Scene Context 記述法の高度化による物体検出の性能向上

神谷　保徳†　　出口　大輔†　　井手　一郎†　　村瀬　　洋†

† 名古屋大学大学院情報科学研究科，〒 464-8603 名古屋市千種区不老町
E-mail: †kamiya@murase.m.is.nagoya-u.ac.jp, ††{ddeguchi,ide,murase}@is.nagoya-u.ac.jp

あらまし　Context の一つである対象物体が存在するシーンの情報を物体検出に利用する．シーンの見た目を記述し，記述したシーン上に物体の出現分布をマッピングする．シーンの見た目は，2 次元空間を仮定し，グリッド状に記述する．本研究では，シーン情報の記述方法について，従来手法の問題点である以下の二点について解決した新たな手法を提案する．1．シーン中には，変動が大きな領域と小さな領域が混在しており，変動が大きな領域はシーンの記述には不適切であるが，従来手法はこのことを考慮していない．2．画像として切り出したシーンの見た目を直接記述しており，撮影位置のずれやズーム度合の変化は考慮されていない．実験は二種類の評価法を用いて行った．上記の問題点二点を持つ従来手法と提案手法とを比較し，提案手法の有効性を確認した．

キーワード　Scene Context，物体出現分布，物体検出

A Novel Scene Context Descriptor for Improving Object Detection Performance

Yasunori KAMIYA†, Daisuke DEGUCHI†, Ichiro IDE†, and Hiroshi MURASE†

† Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
E-mail: †kamiya@murase.m.is.nagoya-u.ac.jp, ††{ddeguchi,ide,murase}@is.nagoya-u.ac.jp

Abstract　We focus on the contextual information for improving object detection peformance, which is the information on the scene where targeted objects exist. The scene appearances are described and the occurrence distributions of targeted objects in the scene are plotted. The scene appearances are assumed to be in 2D space and are described in a grid form. In this paper, we propose a novel method which addresses the following two problems that previous methods have: 1) Certain areas in a scene appear differently among images. Areas with large appearance variations are not suitable for describing scenes differently among images. 2) The previous methods directly describe the scene appearances which are cut out as images; thus the methods are weak for shift of shooting positions or change of zooming level. Experimental results showed the effectivness of the proposed method.

Key words　Scene Context, Object occurrence map, Object detection

1. はじめに

現在，様々な物体検出手法が提案されているが，多くの手法は，物体自身の見た目を如何に記述するか，という点について研究がなされている．しかしながら，物体以外の見た目の情報，例えば物体が存在するシーンの見た目からも，物体の位置と種類を推定することが可能である．例えば，車は灰色で一様な領域上 (道路) に存在しやすい．飛行中の飛行機は全体が水色で少し白色の部分も含む領域中 (空) に小さく存在しやすい．逆に両物体とも緑や茶色のテクスチャ領域中 (森) には存在しない，などである．

本研究では，物体が存在するシーンの情報を，物体検出において使用することを考える．物体が存在するシーンの見た目をモデル化し，モデル化したシーン中のどこに対象物体が出現しやすいかをマッピングし，その出現分布を用いて物体検出の精度を向上させる．シーンの記述には，道路，空，森といったラベル情報は使用しない．純粋にシーンの見た目情報のみを記述する．本研究では，対象物体が存在するのはどの様な見た目のシーンかを扱い，シーン中の各領域がなにであるか，対象物体の他にどの様な物体が含まれるか，ということは扱わない．

物体が存在するシーンの見た目を記述し，検出性能の向上を行った研究はいくつかある．[1] では物体の記述に用いる星形モデルを用いて，物体領域の周囲の領域の見た目を記述し，各モデルの出力を統合することで検出性

手法の学習に用いるものと完全に同一である．従って本手法のためにデータを準備する必要は無く，検出手法の学習用データのみで検出性能の向上が可能である．また，シーンの見た目のモデル化と物体検出手法とは独立であるため，本手法は様々な検出手法に適用可能である．

