大域的特徴としてBoFを導入したCRFによる一般物体認識

奥村　健志†　　滝口　哲也††　　有木　康雄††

† 神戸大学大学院工学研究科　〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
†† 神戸大学自然科学系先端融合研究環　〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
E-mail: †okumura@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし　計算機による一般物体認識の実現は，ロボットビジョンや画像検索など様々な分野で求められており，近年盛んに研究されている．従来研究の一つに Conditional Random Field(CRF) を用いる手法がある．この手法は，画素や領域といった局所領域から抽出された特徴と隣接領域間のクラス共起に基づいて各領域のクラスを認識する．しかし，局所的な特徴と関係のみを用いているので，局所最適な認識結果に陥りやすいという問題がある．この問題に対処するため，本研究では，大域的特徴として Bag of Features(BoF) を CRF に導入して一般物体を認識する方法を提案する．21 クラスの画像データセットによる実験の結果，提案手法により認識率が 6.5%向上した．
キーワード　一般物体認識，画像セグメンテーション，条件付確率場，Bag of Features

Generic Object Recognition using CRF by Incorporating BoF as Global Features

Takeshi OKUMURA†, Tetsuya TAKIGUCHI††, and Yasuo ARIKI††

† Graduate School of Engineering, Kobe University　1–1, Rokkodai, Nada, Kobe, 657–8501 Japan
†† Organization of Advanced Science and Technology, Kobe University　1–1, Rokkodai, Nada, Kobe, 657–8501 Japan
E-mail: †okumura@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract　Generic object recognition by the computer is required in various fields like robot vision and image retrieval in recent years. A conventional method uses Conditional Random Field(CRF) that recognizes the class of each region using the features extracted from the local regions and the class co-occurrence between the adjoining regions. However, there is a problem that it tends to fall into the local optimal recognition result because it uses only a local feature and the relation. To solve this problem, we propose the method that recognizes the generic object by incorporating Bag of Features(BoF) as the global feature into CRF. As a result of the experiment to the image data set of 21 classes, the proposal method has improved the recognition rate by 6.5%.
Key words　generic object recognition, image Segmentation, Conditional Random Field, Bag of Features

1. はじめに

一般物体認識とは，計算機によって制約のない実世界のシーンの画像に含まれる物体を一般的な名称で認識することを指し，コンピュータビジョンの分野において最も困難な課題の一つである．計算機による人間の高次視覚機能の実現という観点から，ロボットビジョンへの応用が期待できる．また，近年デジタルカメラの普及やハードディスクドライブの大容量化に伴い，膨大な動画像の分類や検索が人手では困難となっている．そこで，計算機による動画像の分類や検索が求められており，一般物体認識の重要性がますます高まってきている．

一般物体認識の従来のアプローチは大きく分けて 2 種類存在する．ひとつは，画像単位でクラスを認識するアプローチである．これには，Bag of Features(BoF) [1] という局所的特徴の集合で画像全体を特徴付ける手法がよく用いられる．この大域的な特徴を用いて，機械学習の Support Vector Machine(SVM) や統計的言語処理の probabilistic Latent Analysis(pLSA) により，画像のクラスを認識する．

もう一つは，図 1 のように画像の画素単位でクラスを認識するアプローチである．画素や領域といった局所領域から色特徴やテクスチャ特徴などの低次特徴を抽出し，それに基づき各局所領域のクラスを認識するが，その一つとして Gaussian Mixture Model(GMM) を用いた手法 [2] がある．しかし，この手法は各領域のクラスを

独立して認識するため，特徴が曖昧で認識の難しい領域には対応できず，また，全体としての整合性が取れていない認識結果になりやすいという問題がある．

(a) 入力画像　　(b) 出力結果

図 1　画素単位でのクラスの認識

そこで，より自然で整合性の取れた認識を目指すため，画像内の物体同士には共起の関係が存在するという考えに基づき，グラフィカルモデルである Conditional Random Field(CRF) [3] を用いた手法 [4] [5] が近年注目されている．これは，各局所領域から抽出される特徴に基づくだけでなく，隣接する局所領域間のクラス共起まで考慮した上で各局所領域のクラスを認識するというものである．特徴が曖昧で単体では認識の難しい領域に対して，周辺の領域との関係を考慮することによって，認識を改善することが可能になる．ここで述べるクラス共起とはコンテキスト情報の一種で，例えば,「牛」というクラスは「草」というクラスと同時に存在しやすいが,「車」というクラスとは同時に存在しにくい，などといった情報のことを指している．

前者の画像単位で認識を行うアプローチは，1 枚の画像に 1 クラスの物体のみが存在していると暗に仮定していると言える．これは複数クラスの物体が同時に存在するといった一般的な画像には適したアプローチではない．ゆえに本研究では，画素単位で認識を行う後者のアプローチをとり，認識のためのモデルとして CRF を用いる．

