Vol.2010-CVIM-171 No.25
2010/3/19

# 類似画像検索における部分教師付き特徴次元圧縮

吉田正和[†1]　松本哲也[†1]　大西昇[†1]

近年，ディジタルカメラ，カメラ付き携帯電話の普及により，ディジタル画像に関する知識を全く持たない人々が多量の画像を収集・蓄積できるようになってきた．そのため，ユーザに特別な知識を要求せずに，容易な操作で，画像を分類・検索できるシステムが求められている．ユーザの画像データベースには，多様なカテゴリの画像が混在しており，完全に自動化されたカテゴリ分類は困難である．また，人手のみによるラベル付けは煩雑である．そこで，本研究では，教師となる少数の画像のカテゴリを指定するだけで，高精度なカテゴリ分類を実現することを目的とする．本研究では，カテゴリ学習の手法として，部分教師付き学習手法を採用するとともに，学習精度をより向上させるために，新たに部分教師付き次元圧縮手法を提案した．

自然画像を用いた評価実験の結果，教師なし次元圧縮手法を用いた場合と比較して，識別率が平均して 5.1 ポイント上昇した．

# Semi-supervised Dimensionality Reduction to Content-Based Image Retrieval

Masakazu Yoshida,[†1] Tetsuya Matsumoto[†1] and Noboru Ohnishi[†1]

Recently, it has become usual that computer user who has no knowledge of digital image collect and accumulate a large amount of image data by spreads such as digital cameras and mobile phones with cameras. Therefore, it is necessary to develop the system which can classify or retrieve requested image easily even for novice users. It is difficult to classify images into correct categories automatically, because image databases of users have various categories of images. And, it is not easy to label images by hands as the number of images increase. The goal of this study aims to achieve a highly accurate category classification only by specifying the category of small number of images which will become supervisors. This study examines semi-supervised dimensionality reduction using small number of images which will become supervisors used for content based image retrieval.

By the result of experiments, the classification rate of the proposed method was 5.1 points better than that of the unsupervised method.

## 1. はじめに

近年，パソコン利用者の増加やディジタルカメラ，カメラ付き携帯電話の普及により，ディジタル画像に関する知識を全く持たない人々が多量の画像を収集・蓄積できるようになってきた．そのため，ユーザに特別な知識を要求せずに，容易な操作で，画像を分類・検索できるシステムが求められている[1)2)3)]．

この目的を実現するために，ユーザに対して，高度な専門知識や煩雑な操作を要求せず，画像の分類・検索を自動的に行えることが求められる．しかし，対象とする一般のユーザの画像データベースには，多様なカテゴリの画像が混在しており，完全に自動化されたカテゴリ分類は困難である．また，全ての画像に対して人手でラベル付けを行うことも，画像の枚数が多くなる程，ユーザへかかる負担が増加する．

そこで，本研究では，ユーザから提示された少数の典型的画像（教師）に関する情報を最大限に利用して，各ユーザ特有の画像カテゴリ分類基準を推定するシステムを提案する．これにより，専門知識を持たない一般ユーザでも，少数の画像に対するカテゴリを与えることによって，全画像に対する自動ラベル付けを実現することができる．本研究では主に，特徴次元圧縮の部分に部分教師付き次元圧縮を導入することによって，精度の向上を目指した．

本研究の対象となる画像カテゴリの自動ラベル付けシステムは，以下のような前提を満たすと考えられる．

1. 対象となる画像データベースの画像カテゴリは，あらかじめ想定できない．従って，どの様な画像データベースにも対応できるように，十分な画像特徴量を用意する必要がある．一般に，画像特徴量は高次元で，多くの冗長性を有しており，何らかの特徴量集約（次元圧縮）手法が必要不可欠となる．

2. 類似画像検索の分野では，関連フィードバック（Relevance Feedback）等の手法でユーザから情報フィードバックを得ることは一般的であり，学習のための少数の教師データを得ることには，一般に何ら問題ない．この事から，部分教師付き手法を採用することは，合理的であると言える．

3. 一般に，利用者の意図するカテゴリ分布は，単純な正規分布で近似できる程単純であることは稀であり，システムの有用性を高めるには，より複雑なデータの分布に対応可能な

†1 名古屋大学大学院
Nagoya University Graduate School

GMM(Gaussian Mixture Model) 等のモデリング手法を採用することが望ましい.

上記前提に基づき，本研究では，最終的なデータ分布が GMM でモデル化される場合に，モデル化に必要十分な情報を保存しつつ，画像特徴量の次元を圧縮し，モデル学習の精度を向上できるような部分教師付き手法を提案する．

次元圧縮後のデータの分布は GMM[4] で表し，GMM のパラメータの学習には，EM アルゴリズム[5] を用いた部分教師付き学習を用いる．これによって，少数の画像のカテゴリ情報のみが教師データとして与えられた時に，画像データベース中のすべての画像のカテゴリ分布を精度良く推定する事が可能となる．

## 2. 特徴次元圧縮

### 2.1 定 式 化

$\mathbf{x}_i \in \mathbf{R}^d (i = 1, 2, ..., n)$ を d 次元のサンプルとし，$y_i \in 1, 2, ..., l$ をサンプル $i$ のカテゴリとする．$l$ はカテゴリの数，$n_i$ はカテゴリ $i$ のサンプル数,．$n = \sum_{i=1}^{l} n_i$ はサンプルの数である．$\mathbf{z}_i \in \mathbf{R}^m (1 \leq m \leq d)$ を埋め込み空間とする，$m$ は埋め込み空間の次元数である．

本研究では，$d \times m$ の変換行列 $\mathbf{T}$ を用い，線形次元圧縮を行う．$\mathbf{z}_i$ は以下の式で得られる．

$$\mathbf{z}_i = \mathbf{T}^\top \mathbf{x}_i \tag{1}$$

$\top$ は行列やベクトルの転置を表す.

