あり画像群のもつラベル情報に従って学習することで，目的とするカテゴリ分類器を得ている［7］。近年では画像特徴量の抽出と分類を共に学習する手法として深層学習が注目され，人手に依らない高度な特徴量抽出により高い画像分類精度を達成している。

しかし，教師あり画像を十分に用意することが困難な場合，カテゴリの分類は困難になる。このような場合，特徴空間中でクラスタをつくり，これをカテゴリと対応づける方法が考えられる。しかし，この方法は分類すべきカテゴリとクラスタが1対1になるとは限らない。また，クラスタを構成するサンプルが十分揃わず，特徴空間に対してスパースな場合は，クラスタを構成すること自体が困難である。

そこで本研究では，スパースなデータ空間にある少数の教師あり画像から，その近傍空間を代表する特徴ベクトルを深層学習によって得るための方法について扱う。これによって少数の教師あり画像から行う画像分類の精度向上を目的とする。

❷……画像分類

2.1　一般的な画像分類の手法

一般的な画像分類の手法として教師あり画像群を用いた分類を図2.1に示す。画像を画像特徴空間へ写像し，その空間上で識別面を学習する。この方法で核となるのは画像の画像特徴空間への写像，つまり，画像特徴量を抽出する手法と識別面を学習する方法である。画像を画像特徴空間へ写像し画像特徴量を得る手法としてSIFT（Scale-Invariant Feature Transform）［4］やSURF（Speeded Up Robust Features）［5］などが知られており，よく用いられている。画像特徴量を用い，教師あり画像群のもつラベル情報に従って識別面を学習する手法としてSVM（Support Vector Machine）やニューラルネットワークなどが知られている。これらの組み合わせによって目的のカテゴリ分類器が獲得できる。近年では画像特徴量の抽出と分類を共に学習する手法として深層学習が注目され，人手に依らない高度な特徴量抽出により高い画像分類精度を達成している。しかし上記の画像分類の課題として，教師あり画像が十分な数存在しなければ学習が困難であることが挙げられる。特に深層学習はネットワークの規模が大きくパラメータの自由度が高いために，十分な教師あり画像が必要となる。

2.2　教師あり画像が少数な場合の分類困難性

教師あり画像が少ない場合に画像の持つラベル情報を用いて分類を行うことは上記の通り困難である。このような場合においては，教師なし学習であるk-means法や自己組織化マップ（Self

Organizing Map: SOM）を用いることにより，特徴空間中でクラスタをつくり，これをカテゴリと対応づけることで分類を行う方法が考えられる。教師なし画像を用いた分類を図 2.2 に示す。しかし，クラスタリングは特徴空間中で近い位置に存在するサンプルによってクラスタをつくるため抽出される特徴が分類に適切でないとき，この方法は分類すべきカテゴリとクラスタが 1 対 1 になるとは限らない。また，クラスタを構成するサンプルが十分揃わず，特徴空間に対してスパースな場合は，クラスタを構成することが困難である。よって，教師あり画像が少数な場合において正確にカテゴリの分類を行うことは困難である。

そこで本研究では，スパースなデータ空間にある少数の教師あり画像から，その近傍空間を代表する特徴ベクトルを深層学習によって得るための方法について扱う。具体的には，カテゴリに対応した教師あり画像をもとに一般画像空間から Google 画像検索を用いて類似画像を収集し，疑似的につくられた画像クラスタを用いて深層学習を行うことで，カテゴリに対応した特徴ベクトルを抽出する。

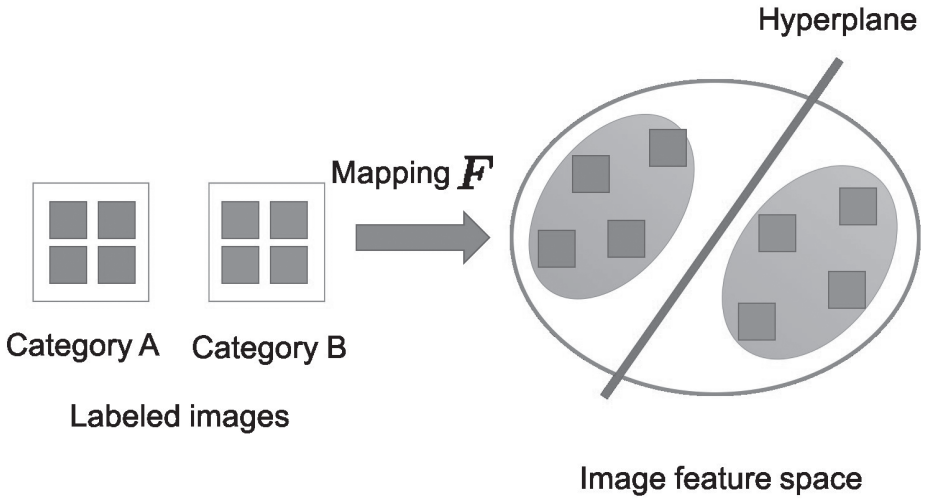


図 2.1　Classification with labeled images.

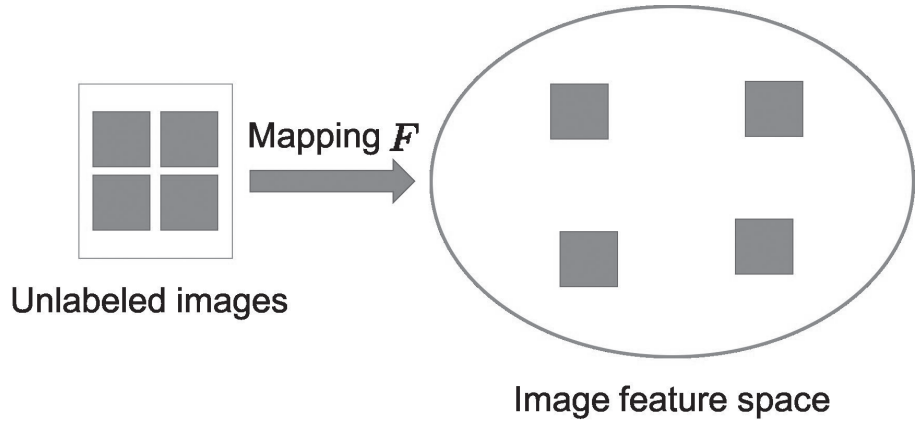


図 2.2　Classification with unlabeled images.

❸……深層学習

深層学習は深い層のネットワーク構造と大量のパラメータ最適化により，様々な分野において高い性能を実現している手法である。深層学習はベースの手法としてニューラルネットワークを用いている。代表的なニューラルネットワークとして多層パーセプトロンがある。

3.1　多層パーセプトロン

多層パーセプトロンとは，ニューロンと呼ばれる脳の神経細胞を模したモデルを入力層，隠れ層，出力層の3層に配置したニューラルネットワークである。多層パーセプトロンを図3.1に示す。入力層に与えられた入力データから隠れ層を通し，重み付けや結合を行うことで出力層で出力を得るモデルとなっている。これは教師あり学習として知られ，入力データと期待されるラベルをもとに隠れ層の重みパラメータなどを学習し，入力に対するパターンを判定する。

3.1.1　ニューロン

ニューラルネットワークを構成するニューロンを図3.2に示す。ニューロンは各要素 $x_i(i = 1,\cdots,n)$ が0～1の値をとる入力ベクトル $x = (x_1,\cdots,x_n)$ に対して，以下の式（3.1）に従って1つの実数値 y を出力する。

$$y = g\left(\sum_{i=1}^{n} w_i x_i + \theta\right) \tag{3.1}$$

式（3.1）において w_i は各入力に対する重みを表し，結合重み，もしくは結合係数と呼ばれる。また θ はバイアスである。$g(x)$ は入力に対してどのような出力を返すかを示しており，活性化関数と呼ばれ，ステップ関数，式（3.2）で表されるシグモイド関数等が用いられる。a は入力に対する係数であり，一般的には1が用いられる。

$$g(x) = \frac{1}{1 + \exp(-ax)} \tag{3.2}$$

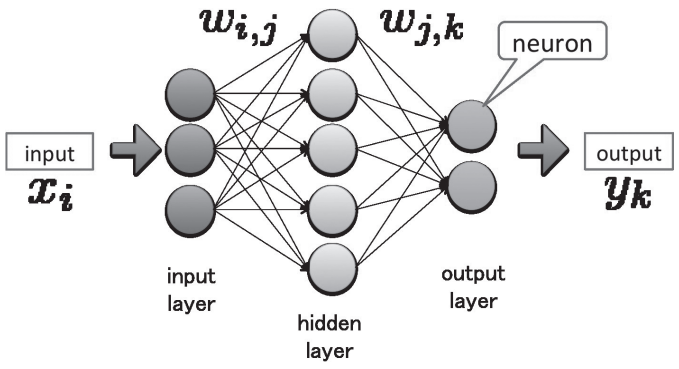


図3.1　Multi layer perceptron.

