計測自動制御学会東北支部 第 317 回研究集会 (2018.7.19)
資料番号 317-3

視覚的顕著性とパーツ特徴量に基づく屋内シーンの意味的認識

Semantic Indoor Scenes Recognition Based on Visual Saliency and Part-Based Features

○徳原匡亮 *，間所洋和 *，佐藤和人 *

○ Kyosuke Tokuhara*, Hirokazu Madokoro*, Kazuhito Sato*

*秋田県立大学大学院 システム科学技術研究科

*Graduate School of Systems Science and Technology, Akita Prefectural University

キーワード： 特徴記述 (Feature Description)，SM (Saliency Map)，SOM (Self-Organizing Map)，CPN (Counter Propagation Network)

連絡先： 〒 015–0055 秋田県由利本荘市土谷字海老ノ口 84–4 秋田県立大学大学院 システム科学技術研究科 機械知能システム学専攻 脳情報工学研究室

徳原匡亮，Tel.: 0184-27-2081，E-mail: M19A014@akita-pu.ac.jp

1. はじめに

シーンの意味的認識はロボットビジョンの難題の 1 つである．特に，ロボットが日常生活の中で人間と共生するためには，視覚情報を用いたシーンの意味的認識が重要となる．近年，計算能力の急速な進歩により，移動ロボットに取り付けられたカメラを視覚センサとして用いることで，高分解能の時系列画像をリアルタイム処理することが可能となった．人間は視覚情報から 10^9 bit/s で膨大な量の情報の中から目立つ情報を選択する凝視メカニズムを持っていることが分かっている [1]．注意を引く視覚的特徴を生理学的知見に基づいてボトムアップに抽出したものを顕著性と呼ぶ [2]．そして，概念モデルとして初めて SMs（Saliency Maps）が提案された [3]．その後，Itti ら [4] が，画像をコンピュータ上で処理することができる計算モデルとして SM を実装した．顕著性モデルを用いたアプリケーションとして，コンピュータビジョン，産業機械，ロボットビジョン，自動車システム，認知システムなどが提案されている [5]．

顕著性に基づく物体認識の先行研究として，Shokoufandeha ら [6] は、ウェーブレット変換を用いて複数のスケールで物体の顕著領域を確保する SMG（Saliency Map Graph）を提案している．Walther ら [7] は，SM に基づいて，自然のシーンの中から物体を検出する生物学的にもっともらしいモデルを提案している．彼らは，物体上の特徴記述に SIFT（Scale-Invariant Feature Transform）[8] を使用した．さらに，複雑なシーンの画像の場合，物体に特徴点の集合を一致させることで複雑さが明らかに減少することを証明しているが，領域選択アルゴリズムが物体を見つけることが可能であるという保証はない．これらの方法は，単純にボトムアップに駆動しているため，意味的な物体認識としての概念は存在しない．

顕著性に基づく特徴は，屋内および屋外のシーン認識において使用されている．屋外のシーン認識には，GPS とセンサを同時に用いることで，正確な位置を特定する手法が提案されている [9]．Quattoni ら [10] は，屋内のシーン認識が難しい理由として，一部の屋内シーンが全体的な空間特性によって特徴付けられる一方，多くの屋内シーンが物体により特徴付けられるということを示した．Fornoni ら [11] は，屋内シーン

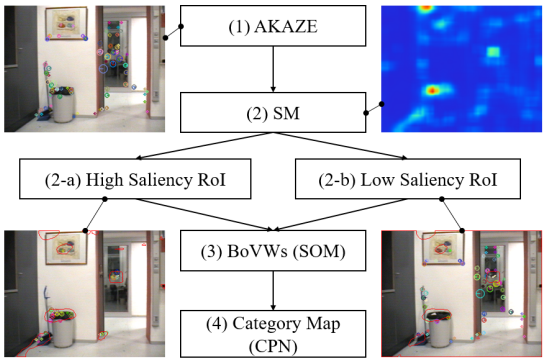

Fig. 1　システムの全体構成

の意味的認識のための顕著性に基づく画像分類手法を提案しており，彼らは特徴記述に SIFT を使用し，分類のために SVMs（support vector machines）[12] を使用している．Botterill ら [13] は，移動ロボットの位置推定のために同一のシーンをリアルタイム検出可能な手法を提案している．彼らは，SURF（speeded-up robust features）[14] に基づく高速記述子と組み合わせた低次元のコードブックを使用している．この手法は，高速に物体抽出と認識を行うだけでなく，動画のフレームごとにおけるリアルタイムでの移動ロボットの位置推定を実現している．