### 2.2 教師なし次元圧縮

#### 2.2.1 PCA(Principal Component Analysis)

PCA は，最もよく知られた次元圧縮手法である．サンプル $\{\mathbf{x}_i\}_{i=1}^n$ が与えられた時，埋め込み空間内のサンプルの分散が最も大きくなるような，変換行列 $\mathbf{T}$ を得ようとする．分散行列 $\Sigma_t$ は以下の式で得られる．

$$\mathbf{\Sigma}_t = \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top \tag{2}$$

$\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$ は全サンプルの平均である．PCA の解は以下の固有ベクトルを得ることによって求められる．

$$\mathbf{\Sigma}_t \mathbf{z} = \lambda \mathbf{z} \tag{3}$$

式 3 の固有値方程式の解を固有値の大きい順に $(\lambda_i, \mathbf{z}_i)(i = 1, ..., n)$ とすると，$\mathbf{T}^\top = (\mathbf{z}_1, ..., \mathbf{z}_m)^\top$ で表される．

### 2.3 教師あり次元圧縮

#### 2.3.1 FDA(Fisher Discriminant Analysis)

FDA は一般的な教師あり次元圧縮手法である．$n'$ 個のラベル付きのサンプルがあるとする．$y_i$ はサンプル $x_i$ に関連付けられたカテゴリのラベルであり，$c$ はカテゴリの数である．$n'_m$ はカテゴリ $m$ のラベル付けされたサンプルの数である．

$\mathbf{\Sigma}_b$ と $\mathbf{\Sigma}_w$ は，それぞれ級間分散行列と級内分散行列であり，以下のように定義される．

$$\mathbf{\Sigma}_b = \frac{1}{n'} \sum_{m=1}^{c} n'_m (\boldsymbol{\mu}_m - \boldsymbol{\mu})(\boldsymbol{\mu}_m - \boldsymbol{\mu})^\top \tag{4}$$

$$\mathbf{\Sigma}_w = \frac{1}{n'} \sum_{m=1}^{c} \sum_{i:y_i=m} (\mathbf{x}_i - \boldsymbol{\mu}_m)(\mathbf{x}_i - \boldsymbol{\mu}_m)^\top \tag{5}$$

$\boldsymbol{\mu}_m = \frac{1}{n'_m} \sum_{i:y_i=m} \mathbf{x}_i$ はカテゴリ $m$ のサンプルの平均である．

FDA は，埋め込み空間内で級間分散を大きく級内分散を小さくする，変換行列 $\mathbf{T}$ を得ようとする．解は以下の一般的化固有値問題の解として求めることができる．

$$\mathbf{\Sigma}_b \mathbf{z} = \lambda \mathbf{\Sigma}_w \mathbf{z} \tag{6}$$

式 6 の固有値方程式の解を固有値の大きい順に $(\lambda_i, \mathbf{z}_i)(i = 1, ..., n)$ とすると，$\mathbf{T}^\top = (\mathbf{z}_1, ..., \mathbf{z}_m)^\top$ で表される．

#### 2.3.2 LFDA(Local Fisher Discriminant Analysis)[6]

LFDA は，FDA を局所的な形へ改良した，教師あり次元圧縮手法である．

$\mathbf{A}_{i,j} (\in [0, 1])$ を $\mathbf{x}_i$ と $\mathbf{x}_j$ の類似度とし，類似度は対称性を持つ．すなわち，$\mathbf{A}_{i,j} = \mathbf{A}_{j,i}$ であるものとする．類似度を決定するいくつかの手法があるが，本研究では局所スケーリング[8] を用いる．

$$\mathbf{A}_{i,j} = exp\Big(-\frac{\| \mathbf{x}_i - \mathbf{x}_j \|^2}{\sigma_i \sigma j}\Big) \tag{7}$$

$\sigma_i = \| \mathbf{x}_i - \mathbf{x}_i^{(k)} \|$ は $\mathbf{x}_i$ 周辺の局所スケールを表し，$\mathbf{x}_i^{(k)}$ は $\mathbf{x}_i$ から $k$ 番目に近いサンプルである．

$\mathbf{\Sigma}_{lb}$ と $\mathbf{\Sigma}_{lw}$ は，それぞれ局所級間分散行列と局所級内分散行列であり，以下のように定義される．

$$\mathbf{\Sigma}_{lb} = \frac{1}{2} \sum_{i,j=1}^{n'} \mathbf{W}_{i,j}^{(lb)} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \tag{8}$$

$$\mathbf{\Sigma}_{lw} = \frac{1}{2} \sum_{i,j=1}^{n'} \mathbf{W}_{i,j}^{(lw)} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \tag{9}$$

$\mathbf{W}^{(lb)}$ と $\mathbf{W}^{(lw)}$ は $n' \times n'$ の行列であり，以下のように表される．

Vol.2010-CVIM-171 No.25
2010/3/19

**表 1** 次元圧縮手法の比較
Table 1 comparison of dimensionality reduction method

| 次元圧縮手法 | 圧縮後の次元数の制限 | 多峰性のデータに対して | 計算速度 |
|---|---|---|---|
| PCA | ○ | × | ○ |
| FDA | × | × | △ |
| LFDA | ○ | ○ | × |

$$\mathbf{W}_{i,j}^{(lb)} = \begin{cases} \mathbf{A}_{i,j}(1/n' - 1/n'_{y_i}) & (y_i = y_j) \\ 1/n' & (y_i \neq y_j) \end{cases} \tag{10}$$

$$\mathbf{W}_{i,j}^{(lw)} = \begin{cases} \mathbf{A}_{i,j}/n'_{y_i} & (y_i = y_j) \\ 0 & (y_i \neq y_j) \end{cases} \tag{11}$$

全ての $i, j$ に対して $\mathbf{A}_{i,j} = 1$ ならば，上記の行列は FDA で用いられるものと同じである．従って，LFDA は FDA の局所的な変更であると言える．

