コンテクストに基づく移動ロボットのための教師なしシーン分類

内海　祐哉† 　間所　洋和† 　佐藤　和人†

† 秋田県立大学大学院システム科学技術研究科　〒 015–0055 秋田県由利本荘市土谷字海老ノ口 84–4
E-mail: †{m12a005,madokoro,ksato}@akita-pu.ac.jp

あらまし　本論文では，自律移動ロボットにおける屋内シーンの意味的な認識を目的として，事前にカテゴリ数の設定を必要としない教師なしシーン分類法を提案する．本手法では，SIFT (Scale-Invariant Feature Transform) の特徴量を用いて作成した VW (Visual Words) と，Gist の特徴量を用いて作成したブロック毎の VW を用いて，SIFT の特徴点における VW とその特徴点が含まれる Gist のブロックの VW を，2 次元ヒストグラムとして投票することにより，BoVW (Bags of VW) を作成する．更に，ART-2 (Adaptive Resonance Theory-2) の追加学習機能を用いることにより，時系列データに対して安定性と可塑性を保ちながらカテゴリの候補となるラベルを生成するとともに，CPN (Counter Propagation Networks) の教師信号として生成されたラベルを用いることにより，教師なし学習によるシーン分類を実現する．また，CPN のカテゴリマップには，シーンの空間的な位相関係が写像されるため，カテゴリ内に含まれる分類シーンの関係性が可視化される．ロボットの位置推定とナビゲーションの評価用として公開されている KTH-IDOL データセットを用いた基礎実験では，キッチンやリビング，廊下といった意味カテゴリ単位の分類精度について評価した．更に，独自に試作開発した移動ロボットを用いて，全方位センサから得られる視野画像列を用いた評価実験では，時系列画像に対する提案手法の教師なしシーン分類における有効性を検証した．

キーワード　シーン分類，ロボットビジョン，コンテクスト，教師なし学習，SIFT，Gist，SOM，ART-2，CPN．

1. はじめに

ロボットが私たちの生活環境で共生するためには，与えられたプログラムに沿って動作するだけでなく，時々刻々と変化する環境や状況に合わせて，自律的に判断し行動する能力が求められる．ロボットの自律移動には，レーザレンジファインダなどの距離情報が得られるセンサを用いて，環境地図を構築しながら同時に自己位置を推定する SLAM (Simultaneous Localization and Mapping) [1] によるアプローチが主流となっている．しかしながら，SLAM のみでは環境内のオブジェクトや壁，障害物までの距離情報しか得られないため，例えば，キッチンやリビングといった意味カテゴリまで認識することは困難である [2]．意味カテゴリの認識と，ロボットの位置を同定しながら地図構築をする SLAM を組合わせることで，ロボットの知能化や知的な自律行動へと結び付くと考えられている [3]．このため，シーンの意味カテゴリの認識は，コンピュータビジョンやロボットビジョンにおいて，興味深いテーマとなっている．

コンピュータビジョン分野では，インターネットを通じて収集された大量のシーン画像から，意味カテゴリを認識するための分類手法が数多く提案されている [4]．しかしながら，分類対象は静止画像でかつ屋外シーンが中心となっている．また，屋外シーンに対する分類法を屋内シーンに対し適用した場合，分類精度が急激に低下することが報告されている [5]．今後，急速な普及が期待されている人間共生型のロボットへの適用を考えると，人間の生活環境である屋内シーンに対する分類精度の向上が期待されている．また，このようなロボットが動作する環境は，人間の活動や生活スタイルに合わせて時々刻々と変化するため，ロボットは学習によって環境に適応する能力が求められる．近年の計算機能力の進歩と相まって，シーンの認識や分類では，機械学習を用いた汎用的かつ適応的な手法が数多く提案されている．

機械学習は教師あり学習と教師なし学習に大別される．教師あり学習では，事前に教師信号を含む訓練用のデータセットを準備しなければならない．このため，ロボットの設計者や利用者にとって，教示に伴う負荷が大きくなる．一方，教師なし学習では，明示的な教師信号は必要とせず，学習によって得られた結果に対して，利用者が意味情報を付与するのみとなる．つまり，ロボット自身が環境から得られた様々なセンシング情報を，カテゴリとして整理し提示してくれるため，教師あり学習を用いる場合と比較してユーザの負荷が軽減される．また，教師なし学習により整理された情報の中から，ロボット自身が知識を発見できる可能性がある．このため，教師なし学習に基づいた手法を用いることで，ロボットと人の高度なコミュニケーションインタラクションが実現できると考えられている [6]．

本論文では，自律移動ロボットにおける屋内シーンの意味的な認識を目的として，コンテクストに基づく教師なしシーン分類法を提案する．本手法では，SIFT (Scale-Invariant Feature Transform) の特徴量を用いて作成した VW (Visual Words) と，Gist の特徴量を用い

て作成したブロック毎の VW を用いて，SIFT の特徴点における VW とその特徴点が含まれる Gist のブロックの VW を，2 次元ヒストグラムとして投票することにより，BoVW (Bags of VW) を作成する．更に，ART-2 (Adaptive Resonance Theory-2) の追加学習機能を用いることにより，時系列データに対して安定性と可塑性を保ちながらカテゴリの候補となるラベルを生成するとともに，CPN (Counter Propagation Networks) の教師信号として生成されたラベルを用いることにより，教師なし学習によるシーン分類を実現する．また，CPN のカテゴリマップには，シーンの空間的な位相関係が写像されるため，カテゴリ内に含まれる分類シーンの関係性が可視化される．ロボットの位置推定とナビゲーションの評価用として公開されている KTH-IDOL データセットを用いた基礎実験では，寝室やリビング，廊下といった意味カテゴリレベルでの分類精度を評価する．更に，独自に試作開発した移動ロボットを用いて，全方位センサから得られる視野画像列を用いた評価実験では，時系列画像に対する提案手法の教師なしシーン分類における有効性を検証する．