顕著性に基づく特徴は一般物体から多く得られるため，一般物体認識とシーンの意味的認識は密接に関連していると言える [15]．全体的なシーンの理解のために，Yao ら [16] は，領域，位置，クラス，および空間的な物体の関係を物体検出器としてピクセル単位で推定する物体分類手法を提案している．彼らは，ピクセル単位で分割された物体を用いてシーンの分類精度を上げている．しかし，現在，一般的に使われるカメラを用いて得られた高解像度画像に適応する場合，ピクセル単位の分類は計算負荷が大きい．

本論文では，教師あり機械学習と教師なし SMs に基づく屋内シーン認識手法を提案する．評価実験を 2 種類の時系列画像で構成されているベンチマークデータセットを用いて行った．一般的な物体からなる顕著性領域に基づき，パーツ特徴量を使用して特徴の組み合わせの基本特性を実証した．

2.　提案手法

2.1　提案手法の全体構成

本研究で提案するシステムの全体構成を図 1 に示す．提案手法では特徴記述に AKAZE（Accelerated KAZE）[17] を用いる．さらに，特徴記述範囲の選択に SMs を用い，BoVWs（bags of visual words）[18] の作成に SOMs（self-organizing maps）[19] を用いる．さらに，CPNs（counter propagation networks）[20] によるカテゴリマップに基づくシーン認識を行う．SOMs と CPNs は我々の先行研究に基づいて組み込んだ [21]．

原画像を I_{org}，AKAZE を記述した画像を I_{aka}，SM によるマスク画像を S とすると，高顕著画像 I_{high}，低顕著画像 I_{low} は次式で定義される．

$$I_{high} = I_{org} \wedge I_{aka} \wedge S, \tag{1}$$

$$I_{low} = I_{org} \wedge I_{aka} \wedge \overline{S}. \tag{2}$$

2.2　特徴量の記述

Gist[22] は，特に山や湖，雲などの一般的な自然環境のシーンを対象にした特徴量である．これは，シーンの構造を特徴量として表現するため，オブジェクトなどの細かな構造の多い屋内シーンには向いていない．部分ベースの特徴量として，SIFT[8] が一般的な物体認識の研究に広く用いられている．その後，非線形スケールスペースを使用した KAZE[23] が提案され，SIFT よりも高い精度を記録した．本研究で用いる AKAZE[17] は，KAZE を高速処理化したもので，特徴記述性能が高いだけでなく，リアルタイム処理を行う際の計算負荷が少ないことが分かっている．

2.3　顕著性マップ

輝度成分 I は次式で定義される．

$$I = \frac{1}{3}(r + g + b). \tag{3}$$

ここで，r，g，b はそれぞれ赤成分，緑成分，青成分を表している．色相成分では，RGB 成分と黄色成分 Y を抽出する．Y 成分は，

$$Y = \frac{1}{2}(r + g) - \frac{1}{2}|r - g| - b, \tag{4}$$

で算出される．方向成分はガボールフィルタを用いて θ=0，45，90，135 deg の 4 方向に対して求める．ガボールフィルタ G は，正弦波とガウス関数の積として定義される．$G(x,y)$ は次式で定義される．

$$G(x,y) = \exp\{-\frac{1}{2}(\frac{R_x^2}{\sigma_x^2} + \frac{R_y^2}{\sigma_y^2})\} \exp(i\frac{2\pi R_x}{\lambda}). \quad (5)$$

$$\begin{bmatrix} R_x \\ R_y \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (6)$$