LFDA は，埋め込み空間内で局所級間分散を大きく，局所級内分散を小さくする．変換行列 $\mathbf{T}$ を得ようとする．解は以下の一般化固有値問題の解として求めることができる．

$$\mathbf{\Sigma}_{lb}\mathbf{z} = \lambda\mathbf{\Sigma}_{lw}\mathbf{z} \tag{12}$$

式 12 の固有値方程式の解を固有値の大きい順に $(\lambda_i, \mathbf{z}_i)(i = 1, ..., n)$ とすると，$\mathbf{T}^\top = (\mathbf{z}_1, ..., \mathbf{z}_m)^\top$ で表される．

FDA に対する LFDA の利点は，以下の 2 つである．FDA は，カテゴリ内に多峰性や外れ値がある時，正しい解が求まらない[7]．しかし，LFDA は，局所的な級内分散を解くことによって，この弱点を解決する．また，FDA では，圧縮後の次元数 $r$ が最大で $c - 1$ である[7]．しかし，LFDA は，類似度 $\mathbf{A}_{i,j}$ によって，圧縮後の次元数を任意の数にすることができる．

図 1 に，人工データを用いたシミュレーション実験の結果を示す．カテゴリ数は 2 で，それぞれの手法によって，2 次元から 1 次元にデータを圧縮する時に射影する軸を示している．

単峰性のデータ（図 1 左）に対しては，FDA と LFDA はともにカテゴリ情報を保持した次元圧縮が行えていて，ほぼ同じ結果であることがわかる．多峰性のデータ（図 1 右）に対しては，LFDA ではカテゴリ情報を保持できるが，FDA では射影後のデータはカテゴリ情報が保持されていないことがわかる．PCA に関しては，教師情報を用いずデータの分散のみに依存するので，図 1 のどちらのデータに対しても，クラス間分離を保存した変換行列が得られるとは限らない．表 1 に，それぞれの次元圧縮手法の特徴をまとめた．

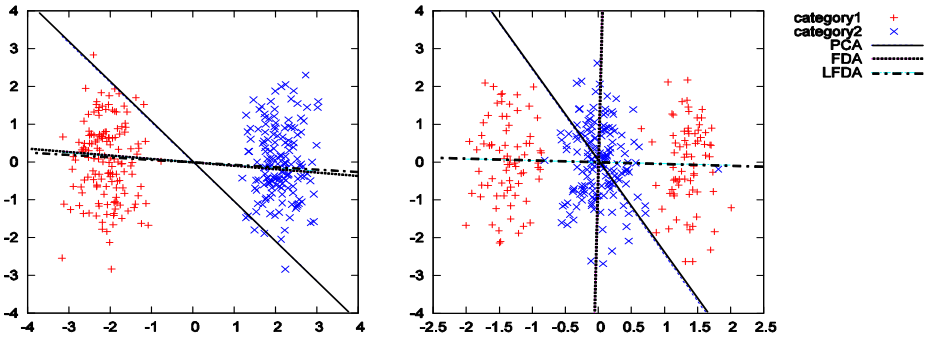


**図 1** 人工データによる実験結果（左：単峰性データ，右：多峰性データ）
Fig. 1 experimental result for artificial data (left : monomodal data, right : multimodal data)

## 3. 部分教師付き次元圧縮

この章では，今回提案する部分教師付き次元圧縮手法について述べる．まず，少数教師の場合に起こる問題点を述べ，その後に，部分教師付き次元圧縮の処理の流れを述べる．

FDA を用いた部分教師付き次元圧縮手法を，α-SFDA(α-Semi-supervised Fisher Discriminant Analysis)，LFDA を用いた部分教師付き次元圧縮手法を，α-SLFDA(α-Semi-supervised Local Fisher Discriminant Analysis) とする．

### 3.1 少数教師による問題点

部分教師付き次元圧縮では，教師の数が少ない場合，級内分散（$\Sigma_w$）が縮退してしまい，式 6 の計算が正しく行えず，正常に次元圧縮が行えない問題が起こる．具体的には，特徴量の次元数が $n$ である時，教師数が $n + 1$ 個以上ないと，分散共分散行列が $n$ 次元のフルランクを満たさないため，縮退が起こる．その問題を解決するために，本研究では，$\Sigma_{b\_estimated}$ と $\Sigma_{w\_estimated}$ を定義する．また，2.3.1 でも述べたが，FDA では圧縮後の次元数が最大で（カテゴリ数 −1）に制限されてしまう．圧縮後の次元数を増やすために，本研究では，$\Sigma_{b\_supervised}$ を定義する．

本研究では，以上の $\Sigma_{b\_estimated}$，$\Sigma_{w\_estimated}$ と $\Sigma_{b\_supervised}$ を用いることにより，部分教師付き次元圧縮を行う．

Vol.2010-CVIM-171 No.25
2010/3/19

### 3.2 FDA の改良版 α-SFDA

3.1 で述べた問題点のために，本研究では FDA をそのまま利用することはできない．そこで，$\Sigma_b$，$\Sigma_w$ の定義を変更した手法を提案する．

**$\Sigma_{b\_estimated}$，$\Sigma_{w\_estimated}$ の定義**

まず，カテゴリ未知の教師なしデータ（ラベルなしデータ）のカテゴリ推定を行う，今回は単純な手法であるが，カテゴリ既知の教師データ（ラベル付きデータ）のみで各カテゴリの「カテゴリ平均」を求め，全ての教師なしデータと「カテゴリ平均」の距離を求め，最も近い「カテゴリ平均」のカテゴリを付与する（図 2）．