しかし，従来の CRF を用いた手法の多くに共通する問題点として，認識結果が局所最適に陥ることが挙げられる．この原因としては，局所的な特徴と関係のみに基づいて認識を行っていて，大域的な観点からの情報が不足していることが考えられる．この問題に対処するため，本研究では，画像単位で認識を行うアプローチでよく用いられており，物体の見え方の変化やオクルージョンに頑健である特徴表現の BoF を，画像全体を見たときにどんな物体が含まれているのかという大域的特徴として CRF に導入し一般物体を認識する手法を提案する．

以下では本稿の構成について述べる．2. 章では，本研究と関連性のある研究について簡単に述べ，本研究との違いを明確にしておく．3. 章では，提案手法である大域的特徴の導入を中心に手法全体について述べる．4. 章では，21 クラスと 7 クラスの 2 つの画像データセットを用いて，それぞれ認識実験を行い，提案手法の有効性を確認すると共に従来手法との比較も行う．また，それらについての考察も行う．5. 章では，問題点と今後の課題についてまとめる．

2. 関 連 研 究

大域的特徴を用いた従来研究としては，He ら [4] の提案した multiscale Conditional Random Field(mCRF) が挙げられる．これは，まず画像の各画素から色特徴やテクスチャ特徴などを抽出し，それに基づく Neural Network(NN) からの出力を画素レベル特徴とする．そして，一定の大きさの正方形領域ごとに見た NN からの出力を統合したものを領域レベル特徴とし，さらに，画像全体から見た NN からの出力を統合したものを画像レベル特徴とする．領域レベル特徴は,「車」は「道路」の上にあるなどの物体同士の境界付近のクラス共起を表現するためのもので，画像レベル特徴は，画素レベル特徴の曖昧性をなるべく解消させるための大域的特徴という位置付けとなる．これら 3 レベルの特徴を CRF の枠組みにおいて統合し，最終的な各画素のクラスの認識を行うというものが mCRF となる．

mCRF と我々の提案手法はどちらも大域的特徴を CRF に導入しているが，大域的特徴の性質に違いがある．mCRF では，大域的特徴と定義はしているが，それは画素レベル特徴という元を辿れば画素単位で抽出された低次特徴に依存しているので，曖昧な特徴になり易く不安定といえる．それに対して，本研究では，不変性を有する局所的特徴の集合に基づいており，物体の見え方の変化やオクルージョンに頑健である BoF を用いることで，安定した大域的特徴を得ることを可能にしている．

3. 提 案 手 法

図 2 に提案手法全体の流れを示す．

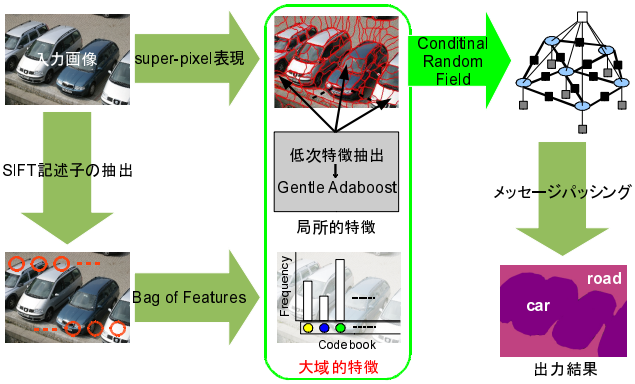

図 2　提案手法の流れ

まず，入力画像を Normalized Cuts [6] を用いた super-pixel 表現 [7] によって，小領域 (super-pixel) に領域分割する．そして，各 super-pixel から色特徴やテクスチャ特徴などの低次特徴を抽出し，それに基づいて認識対象の各クラスに対するスコアを識別器である Gentle Adaboost [8] から算出する．この各 super-pixel から得られ

る特徴ベクトルがグラフィカルモデルである Conditional Random Field(CRF) [3] における局所的特徴となる．

また，入力画像からグリッドサンプリングによって，回転不変性を有する局所的特徴である SIFT(Scale-Invariant Feature Transform) 記述子 [9] を抽出する．そして，Bag of Features(BoF) [1] により，それらをベクトル量子化したもののヒストグラムで画像全体を特徴付ける．これが CRF における大域的特徴となる．

CRF のグラフ構造は，各 super-pixel を頂点 (ノード) として隣接ノード間を辺 (エッジ) で結んだもので表し，各ノードのクラス分布は局所的特徴と大域的特徴の両方に基づいて計算される．最後に，最大周辺化事後確率推定を基準とし，隣接ノード間において，学習時に得られたクラス共起を考慮したメッセージパッシングを行う．これによって，最終的な各ノードのクラス分布が推定され，画素単位のクラスの認識が実現する．

提案手法で用いた各手法について，3.1〜3.4 節でそれぞれ述べる．

3.1 super-pixel 表現による画像の領域分割

本節では，画像の領域分割について述べる．まとまりを持った領域は単一の画素に比べ，情報量が多く冗長性も少なくて済む．しかし，領域分割の際に複数クラスを含んだ領域ができてしまうと，以降の処理でその誤りを訂正することは不可能である．これら利点と問題点を考慮した上で本研究では，Nomalized Cuts [6] を用いた super-pixel 表現 [7] により画像の領域分割を行う．これは，図 3 のように，画像を数百個単位で小領域 (super-pixel) に分割する手法である．この手法には，過剰に分割することにより，分割時の誤りを出来る限り減らし，なおかつ，色やテクスチャが一様な領域を得ることが可能であるいった利点がある．