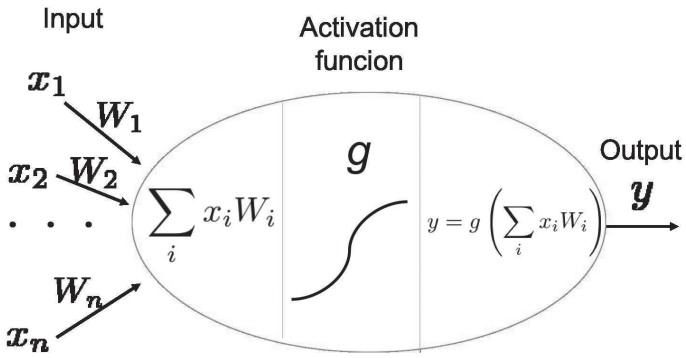


図 3.2　A mathematical model for a neuron.

3.1.2　順方向計算

入力はニューロン間の結合係数によって重み付けがなされ，入力層，隠れ層，出力層に移っていく。そして出力層の計算結果が最終的な出力となる。

入力ベクトル $x^{in} = \left(x_1^{in}, \cdots, x_n^{in}\right)$ を要素ごとに入力層の各ニューロンに与えた時，隠れ層の入力は各結合係数によって重み付けられる。隠れ層における出力を y^{hid}，バイアスを θ^{hid}，i 番目の入力層ニューロンと j 番目の隠れ層ニューロンの結合係数を $\omega_{i,j}$ とすると，隠れ層の出力は式（3.3）で表される。

$$y_j^{hid} = g\left(\sum_{i=1}^{n} \omega_{i,j} x_i^{in} + \theta_j^{hid}\right) \tag{3.3}$$

同様にして出力層における出力を y^{out}，バイアスを θ^{out}，j 番目の隠れ層ニューロンと k 番目の出力層ニューロンの結合係数を $\omega_{j,k}$ とすると，出力層の出力は式（3.4）で表される。

$$y_k^{out} = g\left(\sum_{j=1}^{m} \omega_{j,k} y_j^{hid} + \theta_k^{out}\right) \tag{3.4}$$

3.1.3　誤差逆伝播法

誤差逆伝播法は教師あり学習の一種であり，入力データのラベルと実際の出力との誤差を極小値に向けて結合係数を修正していくように学習を行う。その際に，隠れ層と出力層の間の結合係数を修正した後，入力層と隠れ層の間の結合係数を修正するため，誤差逆伝播法と呼ばれる。また，誤差がない時には修正を行わないため，誤り訂正学習とも呼ばれる。

二乗誤差 E を式（3.5）のように定義する。ラベル $d = (d_1 \cdots d_l)$ と出力 $y^{out} = \left(y_1^{out} \cdots y_l^{out}\right)$ の誤差を減少させるように修正を加える。

$$E = C\sum_{k=1}^{l}(d_k - y_k^{out})^2 \tag{3.5}$$

ここで C は正の定数である。そして最急降下法により修正量 $\Delta\omega_{j,k}$ を式（3.6）で計算する。

$$\Delta\omega_{j,k} = -\alpha \frac{\partial E}{\partial \omega_{j,k}} \tag{3.6}$$

式（3.6）を解く事によって

$$\Delta\omega_{j,k} = -\alpha \frac{\partial E}{\partial y_k^{out}} \frac{\partial y_k^{out}}{\partial x_k^{out}} \frac{\partial x_k^{out}}{\partial \omega_{j,k}} \tag{3.7}$$

$$x_k^{out} = \sum_{j=1}^{m} \omega_{j,k} y_j^{hid} \tag{3.8}$$

となる。よって，以上より $\partial E/\partial y_k^{out} = -(d_k - y_k^{out}), \partial y_k^{out}/\partial x_k^{out} = y_k^{out}(1 - y_k^{out})$ である事を用いて入出力と教師信号より修正値を求めることができる。

$$\Delta\omega_{j,k} = \ -\alpha\delta_{1,k} y_j^{hid} \tag{3.9}$$

$$\delta_{1,k} = \ (d_k - y_k^{out}) y_k^{out}(1 - y_k^{out}) \tag{3.10}$$

同様に，入力層と隠れ層の間の結合係数の修正値を式（3.11）に示す。ここで x^{hid} は入力層と隠れ層の間において式（3.8）と同様に計算される値である。

$$\begin{aligned} \Delta\omega_{i,j} &= -\beta \frac{\partial E}{\partial \omega_{i,j}} \\ &= -\beta\delta_{2,j} x_i^{in} \end{aligned} \tag{3.11}$$

$$\delta_{2,j} = \sum_{k=1}^{l} \delta_{1,k} \omega_{j,k} x_j^{hid}(1 - x_j^{hid}) \tag{3.12}$$

式（3.6），式（3.11）に用いられている α, β を学習率と定義する。この学習率を小さくすると修正量が小さくなり，学習速度は一般的に遅くなる。しかし，大きすぎると学習が収束しなくなる。また，隠れ層のニューロン数も重要なパラメータであり，この数を大きくすると，一般に学習速度は遅くなるが，性能は高くなる。これらのパラメータは適用する問題において，適切に設定される必要がある。

3.2　畳み込みニューラルネットワーク

深層学習は従来のニューラルネットワークよりも多段の層をもち，様々な分野で高い性能を達成している手法である。深層学習は目的タスクの性質などから様々な提案がされているが，本稿では画像認識において最も広く用いられている畳み込みニューラルネットワーク［8］［10］を用いる。

畳み込みニューラルネットワークは局所領域に対して重みを畳み込む処理を行う層をもつ順伝播型ネットワークであり，教師ありデータを用いて逆誤差伝播法により学習を行う。図 3.3 に示すように，局所領域に対して重みの畳み込み計算を行うことで，生物の脳の視覚受容野を模倣しており，画像認識において重要な技術となっている。画像認識でよく用いられる畳み込みニューラルネットワークは，畳み込み層とプーリング層を交互に接続しその後全結合層が続き，最後に各カテゴリの出力をもつ。

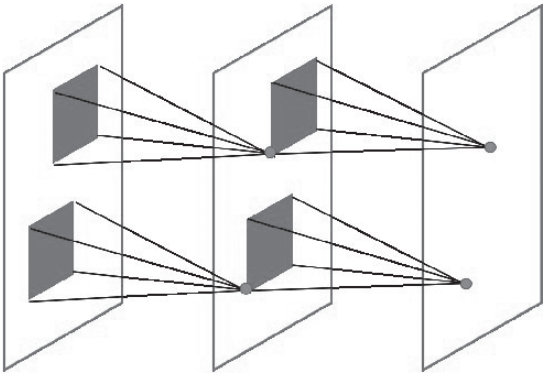


図 3.3　Convolution.

3.2.1　畳み込み層

単一チャネルをもつ画像について，各画素を $x_{i,j}$ とする。また畳み込み処理を行う際の重みの値の集合をフィルタと呼び，そのフィルタのサイズを $H \times H$ とし各フィルタの値を $h_{p,q}$，バイアスを $b_{i,j}$，出力を $u_{i,j}$ とすると，画像の畳み込みは式（3.13）で示される。

$$u_{i,j} = \sum_{p=0}^{H-1} \sum_{q=0}^{H-1} x_{i+p,j+q} h_{p,q} + b_{i,j} \tag{3.13}$$

式（3.13）では縦横方向に 1 画素ずつずらしながら畳み込みを計算しているが，このずらす幅をストライドと呼び，場合によってはストライドを数画素ずつずらすことがある。ストライドを s とすると，ストライドを考慮した畳み込みは式（3.14）で示され，出力画像サイズは約 1/s 倍となる。

$$u_{i,j} = \sum_{p=0}^{H-1} \sum_{q=0}^{H-1} x_{si+p,sj+q} h_{p,q} + b_{i,j} \tag{3.14}$$

カラー画像のような複数チャネルの画像に対しては各チャネルに同様の畳み込み処理を行い，また畳み込むフィルタ数も複数チャネルにする場合がある。入力が K チャネルの画像の時，入力チャネル k，フィルタ m に対する畳み込みは式（3.15）で示される。またバイアス $b_{i,j,m}$ はフィルタごとに共通の値をもつのが一般的である。

$$u_{i,j,m} = \sum_{k=0}^{K-1} \sum_{p=0}^{H-1} \sum_{q=0}^{H-1} x_{i+p,j+q,k} h_{p,q,k,m} + b_{i,j,m} \tag{3.15}$$

畳み込みニューラルネットワークにおける各フィルタの値は同一チャネルの全画素で同じ値を用いており，これは共有重みと呼ばれている。

畳み込みによって得られた $u_{i,j,m}$ に対し，活性化関数を適用することで最終的な出力 $z_{i,j,m}$ が得られ，式（3.16）で示される。

$$z_{i,j,m} = f(u_{i,j,m}) \tag{3.16}$$

様々な活性化関数が提案されているが，近年では式（3.17）に示される正規化線形関数（rectified

linear function: ReL）が一般的に用いられる。正規化線形関数は活性化関数の中でも計算量が少なく，学習の収束が速いことが知られている。

$$f(u) = \max(u, 0) \tag{3.17}$$

3.2.2　プーリング層

畳み込み層の後に繋げるプーリング層では，畳み込み層で得られた出力の中で一部の情報を間引く処理を行うことで画素の位置変化にロバストな特徴を得ている。プーリングには様々な種類があるが，画像認識においては最大プーリング（Maxpooling）が一般的であり，本稿においてもこの手法をとる。最大プーリングの処理を図3.4に示す。プーリングを行う領域のサイズを $H \times H$ としたとき，最大プーリングでは H^2 個の画素値の中で最大値以外の値を間引く処理を行う。

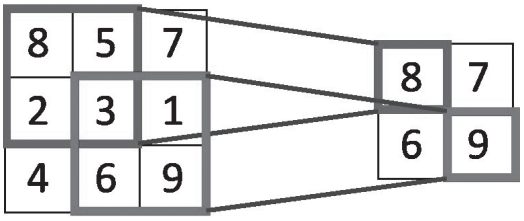


図3.4　Max-pooling.