2. 関 連 研 究

意味カテゴリ認識のためのシーン分類は，Siagian ら [4] によって，オブジェクトベース，領域ベース，コンテクストベースの 3 種類のアプローチに分類されている．

オブジェクトベースのシーン分類では，ランドマークとなるオブジェクトに基づきシーンを分類する．このため，シーン中にランドマークとなる物体が，事前に存在することが必要条件となる [7], [8]．シーンから局所特徴を抽出する手法として，SIFT や SURF (Speeded Up Robust Features) が多く用いられているが，環境変化に対するロバスト性の低さが課題となっている．また，視野範囲が広い環境では，シーン中のオブジェクトは相対的に小さく写るため，オブジェクトを表現している画素の情報が，センサノイズや照明変化に影響を受けやすくなり，特徴点が検出されない場合が生じる．このような問題点を解決するために，Kawewong ら [9] は，照明変化や移動物体などの環境変化に対してロバストな局所特徴量の PIRF (Position-Invariant Robust Features) を提案した．複数枚の時系列画像から SIFT もしくは SURF を抽出して連続的にマッチングを取る PIRF は，すべての連続画像間でマッチングが取れた特徴点のみを抽出するため，画像間の系列性を考慮した特徴抽出法となっている．PIRF を用いて森岡ら [10] は，人の多い環境下での，視覚に基づく安定したナビゲーションを実現している．

領域ベースのシーン分類では，分割された領域とそれらの階層的な配置から，それぞれの位置でのシーンの特徴を形成する．Katsura ら [11] は，移動ロボットを用いて撮影した屋外シーンの画像を，空，建物，樹木の領域に分割することにより，天候や季節の変化に対してロバストなシーンの分類法を提案している．Matsumoto ら [12] は，視覚に基づくナビゲーションのためのシーンの分類法を提案しており，実ロボットを用いて屋内外での走行を実現している．オブジェクトベースの分類法と比較して，領域ベースの分類法は，環境の局所的変動に対してロバストであり，かつ，処理方式が比較的単純であるため，ロボットでの具体的なタスクとして，ナビゲーションへの適用などの実用性に優れている．しかしながら，領域ベースの問題点としては，領域分割の精度に分類結果が左右されるため，実環境において正確な領域分割は困難な課題となっている．Shi ら [13] は，normalized-cut により領域分割の精度を向上させたが，リアルタイム処理を考慮すると，領域分割のための計算コストが課題となっている．

コンテクストベースのシーン分類では，人間がシーンを認識するメカニズムに基づいて，シーン全体を低次元に圧縮して特徴を表現する．このアプローチでは，コンテクストとして大まかにシーン全体の情報が記述できるため，オブジェクトの有無，もしくは領域分割の精度による影響が少ない．コンテクストベースの特徴量としては，Oliva ら [14] によって提案された Gist が最も多く使用されている．Torralba らは Gist を利用したシーン分類法として，HMM (Hidden Markov Models) の状態数をシーンのカテゴリとして割当てたシーン分類法を提案している [15]．しかしながら，Gist のみによる手法では，屋外シーンに対しては高精度に分類できているが，屋内シーンに対しては分類精度が急激に低くなることが，Quattoni らによって報告されている [5]．そこで，彼らは屋内シーンの分類に焦点を当て，関心領域から得られる SIFT と，画像全体の Gist の特徴量を，距離関数を用いて分類することにより，屋内シーンにおける分類精度を向上させる手法を提案した [5]．しかしながら，関心領域を設定するためのアノテーションを手動で行わなければならず，関心領域の探索結果がシーンの分類結果に強く依存している．

3. コンテクストに基づく教師なしシーン分類法

本研究では，自律移動ロボットにおける屋内シーンの意味的な認識を目的として，Gist と SIFT を用いた，事前にカテゴリ数の設定を必要としない教師なしシーン分類法を提案する．以下に，提案手法の全体構成を示し，各処理手順について個別に説明する．

3.1 提案手法の全体構成

本研究で提案するシーン分類法のネットワーク構成を図 1 に示す．処理手順は，

（1） SIFT による特徴点の検出と特徴量の記述，
（2） Gist によるブロック毎の特徴量の記述，
（3） SIFT と Gist の 2 次元ヒストグラムの作成，

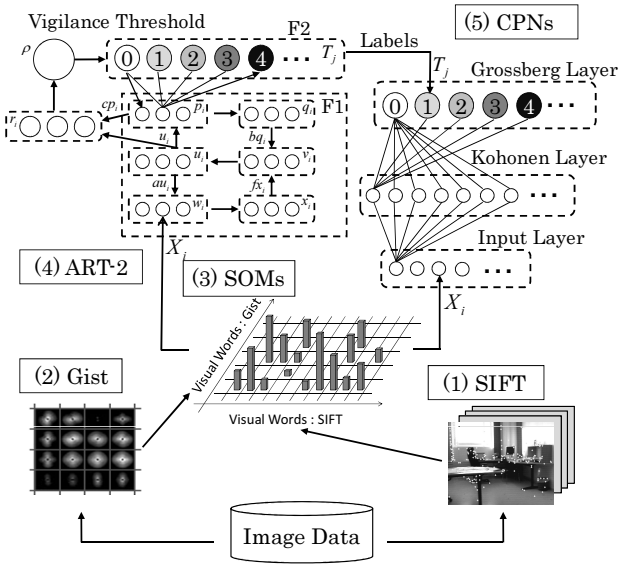


図 1　提案手法のネットワーク構成
Fig. 1　The whole architecture of our method.