ここで，λ，θ，σ_x，σ_y はそれぞれ，波長の余弦成分，ガボール関数の方向成分，水平軸方向のフィルタの大きさ，垂直軸方向のフィルタの大きさを表す．画像上の傾きを持った線分に対して G を作用させると，線分の垂直な方向に積分値が最大となる．そのため，傾きとその周波数成分を検出する．フィルタのサイズを $M \times N$ pixel とすると，フィルタ出力 $Z(x,y)$ とそのサンプル点 $P(x,y)$ は，

$$Z(x,y) = \sum_{i=0}^{N} \sum_{j=0}^{M} G(x+i, y+j) P(x+i, y+j), \quad (7)$$

で表される．Z は複素項を含み，$\sqrt{Rm^2 + Im^2}$ となる．

ガウシアンピラミッドにより得られたスケールの異なるペアから差分を求め，重ね合わせることで注意を引く位置を特定する．この操作を Center-Surround と呼び，演算子 $\ominus$ で表す．差分をとる際は，スケールの大きい側に小さい側を拡大する．スケールを $c, s (c < s)$ と定義すると，大きい側は c=2,3,4, 小さい側は $s = \{c + \delta | \delta \in 3, 4\}$ となる．輝度に対する差分 $I(c,s)$ は次式で定義される．

$$I(c,s) = |I(c) \ominus I(s)|. \quad (8)$$

赤と緑の色相の差を $RG(c,s)$，青と黄の色相の差を $BY(c,s)$ とし，それぞれ次式で定義する．

$$RG(c,s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|, \quad (9)$$

$$BY(c,s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|, \quad (10)$$

方向成分は，各方向毎に差分を計算する．

$$O(c,s,\theta) = |O(c,\theta) \ominus O(s,\theta)|. \quad (11)$$

正規化後，輝度成分，色相成分，方向成分の FMs（feature maps）の重ね合わせ処理を行う．このとき，小さいマップは各ピクセルの加算により拡大される．正規化関数を N とすると，輝度成分 $\overline{I}$，色相成分 $\overline{C}$，方向成分 $\overline{O}$ の線形和は次式で表される．

$$\overline{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=4}^{c+3} N(I(c,s)), \quad (12)$$

$$\overline{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=4}^{c+3} \left(N(RG(c,s)) + N(BY(c,s))\right), \quad (13)$$

$$\overline{O} = \sum_{\theta} N \bigoplus_{c=2}^{4} \bigoplus_{s=4}^{c+3} N\left(O(\theta,c,s)\right). \quad (14)$$

得られたマップを CMs（conspicuity maps）と呼ぶ．FMs の各成分と線形和を正規化すると，SMs は次式で表される．

$$S = \frac{1}{3}(N(\overline{I}) + N(\overline{C}) + N(\overline{O})). \quad (15)$$

最後に，WTA（winner-take-all）により高顕著領域を取得する [3]．

2.4 BoVW

入力データを $x_i(t)$，入力層ユニット i からマッピング層ユニット j への時刻 t における結合荷重を $w_{i,j}(t)$ で表す．ここで，I は入力層，J はマッピング層の総数を表す．学習前に，$w_{i,j}(t)$ はランダムに初期化される．$x_i(t)$ と $w_{i,j}(t)$ のユークリッド距離が最小となるユニットは勝者ユニット c_j として表される．

$$c_j = \underset{1 \leqq j \leqq J}{\operatorname{argmin}} \sqrt{\sum_{i=1}^{I} (x_i(t) - w_{i,j}(t))^2}. \quad (16)$$

勝者ユニット c_j を中心とする近傍領域 $N(t)$ を設定する．

$$N(t) = \lfloor \mu \cdot J \cdot \left(1 - \frac{t}{O}\right) + 0.5 \rfloor. \quad (17)$$

ここで，μ $(0 < \mu < 1.0)$ は $N(t)$ 内で初期サイズであり，O は最大学習回数を表す．また，四捨五入のための係数として 0.5 が付加されている．次に，$N(t)$ の $w_{i,j}(t)$ は更新される．

$$w_{i,j}(t+1) = w_{i,j}(t) + \alpha(t)(x_i(t) - w_{i,j}(t)). \quad (18)$$

ここで，$\alpha(t)$ は学習の経過とともに減少する学習率係数である．

2.5 CPN

$u^i_{n,m}(t)$ 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．ここで，$v^j_{n,m}(t)$ 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が最小となる勝者ユニット $c_{n,m}$ とすると，

$$c_{n,m} = \underset{1 \leqq n \leqq N, 1 \leqq m \leqq M}{\operatorname{argmin}} \sqrt{\sum_{i=1}^{I} (x_i(t) - u^i_{n,m}(t))^2}, \tag{19}$$