❹……既存手法

4.1　畳み込みニューラルネットワークを用いた画像分類

CNNを用いた画像分類の例として，研究［8］がある。この研究では，ImageNet LSVRC-2010［9］において画像を1,000カテゴリに分類するタスクに対し，深層学習を用いて従来よりも優れた分類精度を得ている。ImageNet LSVRC-2010の画像例を図4.1に示す。

図4.1　ImageNet LSVRC-2010.

表 4.1　Alexnet structure.

Name	Patch	Stride	Output map size	Activation function
Input	-	-	224 × 224 × 3	-
Convolution1	11 × 11	4	55 × 55 × 96 (48 × 48)	ReL
Convolution2	5 × 5	1	27 × 27 × 256 (128 × 128)	ReL
Max-pooling1	3 × 3	1	13 × 13 × 384 (192 × 192)	-
Max-pooling2	3 × 3	1	13 × 13 × 384 (192 × 192)	-
Max-pooling3	3 × 3	1	13 × 13 × 256 (128 × 128)	-
Fullyconnected1	-	-	1 × 1 × 4,096	ReL
Dropout	-	-	1 × 1 × 4,096	-
Fullyconnected2	-	-	1 × 1 × 4,096	ReL
Dropout	-	-	1 × 1 × 4,096	-
Fullyconnected3	-	-	1 × 1 × 1,000	-

用いられた CNN の構造を表 4.1 に示す。畳み込み，プーリング層を 5 層，全結合層を 3 層つなげた，計 8 層からなる深い層をもつネットワークにより高い分類精度を実現している。この研究では Output map size が 2 つに分かれており，それぞれ 2 つの GPU を用いて並列に計算することで計算時間の削減を行っている。

この研究のように多段の層をもつ大規模なネットワーク構造の場合，学習するパラメータ数が非常に多いため，大量のデータを用いて学習を行う必要がある。実験では 120 万枚のラベル付き訓練画像，5 万枚のバリデーション画像，15 万枚のテスト画像を用いており，非常に大量のデータによるパラメータ最適化を行っている。

深層学習は非常に多くのパラメータを用いて学習するために，訓練データのみに過剰に反応してしまう過学習の問題が発生しやすく，テストデータに対して十分な精度が得られないことがある。この研究では過学習を防ぐ手法の 1 つとして，学習時にニューロンの一部を無視してパラメータを最適化する Dropout を用いている。

❺……提案手法

本研究ではスパースなデータ空間にある少数の教師あり画像から画像分類を行うため，分類カテゴリに属する教師あり画像を各カテゴリで選択し，代表画像と定義する。各カテゴリの代表画像に対し，一般画像特徴空間における類似画像を収集し学習を行うことで各カテゴリの近傍空間を学習する。提案手法における近傍空間の学習の概略図を図 5.1 に示す。本手法で学習した近傍空間を用いることで，一般画像特徴空間の観点から近似的に画像分類することが可能となる。

スパースなデータ空間において画像特徴空間全体の学習を行うのが困難であるという課題を解決するため，提案手法では一般画像特徴空間上で代表画像と類似した画像群を収集し学習を行う。これによりスパースであったデータ空間中のサンプルを補完し，画像特徴空間全体の学習の不完全性を緩和することが期待できる。

提案手法による画像分類の概略図を図 5.2 に示す。提案手法では学習フェーズと分類フェーズの 2 つのフェーズから成る。

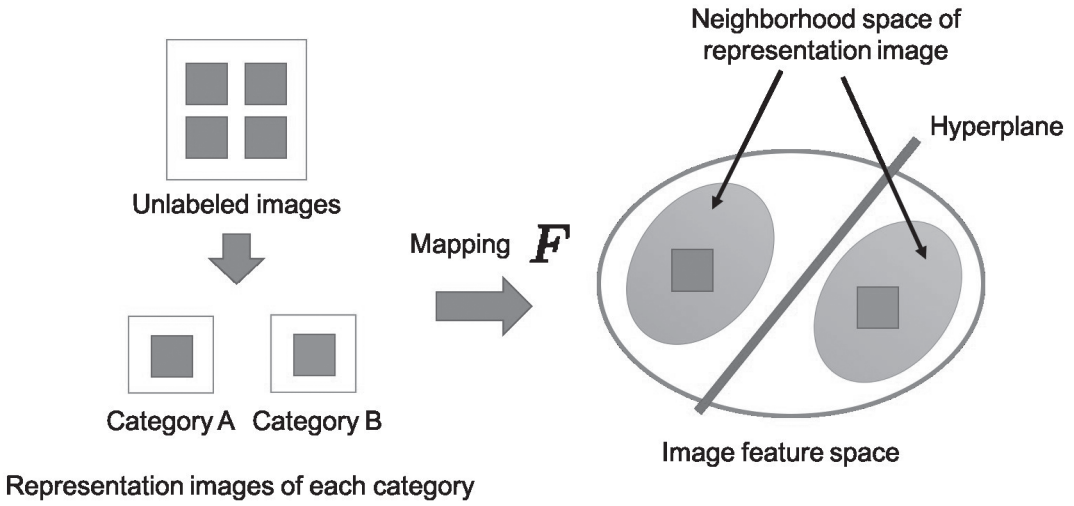


図 5.1　Classification with proposed method.

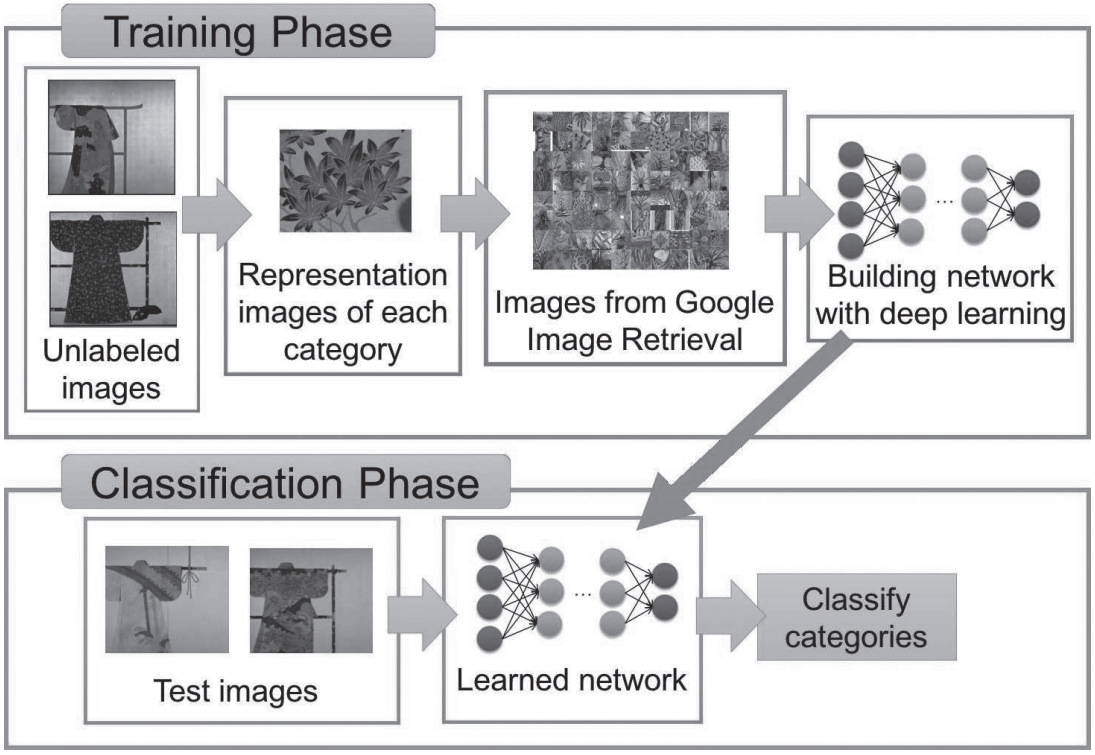


図 5.2　Overview of proposed method.

