論　文

確率的トピックモデルによる文書画像の領域分割

山口　拓真[†]　　丸山　稔[†]

Document Image Segmentation with Probabilistic Topic Model

Takuma YAMAGUCHI[†] and Minoru MARUYAMA[†]

あらまし　本論文では，確率的トピックモデルを用いた文書画像の領域分割について述べる．確率的トピックモデルとして，bag-of-visual words 表現によって画像分類などに応用されている，文書解析手法の pLSA (probabilistic Latent Semantic Analysis) モデルを用いる．本論文では，文字領域と図表等の領域とを分離することのみを目的とするのではなく，文字領域であっても言語の違いなどによって領域分割を行うことを目的としている．画像を単純に分割し，それぞれの部分領域のカテゴリーを推定するといった手段を用いた場合，詳細な領域分割を行うためには，できるだけ部分領域を小さくすることが望ましいが，細かく分割することにより，各部分領域に含まれる情報量が少なくなり，誤識別を招く可能性が高くなる．そこで本論文では，最初から細かく分割するのではなく，十分な情報量をもつ大きさに分割し，それらにモデルを適用して得られたパラメータを用いて，各部分領域を再分割する手法を提案する．提案手法の検証実験の結果，高い識別率で詳細な領域分割が可能であることが示された．

キーワード　確率的トピックモデル，pLSA，bag-of-visual words，文書画像，領域分割

1. まえがき

現在の OCR (Optical Character Recognition) は，十分に実用レベルであり，様々な場面で利用されている．文書画像を認識する際，文字領域や図表領域の判別や，段組などの構造の理解といったレイアウトの解析がはじめに行われる．認識対象領域かどうかもレイアウト解析で判別することになるが [1]，文字認識エンジンを用い，仮の認識結果の信頼度等によって認識対象かどうかを決定することも多い．例えば，文献 [2] では，認識対象の数式領域を文字認識エンジンを用いて抽出している．多くの OCR ソフトウェアでは，レイアウト解析において認識対象でないと識別された領域は，詳細な認識処理は行われずに画像として出力される．写真や図の領域と文字領域の判別はおおむねうまくできるが，対応していない言語の文字領域が入力された場合，それを認識結果の信頼度等でリジェクトするのは困難な場合も多い．例えば，日本語用の OCR にハングル文字の領域を含む文書画像が入力された場合などである．また，文書画像に数式が含まれる場合は，数式に対応していない OCR では，数式部分の認識結果として不適切な文字列が並ぶだけでなく，数式部分の行の高さが他の行と異なることなどから，数式付近のレイアウト解析に影響を及ぼすことも考えられる．

そこで，本論文では認識エンジンに頼らない文書画像の領域分割手法について提案する．文書画像をその内容（言語や数式や図）によって分割することは，レイアウト解析の用途だけでなく，文書画像に含まれる言語や数式領域の分布や，その位置関係などを利用した文書画像検索や分類といった応用も考えられる．これまで，数々の文書画像におけるレイアウト解析に関する研究が行われており，文書領域や図表領域・写真領域などの幾何学的な構造を解析するものや，文書領域においてタイトル部分やキャプション部分などの論理的な構造の理解を目的としたものが挙げられる [3], [4]．本研究の目的は，これらとは異なっており，言語の違いや印刷文字/手書き文字の違いなどを文字認識エンジンを用いずに判別することである．具体的には，文字領域の特徴量に基づいて，各領域を数式・日本語・英語・手書き文字に分類した．文書画像の領域分割手法として，bag-of-visual words 表現 [5] と統計的文書解析のモデルである確率的トピックモデ

† 信州大学工学部情報工学科，長野市
Dept. of Information Engineering, Shinshu University, 4–17–1 Wakasato, Nagano-shi, 380–8553 Japan

ルを用いる．近年の画像検索・画像分類の分野において，画像中の局所特徴の特徴ベクトルをベクトル量子化することによって画像を表現する，bag-of-visual words，bag-of-keypoints [5] などと呼ばれる画像表現が多く用いられている．これは，統計的文書解析における bag-of-words と同様の考え方であり，文書解析において，語順を無視して文書を単語 (word) の集合として表現するのに対し，位置を無視して画像を局所特徴 (visual word) の集合として表現するというものである．この bag-of-visual words 表現によって，統計的文書解析の手法が適用できるようになり，例えば，probabilistic Latent Semantic Analysis (pLSA) モデル [6]〜[8] や Latent Dirichlet Allocation (LDA) モデル [9] などといった確率的トピックモデルが画像解析の分野にも応用されている．bag-of-visual words 表現と確率的トピックモデルを用いた識別手法により，物体やシーンの認識などにおいて，その有効性が高いことが確認されており [10]〜[12]，文書画像の領域分割においても有用であると思われる．

画像の領域分割を行う場合，画像を単純に分割し，それぞれの部分領域のカテゴリーを識別するというアプローチが考えられる．詳細な領域分割を行うためには，できるだけ部分領域を小さくすることが望ましいが，部分領域が小さければ小さいほど，各領域に含まれる特徴量（情報量）が減少するため，従来より識別問題によく用いられている，kNN (k-Nearest Neighbour) 法や SVM (Support Vector Machine) 法 [14] では，識別性能の低下を招く可能性がある．そこで，本論文では，十分な情報量が得られる大きさの部分領域に確率モデルを適用し，そこで得られたパラメータを用いて，情報量の少ない小さな部分領域へのモデルの適用を近似することにより，小領域に対しても十分な精度を有する領域分割方法を提案する．確率モデルとして，pLSA モデルや LDA モデルに代表される確率的トピックモデルを用いる．pLSA モデルと LDA モデルは類似したモデルであるが，pLSA モデルの方が LDA モデルに比べてより単純で実装も容易である．本研究では双方を比較した識別実験を通して，pLSA モデルに基づく識別方法によって LDA モデルと同等の識別性能が得られることを示し，pLSA モデルに基づく領域分割手法を提案する．また，部分領域を表現する visual word は，画像の局所特徴から生成されるため，文書画像の領域分割の精度は，特徴抽出の影響を大きく受けると考えられる．そこで，文書画像の領域分割に適した特徴抽出方法についての比較・検討も行う．