2. 関連研究

本研究は，Context(対象物体と対象物体以外のものとの関係性) を使用した認識に関連が深い. [6] では Context について解説している．Context は大きく分けて，物体間の関係性と，物体と背景 (シーン) との関係性があるが，本研究は後者に相当する. [7], [8] では物体が置かれているシーンの情報が，物体検出に有用であることを述べている．以降，本研究で対象とする Context およびタスクとは異なるが Context として関連する研究について述べる. [9]〜[12] では，物体（道路，空，といったラベルの付いた“ ラベル付き領域 ”を含む）間の共起と位置関係を使用して，画像中の物体領域のセグメンテーションを行っている. [13] では人手で切り出した物体領域のカテゴリの識別に，物体間の共起情報を用いている. [14] では物体検出に対して，物体間の共起，位置関係，見た目の関係性の情報を用いている．与えられた Context がすべて適切になるように検出窓の位置と大きさを調整することで物体検出を行う. [15] では物体間の共起と画像中の存在比率，また手作業で与えた背景のカテゴリを用いて画像中に対象物体が含まれるか含まれないかの識別を行っている. [16] ではカメラからの物体の距離と画像中の物体領域の大きさとの関係性，および物体の周囲領域の情報を用いて物体検出の精度向上を行っている．なお本研究で扱うシーンに関する Context は，物体間の共起など他の Context とは独立であり，併用が可能である．本手法を他の Context と扱う手法と組み合わせることで，さらなる性能向上が可能である．また本研究は,“ 物体 ”と“ ラベル付き領域 ”間の共起を用いた手法に比べると，各領域に対してラベル付けする手間がない分利便性が高い．

本研究は，物体検出手法を歩行者検出や前方車両検出に適用する際に行われる，カメラアングルが固定という条件の下で検出対象画像領域を制限する処理とも関連が深い．本手法を，カメラアングル固定という条件無しでこの制限処理を実行可能にする方法，と捉えることもできる．この様な検出対象領域の制限処理は，シーンの Context を暗に使用していると考えることができ，これらの処理との関連を考えることは非常に重要である．

3. 提案手法

まず提案モデルの構造について述べ，次にモデルの構築法について述べる．最後に，入力画像に対するモデルの適用方法について述べる．

3.1 提案モデルの構造

提案モデルは，複数の“ シーン記述 ”で構成される．学習画像中に含まれる各シーン（例：ビル街のシーン，郊外のシーン，田園地域のシーン）ごとに個別のシーン記述が構築される．シーン記述は擬似的に 2 次元空間で表現されたグリッドベースの表現であり，正方領域 (以降,“ セル ”と呼ぶ) を 2 次元的に連結したものである．このグリッド表現は，画像のシーンカテゴリの推定をタスクとした研究においても用いられており，本研究ではその研究のひとつ [17] で用いられた手法を参考とした．なお本研究のシーン記述の外形は長方形とは限らない．学習画像に含まれるシーンによってその形状は変化する．また，変動の大きな部分領域の除外はセル単位で行う．学習画像は，シーン記述と同様にグリッド表現で記述し学習に用いる．長辺を基準とし，指定セル数に分割，同じセル幅で短辺を分割する．

学習画像及び，シーン記述の各セルで記述に用いる特徴量について述べる．本研究では以下の二種類の特徴量を用いた．

- Gabor フィルタ（8 方向，4 スケール）32 次元
- 代表色（R,G,B）3 次元

Gabor フィルタは [17] で用いられているものと同様のものを用いた．

また，シーン記述中の各セルがそれぞれ保持する値を以下に示す．

- 特徴量 x の平均値 μ（35 個）
- 特徴量 x の標準偏差 σ（35 個）
- x^2 の平均値 $E(x^2)$ (35 個)
- セルのパラメータ更新に寄与した画像枚数 N

各セルは μ で記述される．σ は変動の大きなセルかどうかの判断用である．$E(x^2)$ 及び N は，μ と σ の更新用である．なお更新のためにすでに学習済みの学習画像を保持する必要はなく，上記の値を保持するだけでよい．