と定義される．ここで，式 (17) より，N は勝者ユニット $c_{n,m}$ の近傍領域である．さらに，N の内部の $u^i_{n,m}(t)$ と $v^j_{n,m}(t)$ は Kohonen の学習アルゴリズム，Grossberg の学習アルゴリズムでそれぞれ更新される．

$$u^i_{n,m}(t+1) = u^i_{n,m}(t) + \alpha(t)(x_i(t) - u^i_{n,m}(t)), \tag{20}$$

$$v^j_{n,m}(t+1) = v^j_{n,m}(t) + \beta(t)(t_j(t) - v^j_{n,m}(t)). \tag{21}$$

ここで，$t_j(t)$ は Grossberg 層に提示される教師信号であり，$\alpha(t)$ と $\beta(t)$ は学習の経過とともに減少する学習率係数である．CPN の学習は事前に設定した学習回数 O' だけ繰り返す．

3. KTH-IDOL2 を用いた評価実験

3.1 データセット

KTH-IDOL2 データセット [24] 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公開されている画像データセットである．このデータセットは，解像度 320×240 pixel の時系列画像で構成されており，2 種類の移動ロボットを用いて撮影されている．さらに，MIT Places2[25] のような最新のデータセットとは異なり，位置情報が GT として含まれている．今回は，98 cm のロボットが取得したデータセットを使用する．また，線形補間により 30 fps から 10 fps にダウンサンプリングした．対象とするシーンは，PA（Printer Area），EO（One-person office），BO（Two-persons office），KT（Kitchen），CR（Corridor）の 5 カテゴリから構成されている．本研究では，夜間のデータセットを用いる．

Table 1　実験に使用した SOMs と CPNs の各パラメータの値

ネットワーク	パラメータ	値
SOMs	I	61
	J	256
	O	1,000,000 [回]
CPNs	α	0.80
	β	0.50
	O'	100,000 [回]
	$N \times M$	50×50 [units]

3.2 評価基準とパラメータ

表 1 に，本研究で用いた SOMs と CPNs のパラメータを示す．これらのパラメータは，KTH-IDOL2 データセットを用いた先行研究 [21] と予備実験の結果から決定した．評価基準として，認識率 R_{acc} を次式のように定義する．

$$R_{acc} = \frac{S_{test}}{N_{test}} \times 100, \tag{22}$$

ここで，N_{test} と S_{test} はそれぞれテスト画像の総枚数，正解画像枚数を表す．本研究では，機械学習のアプローチとして，LOOCV（Leave-One-Out Cross-Validation）[26] により評価を行った．

3.3 特徴記述とカテゴリマップ

KT における AKAZE と SMs の画像出力結果を図 2 に示す．I_{aka} の出力画像より，多くの特徴点がゴミ箱，箒，絵画，ドアのフレーム，後方部のドア周辺に割り当てられていることが分かる．SM の出力画像からは，ゴミ箱，箒，絵画の中央及び角の顕著度が高くなっていることが分かる．さらに，後方部のドアに反射した像も顕著度が高くなっていることから，夜間の屋内環境では，反射により出現した像による影響も見られる．I_{high} に着目すると，顕著度が高い領域が赤い線で囲まれており，その領域内に特徴記述が行われていることが分かる．一方，I_{low} は顕著度が低い領域に特徴記述が行われている．