$\Sigma_{b\_estimated}$，$\Sigma_{w\_estimated}$ は，推定されたカテゴリ情報を用いて，以下の式で求める．

$$\boldsymbol{\Sigma}_{b\_estimated} = \sum_{m=1}^{c} \frac{n_{e\_m}}{n}(\boldsymbol{\mu}_{e\_m} - \boldsymbol{\mu})(\boldsymbol{\mu}_{e\_m} - \boldsymbol{\mu})^{\top} \tag{13}$$

$$\boldsymbol{\Sigma}_{w\_estimated} = \sum_{m=1}^{c} \sum_{i:y_{e\_i}=m} \frac{1}{n}(\mathbf{x}_i - \boldsymbol{\mu}_{e\_m})(\mathbf{x}_i - \boldsymbol{\mu}_{e\_m})^{\top} \tag{14}$$

$n_{e\_m}$ はカテゴリが $m$ であると推定されたデータ数，$y_{e\_i}$ はサンプル $i$ の推定されたカテゴリ，$\boldsymbol{\mu}_{e\_m} = \frac{1}{n_{e\_m}} \sum_{i:y_{e\_i}=m} \mathbf{x}_i$ はカテゴリが $m$ であると推定されたサンプルの平均，$\boldsymbol{\mu}$ は全サンプルの平均，$n$ は全てのサンプル数である．推定されたカテゴリ情報を用い，全てのデータを用いるため，大域的な情報を用い圧縮を行うことができる．

$\Sigma_{b\_estimated}$，$\Sigma_{w\_estimated}$ は，1 つのカテゴリの分布が 1 つの正規分布からなると仮定し計算している．また，カテゴリの推定に最も近い「カテゴリ平均」を用いているため，多峰性がある場合は正しく計算できない．$\Sigma_{b\_estimated}$，$\Sigma_{w\_estimated}$ の計算方法を簡単に説明した図を図 2 に示す．

**$\Sigma_{b\_supervised}$ の定義**

与えられた教師データのみを用い，以下の計算を行う．

$$\boldsymbol{\Sigma}_{b\_supervised} = \sum_{i}^{n'} \sum_{j:y_i \neq y_j} \frac{1}{n'}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\top} \tag{15}$$

$n'$ は教師データ数であり，$y_i$ はサンプル $i$ のカテゴリである．教師データのみを用いるため，局所的なデータの分布を考慮し，次元圧縮を行う．

$\Sigma_{b\_supervised}$ は，1 つのカテゴリの分布が複数の正規分布（各カテゴリの正規分布の数が各カテゴリの教師数である）からなると仮定して計算している．それぞれの正規分布の平均は教師データで与えられ，分散共分散行列は同一であるとしている．$\Sigma_{b\_supervised}$ の計算方法を簡単に説明した図を図 2 に示す．

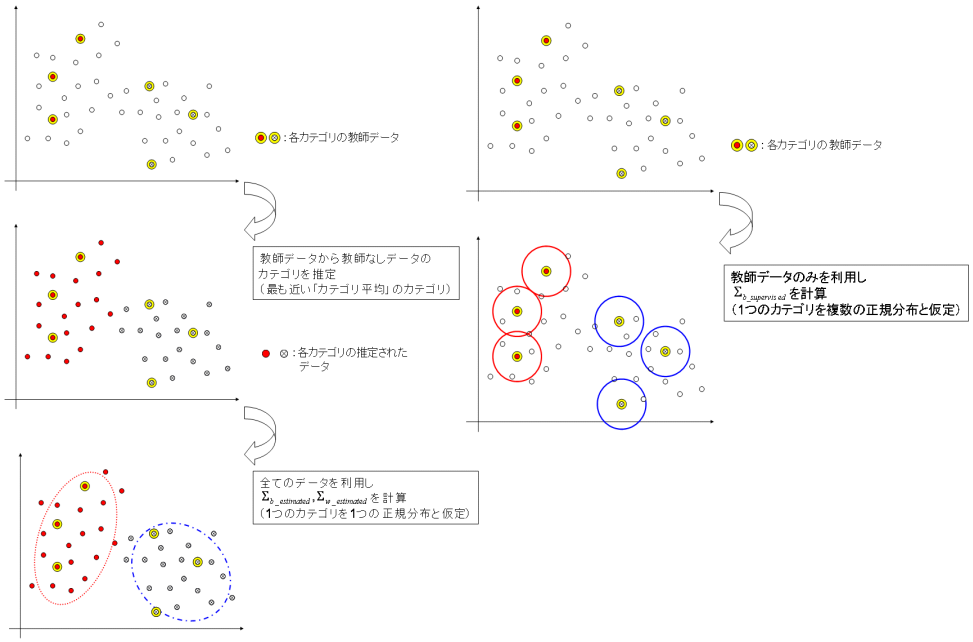


**図 2** $\Sigma_{b\_estimated}$，$\Sigma_{w\_estimated}$，$\Sigma_{b\_supervised}$ の計算方法
Fig. 2 calculation method of $\Sigma_{b\_estimated}$,$\Sigma_{w\_estimated}$,$\Sigma_{b\_supervised}$

**$\Sigma_{b\_semi-supervised}$ の定義**

最終的な級間分散行列 $\Sigma_{b\_semi-supervised}$ は，$\Sigma_{b\_supervised}$ と $\Sigma_{b\_estimated}$ のパラメータ $\alpha$ による重み付け和であり，以下の式で定義される．

$$\Sigma_{b\_semi-supervised} = \alpha \times \Sigma_{b\_supervised} + (1-\alpha) \times \Sigma_{b\_estimated} \tag{16}$$

**α-SFDA の計算方法**

与えられた教師なしデータと教師データを用いて，前述の $\Sigma_{b\_semi-supervised}$ と $\Sigma_{w\_estimated}$ を計算し，以下の一般化固有値問題の解を求める．

$$\boldsymbol{\Sigma}_{b\_semi-supervised}\mathbf{z} = \lambda \boldsymbol{\Sigma}_{w\_estimated}\mathbf{z} \tag{17}$$