物体の出現分布について述べる．物体の出現分布はシーン記述ごとに構築される．物体領域の重心位置 $c = \{x, y\}$，物体領域の大きさ $a = \{w, h\}$ について記述する．$s = \{l, c, a\}$ を一つのサンプルとし，サンプルの集合として出現分布を記述する．なお，l は物体カテゴリのラベルである．検出対象の物体カテゴリが複数でもシーン記述は共通であり，カテゴリ別にシーン記述を構築する必要はない．

3.2 モデル構築法

モデルの構築方法について説明する．モデルの構築は，構築済みのモデルに対して画像を 1 枚ずつ追加していく追加型学習で行われる．モデル中のどれかのシーン記述中に，画像の見た目と類似する位置が存在する場合は，その画像でシーン記述のパラメータを更新する．どのシーン記述とも見た目が大きく異なる画像の場合は，その画像を元に新しいシーン記述を構築，モデルに追加

●定義

◇有効セル（変動の小さなセル）：

- パラメータ更新に寄与した画像枚数 N が一定値未満のセル
- パラメータ更新に寄与した画像枚数 N が一定値以上で，かつ特徴量の全ての σ が閾値未満のセル

◇画像とモデル間の類似位置決定基準：

1. $\frac{\text{類似セル数}}{\text{画像中のセル数}}(\%)$ が最も大きい位置 (シーン記述 ID，画像のオフセット位置とズーム度合) を選択
2. 1. で一意に決まらない場合は，画像とシーン記述のセル間の類似尺度の平均値が小さい位置を選択

但し，$\frac{\text{類似セル数}}{\text{画像中のセル数}}(\%)$ が，閾値パーセントを下回った場合は，類似位置無し，とする．

◇類似セル数：

画像とシーン記述のセル間の類似尺度が閾値未満となった，類似尺度計算の対象となる対象セルの個数

◇画像とシーン記述のセル間の類似尺度：

$$Dist_{grid} = Dist_{gabor} + \alpha \cdot Dist_{color} \tag{1}$$

$$Dist_{gabor}(\mathbf{x}, \boldsymbol{\mu}) = \sum_d^{32} |x_d - \mu_d| \tag{2}$$

$$Dist_{color}(\mathbf{x}, \boldsymbol{\mu}) = \sqrt{\sum_d^{3} (x_d - \mu_d)^2} \tag{3}$$

◇類似尺度計算の対象となるセル：

画像中のセル ∩ シーン記述中の有効セル

◇パラメータ更新処理：

画像中のセル ∩ シーン記述中の全セル，において μ, σ, $E(x^2)$, N を更新する

$$\mu^{new} = \frac{1}{N+1}\{N \cdot \mu + x\} \tag{4}$$

$$\sigma^{new} = \sqrt{E(x^2)^{new} - (\mu^{new})^2} \tag{5}$$

$$E(x^2)^{new} = \frac{1}{N+1}\{N \cdot E(x^2) + x^2\} \tag{6}$$

$$N^{new} = N + 1 \tag{7}$$

●モデル構築手順

以上を踏まえて，以下の手順を実行する．

1. 最初の学習画像をモデルの一つ目のシーン記述とする．
2. 次の学習画像に対して画像とモデル間の類似位置決定基準を用いて類似位置を決定，パラメータ更新処理を行う．もし，類似位置無し，となったら，その画像を新しいシーン記述とする．
3. 学習用画像が無くなるまで，2. を繰り返す．

●シーン推定手順

入力画像とモデルとで，画像とモデル間の類似位置決定基準を用いて類似位置を決定する．

図 3　モデル構築アルゴリズム

する．

類似位置の検索では，シーン記述上での画像のオフセット位置とズーム度合を求める (新規性１．)．その際の位置計算のずらし照合はセル単位で行う．ズーム度合の計算は，画像の長辺のセル数を変化させた複数の特徴量を計算しておき照合を行う．

見た目の変動度合いの計算は，画像のパラメータ更新処理の際に行う．セルを記述する特徴量の標準偏差 σ を計算し，その値の大きなセルは変動が大きいとし無効にする (新規性２．)．ただし，パラメータ更新に寄与した画像枚数 N が一定値未満の場合は，σ の信頼性が低いと見なし，常に有効と設定し学習を促す．