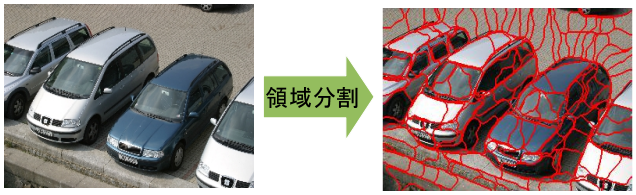

図 3　super-pixel 表現

分割アルゴリズムには，Normalized Cuts を用いる．これは，固有値分解を用いるスペクトラルクラスタリングと呼ばれるクラスタリング手法の一つである．クラスタリングの基準には，同一クラスタ内の類似度は大きく，異なるクラスタ間の類似度は小さくなるような評価関数を定義し，これを最小化する問題を固有値問題に帰着させて解くことでクラスタ分割を行う．Normalized Cuts は本来は高々10〜20 個の領域に分割することを意図した手法であったが，設定パラメータを変更することにより，数百個単位の分割を行うことを可能にしたのが super-pixel 表現である．

3.2 認識に用いる局所的特徴

本節では，認識に用いる局所的特徴について述べる．まず，3.1 節で述べた各 super-pixel から以下の低次特徴を抽出する．

- RGB，HSV，Lab，YCbCr 色空間の各成分
- Gabor filter，LoG のフィルタ応答
- super-pixel の重心座標
- super-pixel の面積

1 つ目は色特徴で，代表的な色空間を用いている．各色空間は 3 つの成分を持つ．色特徴は画素ごとに得られるが，super-pixel ごとに特徴づけを行うため，各 super-pixel 内の画素の色特徴から統計量を求める．用いる統計量は，平均，標準偏差，歪度，尖度の 4 つで，歪度は分布の非対称性の度合いを示し，尖度は分布が平均の近くに密集している度合いを示す統計量である．よって，色特徴の次元数は 4(色空間) × 3(成分) × 4(統計量) の 48 次元となる．

2 つ目はテクスチャ特徴で，2 つのフィルタ関数をそれぞれ画像と畳み込むことによって特徴 (フィルタ応答) が得られる．Gabor filter は，パラメータによって特定の向きや大きさのテクスチャを抽出することが可能なフィルタ関数で，LoG(Laplacian-of-Gaussian) は，画像内の輝度勾配が大きいところほど強く特徴が抽出されるフィルタ関数である．本研究では，Gabor filter からは 36 フィルタ応答，LoG からは 6 フィルタ応答を各画素から得る．色特徴と同様の理由から，統計量を求めるので，テクスチャ特徴の次元数は 42(フィルタ応答) × 4(統計量) の 168 次元となる．

3 つ目は位置特徴，4 つ目は幾何特徴であり，これらは画像のサイズでそれぞれ正規化しておく．位置特徴の次元数は 2 次元，幾何特徴は 1 次元である．まとめると，各 super-pixel からは 48(色特徴) + 168(テクスチャ特徴) + 2(位置特徴) + 1(幾何特徴) の 219 次元の低次特徴ベクトルが抽出されることになる．

次に，各 super-pixel から抽出した低次特徴に基づいて，識別器の Gentle Adaboost により，認識対象クラスごとのスコアを算出する．スコアは実数値を取り，大きいほどそのクラスに近いことを示している．Gentle Adaboost は，多数の弱識別器の重み付き投票で出力を決定する Boosting の一種である Adaboost から派生したものである．通常の Adaboost に比べ，外れ値への頑健性が向上している．

Gentle Adaboost は二値判別の識別器であるので，学習時に認識対象クラス数分だけ識別器を作成し学習させる．学習は，画素単位で正解ラベルの与えられた画像を用いて，各 super-pixel に正解クラスを割り当てることで行う．

学習後，ある super-pixel から抽出した低次特徴を $\mathbf{f}$，認識対象クラス数を C，あるクラス $c \in \{1, 2, \cdots, C\}$ のための識別器の出力を $H_c(\mathbf{f})$ とすると，低次特徴 $\mathbf{f}$ に基づくクラス c のスコア $l_c(\mathbf{f})$ は以下の式で示される．

$$l_c(\mathbf{f}) = \frac{\exp\{H_c(\mathbf{f})\}}{\sum_{i=1}^{C} \exp\{H_i(\mathbf{f})\}} \quad \left(\sum_{i=1}^{C} l_i(\mathbf{f}) = 1\right) \tag{1}$$

ソフトマックス関数を用いて，スコアの総和が 1 になるよう正規化している．こうして各 super-pixel は，抽出した低次特徴に基づく認識対象クラスごとのスコアを並べたベクトル $\mathbf{l} = [l_1, l_2, \cdots, l_C]$ で特徴付けられる．次元数は C 次元である．これが認識に用いる局所的特徴となる．