2. 確率的トピックモデル

2.1 pLSA モデル

本論文では，確率的トピックモデルとして，pLSA (probabilistic Latent Semantic Analysis) モデル [6]〜[8] を用いる．文書は単語の集合として記述されるが，この単語はある確率分布に従って発生する．このとき，単一の確率分布を想定する単純なモデルでは，多様な単語から構成される文書のモデルとしては不十分である．そこで，観測できない隠れ変数（トピック）を導入し，トピックごとに単語発生確率分布が与えられ，これに従って単語が発生するモデルを考える．このとき，pLSA モデルにおいては，この隠れた変数であるトピックの比率（トピック発生確率）は文書ごとに決まっており，各文書についてそのトピック比に従ってトピックが決定され，そのトピックに対応する確率分布に従って単語が発生すると考える．今，単語 w を，語彙 W の要素 ($w \in W = \{w_1, \ldots, w_V\}$) とし，単語 w の集合で表現される文書を d，N 個の文書の集合を $D = \{d_1, \ldots, d_N\}$，T 個の隠れ変数（トピック）を $z \in Z = \{z_1, \ldots, z_T\}$ とする．また，K 個のカテゴリーを $c \in C = \{c_1, \ldots, c_K\}$ で表す．

pLSA モデルのグラフィカルモデルを，図 1 に示す．pLSA モデルでは，各単語 w にトピック z が割り当てられ，各単語の出現確率は $p(w|z)$，各文書 d のトピック分布は $p(z|d)$ で表される．pLSA モデルでは，$p(w|z)$ と $p(z|d)$ に基づいて，以下のように文書が生成される．

1. 文書 d が $p(d)$ に基づいて生成される．
2. 文書中の各単語について，$p(z|d)$ に従ってトピック z が選択される．
3. $p(w|z)$ に従って，単語が生成される．

トピック z について，周辺化を行うことにより，同時確率 $p(w, d)$ が得られる．

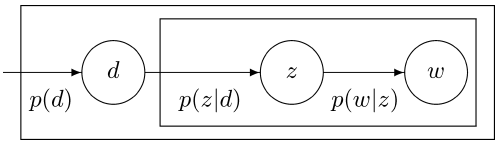


図 1 pLSA モデルのグラフィカルモデル
Fig. 1 Graphical model representation of pLSA.

$$p(w,d) = p(d)\sum_{z\in Z} p(w|z)p(z|d) \tag{1}$$

pLSA モデルのパラメータである $p(w|z)$ と $p(z|d)$ は，EM アルゴリズム [15] を用いて対数ゆう度を最大化することで算出することができる．対数ゆう度は，以下のように表される．

$$\begin{aligned}\mathcal{L} &= \sum_{d\in D}\sum_{w\in W} n(w,d)\log p(w,d)\\ &= \sum_{d\in D}\sum_{w\in W} n(w,d)\log p(d)\\ &\quad+\sum_{d\in D}\sum_{w\in W} n(w,d)\log\sum_{z\in Z} p(w|z)p(z|d)\end{aligned} \tag{2}$$

ここで，$n(w,d)$ は，文書 d に含まれる単語 w の出現頻度を示す．pLSA モデルにおける EM アルゴリズムは，

[E-Step]

$$p(z|w,d) = \frac{p(w|z)p(z|d)}{\sum_{z\in Z} p(w|z)p(z|d)} \tag{3}$$

[M-Step]

$$p(w|z) = \frac{\sum_{d\in D} n(w,d)p(z|w,d)}{\sum_{w\in W}\sum_{d\in D} n(w,d)p(z|w,d)} \tag{4}$$

$$p(z|d) = \frac{\sum_{w\in W} n(w,d)p(z|w,d)}{\sum_{z\in Z}\sum_{w\in W} n(w,d)p(z|w,d)} \tag{5}$$

となる．ここで，オーバフィッティングを緩和するために，ゆう度関数を最大化する代わりに事後確率を最大にするような $p(w|z)$，$p(z|d)$ を求めることを考える．事前分布として以下のような Dirichlet 分布を仮定する．

$$p(w_v|z_t) \sim \mathrm{Dir}(\alpha_1,\ldots,\alpha_V) \tag{6}$$

$$p(z_t|d_i) \sim \mathrm{Dir}(\beta_1,\ldots,\beta_T) \tag{7}$$

この場合の EM アルゴリズムは，E-Step は式 (3) と等しく，M-Step は，

$$p(w|z) = \frac{(\alpha-1)+\sum_{d\in D} n(w,d)p(z|w,d)}{V(\alpha-1)+\sum_{w\in W}\sum_{d\in D} n(w,d)p(z|w,d)} \tag{8}$$

$$p(z|d) = \frac{(\beta-1)+\sum_{w\in W} n(w,d)p(z|w,d)}{T(\beta-1)+\sum_{z\in Z}\sum_{w\in W} n(w,d)p(z|w,d)} \tag{9}$$

となる．ここで，$\alpha=\alpha_1=\cdots=\alpha_V$，$\beta=\beta_1=\cdots=\beta_T$ とした．

未知文書 d' のトピック分布 $p(z|d')$ は，上記手順によって得られた $p(w|z)$ を用いた folding-in アルゴリズムで算出する [6]．pLSA モデルにおける，folding-in アルゴリズムは，

[E-Step]

$$p(z|w,d') = \frac{p(w|z)p(z|d')}{\sum_{z\in Z} p(w|z)p(z|d')} \tag{10}$$

[M-Step]

$$p(z|d') = \frac{(\beta-1)+\sum_{w\in W} n(w,d')p(z|w,d')}{T(\beta-1)+\sum_{z\in Z}\sum_{w\in W} n(w,d')p(z|w,d')} \tag{11}$$

で与えられる．

2.2 LDA モデル

本論文では，pLSA モデルに基づく識別方法の比較対象として，LDA (Latent Dirichlet Allocation) モデル [9] に基づく識別も行うため，ここで LDA モデルの概要について述べる．pLSA モデルにおいては，文書ごとにトピック生成確率（トピック比）が与えられるが，このパラメータは確率変数ではない．このため，文書数が増大すると，推定すべき未知パラメータもそれに伴って増大することになる．LDA モデルはこのトピック生成確率を与えるパラメータを確率変数として扱い，その確率分布（Dirichlet 分布）を導入してデータ生成過程をモデル化するものである．本論文で用いる LDA モデルは，文献 [13] と同じもので，データの生成過程は以下のように表すことができる．ここで，文書 d_i における各単語へのトピックの割当を $\boldsymbol{z}'_{d_i}$ と記述する．

（1） 各トピック $j=1,\ldots,T$ ごとに，トピックに対応する単語生成確率 $\boldsymbol{\phi}^{(j)}=(\phi_1^{(j)},\ldots,\phi_V^{(j)})$ が Dirichlet 分布に従って得られる．

$$p(\boldsymbol{\phi}^{(j)}) \sim \mathrm{Dir}(\beta_1,\ldots,\beta_V) \tag{12}$$

（2） 各文書 $d_i(i=1,\ldots,N)$ ごとに以下を繰り

返す．

(a) 文書 d_i のトピック比を表す $\boldsymbol{\theta}^{(d_i)} = (\theta_1^{(d_i)}, \ldots, \theta_T^{(d_i)})$ が Dirichlet 分布に従って得られる．

$$p(\boldsymbol{\theta}^{(d_i)}) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_T) \tag{13}$$

(b) 文書 d_i 中のそれぞれの単語について，多項分布に従ってトピックが割り当てられる．

$$p(\boldsymbol{z}'_{d_i}|\boldsymbol{\theta}^{(d_i)}) \sim \mathrm{Mult}(\boldsymbol{\theta}^{(d_i)}) \tag{14}$$