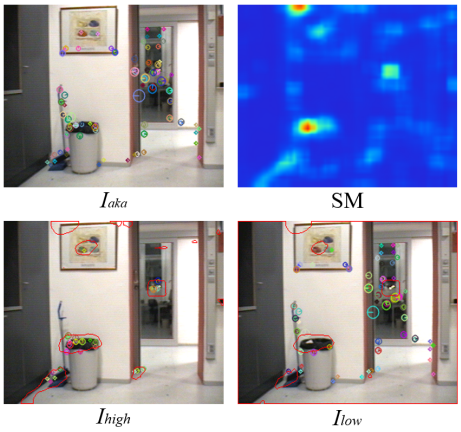

Fig. 2　KT の特徴抽出結果

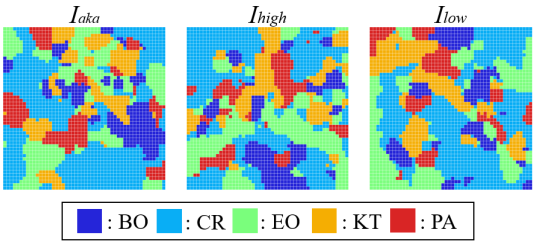

Fig. 3　KTH-IDOL2 のカテゴリマップの出力結果

3 種類の特徴パターンごとのカテゴリマップの出力結果を図 3 示す．ユニットの色はそれぞれのシーンカテゴリのラベルに対応している．出力結果より，シーンのカテゴリがいくつかの独立したクラスタに分かれており，大小さまざまなクラスタが点在していることが分かる．さらに，単一のユニットで構成される小さなクラスタもいくつか存在している．

3.4　認識率の比較結果

カテゴリごとの認識率の比較結果を図 4 に示す．結果より，どのカテゴリにおいても，I_{aka} の認識率が I_{high} と I_{low} の認識率よりも上回っていることが分かる．I_{high} と I_{low} の認識率は類似しているように見えるが，それぞれのカテゴリごとに異なっている．CR は比較的高い認識率が得られたが，BO と EO は他のシーンカテゴリよりも低い．

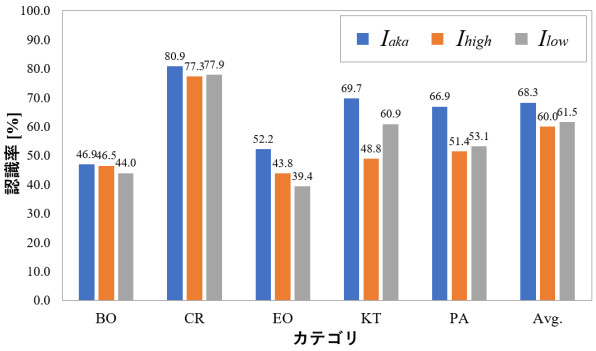

Fig. 4　KTH-IDOL2 における各カテゴリごとの認識率の比較結果

Table 2　KTH-IDOL2 の混同対照表

I_{aka}	BO	CR	EO	KT	PA
BO	**61**	19	19	23	8
CR	22	**338**	23	21	14
EO	24	23	**82**	18	11
KT	16	17	11	**107**	3
PA	9	16	10	9	**88**

3.5　考察

誤認識を分析するために混同対照表を用いる．KTH-IDOL2 の混同対照表を表 2 に示す．この表では，横方向を基準として，識別に成功した画像枚数が太字で対角線上に表示される．それ以外のマス目には，誤認識となった画像枚数が表示され，縦方向のカテゴリ名を参照することで誤認識したカテゴリを特定することができる．このデータセットは，スライド式のガラス製ドアにより物理的に部屋が分けられている．例外として，PA は CR と直接繋がっているが，部屋の機能が異なるため別の部屋として定義している [24]．そのため，CR の誤認識が多く発生すると考えられる．図 4 より，各カテゴリの中でも BO と EO の認識率が低いことが分かる．これは，BO と EO がどちらもオフィスであるため，特徴が似ていることが考えられる．図 3 のカテゴリマップから，多くのユニットが CR にラベル付けされていることが分かる．そのため，CR の誤認識数は他のシーンカテゴリに比べて多いと分かる．