5.1　学習フェーズ

学習フェーズでは，少数の教師あり画像群から各カテゴリに属する代表画像を選択し，一般画像特徴空間上で類似した画像群を収集する。一般画像特徴空間における類似画像は Google 画像検索により収集する。これにより対象タスクのデータセット外からデータを収集することができ，デー

タ空間の補完が可能となる。Google 画像検索によって得られた画像群を用いて学習に必要なクラスタを構成し，得られた特徴量による分類を行う。クラスタの構成，特徴量の抽出は画像認識で広く用いられる畳み込みニューラルネットワークを用いる。

Google 画像検索で収集した類似画像を入力としてネットワークへ与えると，各近傍空間に対する出力が得られる。ここで収集した類似画像は Google 画像検索のクエリに用いた代表画像のラベルをもっているものとみなすことで，各近傍空間の出力に対して誤差を計算することができる。この誤差を用いて誤差逆伝播法によりパラメータの最適化を行う。学習フェーズにおける畳み込みニューラルネットワークの学習を図 5.3 に示す。

5.2　分類フェーズ

分類フェーズでは，未知の画像に対し学習フェーズで構築した畳み込みニューラルネットワークを用いて各近傍空間に対する値を計算する。各近傍空間の値に対して，最も値の大きい近傍空間を入力画像の分類結果として用いる。これにより分類された近傍空間と対応する代表画像のカテゴリに対して最終的に分類される。

分類フェーズにおける未知の画像の分類を図 5.4 に示す。

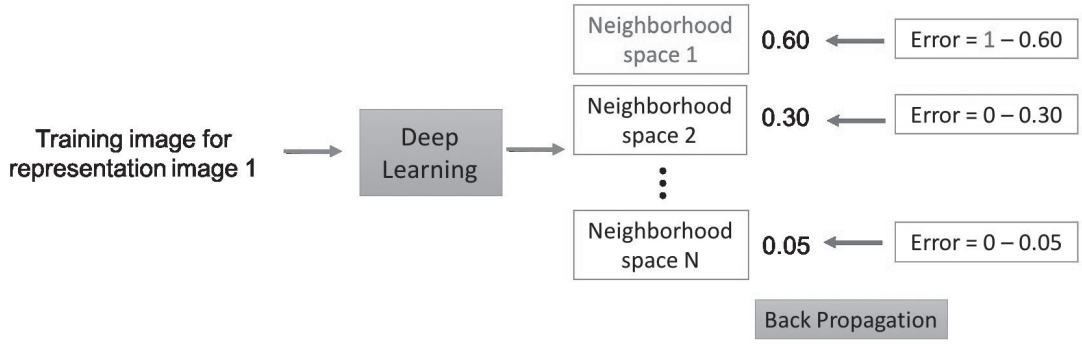


図 5.3　Training Phase.

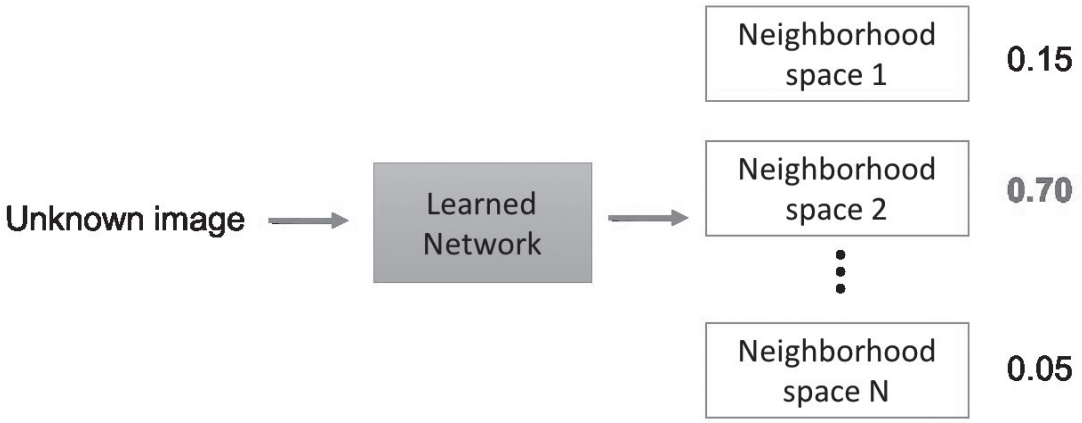


図 5.4　Classification Phase.

❻……実験

提案手法の有効性を確認するため，少数の教師あり画像をデータセットとした場合の画像分類タスクで実験を行った。

6.1　小袖屏風

本研究は国立歴史民俗博物館の共同研究の一環であり，国立歴史民俗博物館に収蔵されている小袖屏風画像［11］における知的構造の創出を目指している。小袖屏風は日本の伝統的衣装である小袖を屏風に貼り付けた歴史資料であり，これをデジタルデータとして保存したデジタルアーカイブの活用が期待されている。小袖屏風の特徴的な要素として図6.1に示すモチーフ（模様や柄）があり，画像中のモチーフの把握は重要である。しかし文化資源の数は有限であるためにデータ数が限られ，モチーフの分類も専門家でなければ困難であることが知られているため，モチーフに関する教師あり画像を十分収集することが困難である。そこで本手法による少数の教師あり画像を用いた画像分類を行い，十分な教師あり画像が収集困難である場合に有効であることを示す。本実験では分類カテゴリとして小袖屏風画像におけるモチーフを扱う。

小袖屏風画像中のモチーフ（模様や柄）を扱うため，小袖屏風全体の画像を分割してデータセットとした。具体的には図6.2に示すように1隻の小袖屏風画像に対し，縦に40，横に30分割して得られた1,200枚のうち，80 × 80画素に満たないサイズの画像を除いた1,141枚を得た。この分割処理を小袖屏風102隻に対して行い，得られた計116,381枚の画像を本実験のデータセットとした。

小袖屏風画像におけるモチーフの代表画像は，国立歴史民俗博物館が一般公開している館蔵野村正治郎衣裳コレクションデータベース［12］のモチーフ情報を基に，小袖屏風分割画像から代表画像を選択した。実験では50種類のモチーフについて分類を行い，各モチーフの代表画像の枚数を変えて実験を行った。代表画像の例を図6.3に示す。

6.1.1　Google画像検索で収集した類似画像

各代表画像をクエリとしてGoogle画像検索により類似画像を収集した。収集した類似画像例として，菊のモチーフにおける代表画像と類似画像を図6.4～図6.8，松のモチーフにおける代表画像と類似画像を図6.9～図6.13に示す。小袖屏風画像中に現れる菊のモチーフは，その描かれ方が小袖屏風の構図や時代背景によって大きく異なり，図6.4～図6.8に示す通り，代表画像によってGoogle画像検索で得られる類似画像が大きく異なっている。松のモチーフについても図6.9～図6.13に示す通り同様である。このため単純に，あるモチーフに対してそれぞれの代表画像を訓練データとして学習しても，代表画像ごとの差が大きいために分類が困難となると考えられる。そこで提案手法において代表画像ごとに類似画像を収集することでデータ空間の補完を行い，画像の分類を行う。

本手法で用いる畳み込みニューラルネットワークは深層学習ライブラリの1種であるChainer［13］を用いて実装した。Chainerは Preferred Infrastructure社によって開発されたPythonベースのライブラリであり，様々なネットワーク構造を柔軟に実装でき，GPUを用いた処理にも対応している。

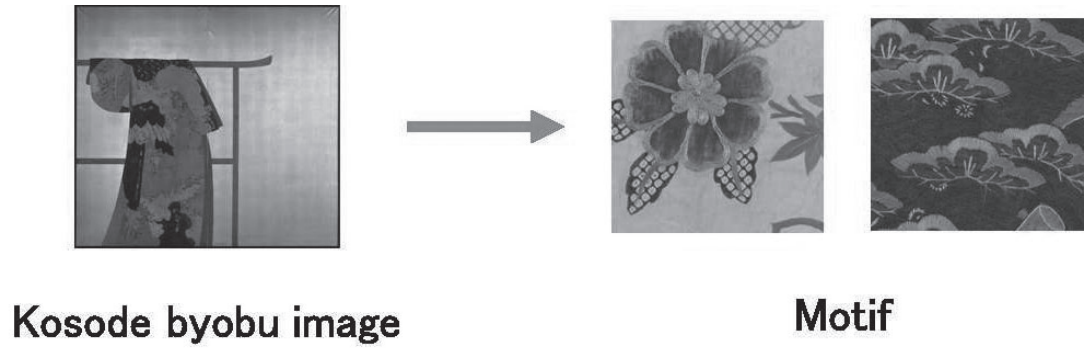


図 6.1　Kosode byobu images.