3.3 認識に用いる大域的特徴

本節では，認識に用いる大域的特徴について述べる．本研究では，Bag of Features(BoF) [1] という特徴表現を用いて，画像全体を特徴付ける．これは，言語処理における Bag of Words(BoW) のアナロジで，BoW が語順を無視した単語の集合として文書を表現する手法であるのに対し，BoF は位置情報を無視した局所的特徴の集合として画像を表現する手法となっている．

BoF は本来，SIFT(Scale-Invariant Feature Transform) 特徴 [9] という拡大・縮小と回転に対して不変な局所的特徴をベースに用いる．拡大・縮小に対する不変性は，DoG(Difference-of-Gaussian) による特徴点とそのスケールの自動検出という処理で得られる．このとき，特徴点は輝度勾配の変化の大きいところに集中して検出され，画像の特徴付けに偏りが生じるという問題がある．

そこで，本研究では DoG による処理は行わず，図 4 のように，格子状に等間隔に抽出するグリッドサンプリングで特徴点を検出し，特徴点のスケールは実験的に複数の値を定める．各円の中心が特徴点で，円の半径の大きさが特徴点のスケールの大きさを表す．このように検出を行うことで，特徴付けに偏りが生じることを防ぐことができる．また，拡大・縮小に対する不変性の欠如には複数スケールを取ることで補っており，総じて安定した特徴付けが実現できる．

図 4　複数スケールを持たせたグリッドサンプリング

特徴点とそのスケールが決定すると，次は各特徴点において特徴量の記述を行う．これは従来通りに SIFT 記述子を用いる．まず，各特徴点に対して代表となるオリエンテーションを求める．ノイズの影響を抑えるため，入力画像 $I(x, y)$ とガウス関数 $G(x, y; \sigma)$ を畳み込んだ平滑化画像 $L(x, y; \sigma)$ を以下の式より求める．

$$L(x, y; \sigma) = G(x, y; \sigma) * I(x, y) \tag{2}$$

$$G(x, y; \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \tag{3}$$

平滑化画像の各画素の勾配強度 $m(x, y)$ とその勾配方向 $\theta(x, y)$ を以下の式より求める．

$$m(x, y) = \sqrt{f_x(x, y)^2 + f_y(x, y)^2} \tag{4}$$

$$\theta(x, y) = \tan^{-1} \frac{f_y(x, y)}{f_x(x, y)} \tag{5}$$

$$\begin{cases} f_x(x, y) = L(x + 1, y) - L(x - 1, y) \\ f_y(x, y) = L(x, y + 1) - L(x, y - 1) \end{cases} \tag{6}$$

以上により得られた勾配強度 $m(x, y)$ と勾配方向 $\theta(x, y)$ を用いて，図 5 に示すような重み付き勾配方向ヒストグラムを以下の式により作成する．

$$h_{\theta'} = \sum_x \sum_y w(x, y) \cdot \delta[\theta', \theta(x, y)] \tag{7}$$

$$w(x, y) = G(x, y; \sigma) \cdot m(x, y) \tag{8}$$

ここで，h_θ は全方向を 36 方向に量子化したヒストグラムである．$w(x, y)$ はある局所領域の画素 (x, y) の重みであり，特徴点のスケールサイズのガウス窓 $G(x, y; \sigma)$ と勾配強度を掛け合わせたものである．このガウス窓による重みを付けることで，特徴点に近い特徴量がより強く反映される．δ は Kronecker のデルタ関数で，勾配方向 $\theta(x, y)$ が量子化した方向 θ' に含まれるとき 1 となる．この 36 方向のヒストグラムの最大値のオリエンテーションを，その特徴点の代表オリエンテーションとして割り当てる．

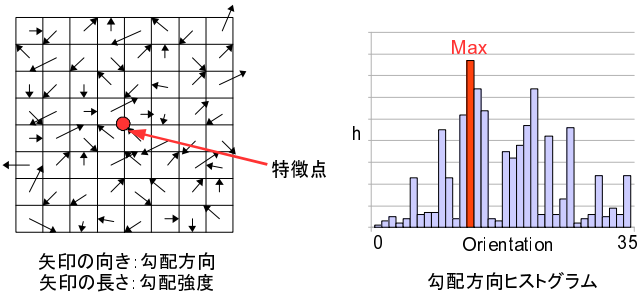

図 5　代表オリエンテーションの割り当て

スケールに基づく特徴点の周辺領域を，代表オリエンテーションを基準とした軸に回転させる．これにより，常に代表オリエンテーションを基準として特徴量が算出されるため，回転に対する不変性が得られる．そして，この特徴点周辺領域を一辺 4 ブロックの計 16 ブロックに

分割し，各ブロック毎に 8 方向の輝度勾配ヒストグラムを作成する．これにより，図 6 のような 4(ブロック) × 4(ブロック) × 8(方向) の 128 次元の特徴量である SIFT 記述子が得られる．こうして，画像から回転不変性を有する局所的特徴 (SIFT 記述子) を偏りなく抽出する．