(c) 選択されたトピックに対応する多項分布に従って，文書 d_i 中の単語 w が生成される．

$$p(w|\boldsymbol{z}'_{d_i}, \boldsymbol{\phi}^{(j)}) \sim \mathrm{Mult}(\boldsymbol{\phi}^{(j)}) \tag{15}$$

本論文で用いた LDA モデルのグラフィカルモデルを図 2 に示す．LDA モデルを式で表現すると以下のようになる．

$$\begin{aligned} &p(D, Z', \Theta, \Phi|\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &\quad = p(\Phi|\boldsymbol{\beta}) \prod_{i=1}^{N} p(\boldsymbol{\theta}^{(d_i)}|\boldsymbol{\alpha}) p(\boldsymbol{z}'_{d_i}|\boldsymbol{\theta}^{(d_i)}) p(d_i|\boldsymbol{z}'_{d_i}, \Phi) \end{aligned} \tag{16}$$

ここで，各変数を $Z' = \{\boldsymbol{z}'_{d_1}, \ldots, \boldsymbol{z}'_{d_N}\}$，$\Theta = \{\boldsymbol{\theta}^{(d_1)}, \ldots, \boldsymbol{\theta}^{(d_N)}\}$，$\Phi = \{\boldsymbol{\phi}^{(1)}, \ldots, \boldsymbol{\phi}^{(T)}\}$，$\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_T\}$，$\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_V\}$ でまとめて表記している．観測データが与えられたときの事後確率を直接求めるのは困難であるため，一般的に変分ベイズ法や MCMC (Markov Chain Monte Carlo) 法による近似解法が用いられる [13]．本論文では変分ベイズ法を用いてパラメータ推定を行う．

LDA モデルは pLSA モデルと異なりパラメータがデータとともに増大することはなく，オーバフィッティングへの耐性が高いことが期待できる．しかしながら，推論のために近似解法（変分ベイズ法・MCMC 法）などを用いる必要があり，推論の精度が同程度であるならば，EM アルゴリズムを適用できる pLSA モデルを用いる方がモデルの簡単さや実装の容易さの点から望ましい考えられる．本研究では，**2.1** に述べた pLSA モデルに MAP 推定を適用するモデルにより，LDA モデルと同等の識別能力が得られることを示す．

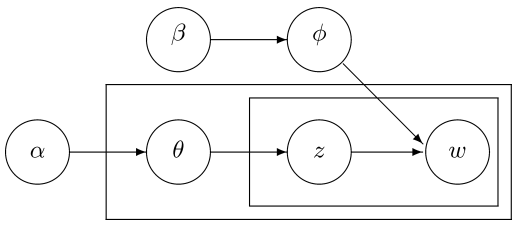


図 2 LDA モデルのグラフィカルモデル
Fig. 2 Graphical model representation of LDA.

3. 確率的トピックモデルの画像データへの適用

3.1 画像の文書表現

確率的トピックモデルは，もとは文書解析用のモデルであるため，画像データに適用するためには，文書データにおける，単語 (word) や文書 (document)，語彙 (vocabulary) に相当するものを，画像データにおいても定義する必要がある．文書データと同様に，文書，語彙は単語の集合として考えることができる．画像データにおける単語 (visual word) は，画像から得られる局所特徴によって表現される．確率的トピックモデルは，語順を無視して，文書を単語の集合として扱う bag-of-words という考え方に基づいており，画像データに適用する際は，bag-of-visual words [5] などと呼ばれ，位置を無視して，文書に相当する画像 (visual document) を visual word の集合として考えることとなる．

3.1.1 visual word の定義

本論文では，文書画像から抽出した局所特徴の特徴ベクトルをベクトル量子化することによって，画像における単語 (visual word) として扱う．本論文では，ベクトル量子化方法として k-means 法を用いた．k-means 法のパラメータであるクラスタ数 k が，visual word の種類数となり，それぞれのクラスタの中心に位置する特徴ベクトルが，語彙 (visual vocabulary) を構成する単語として得られる．

3.1.2 visual document の定義

本論文では，文書画像を複数の部分領域に分割し，それぞれに含まれる局所特徴量を visual word 表現に量子化した部分領域を visual document として扱う．学習用の visual document を生成する場合の文書画像の分割については，単純なグリッド分割と，各部分領域に含まれる特徴点数がほぼ等しくなるように特徴点位置に基づく k-means 法による分割を行った．これは，visual word を生成する場合の特徴ベクトルのベクトル量子化において，カテゴリー間の特徴点数の偏りが識別性能に及ぼす影響を調べるためである．未知画像（テスト画像）の部分領域については，単純にグ

リッドで分割したものを用いる．

3.2 特徴抽出

visual word は，局所特徴の特徴ベクトルを量子化したものであるため，特徴抽出方法は識別精度に大きく影響すると思われ，適した特徴抽出手法について調べる必要がある．本研究では，まず画像から特徴点を検出し，検出された各特徴点において特徴量を算出するという手順で特徴抽出を行う．

3.2.1 特徴点検出方法

本論文では，特徴点検出方法として，一定間隔ごとにサンプリングする方法 (dense detector) と，スケール変化に耐性のある SIFT detector (DoG detector) [16]，スケールだけでなく視点（撮像角度）変化にも耐性のある Harris affine detector [17] を用いて比較検討を行う．SIFT detector 及び Harris affine detector について以下にアルゴリズムの概要を示す．

［SIFT detector (DoG detector)］ 特徴点とスケールの探索方法として，スケールを変化させながら生成した Laplacian of Gaussian (LoG) フィルタを用いた方法 [18] が提案されている．これは，異なるスケールの LoG フィルタを入力画像に適用することで得られるスケールスペースにおいて，その極値を探索することで，特徴点とスケールを決定するというものである．SIFT detector では，LoG の近似である Difference of Gaussian (DoG) を用いて，より効率良く初期の特徴点とスケールの決定を行う．スケールの異なるガウス関数と入力画像とを畳み込んだ平滑化画像の差分が DoG 画像と呼ばれるもので，これを複数の異なるスケール間で求めることでスケールスペースが作成される．画像全体についてスケールスペースの極値を検出し，極値が得られた位置が特徴点の候補点となる．これらの候補点から，ノイズに影響されやすいコントラストが低い点や，開口問題 (aperture problem) の影響を受けやすいエッジ上に存在するような点を取り除いたものが特徴点として扱われる．

［Harris affine detector］ 輝度値の変動に基づく特徴点の検出方法として，最もよく用いられている方法の一つに Harris detector [19] がある．この Harris detector に LoG を用いたスケール探索の概念を取り込み，スケール変化に耐性をもつ Harris Laplace detector [20] が提案されている．Harris Laplace detector では，LoG に等方的なガウス関数が用いられるが，近傍のこう配方向を反映した非等方的なガウス関数を用いることで，視点変化（アフィン変化）にも耐性をもたせたものが，Harris affine detector である．