（4） ART-2 を用いたラベルの生成,
（5） CPN を用いたカテゴリマップの作成,
の 5 ステップから構成されている．ステップ (1)〜(3) が SIFT と Gist を用いたコンテクストに基づく BoVW の作成となり，ステップ (4)〜(5) が教師なし学習に基づくシーンの分類となる．以下に各ステップの詳細な処理手順を記述する．

3.2　SIFT による特徴量の記述

一般物体認識では，局所特徴量の記述方法として，SIFT がよく用いられている [16]．このため，オブジェクトベースのシーン分類においても，SIFT によりシーン内に存在するオブジェクトの特徴を記述している [4]．本手法では，SIFT により抽出される特徴量を，コンテクストを記述するための前景領域の特徴量として用いる．

SIFT の処理は，特徴点の検出と特徴量の記述の 2 段階から構成される [17]．特徴点の検出には，DoG (Different of Gaussians) が用いられる．DoG により注目画素とその周りの 26 近傍を比較し，極値であった場合，その画素を特徴点の候補として検出する．検出された特徴点の候補には，直線上のエッジなどの特徴点が多く含まれるため，絞り込みを行う．続いて，特徴点の周辺領域における勾配強度と勾配方向から重み付方向ヒストグラム算出する．特徴量の記述では，4×4 ブロックの領域に，それぞれ 8 方向ヒストグラムを作成する．よって，128 次元の特徴量が算出される．この 128 次元の特徴量を全ての特徴点に対して算出する．

3.3　Gist による特徴量の記述

Gist とは，シーンに関する意味的なカテゴリ，含まれているオブジェクトのレイアウト，シーンに含まれている少数の主要なオブジェクトに関する属性や知識といった内容の総称である [18]．また，シーンにおける Gist とは，あるオブジェクトが存在する文脈としてのコンテクストと位置づけられている [19]．本手法では，Gist により抽出される特徴量を，コンテクストを記述するための背景領域の特徴量として用いる．

Oliva ら [14] によって提案された Gist 特徴量は，道路や山，建物などの主に屋外シーンの構造的な特徴を記述するための特徴量である．Gist は，画像を $n \times n$ ブロックに分割した領域に対して，フーリエ変換を用いてブロックごとに周波数を解析する．更に，カットオフ周波数によりブロックごとにフィルタリングをする．フィルタリングしたブロックに対し，任意の方向フィルタにおける強度を算出することで特徴量が抽出される．本手法では，ブロック数を n=4，カットオフ周波数を 1, 2, 4 cycle/image とした．方向フィルタは，カットオフ周波数ごとに，8 方向，8 方向，4 方向である．また，RGB カラー画像を用いたため，色空間毎に特徴量が算出される．以上より，本手法における 1 ブロックあたりの Gist の次元数は 60 次元となる．

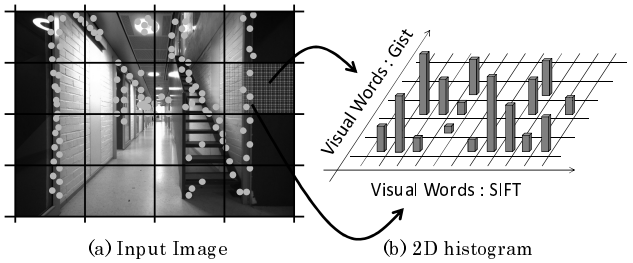


図 2　2 次元ヒストグラムの作成方法
Fig. 2　Generation of 2 dimentional histogram.

3.4　2 次元ヒストグラムの作成

本手法では，SIFT と Gist のそれぞれの特徴量を用いて作成した VW から，BoVW として 2 次元ヒストグラムを作成する．2 次元ヒストグラムの作成方法を図 2 に示す．ヒストグラムの縦軸は Gist の VW，横軸は SIFT の VW を示す．SIFT によって検出された特徴点が位置するブロックの Gist 特徴量が対応する VW と，その SIFT 特徴量が対応する VW のマトリクス上の位置に投票することにより，2 次元ヒストグラムを作成する．これにより，シーン中に含まれる物体のパーツベースでの記述を前景領域，それに対応する大局的な特徴量を背景領域として，コンテクストによる特徴量を記述することができる．

永橋ら [20] は，前景における SIFT 特徴点と，そのキーポイントが持つ 6 倍のスケール領域に存在する背景領域の特徴点を対応付けて投票することにより作成した 2 次元ヒストグラムを用いて，コンテクストとして特徴量を記述している．しかしながら，永橋らの手法では，前景領域と背景領域の境界が既知でなければ，2 次元ヒストグラムを作成することができない．一方，本手法は，

SIFT を前景領域の特徴量，Gist を背景領域の特徴量として位置づけることにより，前景領域と背景領域の境界が未知の画像に対しても適用できる．

VW の作成には一般的に k-means [21] が用いられている [22] が，本手法では Kohonen によって提案された SOM (Self-Organizing Maps) [23] を用いた．寺島ら [24] は，k-means より SOM が教師なしクラス分類において誤認識率が最小となることを示している．本研究においても，予備実験より，VW の作成における SOM の比較優位性を確認している．SOM の学習は，入力データの位相構造を保ちながら，ネットワークの結合荷重を変更する．入力データに最も類似するユニットが発火し，その周辺のユニットが近傍領域を形成する．よって SOM は，学習データの分布と類似する様々なデータを分類することができる．なお，SOM の学習アルゴリズムは，CPN の入力層と Kohonen 層間のアルゴリズム [23] と同じである．