式 17 の固有値方程式の解を固有値の大きい順に $(\lambda_i, \mathbf{z}_i)(i = 1, ..., n)$ とすると，次元圧縮の変換行列は，$\mathbf{T}^{\top} = (\mathbf{z}_1, ..., \mathbf{z}_m)^{\top}$ で与えられる．

Vol.2010-CVIM-171 No.25
2010/3/19

### 3.3 LFDA の改良版α-SLFDA

同様に，3.1 で述べた問題点のために，LFDA もそのまま利用することはできず，$\Sigma_{lb}$，$\Sigma_{lw}$ の定義を変更した手法を提案する．

**$\Sigma_{lb\_estimated}$，$\Sigma_{lw\_estimated}$ の定義**

$\Sigma_{lb\_estimated}$，$\Sigma_{lw\_estimated}$ の教師なしデータのカテゴリ推定は，教師なしデータと全ての教師ありデータの距離を比較し，最も近い「教師データ」のカテゴリを付与する（図 3）．推定したカテゴリ情報を用い，全てのデータに対し，以下の式で計算される．

$$\Sigma_{lb\_estimated} = \frac{1}{2}\sum_{i,j=1}^{n} \mathbf{W}_{i,j}^{(lb\_estimated)}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\top} \tag{18}$$

$$\Sigma_{lw\_estimated} = \frac{1}{2}\sum_{i,j=1}^{n} \mathbf{W}_{i,j}^{(lw\_estimated)}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\top} \tag{19}$$

$\mathbf{W}^{(lb\_estimated)}$ と $\mathbf{W}^{(lw\_estimated)}$ は $n \times n$ の行列であり，以下のように表される．

$$\mathbf{W}_{i,j}^{(lb\_estimated)} = \begin{cases} \mathbf{A}_{i,j}(1/n - 1/n_{e\_y_{e\_i}}) & (y_{e\_i} = y_{e\_j}) \\ 1/n & (y_{e\_i} \neq y_{e\_j}) \end{cases} \tag{20}$$

$$\mathbf{W}_{i,j}^{(lw\_estimated)} = \begin{cases} \mathbf{A}_{i,j}/n_{e\_y_{e\_i}} & (y_{e\_i} = y_{e\_j}) \\ 0 & (y_{e\_i} \neq y_{e\_j}) \end{cases} \tag{21}$$

$n_{e\_y_{e\_i}}$ はカテゴリが $y_{e\_i}$ であると推定されたサンプル数，$y_{e\_i}$ はサンプル $i$ の推定されたカテゴリである．$\mathbf{A}_{i,j}$ は式 7 を用い計算される．$\Sigma_{lb\_estimated}$ は，教師なしデータのカテゴリ情報の付与方法を，最も近い「教師データ」のカテゴリとすることにより，データが多峰性を持つ場合にも対応できる．また，$\Sigma_{lb\_supervised}$ は，異なるカテゴリのデータのみの分散を計算することにより，教師データのみの局所的な分散を考慮している．$\Sigma_{lb\_estimated}$，$\Sigma_{lw\_estimated}$ の計算方法を簡単に説明した図を図 3 に示す．

**$\Sigma_{lb\_supervised}$ の定義**

$\Sigma_{lb\_supervised}$ は，与えられた教師データのみを利用し，以下の式で計算される．

$$\Sigma_{lb\_supervised} = \frac{1}{2}\sum_{i,j=1}^{n'} \mathbf{W}_{i,j}^{(lb\_supervised)}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\top} \tag{22}$$

$\mathbf{W}^{(lb\_supervised)}$ は $n' \times n'$ の行列であり，以下のように表される．

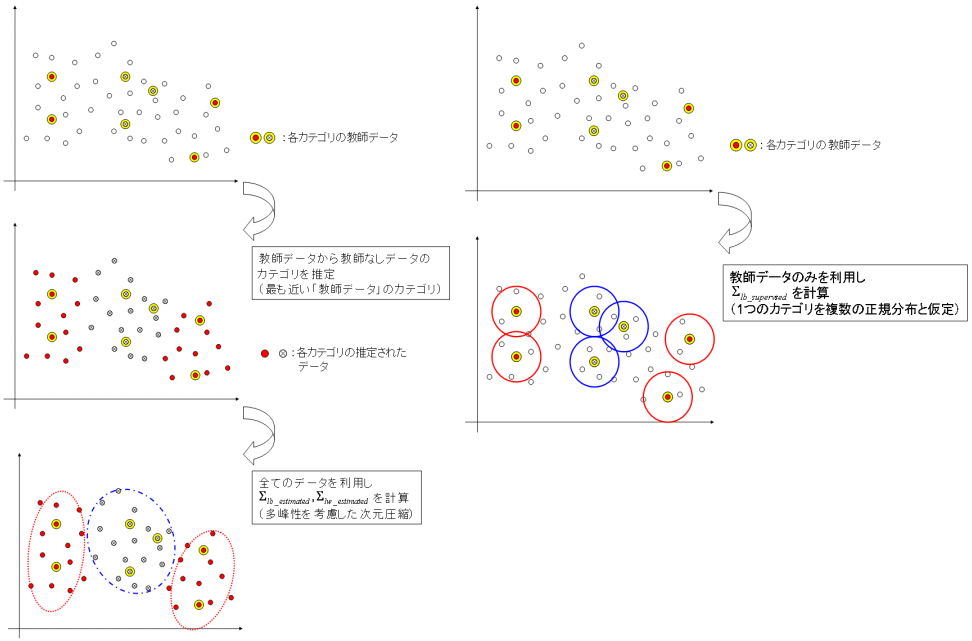


**図 3** $\Sigma_{lb\_estimated}$，$\Sigma_{lw\_estimated}$，$\Sigma_{lb\_supervised}$ の計算方法
Fig. 3 calculation method of $\Sigma_{lb\_estimated}$,$\Sigma_{lw\_estimated}$,$\Sigma_{lb\_supervised}$

$$\mathbf{W}_{i,j}^{(lb\_supervised)} = \begin{cases} 0 & (y_i = y_j) \\ 1/n' & (y_i \neq y_j) \end{cases} \tag{23}$$