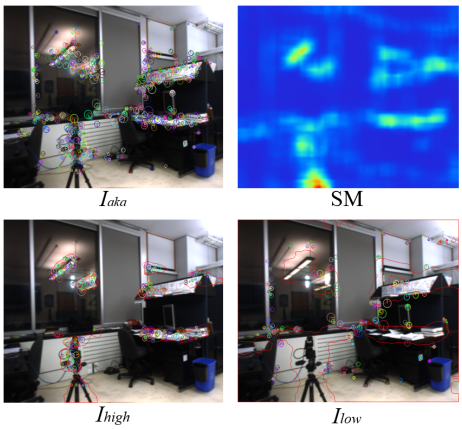

Fig. 5　York 大学 (Lab2) の特徴記述

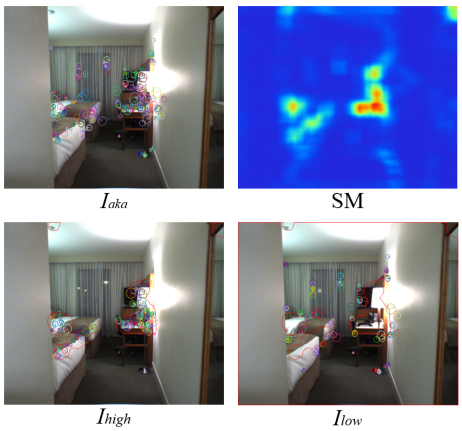

Fig. 6　Coast Capri ホテル (Bed Room) の特徴記述

4. Place Recognition Dataset を用いた評価実験

4.1 データセット

本研究では，追加データセットとして Place Recognition Dataset[27] を使用する．このデータセットは，解像度 640×480 pixel の各シーンカテゴリごとの時系列画像で構成されており，2 種類の移動ロボットを用いて撮影されている．また，KTH-IDOL2 データセットと同様に位置情報が GT として含まれている．フレームレートは 3 fps である．今回は，高さ 117 cm のロボットが取得したデータセットを使用する．対象とするシーンは，York 大学内の Arena，AshRoom，Corridor，Lab2，LivRoom，Lounge，PlantRoom，ProfRoom，SemRoom，WashRoom，WorkPlace の 11 カテゴリ及び，Coast Capri ホテル内の Corridor（CR），Hallway（HW），Dining Room（DR），Bed Room（BR），Conference Room（CN），Lobby（LB）の 6 カテゴリである．

4.2 特徴記述とカテゴリマップ

Lab2 の AKAZE と SMs の画像出力結果を図 5 に，Bed Room の画像出力結果を図 6 に示す．図 5 の I_{aka} の出力画像より，特徴点はカメラと三脚，窓のフレーム，本や紙，ガラスに反射した蛍光灯に割り当てられていることが分かる．AKAZE が記述された部分と，SM で顕著度が高い部分は類似しているように見える．しかし，窓のフレームと椅子周辺は顕著度が低いため，I_{high} の出力結果には AKAZE は記述されない．図 6 の I_{aka} の出力結果では，特徴点はランプ，ベッド，カーテン，机に割り当てられていることが分かる．また，I_{high} の画像から，SM によりカーテンを除く箇所に特徴点が制限されていることが分かる．

York 大学の 3 種類の特徴パターンごとのカテゴリマップの出力結果を図 7 に，Coast Capri ホテルの結果を図 8 に示す．ここで，右のカラーバーは赤から順に，4.1 節に記述したシーンカテゴリの順に対応している．図 7 の結果より，大小さまざまな大きさのクラスタを形成しており，シーンの多様性が見受けられる．図 8 の結果では，ある程度まとまったクラスタを形成しているが，同カテゴリ内で複数のクラスタが存在している．

4.3 認識率の比較結果と考察

York 大学の認識率の比較結果を図 9 に，Coast Capri ホテルの認識率の比較結果を図 10 に示す．また，Coast Capri ホテルの混同対照表を表 3 に示す．図 9 より，すべてのテゴリにおいて，I_{aka} の認識率が I_{high} と I_{low} の認識率よりも上回っていることが分かる．同様に，図 10 の結果も，I_{aka} の認識率が I_{high} と I_{low} の認識率よりも上回っていた．また，平均に注目すると，90%程度の安定した認識率となった．これは，Place Recognition Dataset が KTH-IDOL2 とは異なり，各シーンカテゴリごとの時系列画像で構成されているため，シーンカテゴリ間をロボットが移動していない．そのため，他のシーンカテゴリの