3.5 ART-2 を用いたラベルの形成

Carpenter らによって提案された ART-2 [25] は，時系列データに対して安定性と可塑性を保ちながらラベル形成できる教師なしニューラルネットワークである．ART-2 を用いることにより，追加的にカテゴリの候補となるラベルが生成できる．ロボットから得られる視野画像列は，ロボットの行動とともに時間に沿って変化するため，追加的に時系列データを学習できる ART-2 の適用は，本研究におけるシーン分類において有用と考えられる．

ART-2 のネットワークは，特徴表現の Field 1 (F1) と，カテゴリ表現の Field 2 (F2) から構成されている．F1 は，複数のサブレイヤから構成されており，入力データが各サブレイヤを遷移することによって短期記憶を実現する．また，短期記憶の効果によって，入力データ内のノイズが除去されるとともに，特徴が強調される．F2 には，位相の強弱により長期記憶としてカテゴリが形成される．本研究では，このカテゴリをラベルとして用いる．

ART-2 のアルゴリズムを以下に記す．F1 と F2 は p_i を介して接続されている．入力 $X_i(t)$ に対して F1 伝搬後の F2 の最大活性化ユニット T_j は，

$$T_j(t) = max(\sum_j p_i(t) Z_{ij}(t)), \tag{1}$$

となる．T_j を基準として，トップダウン結合荷重 Z_{ji} とボトムアップ結合荷重 Z_{ij} は次式で更新される．

$$\frac{d}{dt} Z_{ji}(t) = d[p_i(t) - Z_{ji}(t)], \tag{2}$$

$$\frac{d}{dt} Z_{ij}(t) = d[p_i(t) - Z_{ij}(t)], \tag{3}$$

次に，ビジランス閾値 ρ を用いて，カテゴリに属するかを判定する．

$$r_i(t) = \frac{u_i(t) + cp_i(t)}{e + ||u|| + ||cp||}, \qquad \frac{\rho}{e + ||r||} > 1. \tag{4}$$

判定が成立する場合は，選択されたユニットをリセットして再探索する．不成立の場合は，F1 層内の変化率が小さくなるまで伝搬と結合荷重の更新を繰り返す．

3.6 CPN を用いたカテゴリマップの生成

Nilsen によって提案された CPN [26] は，競合学習と近傍学習に基づき，パターンを特定のカテゴリに分類する教師ありニューラルネットワークである．CPN のネットワークは，入力層，Kohonen 層，及び Grossberg 層の 3 層で構成され，学習結果は Kohonen 層のユニットにカテゴリマップとして表現される．本手法では，Grossberg 層のユニットに与える教師データを ART-2 のラベルとすることで，ラベリング処理を自動化することにより，CPN を教師なし学習として用いている．また，ART-2 と CPN を組み合わせることにより，事前にカテゴリ数の設定を必要とせず，追加的にカテゴリの候補となるラベルが生成されるとともに，類似度に基づくカテゴリ間の空間関係を可視化できる [6].

CPN の学習アルゴリズムを以下に記す．$u^i_{n,m}(t)$ は，時刻 t における，入力層ユニット $i(i = 1, ..., I)$ から，Kohonen 層ユニット $(n, m)(n = 1, ..., N, m = 1, ..., M)$ への結合荷重とする．$v^j_{n,m}(t)$ は，時刻 t における，Grossberg 層ユニット j から，Kohonen 層ユニット (n, m) への結合荷重とする．これらの結合荷重は，ランダムに初期化される．$X_i(t)$ は，時刻 t における入力層ユニット i に提示される学習データである．$X_i(t)$ と $u^i_{n,m}(t)$ の間のユークリッド距離 $d_{n,m}$ は次式で計算される．

$$d_{n,m} = \sqrt{\sum_{i=1}^{I} (X_i(t) - u^i_{n,m}(t))^2}. \tag{5}$$

$d_{n,m}$ が最小となるユニットが，勝者ユニット c として定義される．

$$c = argmin(d_{n,m}). \tag{6}$$

$N_c(t)$ は，勝者ユニット c の近傍領域である．$N_c(t)$ の内部の結合荷重 $u^i_{n,m}(t)$ は，Kohonen の学習アルゴリズム [23] を用いて更新される．

$$u^i_{n,m}(t+1) = u^i_{n,m}(t) + \alpha(t)(X_i(t) - u^i_{n,m}(t)). \tag{7}$$

$N_c(t)$ の内部の結合荷重 $v^j_{n,m}(t)$ は，Grossberg のアウトスター学習アルゴリズムで更新される．

$$v^j_{n,m}(t+1) = v^j_{n,m}(t) + \beta(t)(T_j(t) - v^j_{n,m}(t)). \tag{8}$$

ここで，T_j は，Grossberg 層に提示される教師信号（ART-2 で形成されたラベル）である．$\alpha(t)$ と $\beta(t)$ は，学習率係数であり，学習の進行とともに減少する．以上の処理を事前に設定した学習回数 I だけ繰り返す．