この SIFT 記述子は全てグレースケール空間で計算されており，色情報は考慮されていない．そこで，画像の色空間を RGB から HSV に変換した後，各色空間において先と同様の処理を行うことで，色情報を含んだ 128 × 3 の 384 次元の特徴量が得られる．前者を Gray SIFT 記述子，後者を Color SIFT 記述子と名付けておく．

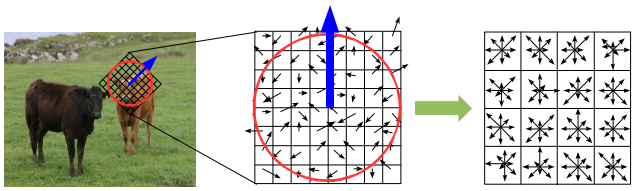

図 6　SIFT 記述子

次に，図 7 のように，全学習用画像から SIFT 記述子を抽出し，k-means 法によってそれらを W 個のクラスタにクラスタリングする．各クラスタの重心ベクトルを Visual Word と呼び，これらは大量に得られた局所的特徴の中の代表的なパターンと考えられる．単語数 W は実験的に決める．後は，各画像から抽出した SIFT 記述子を Codebook(Visual Word の集合) に基づいて，対応する Visual Word にベクトル量子化を行う．こうして，画像全体を Visual Word の出現頻度ヒストグラムで特徴付ける．次元数は W 次元となる．また，ヒストグラムの各ビンの和が 1 になるように正規化を行っておく．

BoF による画像の特徴付けが物体のオクルージョンに頑健であるのは，局所的特徴 (Visual Word) の集合という形で表現するからであり，物体の見え方の変化にも頑健であるのは，Visual Word の元になる SIFT 記述子が回転不変性を有するからである．

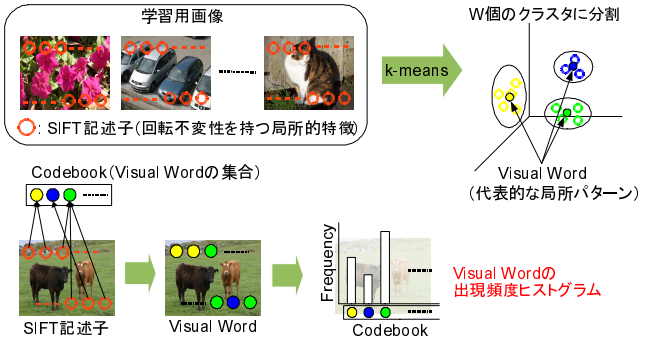

図 7　Bag of Features

3.4　Conditional Random Field による認識

本節では，3.2 節で述べた局所的特徴と 3.3 節で述べた大域的特徴を元に，Conditional Random Field(CRF) [3] を用いて，クラス共起を考慮しながら，画素単位で一般物体のクラスを認識する方法について述べる．

CRF とは言語処理の分野で提案されたグラフィカルな識別モデルで，観測される特徴に基づいて，構造を持つデータのクラスを推定することに用いられる．画像の場合，画素や super-pixel などの局所領域をノード，隣接する領域間をエッジで結んだグラフと見なす．

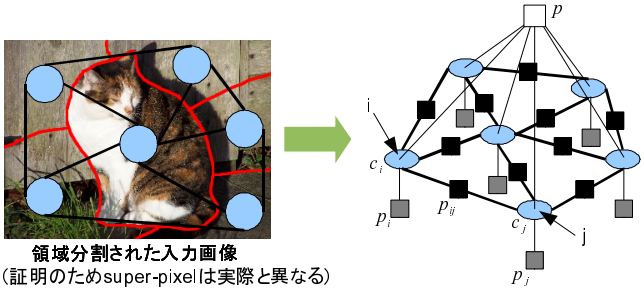

図 8　CRF による画像のグラフ表現

画像 S に対する各 super-pixel を $i \in S$，i に隣接する super-pixel の集合を N_i，観測される特徴である各 super-pixel から抽出される局所的特徴 (Gentle Adaboost からのスコア) を $\mathbf{L} = \{\mathbf{l}_i\}_{i \in S}$，画像全体から抽出される大域的特徴 (Bag of Features によるヒストグラム) を $\mathbf{g}$，また，各 super-pixel において推定されるクラスを $\mathbf{c} = \{c_i\}_{i \in S}$ とすると，CRF のモデル式は条件付分布 $P(\mathbf{c}|\mathbf{L}, \mathbf{g}; \boldsymbol{\theta})$ として以下の式で示される．図で表現すると図 8 のようになる．

$$P(\mathbf{c}|\mathbf{L}, \mathbf{g}; \boldsymbol{\theta}) = \frac{1}{Z} \exp\left[\sum_{i \in S} \{p_i(c_i|\mathbf{l}_i; \boldsymbol{\alpha}) + p(c_i|\mathbf{g}; \boldsymbol{\beta})\} + \sum_{i \in S} \sum_{j \in N_i} p_{ij}(c_i, c_j; \boldsymbol{\gamma})\right] \tag{9}$$

$$p_i(c_i|\mathbf{l}_i; \boldsymbol{\alpha}) = \sum_{k=1}^{C} \alpha_{kc_i} l_{ik} \quad \left(\sum_{k=1}^{C} l_{i,k} = 1\right) \tag{10}$$

