学位申請論文

空間的配置情報を活用した
教師なし物体領域抽出に関する研究

2019 年 3 月

村崎　和彦

要旨

画像に写っている物体が何であるかを計算機によって認識させる画像認識の技術は近年急速に発展しており，人の目による認識精度を超えるほどの結果が報告されている．しかし，そうした高精度な画像認識器を学習するためには大量の正解ラベル付き画像を用意することが必須となっており，認識対象が増加するのに応じて認識器にとっての教師情報である正解ラベルを大量に用意しなければならないという問題が生じている．十分な量の正解ラベルを用意することは多大なコストがかかるが，一方で学習に用いることができる画像データ自体は全世界で爆発的に増えており，物体認識のための大量画像データを取得するコストは日々低下していると言ってよいだろう．

そこで，本論文ではこうした正解ラベルが与えられていない大量の画像データを用いて新たな物体に関する情報を自動で獲得する認識器の実現を目的とし，画像データのみから物体認識を行う教師なし物体領域抽出について提案する．教師なし物体領域抽出においては，多数の画像が与えられたとき，その中に共通する物体を示す画像領域がどこであるかを推測することが解決すべき課題となる．そのようにして同一物体を示す画像領域を集めることができれば，それらを教師情報として画像認識器を学習することができるだろう．この同一物体の領域を見つける手がかりとして，従来研究では局所的な見た目の類似性が主に用いられていた．見た目の類似する領域が同一物体であると仮定することは有効なアプローチではあるが，見た目が類似していても異なる物体である場合や，同一物体であっても見た目が大きく異なる場合もあり，物体の形状変化や画像の撮影条件が大きく変わるような一般的な画像に対しては，適用が難しいという問題があった．これに対して本論文では，局所領域の類似性に加えて局所領域間の空間的配置関係に着目する．空間的配置関係とは，例えば左右に並んでいるといった画像上の配置関係だけでなく，前後に並んでおり一方が他方を遮蔽しているといった 3 次元の意味を含んだ配置関係を意味する．

本論文では空間的配置関係を活用した教師なし物体領域抽出に関して，主に 3 つの提案を述べる．初めに，同一物体の領域を適切に抽出するため，見た目の類似性及び空間的配

置関係を捉えるためのそれぞれの画像特徴について提案する．更に，領域の類似性と空間的配置関係の手がかりを同時に考慮する手法として，生成モデルに基づく物体領域抽出手法を提案する．

本論文は次のように構成される．第 1 章にて本論文の目的について述べる．第 2 章にて教師なし物体領域抽出における関連研究と課題について述べる．第 3 章では，同一物体の局所てきな類似性をより頑健に捉えるため，大量の一般物体を識別するよう学習された深層特徴を用いて物体のカテゴリ分類に特化した領域特徴表現を提案する．教師なし画像集合に対する前景物体検出の実験においてその有効性を示す．第 4 章では，画像特徴によって空間的配置関係を捉えるため，境界線における前後関係を明示的に学習させた境界特徴表現を提案する．画像から物体輪郭となる遮蔽境界線を検出し，物体の前後関係を推定する実験によってその有効性を示す．第 5 章では，従来の画像特徴の類似に加えて空間的配置関係を加味した物体領域抽出のための新たな生成モデルを提案する．各種データセットにおける教師なし物体領域抽出実験によってその有効性を示す．第 6 章では，第 3，4 章で提案する領域特徴表現と境界特徴表現を用いて第 5 章で提案する生成モデルによる教師なし物体領域抽出精度がどう変わるのかを検証する．提案特徴の組み合わせによって領域抽出精度が改善されることを示す．第 7 章では，本論文の提案技術によって達成される教師なし物体領域抽出についてまとめ，その将来性について述べる．

Abstract

In recent years, image recognition technology to recognize what is shown in an image by a computer has developed rapidly, and the recognition accuracy by the state-of-the-art method even exceeds that of human eyes. However, to train such a high-precision image recognizer, it is necessary to prepare a huge number of images with correct labels. The problem remains that a large number of data with correct labels must be prepared additionally each time the number of recognition objects increases. Preparing a sufficient number of correct annotations about images is costly; however, the cost to collect a large number of images showing various objects is decreasing day by day, because image data are increasing explosively worldwide. Therefore, in this study, the aim was to realize a recognizer that automatically acquires appearance information of unknown objects using a large number of images without any correct labels. A novel method is proposed to extract regions of objects by an unsupervised approach.