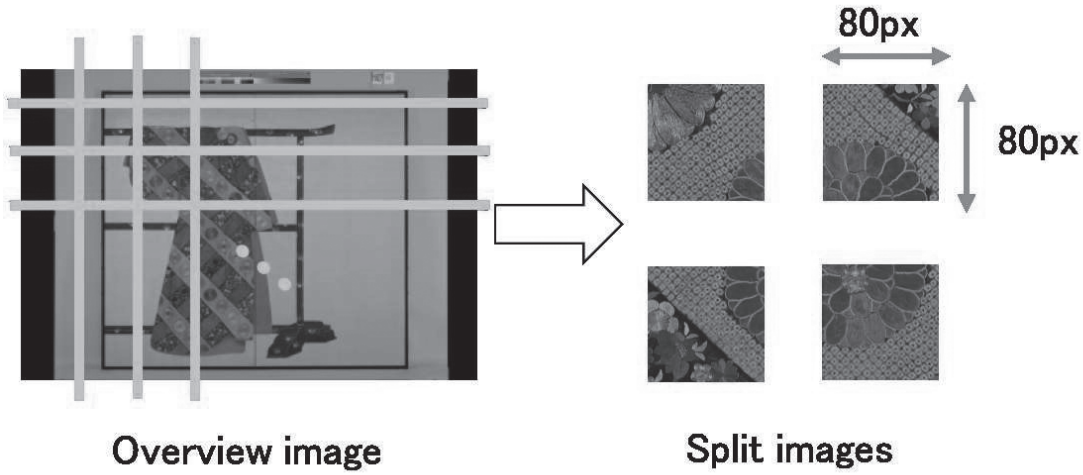


図 6.2　Split images.

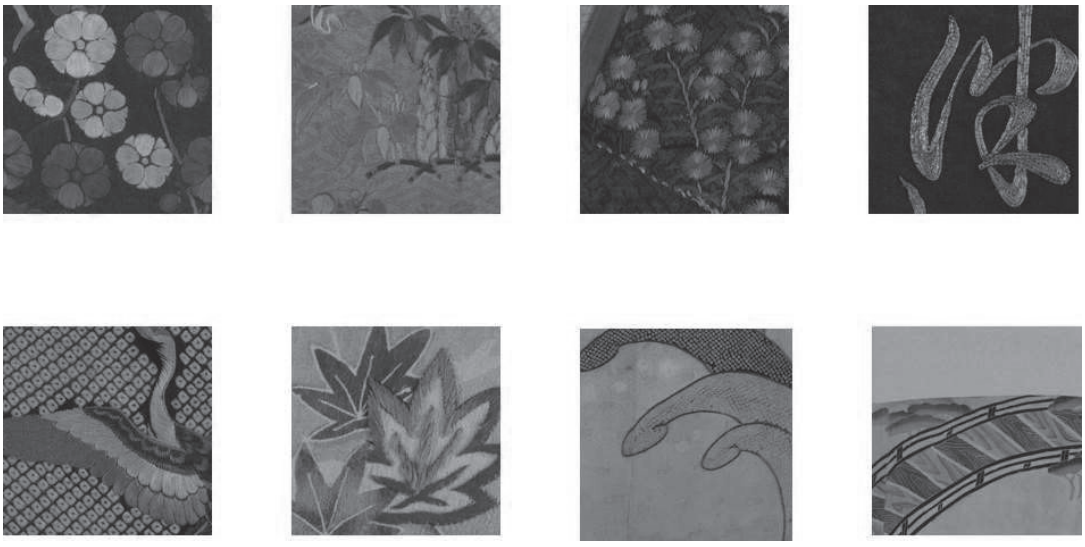
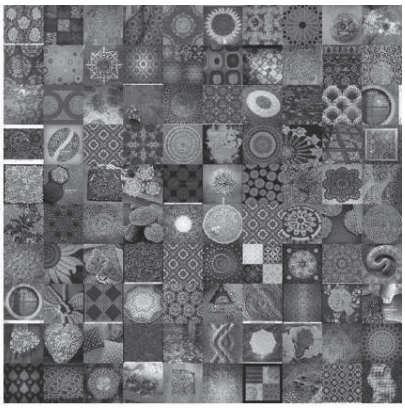


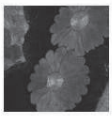
図 6.3　Examples of representation images.

Representation image

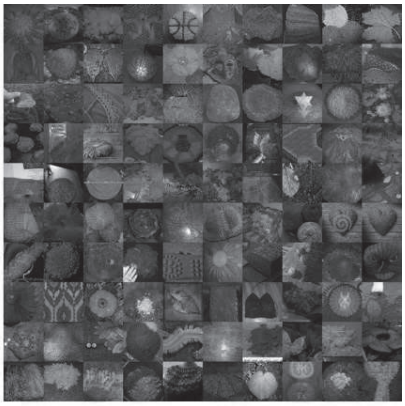


Similar images

図 6.4　Example representation image of kiku 1 and similar images.

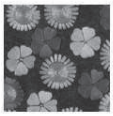


Representation image

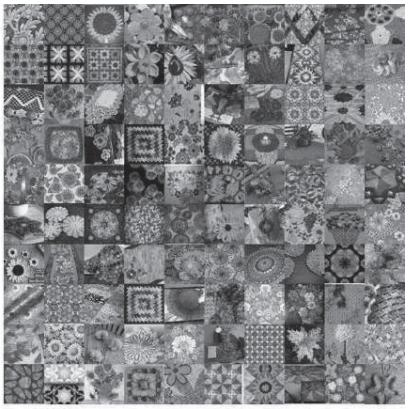


Similar images

図 6.5　Example representation image of kiku 2 and similar images.



Representation image



Similar images

図 6.6　Example representation image of kiku 3 and similar images.

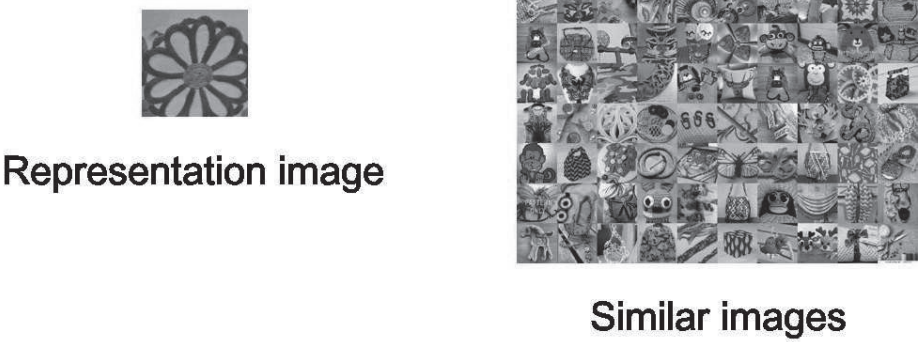


図 6.7　Example representation image of kiku 4 and similar images.

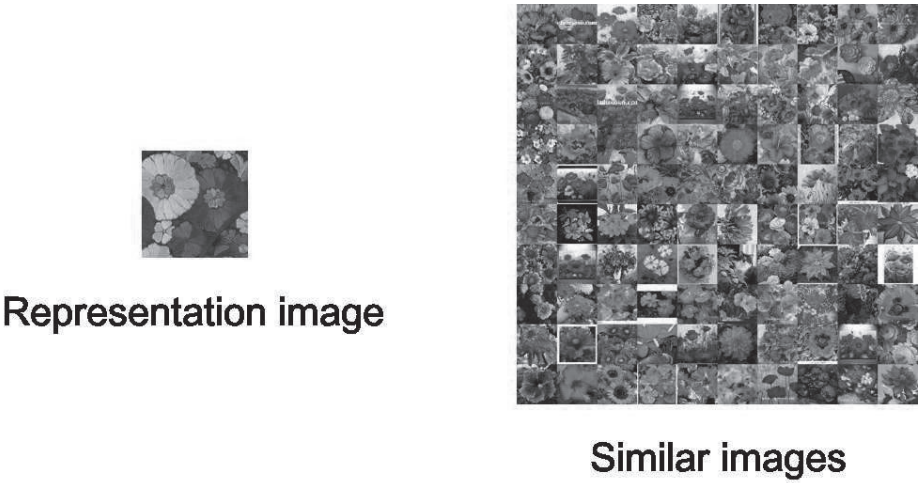


図 6.8　Example representation image of kiku 5 and similar images.