$\Sigma_{lb\_supervised}$ は，$\Sigma_{lb\_supervised}$ の場合と同様に，1 つのカテゴリの分布が複数の正規分布（各カテゴリの正規分布の数が各カテゴリの教師数である）からなると仮定して計算している．それぞれの正規分布の平均は教師データで与えられ，分散共分散行列は同一であるとしている．$\Sigma_{lb\_supervised}$ の計算方法を簡単に説明した図を図 3 に示す．

**$\Sigma_{lb\_semi-supervised}$ の定義**

最終的な級間分散行列 $\Sigma_{lb\_semi-supervised}$ は，$\Sigma_{lb\_supervised}$ と $\Sigma_{lb\_estimated}$ のパラメータ $\alpha$ による重み付け和であり，以下の式で定義される．

$$\Sigma_{lb\_semi-supervised} = \alpha \times \Sigma_{lb\_supervised} + (1-\alpha) \times \Sigma_{lb\_estimated} \tag{24}$$

Vol.2010-CVIM-171 No.25
2010/3/19

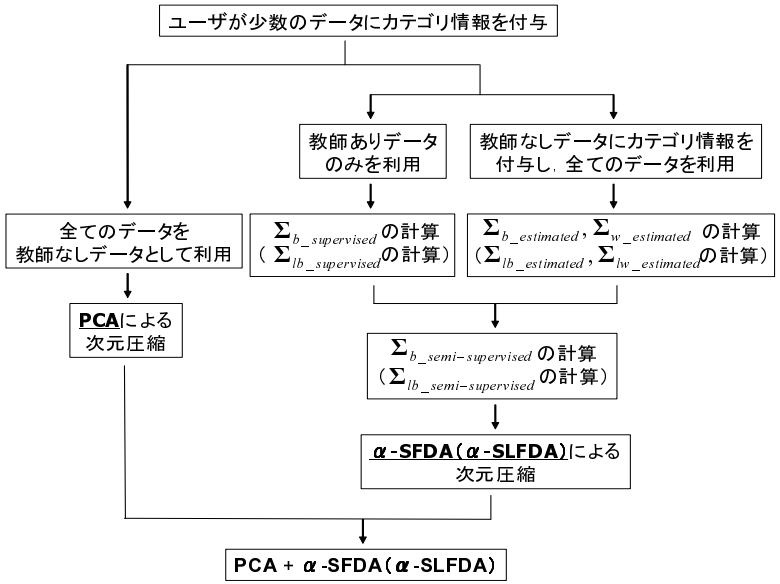


**図 4** 部分教師付き次元圧縮の処理の流れ
Fig. 4 process of semi-supervised dimensionality reduction

### α-SLFDA の計算方法

与えられた教師なしデータと教師データを用いて，前述の $\Sigma_{lb\_semi-supervised}$ と $\Sigma_{lw\_estimated}$ を計算し，以下の一般化固有値問題の解を求める．

$$\Sigma_{lb\_semi-supervised}\mathbf{z} = \lambda\Sigma_{lw\_estimated}\mathbf{z} \tag{25}$$

式 25 の固有値方程式の解を固有値の大きい順に $(\lambda_i, \mathbf{z}_i)(i = 1, ..., n)$ とすると，次元圧縮の変換行列は，$\mathbf{T}^\top = (\mathbf{z}_1, ..., \mathbf{z}_m)^\top$ で与えられる．

### 3.4 部分教師付き次元圧縮の流れ

本研究で提案する部分教師付き次元圧縮手法は，2.2 で述べた教師なし次元圧縮手法 PCA によって得られた低次元特徴空間と，3.2，3.3 で述べた教師情報を利用した次元圧縮手法である α-SFDA または α-SLFDA によって得られた低次元特徴空間をそれぞれ組み合わせることにより，特徴次元の圧縮を行う手法である．

図 4 に，本研究で用いた部分教師付き次元圧縮手法の処理の流れを示す．

## 4. 部分教師付きカテゴリ学習

部分教師付き学習では，多数の教師なし標本から得られる確率分布に関する情報と少数の教師付き標本から得られる情報を統合することにより，効率的な学習が可能となる．

本研究では，GMM（混合ガウス分布モデル）により各カテゴリの母集団をモデル化し，GMM のパラメータ推定問題として，部分教師付きカテゴリ分類を定式化している．GMM による全体の確率密度は，以下の式で表される．

$$p(\mathbf{x}|\lambda) = \sum_{c=1}^{C} p(c)p_c(\mathbf{x}|\lambda_c) \tag{26}$$
$$\lambda = \{\alpha_{cm}, \mu_{cm}, \mathbf{\Sigma}_{cm}\} \quad (c = 1, ..., C\ m = 1, ..., M_c)$$

$p(c)$ はカテゴリ $c$ の生起確率，$p_c(\mathbf{x}|\lambda_c)$ はカテゴリ $c$ の密度関数，$M_c$ はカテゴリ $c$ を構成するガウス分布数，$\alpha_{cm}, \mu_{cm}, \mathbf{\Sigma}_{cm}$ はそれぞれカテゴリ $c$ の $m$ 番目のガウス分布の，混合比，平均ベクトル，共分散行列を表す．

部分教師付き学習で利用する尤度は以下の式で表される．

$$\beta \sum_{n=1}^{N_l} log\, p(c') \sum_{m=1}^{M_{c'}} \alpha_{c'm} N(\mathbf{x}_n|\mu_{c'm}, \mathbf{\Sigma}_{c'm}) + (1-\beta) \sum_{n=1}^{N_u} log \sum_{c=1}^{C} p(c) \sum_{m=1}^{M_c} \alpha_{cm} N(\mathbf{x}_n|\mu_{cm}, \mathbf{\Sigma}_{cm}) \tag{27}$$