To recognize images without supervision, it is difficult to estimate which image regions show the same object among many prepared images. If one can collect image areas showing the same object in some way, the image recognizer can be trained by using these areas as supervision. As a clue to locate the regions belonging to the same object, the local similarity of appearances was mainly used in previous research. Although it is an effective approach to assume that regions with locally similar appearance belong to the same object, it is difficult to apply the approach to images in which the shapes of the objects change or the background varies greatly. Because there are cases where excluded regions belong to different objects, even though they are similar in appearance, or even if they belong to the same object, the appearance is significantly different. To solve this problem, a novel unsupervised image recognition

method is proposed to extract regions of unknown objects based on the spatial relationship between local regions, as well as the similarity of appearances. The spatial relationship means not only the arrangement relationship on the image, such as being arranged on the right and left, but also the relationship including 3D arrangement, such as being arranged on the front and back.

In this paper, the unsupervised image recognition method based on the spatial relationship mainly consists of three proposals. First, to extract regions belonging to the same object appropriately, two types of image features are proposed to capture the apparent similarities and to capture the spatial relations. Furthermore, an unsupervised image segmentation method based on a generative model that simultaneously considers the similarities of appearance and the spatial relationships is proposed.

This paper is organized as follows. In Chapter 1, the purpose of this paper is described. In Chapter 2, related works and remaining problems in unsupervised image segmentation are described. In Chapter 3, to capture more robustly the similarity of local appearance of the same object, feature representation specialized for category classification of objects by using deep features trained to classify a large number of general object categories is proposed. Its effectiveness is shown in an experiment of foreground object detection from an image set without any supervision. In Chapter 4, a feature representation about boundaries that is explicitly trained to estimate the occlusion relation on the boundary line to capture the spatial relationship is presented. Its effectiveness is shown by detecting the occlusion boundary, which is the object outline, in the image and estimating which side is front or back along the boundary. In Chapter 5, a new generative model for unsupervised image segmentation with a spatial relationship clue in addition to conventional similarity of appearance is proposed. Its effectiveness is shown by extracting object regions without supervision from various image sets. In Chapter 6, the accuracy of unsupervised image segmentation by the generative model proposed in Chapter 5 with feature representations proposed in Chapters 3 and 4 is examined. It is shown that image segmentation accuracy is improved by the combination of the proposed features. Chapter 7 summarizes the unsupervised image segmentation achieved by the proposed features and generative model, and future works are described.

目次

図目次

表目次

第 1 章

序論

1.1 本研究の目的

近年，インターネットにおける情報共有手段が発達し，スマートフォンなどの撮影機材が広く普及したことで世界中で撮影・記録される画像データは爆発的な増加を続けている．増え続ける大量の画像データを快適に参照するためには，画像データの内容を確認し，それに応じて画像データを整理する必要があるが，大量の画像データを人が目視で確認するには時間がかかるために，画像データの内容をコンピュータによって自動的に認識する技術が求められている．

画像データは様々な目的で撮影され，様々な意味を持つが，人間にとって最も基礎的な情報は，被写体が何であるか，であろう．被写体が何であるかをコンピュータによって自動認識する技術は一般物体認識技術（Object recognition）と呼ばれ，画像認識における最も基礎的な技術として盛んに取り組まれている．しかし，画像に写る被写体は 1 つと限らない．多くの画像はその中に多数の物体が写っており，その数やそれぞれの物体の配置も様々である．多数の物体が写る画像に対して，どのような物体がどこに写っているかを認識することは単純な一般物体認識ではなく，物体検出（Object detection）や物体領域抽出（Semantic segmentation）と呼ばれる問題設定として知られている．物体検出技術は画像に写るある特定カテゴリの物体位置を矩形で示す技術であり，画像に写る物体のカテゴリとその個数，大まかな位置を知る技術である．それに対して物体領域抽出技術は画像に写る物体の領域を画素単位で抽出する技術であり，画像に写る物体のカテゴリとその詳細な位置や形状を知ることができる．物体検出は被写体を 1 つずつ検出する技術であるため，1 つの物体としてのまとまりが明確なものが対象となるが，物体領域抽出は画素単位に意味付けを行う技術であるため，物体としてのまとまりが不明確な物も認識対象とす