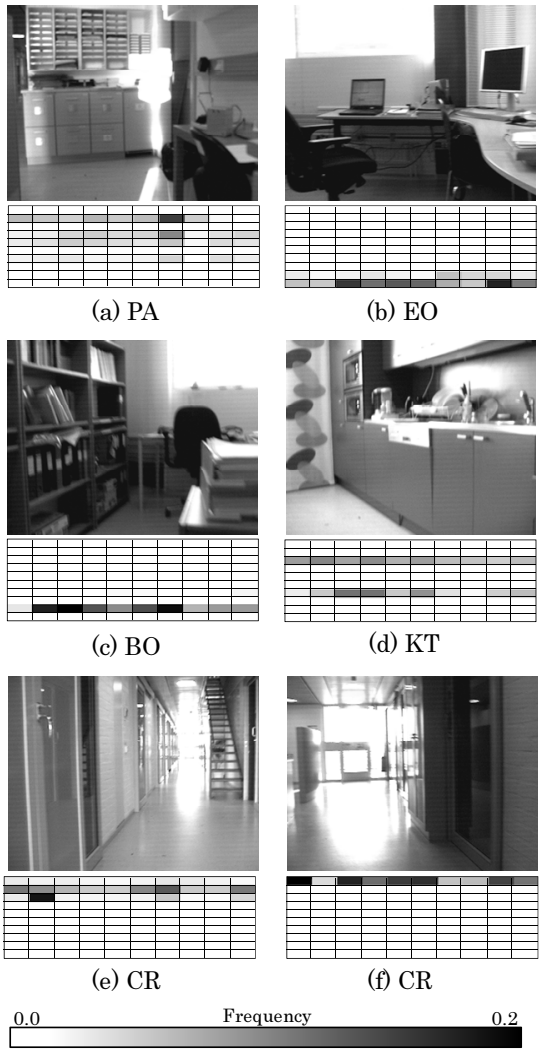


図 3　各カテゴリの画像と生成された特徴量.
Fig. 3　2D histogram.

4. KTH-IDOL を用いた基礎実験

KTH-IDOL (Image Database for rObot Localization) データセット [27] は，屋内環境におけるロボットのナビゲーション及び位置推定用として，インターネットを通じて公開されている画像データセットである．また，KTH-IDOL データセットは，Image CLEF (Cross Language Evaluation Forum) 2009 の一部として提供されている [28]．本実験では，KTH-IDOL データセットを用いたシーン分類実験を行い，寝室やリビング，廊下といった意味カテゴリでの分類精度を評価する．

4.1 実 験 条 件

KTH-IDOL データセットは，異なる高さの 2 台のロボット（Dumbo と Mannie）を用いて，曇り，夜，快晴の 3 種類の条件下で撮影された時系列画像から構成されている．本実験では，人間の身長と同程度の高さを有する Mannie から撮影された快晴時の画像を用いた．対象とするシーンは，印刷室（Printer Area: PA），個人執務室（One-person office: EO），2 人執務室（Two-persons office: BO），台所（Kitchen: KT），及び廊下（Corridor: CR）の 5 カテゴリから構成されている．

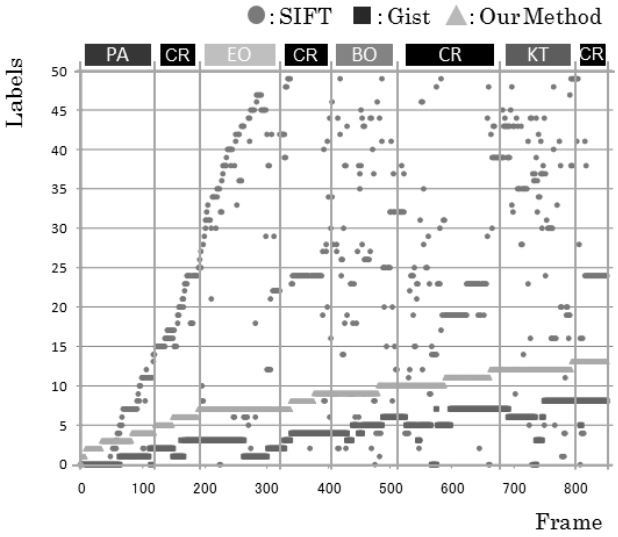


図 4　ART-2 によるラベル形成結果.
Fig. 4　Resutls of generated labes using ART-2.

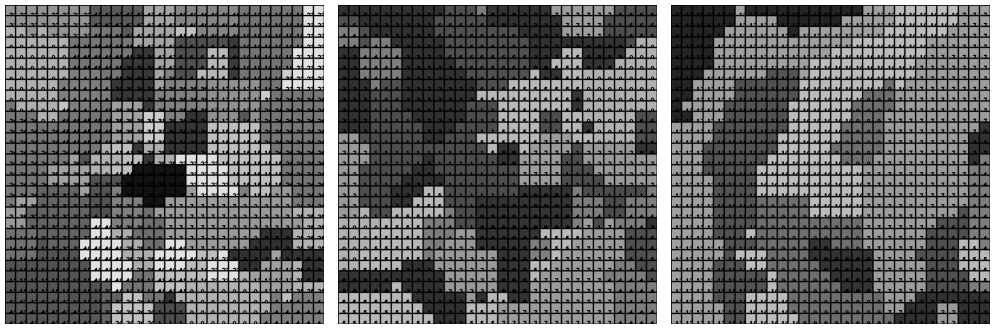
表 1　実験に用いた ART-2 と CPN の各パラメータの値
Table 1　Settings values of parameters on ART-2 and CPN using experiment.