$$p(c_i|\mathbf{g}; \boldsymbol{\beta}) = \sum_{k=1}^{W} \beta_{kc_i} g_k \quad \left(\sum_{k=1}^{W} g_k = 1\right) \tag{11}$$

$$p_{ij}(c_i, c_j; \boldsymbol{\gamma}) = \gamma_{c_i c_j} \tag{12}$$

ここで，C は認識対象のクラス数，W は Visual Word の単語数，Z は正規化項で分配関数と呼ばれる．$\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$ は CRF のモデルパラメータである．本研究では，画素単位で正解クラスが与えられた全ての学習用画像を用い，最大事後確率推定を基準として，このモデルパラメータを学習する．この基準は，過学習を防ぐ

ペナルティ項を加えた上で，T 枚の学習用画像の対数尤度の和を最大化するパラメータ $\boldsymbol{\theta}^*$ を求めるというものであり，以下の式で示される．

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{t=1}^{T} \log P(\mathbf{c}^t|\mathbf{L}^t, \mathbf{g}^t; \boldsymbol{\theta}) - \frac{R}{2}||\boldsymbol{\theta}||^2 \right\} \quad (13)$$

この $\boldsymbol{\theta}^*$ は，非線形計画法の L-BFGS 法 [10] を用いて解析的に求められる．ただし，グラフにループ構造があるため，分配関数 Z の計算が困難になっているので，擬似尤度 [11] を用いて式 (9) の分配関数 Z の近似を行っている．

学習されたモデルパラメータ $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$ それぞれの役割について述べる．$\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ は，局所的特徴と大域的特徴の各次元に対して，認識対象のクラスごとに最適な重みを与えるパラメータ行列である．$\boldsymbol{\gamma}$ は，学習用画像から得られたクラス共起を表す $C \times C$ のパラメータ行列である．具体的には，学習用画像において，頻繁に隣接して現れたクラス同士は強く共起しているということで，該当する行列の要素の値は高くなり，その逆の場合は低くなっているような行列となっている．

つまり，式 (9)(10)(11)(12) において，$p_i(c_i|\mathbf{l}_i;\boldsymbol{\alpha})$ は各ノードにおける局所的特徴に基づくクラス分布，$p(c_i|\mathbf{g};\boldsymbol{\beta})$ は各ノードにおける大域的特徴に基づくクラス分布，$p_{ij}(c_i, c_j;\boldsymbol{\gamma})$ は，隣接ノード間におけるクラス共起を表している．図で表現すると図 9 のようになる．

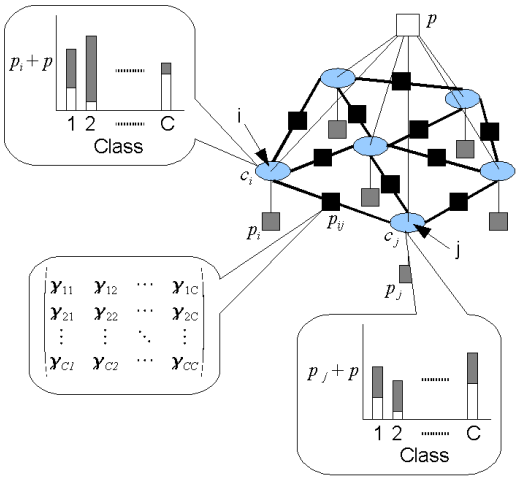

図 9　クラス分布とクラス共起

大域的特徴に基づくクラス分布は全ノードにおいて共通としているが，これは画像中に含まれている一般物体のクラスを大域的に捉えることを意図したもので，バイアスのような働きをしていると言える (図 9 の白い棒グラフ)．これに局所的特徴に基づくクラス分布 (図 9 の黒い棒グラフ) を加える形で定式化することで，局所最適に陥ることを防いでいる．

本研究の目的である正解クラスの与えられていないテスト用画像の認識には，式 (9) を最大化するように各ノードのクラスを推定する必要がある．$p_{ij}(c_i, c_j;\boldsymbol{\gamma})$ の存在から，最大化には隣接ノード間のクラス共起も考慮されることになる．

推定の基準には最大周辺事後確率推定を用いる．これは，各ノード $i \in S$ における周辺事後分布 $P(c_i|\mathbf{L}, \mathbf{g}; \boldsymbol{\theta})$ を最大化するクラス c_i^* を求めるもので，以下の式で示される．

$$\begin{aligned} c_i^* &= \arg\max_{c_i} P(c_i|\mathbf{L}, \mathbf{g}; \boldsymbol{\theta}) \\ &= \arg\max_{c_i} \sum_{\mathbf{c}\backslash c_i} P(\mathbf{c}|\mathbf{L}, \mathbf{g}; \boldsymbol{\theta}) \end{aligned} \quad (14)$$