ることができる．例えば，画像に写る空や海のような領域は物体検出の対象としては適さず，領域抽出によって捉えるべき対象と言える．画像に写る被写体を，その対象を問わず詳細に認識する場合，物体領域抽出による解析が有効となる．

これまで画像の物体領域抽出について多数の研究がなされており，大量の教師データに基づく機械学習による手法 [4] が特に成功を収めている．十分なデータを学習できれば，撮影条件や物体の変化などの影響を受けることなく多様な画像に対して物体領域抽出ができることが報告されている．しかし，一般的に精度よく物体領域抽出を行うには大量の教師データが必要であり，一方で画素単位のラベル付けを行うための画素単位の教師データを用意することは作業コストが高く容易でないことが問題となっている．こうした問題があるため，日々生成される多様な画像データに対して，認識可能な対象物体を日々追加していくことは簡単ではない．

本研究では，煩わしい教師データ作成の手間なく多様な画像を理解することを目的として，教師情報を用いない画像の物体領域抽出に取り組む．現実の画像データを認識する場合においては，全く教師情報が得られないことは希であるが，足りない教師情報を補いながら学習することを想定した最も困難な状況として完全な教師なしの条件の下物体領域抽出を試みる．教師なし問題において有用な領域抽出のアプローチ及び画像特徴の検討により，これらは部分的な教師あり学習においても有用な知見となると考える．

1.2 物体領域抽出技術の現状と課題

物体領域抽出技術とは上述の通り，画像に対して画素毎に何らかの意味づけを行う技術である．物体領域抽出技術に関連して，用いる手がかりと目的に応じて様々な問題設定が定義されており，大きく分けて 3 種類の取り組みが存在している．

1. 画像分割
   1 枚の画像から隣接する画素をクラスタリングし，物体毎の領域に画像を分割する問題である．このような問題においては，分割された各領域がどのような物体であるかは推定しないため，推定結果から直接的に画像を理解することは難しいが，分割された領域毎の物体認識などを行う前処理などに使われる．
2. 画像領域への意味ラベル付与
   画像に写る物体を認識し，画素毎に事前に学習された意味ラベルを付与する問題である．画像特徴と意味ラベルとの関連性を事前に獲得する必要があり，一般的に教師データを用いて機械学習によって認識を行うアプローチがとられる．

3. 画像集合からの同一物体領域抽出

   多数の画像を含む画像集合に対して同一物体を示す画素をクラスタリングし，多数の画像に含まれる同一物体領域を抽出する問題である．これは上述の 2 種の問題を合わせた問題設定と言える．同一物体を示す画素を抽出することは 1 つ目の問題設定と同様であるが，画像 1 枚を分割する場合には，一般に画像内での連続する領域が 1 つの物体を成すため，連続する領域の切れ目を探す問題に帰着する．一方で，多数の画像から同一物体を抽出する場合には，異なる画像間で同一物体を示す領域を見つける必要があるため，領域の連続性だけでなく物体固有の画像特徴の類似性を考慮した領域抽出が必要となる．また，そうした物体固有の画像特徴を用いることは 2 つ目の問題設定と同様であるが，この問題においては画素のクラスタリングが目的であり，画素に対応する意味ラベルを明示的に示すことはない．最終的に画像を理解するための意味ラベルを付与するためには，同一物体の抽出結果に応じて人が意味ラベルを設定するなどの後処理が必要となる．