		SIFT	Gist	Our Method
ART-2	θ	0.1	0.1	0.1
	ρ	0.80	0.80	0.95
CPN	α	0.5	0.5	0.5
	β	0.5	0.5	0.5
	I	10,000	10,000	10,000

本実験では，1) SIFT のみによる BoVW，2) Gist のみによる BoVW，3) SIFT と Gist を用いた提案手法による BoVW の 3 種類の特徴表現方法による分類精度を比較した．本実験で使用したこれら 3 種類の特徴表現方法における ART-2 と CPN のパラメータを表 1 に示す．CPN の学習率係数の初期値の α と β，及び学習回数の I は共通の値を用いた．ART-2 のノイズ除去に関連するパラメータ θ は共通の値としたが，カテゴリの粒度を支配する ρ は，予備実験の結果から決定した．

4.2 カテゴリ生成結果

図 3 に各カテゴリの画像と生成された特徴量を示す．各特徴表現における ART-2 でのラベル形成結果を図 4 に示す．横軸にはフレーム数，縦軸には ART のラベルを示す．また，図の上には各局所シーンの GT (Ground Truth) を示す．ART-2 で形成されたラベル数は，SIFT が 50 ラベル，Gist が 9 ラベル，提案手法が 14 ラベルであった．提案手法による結果は，シーンの意味カテゴリに沿って，段階的にラベルが生成されている．SIFT のみによるラベル形成結果では，全体的にラベルが冗長となっており，特に，BO と KT に対応するラベルが重複



(a) SIFT　　(b) Gist　　(c) Our Method

図 5　生成されたカテゴリマップの比較結果.
Fig. 5　Comparision of results using SIFT, Gist, and our method for creating category maps.

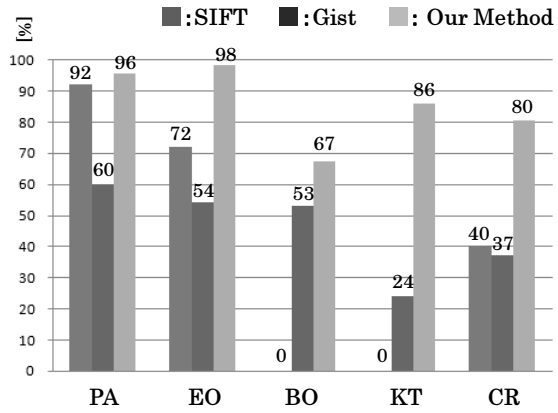


図 6　分類率の比較結果.
Fig. 6　Comparison results of recognition accurary.

している．Gist のみによる結果は，提案手法よりも生成されたラベル数が少ないにもかかわらず，重複するラベルが多数発生している．これは，Gist のみでは大局的な記述となるため，屋内シーンの特徴を捉えきれていないといえる．

各特徴表現におけるカテゴリマップを図 5 に示す．カテゴリマップのサイズは，20×20 ユニットとした．SIFT のみによる結果では，生成されたラベル数が多いことから，ラベルが混在したカテゴリマップとなっている．Gist のみによる結果では，ラベル数が少なかったものの，同一ラベルが与えられているシーン画像が，複数の領域に分布しており，混在した分布となっている．一方，提案手法による結果では，意味カテゴリに基づく局所的なシーンに沿ってカテゴリマップが形成されている．また，独立したラベルや混在するラベルは発生していない．

4.3　分 類 精 度

分類性能を定量的に評価するために，以下に示す分類率をカテゴリごとに算出した．

$$(\text{分類率}) = \frac{(\text{正解枚数})}{(\text{カテゴリの総枚数})} \times 100. \qquad (9)$$

分類率の比較結果を図 6 に示す．いずれの結果においても，提案手法が他の特徴表現法よりも分類率が高くなっている．特に，PA と EO はそれぞれ 96%(113 frame/118 frame) と 98%(120 frame/122 frame) となり，高い分類結果が得られている．BO と CR は，分類率がそれぞれ 67%(66 frame/98 frame) と 80%(307 frame/385 frame) に留まっているものの，他手法と比べ高い分類率が得られている．各特徴表現における平均分類率は，SIFT が 41%，Gist が 46%，提案手法が 85%となり，提案手法は SIFT のみによる結果と比較して 44%，Gist のみによる結果と比較して 38%の優位性が得られた．

5.　移動ロボットを用いた評価実験

標準的なデータセットに加えて，本手法の適応範囲を拡大するために，独自に試作開発した移動ロボットを用いて取得した時系列画像から，シーン分類を行った．また，KTH-IDOL データセットでは部屋や廊下の単位で GT が割り振られているが，部屋の中でも例えばデスクスペースや水廻り，廊下においてもエレベータ付近や各部屋に通じるドア付近などの局所的な空間が存在する．そこで，本実験ではより粒度の細かいシーン分類を対象とする．

5.1　実 験 環 境

我々は，視覚に基づく自律移動とシーン分類のプラットフォームとして，図 7(a) に示す視覚機能を搭載した移動ロボットを試作開発した．設計コンセプトとしては，人間と同じ目線で環境の理解が実現できるように，ロボットの全高を成人男性の平均身長と同程度の 170cm にした．本ロボットには，ステレオカメラが 1 セットと，図 7(b) に示す全方位センサが 1 台搭載されている．全方位センサは，ミラー径が 30ϕ，視野角が上方 15 度，下方 55 度，カメラ素子には 1/3inch インターレース CCD が搭載されており，解像度が 640 × 480pixel の画像を 30fps で取得できる．ロボットの駆動系は，2 軸独立駆動であり，安定性を考慮して補助輪は 2 輪とした．重量は 10.0kg，移動速度は 2.4km/h である．

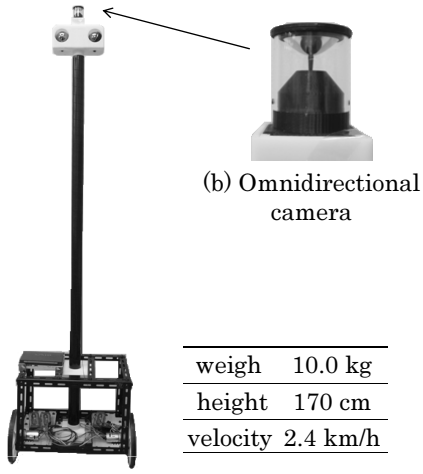


図 7　試作開発した移動ロボットと全方位センサ.
Fig. 7　Prototype of our developed mobile robot and omni-

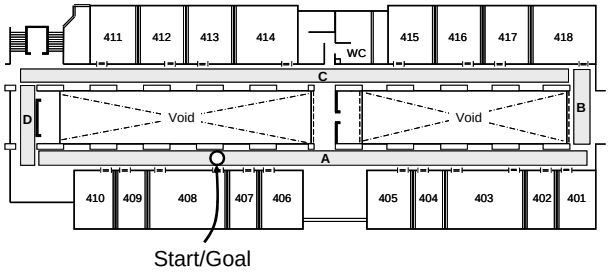


図 8　実験環境（区間 A〜D）.
Fig. 8　Experimental Enviroment (Zone A–D).