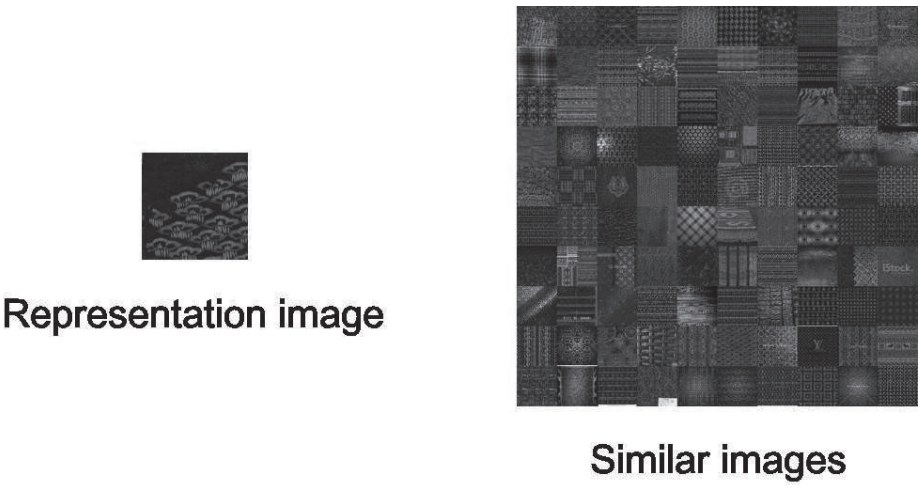
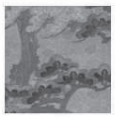


図 6.9　Example representation image of matsu 1 and similar images.

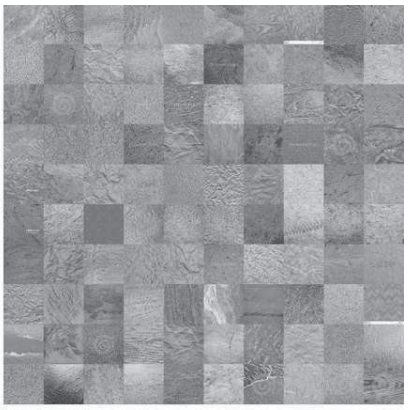


Representation image

Similar images

図6.10　Example representation image of matsu 2 and similar images.

Representation image



Similar images

図6.11　Example representation image of matsu 3 and similar images.

Representation image

Similar images

図6.12　Example representation image of matsu 4 and similar images.

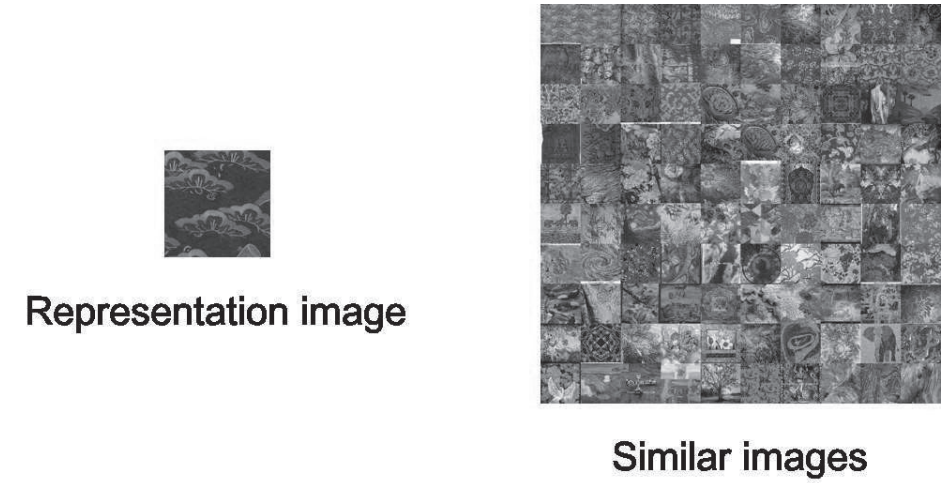


図 6.13　Example representation image of matsu 5 and similar images.

6.2　畳み込みニューラルネットワークの詳細

本手法で用いる畳み込みニューラルネットワークは確率的勾配降下法（Stochastic Gradient Descent）によりパラメータを学習した。確率的勾配降下法は訓練データの一部のみを用いてパラメータを更新する。パラメータ w を更新する際，訓練データからサンプルしたデータ n の誤差関数 $E_n(w)$ の勾配を ∇E_n とすると，式（6.1）で計算される。

$$w^{t+1} = w^t - \epsilon \nabla E_n \tag{6.1}$$

確率的勾配降下法を用いることで訓練データの誤差勾配がパラメータ更新ごとに異なるため，同じ局所解に陥るリスクを減らすことができる。また学習に用いるサンプル集合をミニバッチと呼び，学習時のミニバッチの数をバッチサイズとして実験ごとに設定した。各実験のバッチサイズは各実験設定に記載した通りである。

6.2.1　モーメンタム

モーメンタムはパラメータ更新の収束性能を向上させる手法の 1 つであり，前回のパラメータ修正量を一定の割合で現在のパラメータ修正量に加算する。ミニバッチ $t-1$ に対するパラメータ修正量を $\Delta w^{(t-1)}$ とすると，ミニバッチ t に対する更新を式（6.2）に示す。

$$w^{(t+1)} = w^{(t)} - \epsilon \nabla E_t + \mu \Delta w^{(t-1)} \tag{6.2}$$

本実験では研究［3］で提案されている Nesterov's Gradient Descent（NAG）を用いた。t 回目の学習時におけるパラメータを w_t，パラメータ更新量を v_t，モーメンタム項を $\mu_t v_t$，誤差勾配項を $\epsilon \nabla E(w)$ としたとき，NAG によるパラメータ更新を式（6.3）に示す。

$$v_{t+1} = \mu_t v_t - \epsilon \nabla E(v_t + \mu_t v_t) \tag{6.3}$$

各実験の学習率 ϵ，モーメンタム係数 μ は各実験設定に記載した通りである。

6.2.2　重み減衰

畳み込みニューラルネットワークは扱うパラメータ数が非常に多く自由度が高いため局所解に陥りやすく，過学習を引き起こしやすい。そのため過学習を緩和する手法の1つとして重み減衰を行った。重み減衰は学習時の誤差関数にパラメータ w の2乗ノルムを加えることによりパラメータの自由度に制約を設け，過学習を緩和する。この時の勾配降下法のパラメータ更新式を式（6.4）に示す。

$$w^{(t+1)} = w^{(t)} - \epsilon \left(\nabla E_n + \lambda w^{(t)} \right) \tag{6.4}$$

各実験の重み減衰係数 λ は各実験設定に記載した通りである。

6.2.3　Dropout

畳み込みニューラルネットワークにおける過学習を緩和する手法として，重み減衰と共にDropoutも用いた。Dropoutは学習時に一部のニューロンを無効化して残りのニューロンのみで学習を行う手法である。この手法により多数のネットワークを独立に学習し平均化した結果と同等の性能が得られると考えられており，過学習の緩和に有効である。Dropoutにおいて無効化するニューロンの割合 p は各実験の実験設定で記載した通りである。

6.3　実験1：近傍空間の学習精度に対する実験

1つ目の実験では，画像特徴空間上の近傍空間の学習精度を確認するため，収集した類似画像により学習した近傍空間の学習精度を評価した。小袖屏風画像における50種類のモチーフに対してそれぞれ1枚ずつ代表画像を選択し，Google画像検索により類似画像を収集した。収集した画像群の一部をバリデーションデータとして分離し，バリデーションデータに対する分類精度を求めた。

実験1の概要図を図6.14に示す。

6.3.1　実験設定

各カテゴリの代表画像をクエリとして得られるGoogle画像検索の結果から検索結果順に画像を収集し，そのうち90%を占める画像を訓練画像，残りの10%を占める画像をバリデーションデー

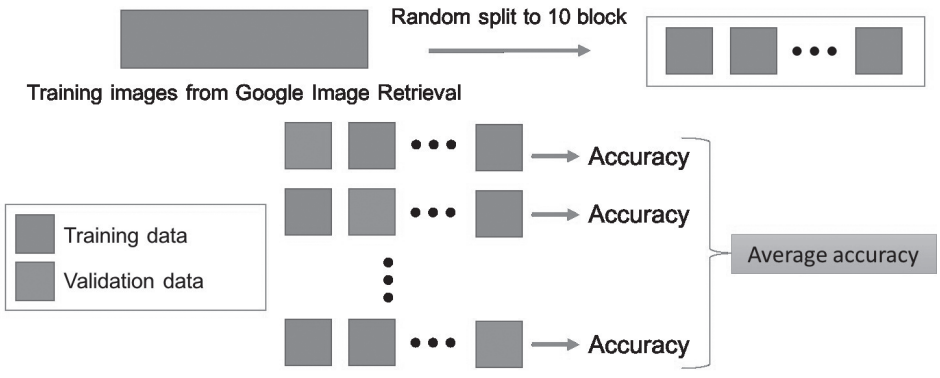


図6.14　Experiment 1.