3 種の問題設定それぞれにおいて技術の現状について以下に述べる．

1.2.1 画像分割

1 枚の画像を多数の領域に分割する問題については，画像処理技術の主要な問題として古くから多数の取り組みがなされている [5, 6, 7, 8, 9, 10]．分割の対象となる被写体を限定せず，多くの自然画像において共通的に適用できることを目的とした研究がなされており，対象物に依存しない汎用的な手がかりに基づく領域抽出が主な課題となる．

境界線検出

画像分割における最も有効な手がかりの 1 つに境界線があげられる．物体領域の周囲に発生する境界線を抽出することで，境界に沿ったきれいな領域抽出を実現することが期待される．境界線を用いた最もシンプルな手法として Sobel フィルタや Canny エッジ検出器 [11] のような勾配情報を用いて強いエッジを検出し，そのエッジに沿って領域分割を行うアプローチが考えられる．設定された閾値にしたがって境界線を決定することで，領域分割を行うことができる．また，強いエッジに囲まれた領域のみを採用するため，エッジ強度が極大となる箇所を境界線として採用する Watershed 変換 [7] が提案されている．フィルタ処理のような画素値の勾配に基づくエッジ抽出では，物体の中にテクスチャとして含まれるエッジと物体の境界に現れるエッジとを区別することができないが，近年

では，機械学習によってこれを解決する手法も提案されている．例えば，Arbelaez らの gPb [12] では勾配方向ヒストグラムに基づく機械学習によって画素毎の境界線となる確率を算出しており，単純なエッジ抽出のアプローチに比べて高い境界線検出性能を実現している．機械学習ベースの境界線検出は本来物理的には太さを持たない境界線を様々な形に変換して教師情報として与えることで発展を続けている．Lim らの Sketch Tokens [13] や Dollar らの Structured Edge [14] では，境界線形状のカテゴリ分けという問題設定を学習することで gPb よりも高精度な認識を実現しており，また深層学習に基づく HED [15] では，境界線からの距離に応じて段階的に教師信号を変えることで頑健な学習を実現している．

類似画素抽出

画像分割においてよく用いられるもう 1 つの手がかりとして，同一領域内の画素値の類似性があげられる．これを用いた最もシンプルな手法として画素値と座標値を用いたクラスタリングによる領域分割が考えられる．Mean shift [5] では，座標値と画素値のユークリッド距離が近い画素を逐次的にクラスタリングすることで色と出現箇所の類似する領域を簡易に抽出することができる．平均値からのユークリッド距離に基づく領域内画素のクラスタリングではなだらかに色が変化する領域を同一のものとして捉えるのが難しいが，Felzenszwalb ら [6] は隣接画素を画素値の変化によって重みづけされたエッジで結ぶグラフを考え，そのグラフの最小スパニングツリーが持つ重みが小さければ，同一の領域らしいとして領域抽出を行うことで輝度値がなだらかに変化する領域を抽出している．また，入力画像の画素値の距離に基づいてクラスタリングするのでなく，類似性を強調するような何かしらの特徴量変換の後にクラスタリングを行うアプローチも提案されている [8].

抽出する領域のモデリングに基づく領域抽出手法としてインタラクティブに画像分割を行うアプローチも提案されている．これは，抽出したい領域のヒントとなる箇所を人手でいくつか与え，それに基づいて領域分割を行う手法である．GrabCut [16] では，抽出したい対象を囲むような領域を人手で設定することで，枠の外側の色情報を混合ガウス分布によってモデリングし，背景領域を除外している．

大域最適化

多くの画像分割手法は，境界線による手がかりと領域内の類似性による手がかりを組み合わせ，画像全体で最適化を行うことで領域抽出結果を得る．グラフカットに基づく画像分割手法 [17, 9] では，画像全体を画素単位の確率変数によって構成されるマルコフ確率場（Markov Random Field: MRF）であると定義し，各画素がある領域に含まれる確率