実験環境を図 8 に示す．実験は，本学の建物の 4 階廊下で実施した．廊下の幅は約 1.8m，長手方向（区間 A 及び C）の全長は約 13m，短手方向（区間 B 及び D）の全長は約 74m となってる．廊下の内側は，2 階から最上階の 6 階までの吹き抜けとなっている．シーン分類の目安として，長手方向と短手方向の廊下を区間 A から区間 D までを GT とした．区間 A 及び C は，廊下の外側に沿って，研究室と教員室が並んでいる．区間 B 及び D は，エレベータや階段，連絡通路となっている．なお，区間 A 及び C の中間地点には渡り廊下が架かっているが，本実験では移動対象外とした．

5.2 分類結果

実験で用いたパラメータは，ART-2 の θ が 0.1，ρ が 0.980，CPN の α および β の初期値が 0.5，学習回数が 10,000 回とした．画像撮影時のサンプリングレートを 1fps として，ロボットが実験環境を時計回りに 1 周させることにより，視野画像列を取得した．

ART-2 によるラベルの生成結果を図 9 に示す．横軸にはフレーム数，縦軸には ART のラベル，上部に GT を示す．本データセットに対して，ART-2 は 20 ラベルを生成した．全てのラベルには重複がなく，ロボットの移動に伴い段階的にラベルが生成されている．なお，GT

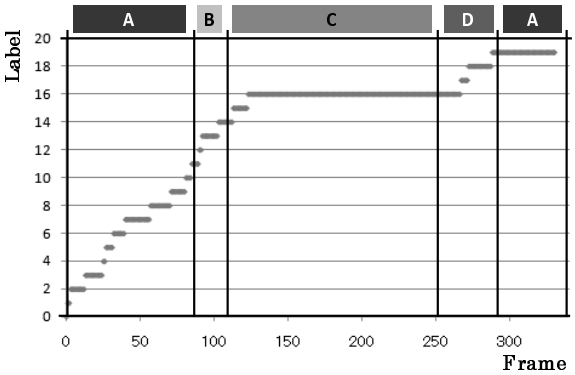


図 9　ART-2 によるラベル形成結果.
Fig. 9　Labeling result using ART-2.

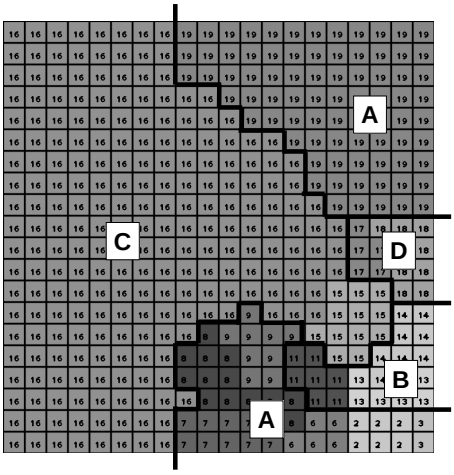


図 10　カテゴリマップの形成結果．各区間に対応するカテゴリの境界を太線で示す.
Fig. 10　Formation result of category map. Thick lines show category boundaries in each zone.

の境界付近でラベルの切り替えが前後するのは，ロボットがコーナを曲がる際において，シーンの見え方が類似する区間であると考えられる．

本ラベルを教師信号として生成したカテゴリマップを図 10 に示す．カテゴリマップのサイズは，20×20 ユニットとした．カテゴリマップ上には，GT と対応するように境界線を追加している．カテゴリマップの生成結果より，独立したラベルのユニットは存在せず，類似したシーンを隣接したユニットに写像できている．区間 A に関しては，移動の開始と終了地点を挟んで異なる領域に写像されている．また，区間 A ではラベル数が冗長になったものの，カテゴリマップ上では対応付く画像枚数の少ないラベルが淘汰されている．多数の画像が対応付いた区間 C のラベル 16 に関しては，カテゴリマップ上において対応付くユニット数が多くなっている．以上の結果から，本手法では，廊下の中でもシーンの特徴に応じて局所的な分類が実現できている．

6. ま と め

本論文では，自律移動ロボットにおける屋内シーンの

意味的な認識を目的として，SIFT と Gist の 2 種類の特徴量をコンテクストとして利用した教師なしシーン分類法を提案した．本手法では，ART-2 のラベル生成によって時間軸を考慮し，CPN のカテゴリマップによってそれらを隣接するユニットに写像することで，カテゴリ内の空間的な関係性を表現した．KTH-IDOL データセットを用いたシーン分類実験では，SIFT と Gist の 2 次元ヒストグラムによる本手法が，SIFT と Gist を単独で用いた場合よりも良好な結果が得られた．また，試作開発した移動ロボットを用いて取得した時系列画像を用いた実験では，シーンの見え方に応じて局所的なカテゴリを形成することができた．よって，提案手法によるコンテクストに基づく特徴量の表現，及び ART-2 の追加学習と CPN の自己写像学習を用いたカテゴリ表現は，ロボットビジョンのための屋内シーンの分類に有効であるといえる．

今後の課題は，CPN のカテゴリマップから分類されたカテゴリの境界を自動抽出し，適切なカテゴリ数を決定することが挙げられる．また，シーンがダイナミックに変化する環境や，多数の人間が行き交う環境等への適用範囲の拡大を目指したい．

文　献

[1] G. Dissanayake, P. Newman, S. Clark, H.F. Durrant-Whyte, and M. Csorba, “An experimental and theoretical investigation into simultaneous localisation and map building (SLAM),” Lecture Notes in Control and Information Sciences: Experimental Robotics VI, Springer, 2000.