タとして学習し分類した。バリデーションデータはGoogle画像検索の検索結果順に関わらずランダムに選択し，実験毎に重複がないようにした。それぞれバリデーションデータを選択して実験を行い，10回の実験の平均分類精度を求めた。詳細な実験設定を表6.1に示す。

実験に用いた深層学習のネットワークを表6.2に示す。

6.3.2 実験結果・考察

実験結果を図6.15に示す。グラフの横軸は学習回数，縦軸は分類精度を示す。結果から，学習が進むにつれて訓練データに対する分類精度は向上するが，バリデーションデータに対する精度は40%程度で収束することが明らかとなった。訓練データとバリデーションデータにおいて精度の差が生じたことから過学習が発生したことがわかり，差が50%以上開いたことから，汎化性能が不十分であることがわかる。

6.4 実験２：モチーフの分類精度に対する実験

2つ目の実験では，提案手法による画像分類の精度を確認するため，小袖屏風画像に対してモチーフの分類を行った。ここでは分類を行う小袖屏風画像が未知の画像であることを想定し，これらをテストデータと呼ぶ。実験1と同様に50種類のモチーフに対してそれぞれ1枚ずつ選択した代表画像により類似画像を収集し，近傍空間を学習した。比較する従来手法として教師あり画像群のサンプル数が少ない場合を想定し，各モチーフの代表画像を学習データとした。また各モチーフに分類された画像を確認することで誤分類した場合の原因について考察した。モチーフの分類精度に対する実験の概要図を図6.16に示す。

6.4.1 実験設定

分類に用いるテストデータは各モチーフを画像中に含む291枚の小袖屏風分割画像である。モチーフの分類精度は全入力画像のうち，正しいモチーフに分類された画像数の割合で求めた。詳細な実験設定を表6.3に示す。実験に用いた深層学習のネットワークは表6.2と同様である。

6.4.2 実験結果・考察

モチーフの分類精度を図6.17に示す。従来手法では学習の進行に関わらず2%程度の精度で収束した。仮に50カテゴリに対してランダムに分類を行った場合，理論上の分類精度は1/50=0.02(2%)であるため，従来手法はランダムな分類と同等の結果となった。これに対し提案手法では学習が進むと精度が若干向上し，最終的に15%程度の精度が得られ，従来手法よりも分類精度の向上がみられた。

上記の分類精度はモチーフ全てに対する分類結果の平均としたため，実際に分類された画像が少数のモチーフに偏っている可能性が考えられる。そこで上記実験に対し，各モチーフの正解データ数，不正解データ数を確認した。各モチーフに分類された画像の分布を図6.18に示す。横軸が各モチーフであり，縦軸がそのモチーフに分類した画像の中での正解数，不正解数を示している。結果を見ると，分類された画像数は各モチーフに差が生じているものの，1枚も分類されていないモ

表6.1　Experiments setting 1.

Classification motif size	50
Representation image size of 1 motif	1
Similar images from each representation images	300
The percentage of validation data	10%
Batchsize	50
Learning rate ϵ	0.01
Momentum coefficient μ	0.99
Weight decay coefficient λ	0.005
The percentage of vanishing neurons p	50%

表6.2　Network structure.

Name	Patch	Stride	Output map size	Activation function
Input	-	-	80 × 80 × 3	-
Convolution1	5 × 5	1	76 × 76 × 15	ReL
Max-pooling1	2 × 2	1	38 × 38 × 15	-
Convolution2	5 × 5	1	34 × 34 × 30	ReL
Max-pooling2	2 × 2	1	17 × 17 × 30	-
Fullyconnected1	-	-	1 × 1 × 8,000	ReL
Dropout	-	-	1 × 1 × 8,000	-
Fullyconnected2	-	-	1 × 1 × 50	-

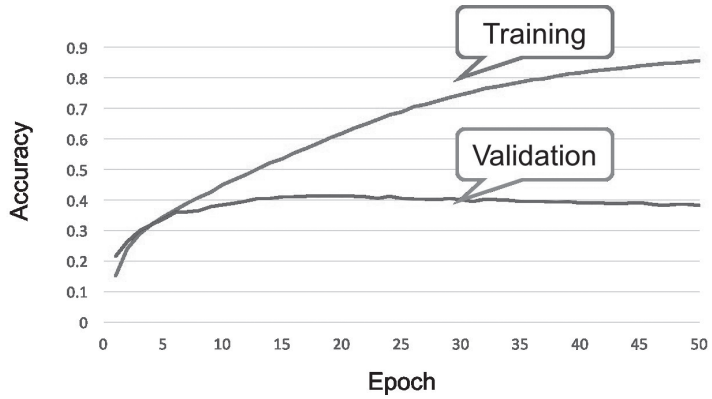
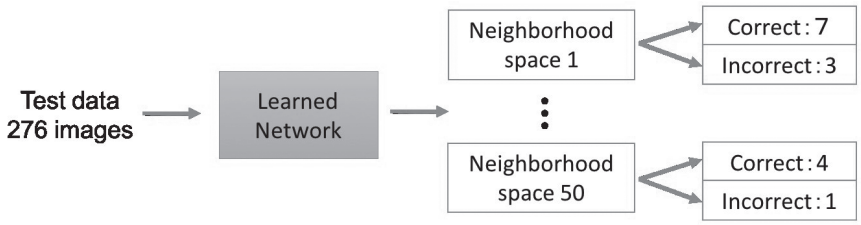


図6.15　Accuracy of neighborhood space.



$$\text{Classification accuracy} = \frac{\text{Total correct in each neighborhood space}}{\text{Number of input data}}$$

図6.16　Experiment 2.

表 6.3　Experiments setting 2.

Classification motif size	50
Representation image size of 1 motif	1
Similar images from each representation images	300
Test data size	291
Batchsize	50
Learning rate ϵ	0.01
Momentum coefficient μ	0.99
Weight decay coefficient λ	0.005
The percentage of vanishing neurons p	50%

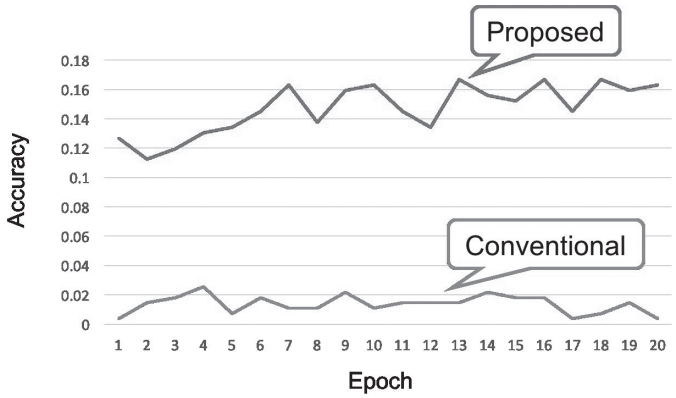


図 6.17　Accuracy of motif classification.

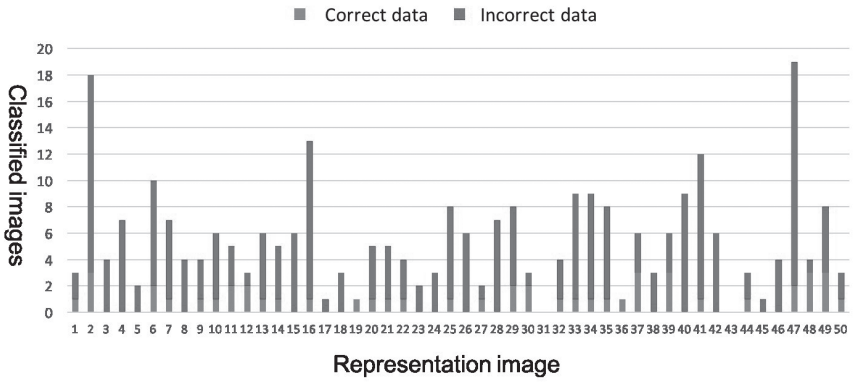


図 6.18　Classification distribution.

チーフは 50 モチーフ中 2 モチーフしかなく，モチーフの分類が極端に偏っていることは確認されなかった。また各モチーフで正解データ数の割合は差が生じており，モチーフによっては最大で 80%程度の精度が得られた。

学習済みの畳み込みニューラルネットワークに対して各代表画像のモチーフと分類された小袖屏風分割画像の例を図 6.19 ～図 6.21 に示す。それぞれ左の画像が代表画像，右の画像が代表画像のモチーフと分類された小袖屏風分割画像群である。図 6.19 より，代表画像に対して画像特徴の類似した画像が同一モチーフであると分類された。具体的には，植物という点では桐と同じ物体が見られる

Representation image

Classified images

図6.19　Examples of motif classification 1.

Representation image

Classified images

図6.20　Examples of motif classification 2.

Representation image

Classified images

図6.21　Examples of motif classification 3.