なお構築アルゴリズムの詳細は図 3 に示す．また，実際に構築したシーン記述の例を図 6 に示す．

シーンの見た目の学習を考える際に学習画像に関して問題となるのが，学習画像から物体領域を除外するかどうか，ということである．物体領域を除外することでより精度の高いシーン記述が可能になるだろうという考えから来る問題である．それに対して我々は，「変動の少ない静的な領域であれば，物体領域，背景領域関係なくシーン記述に適している」と考えている．従って学習時に物体領域を除外せず，そのまま学習を行う．物体が写り込んでいる領域は，もしその物体が常にその領域に存在するのであれば，物体検出に役立つためそのままシー

ンとして記述する．逆に，一カ所に留まっていることはまれで，物体が存在したりしなかったりする領域は，新規性１．変動の大きな部分領域はシーン記述に用いない，によってその領域は記述されない．新規性１．により，シーン記述を行う際に，画像中の各領域が，物体なのか背景なのか，という区別を考える必要が無くなり，また，有用な物体領域であればシーン記述に用いることが可能となる．

3.3 モデルの適用方法

構築したモデルの物体検出への適用方法について述べる．まず入力画像に対してシーン推定を行い，次に推定された物体出現分布を検出結果に適用する．シーン推定は，モデル構築時に用いた，学習画像が最も類似するモデル上の位置の計算方法を用いる．得られた位置 (シーン記述 ID，画像のオフセット位置とズーム度合) が推定結果となる．検出結果への適用は，検出器によって出力された検出領域ごとに行う．入力画像のシーン推定の後，得られたシーン記述 ID のシーン記述の出現分布に対し，画像のオフセット位置とズーム度合を考慮して，各検出結果に対するシーン情報による信頼度を，以下の式で計算する．

$$Conf = S \cdot \left(1 - \frac{Dist_K^2}{D^2}\right) \tag{8}$$

$Dist_K$ は，出現分布を構成する各物体領域サンプル s と検出結果領域 s' との距離 $Dist_{object}$ が最も小さいものから K 個を取り出し平均を取った値を表す．なお，検出器が対象とする物体カテゴリのラベル l が付いた物体領域サンプル s のみが計算の対象となる．

$$Dist_{object}(s, s') = Dist_{position}(c, c') + 0.5 \cdot Dist_{size}(a, a') \tag{9}$$

であり，$Dist_{position}$，$Dist_{size}$ 共にユークリッド距離である．物体領域の大きさは，物体領域の位置に依存するので，物体領域の位置に対して物体領域の大きさの重みを小さくしている．D は信頼度 $Conf$ が 0 となる $Dist_K$ の値，S は信頼度のスケールを表す．

検出領域ごとに計算されたシーン情報による信頼度と検出器による信頼度を考慮して，最終的な検出結果として各検出領域を出力するかどうかを決定する．

$$\text{検出器による検出領域の信頼度} + \text{シーン情報による検出領域の信頼度} > \theta \tag{10}$$

なら，その検出領域は有効，そうでなければ無効とする．なお θ は閾値である．

4. 実　　験

提案手法の有効性検証のため，二つの評価実験を行う．評価実験では，本手法の二点の新規性を持たない基準手法を設定し，比較を行う．

4.1 実験条件

本手法における二点の新規性を持たない手法として k-means を用いた手法を設定し，比較対象とする．図 3 に示すモデル構築アルゴリズムの代わりに，k-means を用いてモデル構築を行う．提案手法によるモデル構築手順はクラスタリング手法の一種と考えることができるからである．セルの特徴量次元数 (35) ×画像の長辺のセル分割数×画像の長辺のセル分割数，の次元数の特徴ベクトルで全学習画像を記述し，k-means を行う．特徴ベクトルの計算では，学習画像は中心位置で位置合わせを行い，また余白のセルの特徴量の値は全て 0 とした．k-means で得られる各クラスタがシーン記述に相当し，各クラスタのプロトタイプベクトルが，本手法における μ に相当する．シーン推定では，入力画像の特徴ベクトルとの距離が最も小さいシーン記述を推定結果とする．物体の出現分布は，各クラスタに属する学習画像中の物体領域を用いて構築する．出現分布の構築方法については本手法と同様の方法を用いる．