3.2.2 特徴表現方法

3.2.1 で挙げた手法によって抽出された各点の特徴量表現として，Haar wavelet [21]，SIFT descriptor [16]，輝度のこう配方向のヒストグラム (Gradient histogram) を用いる．それぞれについて以下にアルゴリズムの概要を示す．

［Haar wavelet］ Haar wavelet を用いた特徴量表現は，物体の検出などによく用いられている [22]．本研究では，各特徴点を中心とした 16×16 ピクセルの局所領域について，式 (17) に示すスケーリング関数，及び，式 (18) に示すウェーブレット関数で定義される Haar wavelet 変換を行い，得られた 256 次元の係数を特徴ベクトルとして用いる．

$$\phi(x) = \begin{cases} 1 & \text{for } 0 \leq x < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

$$\psi(x) = \begin{cases} 1 & \text{for } 0 \leq x < 1/2 \\ -1 & \text{for } 1/2 \leq x < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

［SIFT descriptor］ SIFT descriptor では，各特徴点について，その方向（オリエンテーション）を求め，このオリエンテーションに基づいて特徴点周辺の領域を回転させて特徴量を算出するため，回転に不変な特徴量が得られる．オリエンテーションを求めるには，特徴点周辺の画素のこう配方向とこう配強度を求めて，ガウシアン重み付きの方向ヒストグラムを作成し，それを 36 方向に量子化したヒストグラムを算出する．このヒストグラムの最大値から 80%以上となる要素（方向）を特徴点のオリエンテーションとして用いる．特徴量表現はオリエンテーションによって正規化された周辺領域を $4 \times 4 = 16$ ブロックに分割し，ブロックごとにこう配方向とこう配強度に基づく 8 方向のヒストグラムを作成することで，$4 \times 4 \times 8 = 128$ 次元の特徴ベクトルが得られる．

［輝度のこう配方向のヒストグラム］ SIFT descriptor では，各特徴点についてオリエンテーションを求め，それに基づいて回転を加えた特徴点の周辺領域から特徴量を算出するが，本研究で用いた輝度こう配方向のヒストグラムは，そのオリエンテーションに基づく回転処理を省略したものである．特徴量の表現方法は，SIFT descriptor と同じであり 128 次元の特徴ベクトルとなる．これは，回転に不変な特徴量の必要性に関

して調べるためのものである．

4. カテゴリーに基づく文書画像分割

4.1 確率的トピックモデルによる分類方法

確率的トピックモデルは，教師なしの学習モデルであるため，未知文書 d' のカテゴリー c の推定について述べる．本論文では，画像特徴 (visual word) を生成する確率分布を決定する隠れ変数（トピック）が存在し，観測できないその値によって画像特徴が発生するというモデルを想定している．これを用いて文書をその内容に従って，数式・日本語・英語・手書き文字に分類することを考える．このためには pLSA モデルを適用した推論によって得られたトピック値とカテゴリーを 1 対 1 に対応させる，すなわちトピック=カテゴリーとみなすことにより，トピック値からカテゴリーを推定し，これによって領域分割を行うことがまず考えられる．しかしながら，各カテゴリーに含まれるシンボルは多様で類似画像特徴を発生させるものも多く存在すると考えられ，各カテゴリーに対して単一の固定されたトピックを割り当てるモデル化方式では，カテゴリーのもつ多様性や不確定性を表現するには十分ではないと考えられる．そこで本論文においては，各カテゴリーとトピックを 1 対 1 に対応させるのではなく，トピックが与えられたとき，その値に従ってカテゴリー発生確率が決定されるというモデルを想定し，この条件付きカテゴリー発生確率を導入することによって各領域のカテゴリー確率を決定する以下のようなモデルを想定する．未知文書 d' のカテゴリーとして，以下の式で与えられる $p(c|d')$ が最大となるカテゴリー c を選択する．

$$p(c|d') = \sum_{z \in Z} p(z|d')p(c|z) \tag{19}$$

ここで，$p(c|z, d') = p(c|z)$ とした．$p(z|d')$ は式 (11) によって得られ，$p(c|z)$ は，ベイズの定理を用いて，

$$p(c|z) = \frac{p(z|c)p(c)}{\sum_{c \in C} p(z|c)p(c)} \tag{20}$$

として算出できる．$p(c)$，$p(z|c)$ については，

$$p(c) \approx \frac{N_c}{N}, \tag{21}$$

$$p(z|c) \approx \frac{1}{N_c} \sum_{\{i|category(d_i)=c\}} p(z|d_i) \tag{22}$$

とした．ここで，N は文書数を表し，N_c はカテゴリー c に含まれる文書数を意味している．また，この分類方法は，pLSA モデルだけでなく，LDA モデルにも適用可能である．

4.2 visual document の再分割

未知画像の詳細な領域分割を行うには，部分領域をできるだけ小さくすることが望ましいが，細かく分割することにより，各 visual document に含まれる visual word 数（情報量）が少なくなり，誤識別を招く可能性が高くなる．そこで，最初から細かく分割するのではなく，まず，十分な情報量をもつ大きさに分割し，各 visual document にモデルを適用して，そこで得られたパラメータを用いて，各部分領域を再分割することによって，細かく分割された部分領域のカテゴリーを推定する方法を提案する．

ここで，十分な情報量をもつ未知の visual document d' を再分割した visual document d'' のカテゴリーを推定することを考える．$p(c|d'')$ を算出することによってカテゴリーを選択するが，d'' に folding-in アルゴリズムを適用して算出した $p(c|d'')$ によってカテゴリーを推定すれば，最初から細かく分割した場合と同等であり，情報量不足による識別精度の低下のおそれがある．そこで，以下の式によって $p(c|d'')$ を算出することを考える．

$$p(c|d'') \propto \sum_{w \in W} n(w, d'')p(c|w, d'') \tag{23}$$

ここで，$p(c|w, d'')$ は，式 (10) 及び式 (20) を利用して，

$$p(c|w, d'') = \sum_{z \in Z} p(c|z)p(z|w, d'') \tag{24}$$

$$\approx \sum_{z \in Z} p(c|z)p(z|w, d') \tag{25}$$

として算出する．d'' のカテゴリーとして，$p(c|d'')$ が最大となるカテゴリー c が選択される．これは，$p(z|w, d'')$ を直接推定するのではなく，d' は d'' を含む visual document であることから，$p(z|w, d'') \approx p(z|w, d')$ と近似できるはずであるという考えに起因するものである．これにより，d'' に含まれる visual word 数が少ない場合でも，安定した $p(z|w, d'')$ が得られることが期待できるため，識別率の低下を避けられると考えられる．