グラフにループ構造があるので，この厳密推定は困難となっている．そこで，本研究では，ループ有り確率伝播法の一つである loopy max-product [12] アルゴリズムで近似的に推定する．この手法は，まず，隣接ノード間において局所的にメッセージを繰り返し伝播させ，メッセージを更新する．メッセージの更新には，自身と隣接ノードのクラス分布の情報を含み，クラス共起も考慮される．メッセージの更新が終了後，各ノードのクラス分布は隣接ノードからのメッセージで周辺化される．最後に，各ノードにおいて周辺化されたクラス分布から，最大値をとるクラスを選ぶことで推定が完了する．

こうして，局所的特徴と大域的特徴と局所的なクラス共起に基づき，各ノード，すなわち super-pixel のクラスが全て推定される．

4. 評価実験

4.1 実験概要

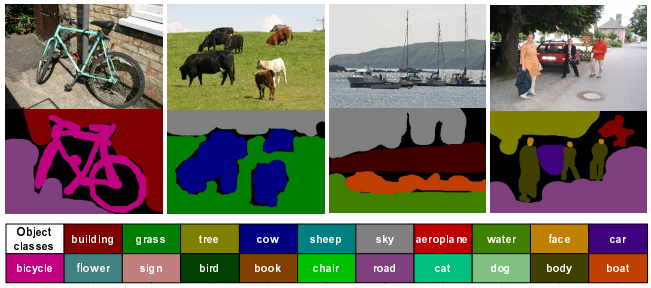

図 10　MSRC21 データセットの例

2 つの画像データセットを用いて認識実験を行う．一つには，Microsoft Reseach Cambridge 21 Dateset(MSRC21) という画像データセットを用いる．Fig.10 に示すように 21 クラスの物体が含まれる 591 枚からなるデータセットで，各画像には画素単位で正解クラスが与えられている．ただ，黒色の領域は正解クラスが与えられていないことを意味し，その領域は学習とテストには用いられない．画像のサイズはほぼ 320 × 240 画素で統一されている．

もう一つは，関連研究 [4] との比較のために用いる画像データセットとして，Corel Dataset を用いる．これ

は，7クラス (“Hippo/Rhino”，“Polar Bear”，“Water”，“Snow”，“Vegatation”，“Ground”，“Sky”) の物体が含まれる 104 枚からなるデータセットで，MSRC21 と同様に各画像には画素単位で正解クラスが与えられており，こちらには正解クラスなしの領域は存在しない．画像のサイズは 180 × 120 画素で統一されている．

評価方法は，画素単位で正解・不正解を数え上げて，以下の式のようにクラスごとの認識率を平均した Accuray を用いる．

$$\text{Accuracy}[\%] = \frac{\text{各クラスの認識率の和}}{\text{クラスの種類数}} \qquad (15)$$

MSRC21 での実験では，学習用画像には 295 枚，テスト用画像には 296 枚をランダムに選択し，これを 3 セット作りそれぞれの Accuray を平均したもので評価する．また，super-pixel の数は 1 画像あたり約 200 個，グリッドサンプリングの間隔は 10 画素，SIFT 記述子のスケールは $\{4, 8, 12, 16\}$ とした．

Corel での実験では，関連研究 [4] と同様に，学習用画像には 60 枚，テスト用画像には 44 枚をランダムに選択し，これを 3 セット作りそれぞれの Accuracy を平均したもので評価する．また，super-pixel の数は 1 画像あたり約 100 個，グリッドサンプリングの間隔は 5 画素，SIFT 記述子のスケールは $\{2, 4, 6, 8\}$ とした．

どちらのデータセットの実験においても，提案手法である大域的特徴としての Bag of Features(BoF) の導入の有無による認識率の変化を調べるが，導入の有無だけでなく，BoF についても以下の 3 パターンで作成したものでそれぞれ実験を行い，認識率の変化を調べる．

- グリッドサンプリング＋ Color SIFT 記述子
- グリッドサンプリング＋ Gray SIFT 記述子
- DoG による自動検出＋ Gray SIFT 記述子 [9]

1 つ目のパターンの BoF を用いた提案手法を Grid-Color，2 つ目のを Grid-Gray，3 つ目のを Dog-Gray と名付けておく．また，Visual Word の単語数についても 100〜1000 の間で 100 刻みに変化させて実験を行う．以上をまとめた実験結果を次節で示す．

4.2 実験結果と考察

表 1 と表 2 に各データセットを用いたときの実験結果を示す．大域的特徴を導入した手法については，Visual Word の単語数を変化させた中で最も高かった認識率と括弧内にそのときの単語数を示している．

表 1　MSRC21 での実験結果

	Accuracy[%]
Grid-Color	64.6(300word)
Grid-Gray	65.5(500word)
DoG-Gray	62.3(600word)
大域的特徴無し	59.0

表 2　Corel での実験結果

	Accuracy[%]
Grid-Color	74.3(600word)
Grid-Gray	73.0(400word)
DoG-Gray	68.0(800word)
大域的特徴無し	73.0
関連研究 [4]	80.9