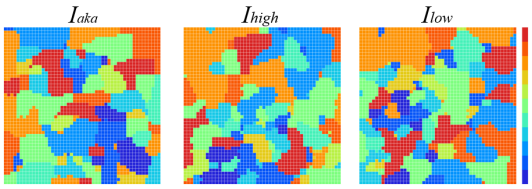

Fig. 7　York 大学のカテゴリマップの出力結果

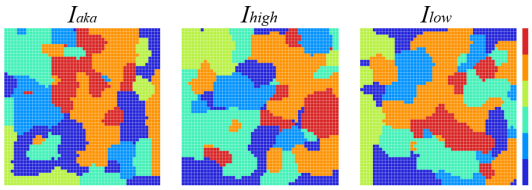

Fig. 8　Coast Capri ホテルのカテゴリマップの出力結果

特徴が取得されないことにより誤認識が比較的少なく，90%程度の安定した認識率が得られたと考えられる．さらに，図 10 の Corridor の認識率が他のシーンカテゴリと比べて低いことが分かる．表 3 より，Corridor の誤認識は，Conference Room，Lobby に多く発生していることが分かる．これらのカテゴリは，廊下のように屋内シーンを特徴付けるための物体が少なく，奥行きの多いシーンが一部見られ，そのシーンで誤認識が発生したと考える．

5.　結論

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認識を目的として，AKAZE，SM，SOMs，CPNs の各手法を用いて教師あり機械学習と教師なし SMs に基づく屋内シーン認識を行った．評価実験として，KTH-IDOL2 と Place Recognition Dataset の 2 種類のデータセットを用いて特徴の組み合わせの基本特性を実証した．LOOCV を用いて得られた認識率の結果から，高顕著領域または低顕著領域の特徴点を用いるよりも，画像全体の特徴点を用いた認識率の方が高いことを明らかにした．

今後の課題は，顕著性に基づいて選択した前景領域と背景領域のパーツ特徴量を使用し，コンテクストに基づくシーン認識を行うことが挙げられる．また，学習方法の変更として，深層学習の導入を検討したい．

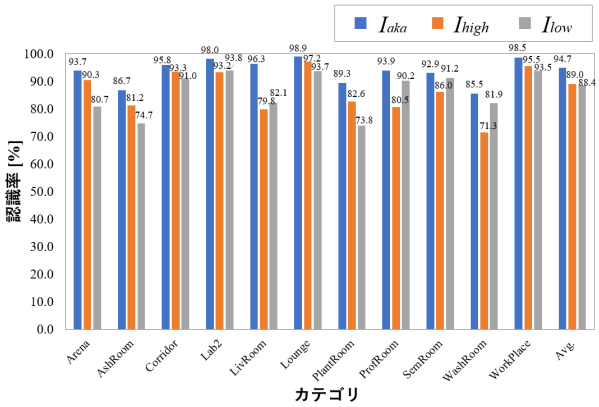

Fig. 9　York 大学における各カテゴリごとの認識率の比較結果

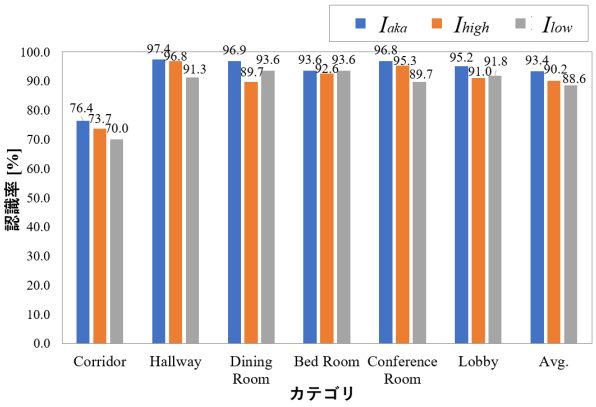

Fig. 10　Coast Capri ホテルにおける各カテゴリごとの認識率の比較結果

Table 3　Coast Capri ホテルの混同対照表

I_{aka}	CR	HW	DR	BR	CN	LB
CR	**275**	12	11	3	26	32
HW	0	**336**	2	0	4	3
DR	0	5	**522**	0	2	11
BR	1	1	0	**192**	3	7
CN	1	2	2	0	**568**	14
LB	0	1	7	0	14	**310**

参考文献

1) K. Koch, J. McLean, R. Segev, M.A. Freed, M.J. Berry, V. Balasubramanian, and P. Sterling, "How Much the Eye Tells the Brain," *Current Biology*, vol.16, no., pp.1428–1434, 2006.