5. 評価実験

5.1 データ

本論文では，識別対象として，数式，活字日本語，活字英語，手書きの 4 カテゴリーの分類を行う．数式データは論文などから数式領域のみを切り出して作成したものを使用し，活字の日本語・英語データは学術論文を用い，手書きデータは，手書きメモや手書き論文を収集したものである．手書きのメモは，2 名の書き手によってレイアウトなどは気にせずに書かれたもので，手書き論文は，ワープロを使わずに論文を投稿していた時代に投稿先のレイアウトの規定に従って，丁寧に書かれた 2 名の書き手のものである．その手書き論文のレイアウトは現在のワープロを用いたものとほとんど変わらない．これらはすべて，A4 サイズのもので 300 dpi の画像として取り込んだものである．また，各文書画像には，対象の言語のみが含まれるように編集を行っている．具体的には，各カテゴリーの文書は表 1 に示すような文字種で構成される．手書きのメモを除き，それぞれのカテゴリーのデータは単一の学術論文から作成したためフォントの種類やサイズの変動は小さく，レイアウト（横書き）も比較的整ったデータであるといえる．これらの画像を分割して学習・テスト用データを作成するが，学習用として，300×300 ピクセルに分割したものと，含まれる特徴点数ができるだけ均等になるように特徴点位置に基づく k-means 法を用いて分割したものを利用した．このとき，k-means 法のパラメータ k は画像ごとに異なり，各局所領域の特徴点数の平均が 500 となるように設定した．テスト用として，240×240 ピクセルに分割したものと，それを更に 60×60 ピクセルに再分割したものを用いた．ここで，各局所領域において含まれる特徴点数が 25 未満の場合は，その領域をリジェクトしている．各文書画像の枚数と，特徴点検出方法として DoG detector を用いたときに各文書画像を分割した場合の局所領域数を表 2 に示す．また，本論文における実験結果は 5 回のランダムサンプリングによって作成した 5 種類のデータセットから得られた結果の平均値を表している．

表 1　各カテゴリーに含まれる文字種
Table 1　Character types contained in each category.

カテゴリー	含まれる文字種
数式文書	数式記号,（数式に含まれる）英字・数字
日本語文書	日本語,（日本語文書中の）数字
英語文書	（英語文書中の）英字・数字
手書き文書	手書き文字

5.2 pLSA モデルによる識別

5.2.1 pLSA モデルのパラメータの選択

学習データとして，k-means 分割によって得られた局所領域 (visual document) から，各カテゴリーについて，それぞれ 100 領域をランダムサンプリングした 400 領域を用いる．visual word の生成には，この 400 領域に含まれる特徴ベクトルのすべてを用いる．また，ここでは，特徴抽出方法として，DoG detector と SIFT descriptor を利用する．テストデータは，240×240 ピクセルに分割して得られたすべての局所領域（8,271 領域）を識別対象とする．pLSA モデルによって識別を行う場合には，visual word の種類数，トピック数を与える必要がある．そこで，visual word の種類数とトピック数を変化させた場合の識別結果を図 3，図 4 に示す．これらの結果より，visual word の種類数及びトピック数は大きい方が識別率が高いことが分かる．トピックモデルを用いて教師なしの画像分類を行う場合，トピック数とカテゴリー数を同一にしている報告もあるが，文書画像を対象とした場合，トピック数が少ない場合の識別率は低くなっており，少なくともカテゴリー数の 2 倍以上のトピック数が必要と思われる．これらの結果から，visual word の種類数は 300，トピック数は 20 で十分であると思われ，以降の実験では，これらのパラメータを用いる．

5.2.2 学習用データの生成方法の比較

次に，学習用の visual document の生成について考える．visual document の生成方法（領域分割方法）として，k-means 法を用いる場合と，単純にグリッ

表 2　各カテゴリーの画像枚数と局所領域数
Table 2　The number of images and local regions of each category.

	# of images	k-means	300×300 pixel grid	240×240 pixel grid	60×60 pixel grid
mathematical formula	12	853	1,014	1,514	7,530
printed Japanese	34	2,599	1,408	2,170	20,175
printed English	15	1,797	797	1,193	15,053
hand written	42	2,779	2,332	3,394	22,233
	103	8,028	5,551	8,271	64,991

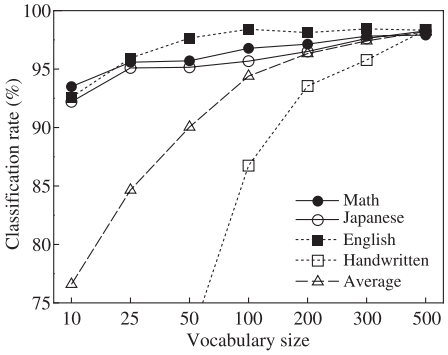


図 3　visual vocabulary のサイズと識別結果の推移
Fig. 3　Comparison of the classification rates with respect to the size of visual vocabulary.

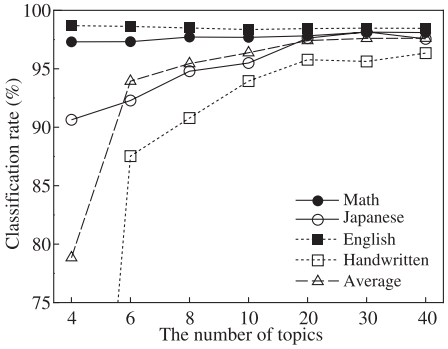


図 4　トピック数と識別結果の推移
Fig. 4　Comparison of the classification rates with respect to the number of topics.

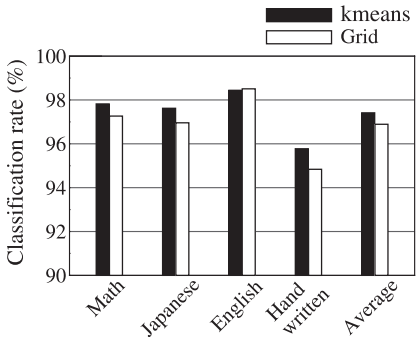


図 5　k-means 法による分割とグリッド分割の比較
Fig. 5　Comparison of the classification rates based on k-means- and grid-based segmentation.