$N_l$ はカテゴリが既知であるデータの数，$N_u$ はカテゴリが未知であるデータの数を表す．EM アルゴリズムを用い，$\alpha, \mu, \mathbf{\Sigma}$ を更新していき，式 27 の値の変化が閾値を下回った時に学習を終了する．

## 5. 評 価 実 験

実際の自然画像からなるデータベースを用いて，今回提案した部分教師付き次元圧縮手法の評価実験を行った．

実験には，建物，植物，動物の 3 種類のカテゴリの画像をそれぞれ 80 枚ずつ（合計 240 枚）用いた．画像データは，色相のヒストグラムから 12 次元，HSV 各成分ごとの Haar ウェーブレット変換により，5 レベルの多重解像度解析を行って得られた低周波成分 48 次元，3 レベルまでの高周波成分について画素値の絶対値が閾値を超えた割合 27 次元の計 87 次元の特徴ベクトルに変換した．各カテゴリの教師数は，4，8，12 で，教師によって結果が大きくことなるため，各教師数毎に 5 組のサンプルを用意し，その平均を結果として示す．

式 16，式 24 で用いる重み $\alpha$ は，特に記述がない限り $\alpha = 0.1$ とする．また，圧縮後の

特徴量の次元数は 10 次元であり，α-SFDA，α-SLFDA を 2 次元，PCA を 8 次元とする．LFDA の局所スケールの計算に用いる $k$ の値は，$k = 80$ とする．

評価尺度には，級間分散・級内分散比と識別率を用いた．級間分散・級内分散比は，特徴次元圧縮後のデータの分布がカテゴリ間でどれだけ離れているかを知る指標であり，識別率は，カテゴリ学習により構成された識別器がどれほどの精度で画像を分類できるかを示す指標である．表の中では，左に平均値，右上に最大値，右下に最小値を示した．

### 5.1 部分教師付き次元圧縮と教師なし次元圧縮の比較

部分教師付き次元圧縮と教師なし次元圧縮の比較を行った結果を表 2（級間分散・級内分散比）と表 3（識別率）に示す．GMM の学習にはどちらの場合も部分教師付き学習を用いた．部分教師付き次元圧縮は，$\alpha$ の値が，0.0（$\Sigma_{b\_estimated}$ のみ利用），0.1，1.0（$\Sigma_{b\_supervised}$ のみ利用）の 3 種類の結果を示す．理想的な値として，全てのデータのカテゴリを既知として FDA を行い（$\Sigma_b$，$\Sigma w$ ともに本来の定義を用いた），PCA と組み合わせた時の結果も示す（※）（ただし，学習には部分教師付き学習を利用し，教師は比較する部分教師付き次元圧縮と同じ教師を用いた）．級間分散・級内分散比は，圧縮し学習に用いた 10 次元のデータを用い計算した．

表 2 から，教師なし次元圧縮に比べ，部分教師付き次元圧縮は級間分散・級内分散比が大きくなっていることがわかる．識別率（表 3）を比べた場合も，部分教師付き次元圧縮の方が良い結果であることがわかる．$\alpha = 0.1$ の部分教師付き次元圧縮では，教師なし次元圧縮に比べ，平均値を用いた場合平均して 5.1 ポイント識別率は上昇している．どの教師を選択するかによって，結果に大きな差が生じた．

識別率で比べた場合（表 3），部分教師付き次元圧縮の $\alpha$ の値の異なる 3 つを比較すると，わずかな差しか見られない．級間分散の計算に $\Sigma_{b\_estimated}$ と $\Sigma_{b\_supervised}$ を組み合わせた $\Sigma_{b\_semi-supervised}$（$\alpha = 0.1$）を用いても，$\Sigma_{b\_estimated}$（$\alpha = 0.0$）と $\Sigma_{b\_supervised}$（$\alpha = 1.0$）を単独に用いた場合と比べ優位であるとは言えない．

表中の※で表した値は，全てのデータを教師データとして FDA を行った場合の実験結果である．次元圧縮の結果を改善できれば，理想的には，部分教師付き学習でも表 3 の※に近い値を取りうると思われる．

### 5.2 α-SFDA と α-SLFDA の比較

教師あり圧縮に α-SFDA を用いた時と，α-SLFDA を用いた時の比較を行った結果を表 4，表 5 に示す．圧縮後の次元数は，α-SFDA，α-SLFDA が 2 次元，PCA が 8 次元である．級間分散・級内分散比は，圧縮後の 10 次元のデータを用い計算した．

**表 2** 部分教師付き次元圧縮と教師なし次元圧縮の比較（級間分散・級内分散比）
Table 2 compare semi-supervised with unsupervised dimensionality reduction (ratio of between-class variance to within-class variance)

| 圧縮手法 | PCA | $\alpha = 0.0$ | $\alpha = 0.1$ | $\alpha = 1.0$ | ※ |
|---|---|---|---|---|---|
| 教師数 4 | 0.1218 | 0.2205 | 0.2410 | 0.2029 | 0.3215 |
| 教師数 8 | | 0.2469 | 0.2405 | 0.2235 | |
| 教師数 12 | | 0.2384 | 0.2379 | 0.2243 | |

**表 3** 部分教師付き次元圧縮と教師なし次元圧縮の比較（識別率）
Table 3 compare semi-supervised with unsupervised dimensionality reduction (rate of classification)

| 圧縮手法 | PCA | | $\alpha = 0.0$ | | $\alpha = 0.1$ | | $\alpha = 1.0$ | | ※ | |
|---|---|---|---|---|---|---|---|---|---|---|
| 教師数 | 70.1 | 73.8 | 75.6 | 78.8 | 75.8 | 79.6 | 78.8 | 80.0 | 89.5 | 92.1 |
| 4 | | 68.3 | | 71.7 | | 71.0 | | 77.1 | | 83.3 |
| 教師数 | 75.6 | 79.1 | 81.7 | 83.3 | 81.3 | 83.3 | 81.5 | 82.9 | 92.7 | 93.3 |
| 8 | | 74.2 | | 78.8 | | 80.0 | | 78.3 | | 91.7 |
| 教師数 | 78.8 | 81.7 | 80.7 | 86.3 | 82.8 | 87.5 | 81.8 | 85.0 | 92.1 | 93.3 |
| 12 | | 73.8 | | 75.4 | | 80.0 | | 78.8 | | 90.4 |