[2] J. Wu and J.M. Rehg, “CENTRIST: A Visual Descriptor for Scene Categorization,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2010.

[3] J. Wu, H.I. Christensen, and J.M. Rehg, “Visual Place Categorization: Problem, Dataset, and Algorithm,” *Proc. IEEE/RSJ Int’l Conf. Intelligent Robots and Systems*, 2009.

[4] C. Siagian and L. Itti, “Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attentinon,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 300–312, Feb. 2007.

[5] A. Quattoni and A. Torralba, “Recognizing Indoor Scenes,” *Proc Computer Vision and Pattern Recognition*, 2009.

[6] M. Tsukada, Y. Utsumi, H. Madokoro, and K. Sato, “Unsupervised Feature Selection and Category Classification for a Vision-Based Mobile Robot,” *IEICE Trans. Inf. & Sys.*, vol. E94-D, no. 1, pp. 127–136, Jan. 2011.

[7] S. Thrun, “Finding Landmarks for Mobile Robot Navigation,” *Proc. IEEE Int’l Conf. Robotics and Automation*, pp. 958–963, 1998.

[8] S. Maeyama, A. Ohya, and S. Yuta, “Long Distance Outdoor Navigation of an Autonomous Mobile Robot by Playback of Perceived Route Map,” *Proc. Fifth Int’l Symp. Experimental Robotics*, pp. 185–194, 1997.

[9] A. Kawewong, S. Tangruamsub, and O. Hasegawa, “Position-invariant Robust Features for Long-term Recognition of Dynamic Outdoor Scenes,” *IEICE Trans. Information and Systems*, vol.E93-D, no.9, pp.2587–2601, 2010.

[10] 森岡博史, 李想揆, 長谷川修, “人の多い混雑な環境下での SLAM による移動ロボットのナビゲーション”, 第 28 回日本ロボット学会学術講演会予稿集, 2010.

[11] H. Katsura, J. Miura, M. Hild, and Y. Shirai, “A View-Based Outdoor Navigation Using Object Recognition Robust to Changes of Weather and Seasons,” *Proc. IEEE/RSJ Int’l Conf. Intelligent Robot and Systems*, pp. 2974–2979, Oct. 2003.

[12] Y. Matsumoto, M. Inaba, and H. Inoue, “View-Based Approach to Robot Navigation,” *Proc. Int’l Conf. Intelligent Robots and Systems*, pp. 1702–1708, 2000.

[13] J. Shi and J. Malik, “Normalized Cut and image Segmentation”, IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 8, pp. 8881–905, 2000.

[14] A. Oliva and A. Torralba, “Building the gist of a scene: the role of global image features in recognition,” *Visual Perception, Progress in Brain Research*, vol. 155, 2006.

[15] A. Torralba, K.P. Murphy, W.T. Freeman, and M.A. Rubin, “Context-Based Vision System for Place and Object Recognition,” *Proc. IEEE Int’l Conf. Computer Vision*, pp. 1023–1029, Oct. 2003.

[16] 柳井啓司, “一般物体認識の現状と今後,” 情報処理学会研究報告 CVIM, pp. 121–134, Sep. 2006.

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

[18] A. Torralba, “How many pixels make an image?”, *Visual Neuroscience*, vol. 26, pp. 123–131, 2009.

[19] 竹内龍人, “シーンの認識と探索にかかわる視覚のメカニズム,” 映像情報メディア学会技術報告, vol. 33, no. 24, pp. 7–14, June 2009.

[20] 永橋知行, 伊原有仁, 藤吉弘亘, “画像分類における Bag-of-features による識別に有効な特徴量の傾向,” 情報処理学会研究報告, vol. 2009-DVIM-169, no. 3, pp. 13–20, Nov. 2009.

[21] J. McQueen, “Some Methods for Classification and Analysis of Multivariate Observations,” *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[22] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual Categorization With Bag of Keypoints,” *ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

[23] T. Kohonen, Self-Organizing Maps, Springer Series in Information Sciences, 1995.

[24] 寺島 幹彦, 白谷 文行, 山本 公明, “自己組織化特徴マップ上のデータ密度ヒストグラムを用いた教師なしクラスタ分類法,” 電子情報通信学会論文誌, D-II vol. J79-D-II, no. 7, pp. 1280–1290, Jul. 1996.

[25] G.A. Carpenter and S. Grossberg, “ART 2: Stable Self-Organization of Pattern Recognition Codes for Analog Input Patterns,” *Applied Optics*, vol. 26, pp. 4919–4930, 1987.

[26] R. Hetch-Nielsen, “Counterpropagation networks,” *Proc. of IEEE First Int’l. Conf. on Neural Networks*, 1987.

[27] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt, “The KTHIDOL2 database,” *Technical Report CVAP304, Kungliga Tekniska Hoegskolan, CVAP/CAS*, Oct. 2006.

[28] A. Pronobis, L. Xing, and B. Caputo, “Overview of the CLEF 2009 Robot Vision Track,” *Proc. 10th international conference on Cross-language evaluation forum: multimedia experiments*, 2010.