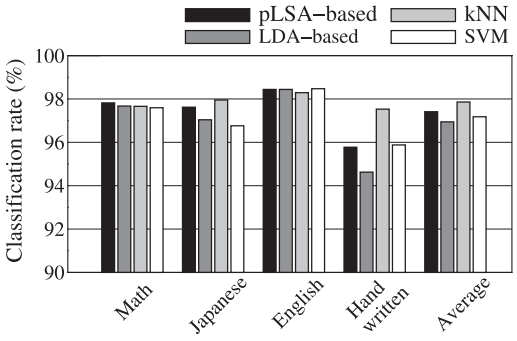


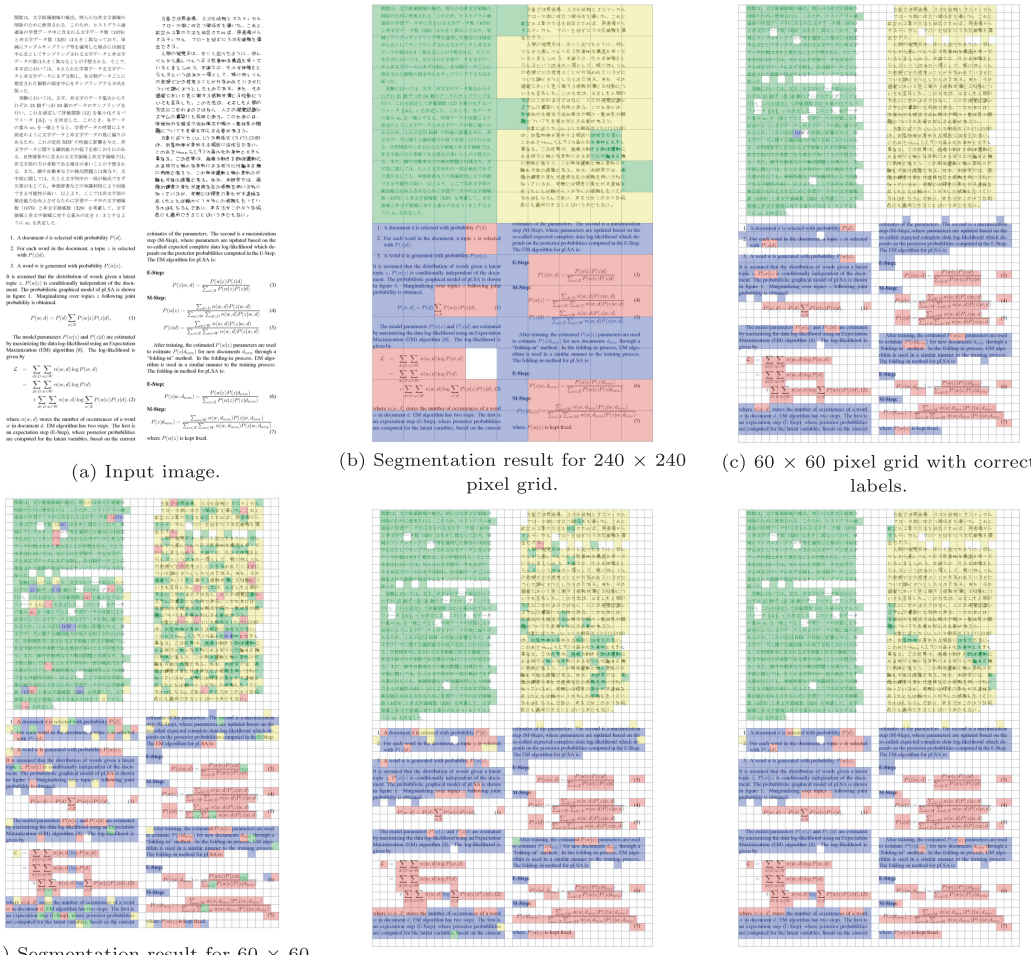
図 6　識別器ごとの識別結果
Fig. 6　Classification rates of each classifier.

ドで分割する場合を考える．各カテゴリーにおける visual document 数を同一にした場合，グリッド分割では各領域の特徴点数に大きな差が出る可能性があり，visual word の生成の際に用いるデータの特徴点数にカテゴリー間で偏りが生じる可能性がある．一方，k-means 法による分割では，各領域の特徴点数はグリッド分割に比べ安定するため，そういった偏りは小さくなる．図 5 に，k-means 法による分割とグリッド分割の場合の識別結果を示す．実験では，学習データとして k-means 法による分割とグリッド分割で得られた visual document を各カテゴリー 100 領域ずつサンプリングした 400 領域を用いている．これらの結果より，わずかではあるが k-means 法による分割の方が識別率が高く，効果があることが分かる．ゆえに，visual word 生成の際，各カテゴリー間の特徴点数の偏りは小さい方が望ましいと考えられる．

5.2.3　他の識別手法との比較及び複数のカテゴリーが含まれる文書画像への適用

pLSA モデルを用いた識別方法の識別能力を調べるために，LDA (Latent Dirichlet Allocation) モデルを用いた識別法，kNN (k-Nearest Neighbour) 法，SVM (Support Vector Machine) 法との比較を行った．各識別器による識別結果を図 6 に示す．kNN 法が最も識別率が高かったが，識別率の平均の差は 1%未満であり，大きな差はないといえる．また，これらの実験結果より，MAP 推定を用いた pLSA モデルと LDA モデルは識別精度に大きな差はなく，確率的トピックモデルとしては，より実装が容易な pLSA モデルで十分であると考えられる．

ここで，四つのカテゴリーが含まれる文書画像（図 7 (a)）を作成し，この文書画像の領域分割を行う．これを 240×240 ピクセルに分割し，それぞれの領域のカテゴリーを pLSA モデルに基づく識別方法によって推定すると図 7 (b) のような結果が得られた．ここで，赤・緑・青・黄は，それぞれ数式・日本語・



(a) Input image.

(b) Segmentation result for 240 × 240 pixel grid.

(c) 60 × 60 pixel grid with correct labels.

(d) Segmentation result for 60 × 60 pixel grid.

(e) Proposed method (i).

(f) Proposed method (ii).

図 7　文書画像への適用結果
Fig. 7　Segmentation results of a document image.

英語・手書き領域を表している．この結果を見ると，240 × 240 ピクセルに分割したのでは，一つの局所領域に複数のカテゴリーが含まれている場合もあり，この入力画像の領域分割を行うには分解能が不足していることが分かる．よって，実用性を高めるにはより細かく分割する必要があると思われる．次に，文書画像を 60 × 60 ピクセルに分割して，手動でカテゴリーを付与したものを図 7 (c) に示す．ここで，カテゴリーは表 1 に従って付与し，領域内に複数のカテゴリーが含まれる場合，及び，特徴点数が 25 に満たない領域は，白く表示している．240 × 240 ピクセルに分割した場合と比べると，細かく領域分割が行える可能性があることが分かる．そこで，表 2 のデータにおいて，60 × 60 ピクセルに分割して作成した 64,991 領域に対し，それぞれの領域のカテゴリーを推定した際の結果を図 8 に示す．240 × 240 ピクセルに分割した場合の図 6 の結果に比べ識別率が悪化しているのが分かる．また，図 7 (a) に示す画像に対し，60 × 60 ピクセルで領域分割を行ったところ，図 7 (d) が得られ，カテゴリーの推定が不安定であることが分かる．

5. 2. 4　部分領域の再分割

図 6 では比較的高い識別結果が得られているが，これは局所領域に単一のカテゴリーのみ含まれる場合のことである．本論文では，ある領域に一つのカテゴリーラベルのみを付与することを前提としているため，局所領域に複数のカテゴリーが含まれていては，そも

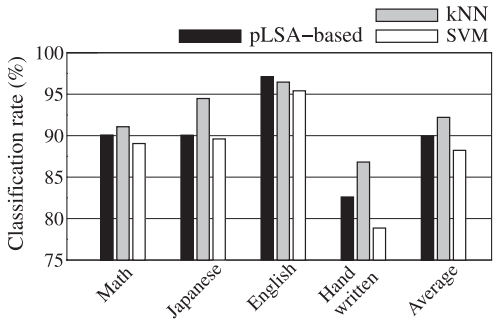


図 8　60 × 60 ピクセルのグリッドに分割した際の識別率
Fig. 8　Classification results for 60 × 60 pixel grids.