2) A.M. Treisman and G. Gelade, "A Feature-Integration Theory of Attention," *Cognitive Psychology*, vol.12, no.1, pp.97–136, 1980.

3) C. Koch and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," *Human neurobiology*, vol.4, no.4, pp.219–227, 1985.

4) L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

5) A. Borji and L. Itti, "State-of-the-Art in Visual Attention Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.35, no.1, pp.185–207, 2013.

6) A. Shokoufandeha, I. Marsicb, S.J. Dickinsona, "View-Based Object Recognition Using Saliency Maps" *Image and Vision Computing*, vo.17, pp.445–460, 1999.

7) D. Walthera and C. Koch, "Modeling Attention to Salient Proto-Objects," *Neural Networks*, vol.19, no.9, pp.1395–1407, 2006.

8) D.G. Lowe, "Object Recognition from Local Scale-Invariant Features," *Proc. IEEE International Conference Computer Vision*, vol. 2, pp. 1150–1157, 1999.

9) M. Agrawal and K. Konolige, "Real-time Localization in Outdoor Environments using Stereo Vision and Inexpensive GPS," *Proc. 18th International Conference on Pattern Recognition*, pp.1063–1068, 2006.

10) A. Quattoni and A.Torralba, "Recognizing Indoor Scenes," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.413-420, 2009.

11) M. Fornoni and B. Caputo, "Indoor Scene Recognition using Task and Saliency-driven Feature Pooling," *Proc. British Machine Vision Conference*, 2012

12) C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol.20, no.3, pp.273–297, 1995.

13) T. Botterill, S. Mills, and R. Green, "Speeded-up Bag-of-Words algorithm for robot localisation through scene recognition," *Proc. 23rd International Conference Image and Vision Computing*, pp.1–6, 2008.

14) H. Bay, T. Tuytelaars, and L.V. Gool, "Surf: Speeded Up Robust Features," *Proc. European Conference on Computer Vision*, pp.404–417, 2006.

15) T. Liu J. Sun, N.N. Zhen, X. Tang, and H.Y. Shum, "Learning to Detect a Salient Object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.33, no.2, pp.353–367, 2011.

16) J. Yao, S. Fidler and R. Urtasun, "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.702–709, 2012.

17) P.F. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces," *British Machine Vision Conference*, 2013.

18) L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol.2, pp.524–531, 2005.

19) T. Kohonen, "Self-Organized Formation of Topologically Correct Feature Maps," it Biological Cybernetics, vol.43, no.1, pp.59–69, 1982.

20) R. H. Nielsen, "Counterpropagation networks," *Applied Optics*, vol.26, pp.4979–4983, 1987.

21) H. Madokoro, N. Shimoi, and K. Sato, "Adaptive Category Mapping Networks for All- Mode Topological Feature Learning Used for Mobile Robot Vision," *Proc. 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pp.678–683, 2014.

22) A. Oliva and A. Torralba, "Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope," *International Journal in Computer Vision*, vol. 42, no. 145–175, 2001.

23) P.F. Alcantarilla, A. Bartoli and A.J. Davison, "KAZE Features," *Lecture Notes in Computer Science*, vol. 7577, pp. 214–227, 2012.

24) J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt, "The KTHIDOL2 Database," *Technical Report CVAP304, KTH Royal Institute of Technology, CVAP/CAS*, 2006.

25) B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba , "Places: A 10 million Image Database for Scene Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (Epub ahead of print)*, 2017.

26) R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Proc. Fourteenth International Joint Conference on Artificial Intelligence*, vol.2, no.12, pp.1137–1143, 1995.

27) Raghavender Sahdev, John K. Tsotos , "Indoor Place Recognition System for Localization of Mobile Robots," *2016 13th Conference on Computer and Robot Vision (CRV)*, pp.53–60, 2016.