実験用画像は，LabelMe [18] から得た画像を用いて二つのデータセットを構築し，それらを用いて行う．一つはシーンと対象物体の見た目がおおよそ類似する様に選択した 140 枚の画像で構成されるデータセット (以降，Selected)，もう一つが無作為に選んだ 500 枚の画像からなるデータセット (以降，Random) である．検出対象物体は“ 車 ”とした．物体領域は，それぞれ 554 領域と 3650 領域存在する．画像例をそれぞれ図 4，図 5 に示す．

検出器には，Bag of Feature をベースとした手法を用いる．物体領域を 4×4 領域に分割し各領域を Bag of Feature による特徴ベクトルで記述，16 個の特徴ベクトルを連結した特徴ベクトルで物体領域全体を記述，SVM を用いて物体領域か背景領域かを分類，スライディングウインドウにより検出を行う．信頼度は $g(x) = \sum_i w_i K(x, x_i) - b$ とした．

学習画像のセル分割数は，長辺 20 分割を基準とし，また各ズーム度合におけるの特徴量の計算のため ±5 セルの分割数（全部で 11 通り）について計算を行った．また学習パラメータは経験的に設定した．なお，k-means で用いた長辺の分割数は 20 である．

4.2 モデル構築結果

構築したモデル中の幾つかのシーン記述を図 6 に示す．代表色セル，Gabor フィルタのセル，物体の出現分布について図に示す．また代表色セルの図において，変動が大きく無効となったセルには緑ドットを付けた．どのシーン記述の大きさ（セル数）も学習画像以上となり，撮影領域の変化を考慮したシーン記述となっていることがわかる．また変動が大きく無効となっているセルも所々見られ，変動の大きな領域と小さな領域が実際にシーン中に混在している事が確認できる．また物体の出現分布では，シーンの上方には物体がほとんど存在しな

図 4　実験画像例 (Selected)

図 5　実験画像例 (Random)

い事，物体領域が大きくなるほど領域の重心位置は大まかに下方に移動する事が分かる．これは，物体との距離（領域の大きさ）と物体の重心位置に関して遠近法を反映してるためであり，物体の出現分布が適切に記述できている事を示している．

また，シーン記述の数と学習の進行度（使用学習画像枚数）に関する遷移を図 7 に示す．初期はシーン記述の数の増加が著しいが，Random では学習画像が約 100 枚の時点からはシーン記述の数が殆ど増加しなくなっている．Selected でも，30 枚ごろから増加が緩やかになっている．これは，学習画像中のシーンの見た目が，学習が進むにつれ既にモデル化されている場合が多くなるからである．

4.3　評価実験 1

未学習画像における物体領域と非物体領域の，$Dist_K$ の平均値の差を評価する．$Dist_K$ は物体の出現分布を構成する物体領域サンプルとの距離であり，この差が大きいほど物体の出現分布を精度良く推定できていることを意味する．非物体領域は，物体領域に被らない領域を物体領域とほぼ同数ランダムに選択した．データセット中の半分の画像を用いてモデルの構築を行い，残りの画像に対し $Dist_K$ を計算し，その平均値を求め，差を計算した．結果を表 1，表 2 に示す．

結果は，本手法の方が k-means による手法に比べ，差の値が大きかった．このことから，提案手法の方がより精度良く物体の出現分布を推定できている事が確認できる．

4.4　評価実験 2

実際に物体検出を行い検出精度を比較することで提案手法の有効性を確認する．画像の半分を用いてモデルの構築及び検出器の学習を行い，残りで評価を行った．検出領域 A_d と正解領域 A_t において，$\frac{A_d \cap A_t}{A_d \cup A_t} > 0.3$ となっ