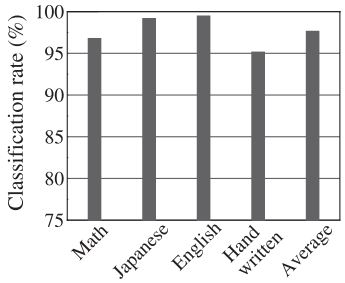


図 9　240 × 240 ピクセルのグリッドを 60 × 60 に再分割した際の識別率
Fig. 9　Classification results by re-segmenting 240 × 240 pixel grid into 60 × 60 pixel grid.

そも正確な識別は不可能である．また，こういった複数のカテゴリーからなる局所領域が得られることをできるだけ避けるために，単純に局所領域を細かくしたのでは識別率は大幅に悪化してしまう．そこで，**4.2** で述べた部分領域の再分割を行う．表 2 に示す 240 × 240 ピクセルに分割したテストデータに対し pLSA モデルを適用し，得られたパラメータを用いて，60 × 60 ピクセルに再分割した場合の識別結果を図 9 に示す．単純に細かく分割した図 8 の結果に比べて識別率が改善されていることが分かる．また，この手法を図 7 (a) に適用した場合の，結果を図 7 (e) に示す．これらの結果より，最初から細かく分割するよりも，大まかに分割しておき，それを再分割した場合の方が高い識別率が得られていることが分かる．これは，細分化した領域 d'' のカテゴリーを推定する際，pLSA モデルを d'' に適用してカテゴリーを推定するのではなく，大まかに分割した各領域 d' に pLSA モデルを適用して得られた $p(z|w, d')$ を用い，$p(z|w, d'') \approx p(z|w, d')$ と近似することで，情報量の少ない細分化された領域のカテゴリーを高い精度で推定できることを示している．

次に，再分割された 60 × 60 ピクセルのグリッドについて，周囲の visual word も利用して，更にカテゴリーの推定を安定させることを考える．本論文では，60 × 60 ピクセルのグリッドを中心とする 120 × 120 ピクセルのグリッドに含まれる visual word を対象の visual document を構成する visual word として考え，それに対してカテゴリーの推定を行った．その結果を図 7 (f) に示す．この実験結果から全体的に識別精度は向上しているように見える．ただし，あまり領域を拡大しすぎると再分割の意味が薄れてしまう可能性はある．

5.3　特徴抽出手法の選択

文書で用いられるフォントサイズは一般には極めて多様である，またスキャナ等により電子化された文書の中には，入力時に回転のゆがみが含まれている場合も考えられる．以下では，学習に用いられた文書データとは異なるサイズの文書や，文書に回転のゆがみが含まれている場合の耐性を見るために，入力画像にスケール変化及び回転を加えた場合の各特徴抽出手法による識別率の違いについて調べる．データは，スケールを変化させたものとして画像を 0.75 倍，1.25 倍したものを用い，回転を加えたものとして画像を 5°〜20° 回転させたものを作成しテストデータとした．なお，学習用データは，スケール変化・回転を加えていない画像を用いて作成している．データ数は，学習用として各カテゴリー 100 領域の 400 領域，テスト用として各カテゴリー 500 領域の 2,000 領域である．

まず，スケール変化と識別率の推移について述べる．ここでは，スケール変化に耐性のある DoG detector と Harris affine detector，スケール変化に耐性のない一定間隔（8 ピクセル）のサンプリング (Dense detector) を特徴点検出方法として用い，画像表現については SIFT descriptor を用いた．図 10 にスケール変化を加えた画像の識別結果を示す．識別率はスケール変化に耐性をもつ特徴点検出方法の方が高いことが分かる．表 3 にスケール変化を加えた場合の各特徴抽出手法の識別率の変化率を示す．ここで，スケール変化を加えない画像の識別率を 1.0 に正規化して表している．スケール変化を加えた場合の識別率の変化率（悪化の程度）についてもスケール変化に耐性のある特徴点検出方法の方が小さいこと分かる．したがって，画像間や，学習データと未知データの間でフォントサイズに差がある可能性がある場合には，スケール変化に耐性のある特徴点検出方法を用いる必要があると思わ

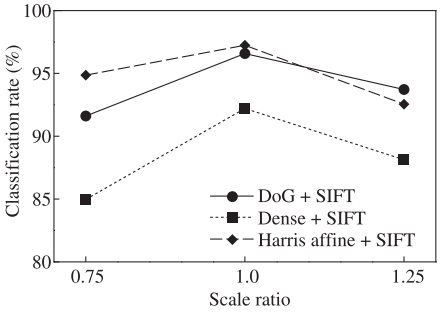


図 10　スケール変化を加えた場合の識別結果
Fig. 10　Classification results for scale changed data.

表 3　識別率の変化率
Table 3　Comparison of classification rates by scale change.

scale	×0.75	×1.0	×1.25
DoG + SIFT	0.957	1.0	0.971
Harris affine + SIFT	0.99	1.0	0.955
Dense + SIFT	0.939	1.0	0.933

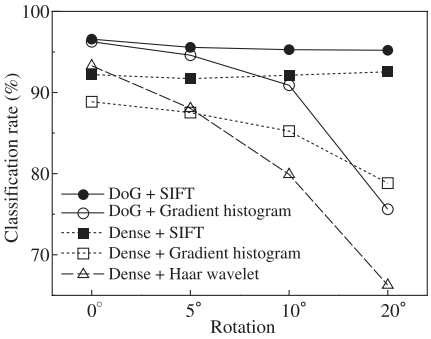


図 11　回転を加えた場合の識別結果
Fig. 11　Classification results for rotated data.