まず，MSRC21 での実験結果について考察を述べる．表 1 より，大域的特徴の導入により，一定の認識率の向上が見られ，最大 6.5%向上した．特に，DoG による偏りのある特徴付けよりも，グリッドサンプリングを用いた特徴付けの方が向上の幅が大きくなっていることが確認できる．色情報を用いていない Grid-Gray の方が，色情報を用いている Grid-Color よりも認識率が高い理由としては，認識対象が 21 クラスの一般物体であるので，クラス内の色の分散が大きくなりやすく，色情報では各クラスを特徴付けにくいということが考えられる．

大域的特徴の導入の有無によって，認識結果に差が表れたものを図 11 に示す．

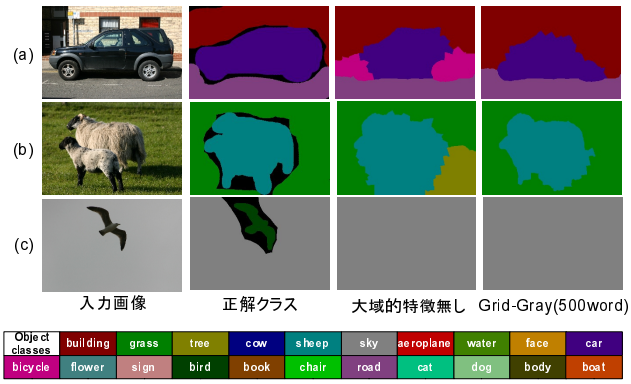

図 11　MSRC21 での実験における認識結果

(a) と (b) は大域的特徴の導入によって，局所最適な認識結果に陥る状態から改善できた場合といえる．(c) は非常に小さく物体が写っていて特徴がほとんど得られず，大域的特徴を導入しても認識は改善されない．対処法としては，物体検出器 (この場合，鳥検出器) を用意して，画像内を走査した結果を新たな特徴として CRF に導入することが考えられる．

次に，Corel での実験結果について考察を述べる．表 2 より，大意的特徴の導入による認識率の改善効果は乏しいことがわかる．1 番の原因としては Corel 画像の解像度の低さが挙げられる．SIFT 記述子は輝度勾配ベースの局所パターンを抽出するのだが，画像が潰れてしまっていて上手く抽出できないと考えられる．DoG-Gray が大きく認識率を下げていることからも，このように考えるのが妥当といえる．

Grid-Color の場合だけ，少し認識率が向上しているが，これは Corel 画像のクラスの多くが MSRC21 と比べて，

クラス内の色の分散が小さいためと考えられる．

関連研究 [4] との比較では，提案手法の認識率の方が低くなり，画像セットに対して適用限界があることがわかった．

今後の課題として，解像度の低い画像に対しても安定して抽出可能な特徴を検討する必要がある．また，クラス内分散が大きいことから，画像から抽出される特徴のみでは一般物体の認識は難しく，クラス共起以外にも有用なコンテキスト情報がないか検討する必要がある．

5. ま と め

本稿では，super-pixel という局所領域から抽出した低次特徴に基づく Gentle Adaboost のスコアを局所的特徴，画像全体からグリッドサンプリングによって抽出した SIFT 記述子に基づく Bag of Features を大域的特徴とし，Conditional Random Field により，これらの特徴と隣接領域間のクラス共起に基づいて，画素単位で一般物体を認識する手法を提案した．BoF を大域的特徴として導入することで，局所最適な認識結果に陥ることを防ぎ，一定の認識精度の向上が見られた．今後の課題としては，より頑健な特徴とクラス共起以外の有用なコンテキスト情報を検討する予定である．

文　　献

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

[2] K. Barnard, and D. Forsyth, “Learning the Semantics of Words and Pictures, ” Proc. IEEE International Conference on Computer Vision, pp. 408–415, 2001.

[3] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data, ” Proc. International Conference on Machine Learning, 2001.

[4] X. He, R. S. Zemel, and M.*Á*. Carreira-Perpi*ñán*, “Multiscale conditional random fields for image labeling, ” Proc. IEEE Computer Vision and Patern Recognition, pp.695-702, 2004.

[5] J. Shotton, J. Winn, C. Rother, and A. Criminisi, “Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation, ” Proc. IEEE European Conference on Computer Vision, pp.1-15, 2006.

[6] J. Shi, J. Malik, “Normalized cuts and image segmentation, ” Proc. IEEE Computer Vision and Pattern Recognition, pp731-737, 1997.

[7] X. Ren, J. Malik, “Learning a classification model for segmentation, ” Proc. IEEE International Conference on Computer Vision, 2003.

[8] J. Friedman, T. Hastie, R. Tibshirani, “Additive logistic regression: a statistical view of boosting, ” Technical Report, 1998.

[9] D. G. Lowe, “Object recognition from local scale-invariant features, ” Proc. IEEE International Conference on Computer Vision, pp.1150-1157, 1999.

[10] J. Nocedal, “Updating Quasi-Newton Matrices With Limited Storage, ” Mathematics of Computation, pp.773-782, 1980.

[11] J. Besag, “Statistical analysis of non-lattice data, ” The Statistician, 24, pp.179-195, 1975.

[12] C. M. Bishop, “Pattern Recognition and Machine Learning, ” Springer, Chapter.8, 2006.