表 4 より，級間分散・級内分散比で比較すると，α-SFDA を用いた場合の方が良い結果となった．識別率（表 5）で比較しても，わずかに α-SFDA を用いた方が良い結果であった．本研究で提案した α-SFDA と α-SLFDA は，圧縮後の次元数が（カテゴリ数 −1）に制限されないので，α-SFDA と α-SLFDA の圧縮後の次元数を変え実験を行うと，α-SLFDA の方が良い結果となる場合もあった．

今回の実験では，α-SFDA と α-SLFDA に大きな差はないことがわかる．その原因は今回の実験データの特徴量の分布が比較的単純な分布であり，多峰性などを考慮しなくても，α-SFDA で比較的容易にカテゴリの分離が行えたからであると考えられる．

**表 4** α-SFDA と α-SLFDA の比較（級間分散・級内分散比）
Table 4 compare $\alpha$-SFDA with $\alpha$-LFDA (ratio of between-class variance to within-class variance)

| 圧縮手法 | α-SFDA+PCA | α-SLFDA+PCA |
|---|---|---|
| 教師数 4 | 0.2410 | 0.2076 |
| 教師数 8 | 0.2405 | 0.2188 |
| 教師数 12 | 0.2379 | 0.2248 |

Vol.2010-CVIM-171 No.25
2010/3/19

**表 5** α-SFDA と α-SLFDA の比較（識別率）
Table 5 compare $\alpha$-SFDA with $\alpha$-LFDA (rate of classification)

| 圧縮手法 | α-SFDA+PCA | | α-SLFDA+PCA | |
|---|---|---|---|---|
| 教師数 | 75.8 | 79.6 | 77.2 | 82.1 |
| 4 | | 71.0 | | 69.6 |
| 教師数 | 81.3 | 83.3 | 80.9 | 83.3 |
| 8 | | 80.0 | | 79.2 |
| 教師数 | 82.8 | 87.5 | 82.5 | 85.8 |
| 12 | | 80.0 | | 76.3 |

## 6. まとめ・今後の課題

本研究では，ユーザから入力された少数の教師に基づき，学習を行う部分教師付き学習の精度の向上を目的とし，次元圧縮にも教師データを利用する，部分教師付き次元圧縮手法を提案し，実装・評価を行った．

実験により，教師なし次元圧縮を利用した場合より，部分教師付き次元圧縮を用いた方が，識別率が平均して 5.1 ポイント上昇した．しかし，今回提案した，$\Sigma_{b\_semi-supervised}$ は，$\Sigma_{b\_estimated}$ と $\Sigma_{b\_supervised}$ のみを利用した場合と比較して，識別率の向上は見られなかった．

全データを教師として次元圧縮を行った場合に比べると（GMM の学習にはどちらの場合も部分教師付き学習を利用），識別率はかなり小さいため，次元圧縮部分の更なる改良は可能であり，必要であると考えられる．

今後の課題は，今回用いた次元圧縮は，全て線形次元圧縮であったため，より複雑なデータへ対応するためには，非線形の次元圧縮手法を導入する必要がある．また，今回の実験では LFDA の有効性を確認できなかったので，LFDA の更なる調査と部分教師付き次元圧縮（α-SLFDA）への適用方法を検討する必要がある．

また，部分教師付き次元圧縮では，用いる教師によって結果に大きな差がある．そこで，次元圧縮されたデータを用い，GMM の学習まで行い，ラベル付けされたデータの尤度を用いる方法が考えられる．その尤度を用いることにより，教師の信頼度を考慮できる．今後は，GMM による分布の学習まで行い，その結果を利用してカテゴリ情報を推定する方法も検討する．

## 参 考 文 献

1) Gustavo Cerneiro, Antoni B. Chan, Pedro J. Moreno, and Nuno Vasconcelos, “Supervised Learning of Semantic Classes for Image Annotation and Retrieval”, IEEE, Vol.29, No.3, March 2007.
2) Nizer Grira, Michel Crucianu and Nozha Boujemaa, “ACTIVE SEMI-SUPERVISED CLUSTERING FOR IMAGE DATABASE CATEGORIZATION”, CBMI’05,Riga,Latvia,June 2005.
3) Ye Lu, Chunhui Hu, Xingquan Zhu, HongJiang Zhang, and Qiang Yang, “A Uni ed Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems”, Proceeding of the 8th ACM Multimedia International Conference Los Angeles, CA, USA (October 2000), pp.31-37.
4) Olivier Chapelle,Bernhard Scholkopf,Alexander Zien, “Semi-supervised Learning (Adaptive Computation and Machine Learning)”, Mit Pr,2006.
5) A.P.Dempster, N.M.Laird, and D.B.Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm”, Journal of the Royal Statistical Society. Series B (Methodological), Vol.39, No.1(1977), pp. 1-38.
6) Masashi Sugiyama, “Local Fisher Discriminant Analysis for Supervised Dimensionality Reduction”, Proceedings of the 23rd international conference on Machine learning(2006), vol.148, pp.905-912.
7) Fukunaga K, “Introduction to statistical pattern recognition”, Boston : Academic Press, 1990.
8) Zelnik-Manor and Perona, “Self-Tuning Spectral Clustering”, Advances in neural information processing systems 17, 1601-1608, 2005, Cambridge, MA:MIT Press.