れる．本実験で用いたデータはフォントサイズの変動は小さいものであるが，スケール変化に耐性のある特徴点検出方法を用いることでフォントサイズの異なる場合にも対応できると考えられる．

次に，回転を加えた場合の識別率の推移について述べる．実験では特徴点検出方法として，DoG detector と Dense detector を用いた．画像表現方法として，回転に耐性のある SIFT descriptor と，回転に耐性のない輝度こう配の方向ヒストグラム (Gradient histogram) と Haar wavelet を用いた．図 11 に，回転を加えた画像に対する識別結果を示す．特徴点検出方法によらず，回転への耐性をもつ SIFT descriptor の優位性が確認できる．スキャナで文書データを取り込む際，10° 以上も傾いた画像が得られることは考えにくいが，手書き文字画像を含む場合，書く時点で傾きが含まれていることも考えられるため，回転に耐性のある画像表現を用いることが望ましいと思われる．

5.4　画像の分割形状について

本論文では，画像のセグメンテーションをグリッド分割に基づく単純な方法で行っている．また，実験に用いたデータは主に学術論文であり，そのレイアウトは比較的整っているといえる．このような整ったレイアウト構造の文書については，提案手法は良好な識別能力を有することが示された．行のベースラインが曲線の場合や，各行が平行に並んでいないような複雑なレイアウトに対しては，単純なグリッドでは対応できない場合が多くなる可能性がある．提案手法では，位置情報を無視した Bag-of-visual words 表現を用いているため，局所領域は単純なグリッドに限らず任意の形のものを利用することができる．したがって，単純なグリッド分割を用いる代わりに，既存のレイアウト解析手法によって構造解析を行い，その領域分割結果に対して提案手法を適用することも考えられる．そうすることで，文書領域を手書き文字部分と印刷文字部分，更には記述言語によって分離するなど，より詳細なレイアウト解析が可能となる．また，確率的トピックモデルを用いた画像分類に関する研究が盛んに行われており，その有効性も高いことから，写真領域と文字領域の判別も可能と思われ，これまでのレイアウト解析の補助手段としても利用できると思われる．

6. む す び

本論文では，確率的トピックモデル (pLSA) を用いた文書画像の領域分割について述べた．文書画像の領域分割を行う際，できるだけ細かく分割したいが，最初から細かく分割するのではなく，まず，十分な情報量をもつ大きさに分割し，それらにモデルを適用し，そこで得られたパラメータを用いて，各部分領域を再分割する手法を提案した．提案手法によって，識別率を損なわずに詳細な領域分割が行えることが検証実験によって示され，提案手法の有効性が確認された．また，文書画像の領域分割に適した特徴抽出方法についても比較・検討を行い，一般的な画像分類などと同様に，スケール変化及び，回転に強い特徴抽出手法を用いることが望ましいことも分かった．本論文では visual words 抽出と bag-of-words 方式のモデル化に基づく手法を用いたが，確率モデルに基づく文書画像解析手法としては MRF や CRF を用いる手法も提案

されており [23]，これらの手法との比較なども今後の課題である．

文　　献

[1] 朴　栄碩，老名　毅，伊藤　昭，“汎用的な文書画像の階層的領域分割と識別法，” 信学論（D-II），vol.J75-D-II, no.2, pp.246–256, Feb. 1992.

[2] 能隅進一，福田亮治，玉利文和，鈴木昌和，“絞り込み法による数式文字認識とその日本語/数式領域切出しへの応用，” 信学論（D-II），vol.J83-D-II, no.3, pp.895–906, March 2000.

[3] R. Cattoni, T. Coianiz, S. Messelodi, and C.M. Modena, “Geometric layout analysis techniques for document image understanding a review,” Technical Report, ITC-IRST, Trento, Italy, 1998.

[4] S. Mao, A. Rosenfeld, and T. Kanungo, “Document structure analysis algorithms: A literature survey,” Proc. SPIE Electronic Imaging, vol.5010, pp.197–207, 2003.

[5] G. Csurka, C. Bray, C. Dance, and L. Fan, “Visual categorization with bags of keypoints,” Proc. ECCV Workshop on Statistical Learning in Computer Vision, pp.1–22, 2004.

[6] T. Hofmann, “Probabilistic latent semantic indexing,” Proc. Special Interest Group on Information Retrieval (SIGIR), pp.50–57, 1999.

[7] T. Hofmann, “Probabilistic latent semantic analysis,” Proc. Uncertainty in Artificial Intelligence (UAI), pp.289–296, 1999.

[8] T. Hofmann, “Unsupervised learning by probabilistic latent semantic analysis,” Mach. Learn., vol.42, pp.177–196, 2001.

[9] D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet allocation,” J. Machine Learning Research, vol.3, pp.993–1022, 2003.

[10] L. Fei-Fei and P. Perona, “A Bayesian hierarchical model for learning natural scene categories,” Proc. Computer Vision and Pattern Recognition (CVPR), pp.524–531, 2005.

[11] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, “Discovering objects and their location in images,” Proc. International Conference on Computer Vision (ICCV), 2005.

[12] A. Bosch, A. Zisserman, and X. Munoz, “Scene classification via pLSA,” Proc. European Conference on Computer Vision (ECCV), 2006.

[13] T. Griffiths and M. Steyvers, “Finding scientific topics,” Proc. National Academy of Science, vol.101, pp.5228–5235, 2004.

[14] V. Vapnik, The nature of statistical learning theory, Springer, 1995.

[15] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” J. Royal Statistical Society B, vol.39, no.1, pp.1–38, 1977.

[16] D. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis. (IJCV), vol.60, no.2, pp.91–110, 2004.

[17] K. Mikolajczyk and C. Schmid, “Scale & affine invariant interest point detector,” Int. J. Comput. Vis. (IJCV), vol.60, no.1, pp.63–86, 2004.

[18] T. Lindeberg and J. Garding, “Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure,” Image Vis. Comput., vol.15, no.6, pp.415–434, 1997.

[19] C. Harris and M. Stephens, “A combined corner and edge detector,” Alvey Vision Conference, pp.147–151, 1988.

[20] K. Mikolajczyk and C. Schmid, “Indexing based on scale invariant interest points,” Proc. International Conference on Computer Vision (ICCV), pp.525–531, 2001.

[21] E. Stollnitz, T. DeRose, and D. Salesin, “Wavelet for computer graphics: A primer, part 1,” IEEE Comput. Graph. Appl., vol.15, no.3, pp.76–84, 1995.

[22] B. Heisele, T. Serre, S. Prentice, and T. Poggio, “Hierarchical classification and feature reduction for fast face detection with support vector machines,” Pattern Recognit., vol.36, no.9, pp.2007–2017, 2003.

[23] S. Nicolas, J. Dardenne, T. Paquet, and L. Heutte, “Document image segmentation using a 2D conditional random field model,” Proc. International Conference on Document Analysis and Recognition (ICDAR), vol.1, pp.407–411, 2007.

（平成 20 年 8 月 18 日受付，12 月 26 日再受付）

山口　拓真　（学生員）

平 14 信州大・工・情報卒．平 16 同大大学院工学系研究科博士前期課程了．同年，メディアドライブ（株）入社．現在，日本信号（株）勤務，及び，信州大学大学院総合工学系研究科博士後期課程在学中．主に，パターン認識の研究に従事．平 20 IAPR-DAS2008 Best Paper Award Honorable Mention，スケジューリング学会技術賞各受賞．

丸山　　稔　（正員）

昭 57 東大・工・計数卒．同年三菱電機（株）入社，先端技術総合研究所勤務．平 2〜3 マサチューセッツ工科大学人工知能研究所客員研究員．平 8 信州大学工学部情報工学科助教授．工博．三次元物体認識，学習等の研究に従事．情報処理学会，IEEE，ACM 各会員．平 20 IAPR-DAS2008 Best Paper Award Honorable Mention 受賞．