が, モチーフの観点から見ると松や菊などのモチーフが桐と誤分類されている。図6.20, 図6.21も図6.19と同様に画像特徴の類似した画像が分類されているが，異なるモチーフの画像を含んでいる。

学習が進むと精度が収束したが，更なる精度向上が達成されなかった要因として，モチーフ間の画像特徴が類似していることが考えられる。提案手法では画像特徴空間における類似性のみを考慮して近傍空間を学習しているため，画像特徴が類似していてもモチーフが異なる場合には分類が困難となる。具体的には，小袖屏風中に含まれるモチーフは植物や山，滝といった自然物に偏りがあるため，画像特徴を考慮した近傍空間同士で重なりが発生しやすい傾向があると考えられる。さらに提案手法において他のカテゴリの空間を考慮せずに近傍空間の学習を行っている点についても，近傍空間の重なりが発生する要因として考えられる。分類画像例より，誤分類した画像でも画像特徴の観点から類似した画像が分類されていたため，画像特徴空間における近傍を考慮した学習となっている。本稿では画像特徴の類似性に対する評価が主観的であるため，さらに定量的な指標を加えることで提案手法に対する客観的評価が行える。

6.5 実験３：各モチーフで複数の代表画像を用いた実験

前述の実験ではモチーフごとに代表画像を1枚ずつ選択して用いていたが，各モチーフで複数の代表画像を用いた場合についても実験を行った。本実験では各モチーフで5枚ずつ代表画像を選択し，各代表画像から類似画像を収集して分類を行った。

6.5.1 実験設定

前述の実験と同様に，50種類のモチーフに対して分類を行った。訓練画像の枚数は1枚の代表画像に対して60枚ずつ収集し，モチーフごとの訓練画像数が実験2と同じ300枚となるようにして学習を行った。

本実験のテストデータは教師データ数の多い菊，松，桜の3種類のモチーフに属する画像を用いた。これは小袖屏風画像のモチーフ画像がそもそも少なく，各モチーフから5枚選択している代表画像が実験2におけるテストデータと多く重複しており，代表画像を訓練画像としている比較手法に対して汎化性能を正しく評価できないためである。テストデータは代表画像を菊, 松, 桜のモチーフの中で，代表画像に用いていない画像51枚とした。詳細な実験設定を表6.4に示す。

実験に用いた深層学習のネットワークは表6.2と同様である。

表 6.4 Experiments setting 3.

Classification motif size	50
Representation image size of 1 motif	5
Similar images from each representation images	60
Test data size	51
Batchsize	50
Learning rate ϵ	0.01
Momentum coefficient μ	0.99
Weight decay coefficient λ	0.005
The percentage of vanishing neurons p	50%

6.5.2　実験結果・考察

実験結果を図6.22に示す。結果から比較手法より提案手法のほうは精度が向上しているものの，実験2に比べて提案手法の精度が低下していることが分かる。比較手法において実験2と比較すると，実験3のほうは精度が最大で4%ほど向上しており，学習に用いる訓練画像数が増加したため妥当な結果であると言える。しかし提案手法の精度が前述の実験より低下した原因として，実験2での考察と同様に近傍空間の重なりが考えられる。実験3は実験2に対して1つのモチーフあたりの代表画像の数を増やしているため，学習する近傍空間が広がることが考えられ，モチーフ間の近傍空間が重なってしまったと考えられる。

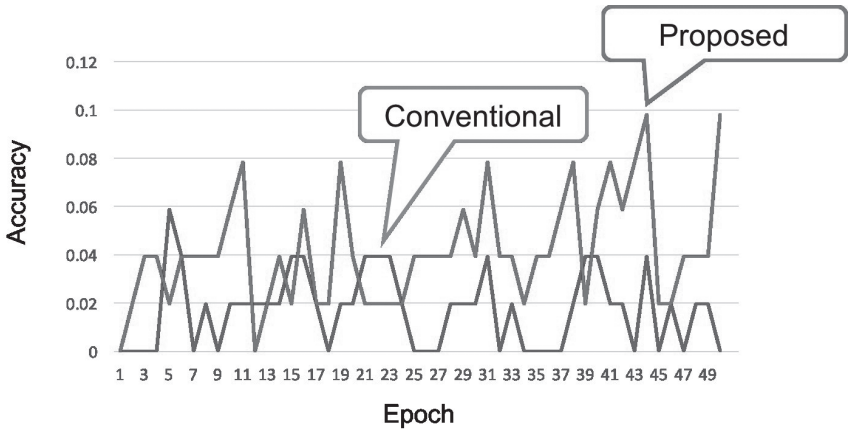


図6.22　Accuracy of motif classification with each 5 representation images.

❼……結論

画像分類を行う際には，分類カテゴリのラベルをもつ教師あり画像による学習が有効であるが，一般にはラベルをもたない画像は圧倒的に数が多く，教師あり画像が十分な数得られない場合に学習が困難となる。さらに近年画像認識分野において高い精度で注目されている深層学習は特に多くのデータ数を必要とし，この課題は顕著となる。

そこで本研究では少数の教師あり画像群から代表画像を選択し，Google画像検索により一般画像特徴空間における類似画像を収集することにより近傍空間を学習する手法を提案した。本稿では少数の教師あり画像を用いた画像分類として，小袖屏風画像におけるモチーフ分類を対象とし実験を行った。実験の結果よりデータ空間のスパース性を緩和することができ，従来では困難であった深層学習の画像分類を可能としたことが明らかとなった。

しかし，画像特徴の類似したカテゴリ間における分類では，学習した近傍空間の重なりが生じやすく，近傍空間の重なる領域では正しくカテゴリ分類するのが困難となる課題が存在する。

そのため，各カテゴリの近傍空間の重なりを考慮することによって近傍空間の重なりが緩和され，分類精度の向上が見込まれる。また本実験では小袖屏風画像を1,200分割したが，これは厳密にモチーフの領域を基にした分割とはなっていないため，分割方法が分類精度に影響を与えてしまう。そこでモチーフ領域に対する適切なセグメンテーションにより分類精度の向上が見込まれる。

参考文献

［1］Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh: “A fast learning algorithm for deep belief nets.”, Neural Computation, vol.18, no.7, pp.1527–1554, 2006.

［2］Quoc V. Le, Marc’ Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeff Dean, and Andrew Y. Ng: “Building high-level features using large scale unsupervised learning.”, International Conference on Machine Learning, 2012.

［3］Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton.: “On the importance of initialization and momentum in deep learning”, Journal of Machine Learning, pp.1139–1147, 2013.

［4］David G. Lowe.: “Object recognition from local scale-invariant features.”, The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol.2, pp.1150–1157, 1999.

［5］Herbert Bay, Tinne Tuytelaars, and Luc V. Gool.: “SURF: Speeded Up Robust Features.”, The Proceedings of the Ninth European Conference on Computer Vision, pp.404–417, 2006.

［6］Stuart Russell, and Peter Norvig: “Artificial Intelligence, A Modern Approach, Second Edition”, Pearson Education, 2003.

［7］Google 画像検索，https://www.google.co.jp/imghp

［8］Alex Krizhevsky, Ilya Sutskever, and Geoffrey E.Hinton: “ImageNet Classification with Deep Convolutional Neural Networks”, Advances in Neural Information Processing Systems, pp.1097–1105, 2012.

［9］Large Scale Visual Recognition Challenge 2010（ILSVRC2010）, http://imagenet.org/challenges/LSVRC/2010/

［10］岡谷貴之：“機械学習プロフェッショナルシリーズ深層学習”，株式会社講談社，2015.

［11］国立歴史民俗博物館：“国立歴史民俗博物館資料図録 2 野村コレクション小袖屏風”，2002.

［12］データベースれきはく，http://www.rekihaku.ac.jp/doc/t-db-index.html

［13］Chainer: A flexible framework of neural networks, http://chainer.org/

（横浜国立大学大学院工学研究院，国立歴史民俗博物館共同研究員）

（2019 年 3 月 14 日受付，2019 年 8 月 5 日審査終了）

Bulletin of the National Museum of Japanese History
Vol.220 March 2020

A Neighbor Space Learning by Similar Image for the Easing of Sparsity

HAMAGAMI Tomoki

With the development of advanced image classification applications, the improvement of learning accuracy is more required. Recently, the feature extraction and learning methods for improving classification accuracy are progressing. In especially, deep learning approaches attract significant attention.

The CNN (convolutional neural network) is widely used in deep learning for image data. When learning, it needs to use a massive number of image data with labels as a supervisor. However, if the number of labeled data is not sufficient, it fails to classify unknown data because it cannot cover the whole space of data. Especially the labeling works which require specialist skill takes high cost; it cannot expect to prepare an enough dataset. To overcome the limitation, a high-precision classification method from the small dataset is required.

This study proposes a new learning method with small labeled data and its neighbor space. This method collects the neighbor samples with representative data in labeled data and learn its neighbor space as a cluster. The experiment results show the technique enables to learn sparse space as a cluster space effectively and can allow classification of the motif of Kosode byobu image.

Key words: CNN, similar image retrieval, image classification, sparse space