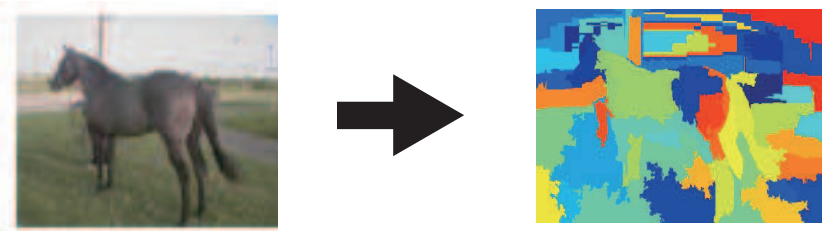


図 1.1　Superpixel 分割結果の例

と隣接画素間で領域ラベルが遷移する確率を定め，それらの同時確率を画像全体で最大化するような最適化を行う．2 種の手がかりがどのような形であったとしても，MRF による最適化に組み込むことができるため，よく用いられる．また，同様の最適化関数に対して領域ラベルを連続値とみなすことでより滑らかな領域抽出が可能な最適化手法も提案されている [18]．グラフカットと同様に良く用いられるのが，スペクトラルクラスタリング [19] に基づく最適化である．Normalized Cut [8] では画像を分割する境界線において分けられる隣接画素の類似度を最小化するような最適化関数を考えるが，これを抽出される領域内の画素間の類似度によって正規化するアプローチを提案している．この関数の最適化には隣接画素の類似度を重みとするグラフに対するスペクトラルクラスタリングを用いることができる．このアプローチは画素間の類似度を示す指標のみを用いて境界線らしさと領域内の類似性とのバランスを取るような出力を得ることができ，領域内画素のモデリングが得られない場合に有効であると言える．

Superpixel 分割

ここまで画像分割へのいくつかのアプローチについて述べたが，境界線や画素値の類似などの原始的な手がかりに基づく画像分割手法を実際に何らかの画像理解の手段として直接的に用いることは難しい．実用上は領域分割の後に何らかの意味付けを行うことが多い．このため，画像分割手法には後に続く認識処理での有用性を意識したアプローチが存在する．Superpixel 分割と呼ばれる問題設定として知られており，抽出される領域が可能な限り複数の物体領域をまたがないようにする一方で，1 つの物体領域が多数の領域に分割されてしまう過分割をある程度許容するという方針に基づく．Superpixel 分割の例を図 1.1 に示す．抽出される領域を Superpixel と呼び，後段の認識処理においてこの Superpixel を単位として意味づけを行うことを想定している．Mean shift のような画素値と座標値に基づくシンプルな手法は過分割に陥ることが多いため，Superpixel 分割として用いることができる．前処理であることから，処理時間が短いことを利点とする手法

図 1.2　画像領域への意味ラベル付与の例

が多い [6, 20]．また，後段の処理において入力となる Superpixels の数を固定にしたいことがあるため，あらかじめ決められた数となるように分割領域のシードを設定し，周辺の類似する画素を Superpixel として出力する手法も提案されている [10, 21]．

1.2.2　画像領域への意味ラベル付与

画像領域抽出において抽出される領域が何であるかを認識することは，画像を理解するためには非常に重要な問題である．この問題は Semantic Segmentation とも呼ばれ，近年，特に良く研究されている分野である．出力される意味ラベルについては事前に設定する必要があるため，基本的に機械学習によるアプローチが採用される．画像特徴と学習されたラベルとの関係に基づいて画素を単位として意味ラベルが付与される．画像領域に意味ラベルが付与されているイメージを図 1.2 に示す．

画像分割との連携

畳み込みニューラルネットワーク（CNN）による画像認識が一般的になる以前には，意味ラベルを付与した画像領域抽出を直接的に学習することは難しく，意味ラベルに依らない画像分割手法との連携による手法が用いられた．Carreira ら [22] は Superpixel 分割を事前に行い，抽出された Superpixel 毎に意味ラベルを推定している．そのための画像特徴として，SIFT 特徴 [23] や HOG 特徴 [24] などの勾配に基づく特徴表現を用いており，Superpixel 単位の特徴抽出及び認識を行っている．また，Tighe ら [25] は物体検出の結果に基づき，矩形で表現される物体検出結果を手がかりとして画像分割を行っている．具体的にはおおまかな物体領域を検出結果に基づいて設定し，MRF に基づく境界線位置の最適化によってエッジに沿った領域分割結果を得ている．このようにして意味ラベルの

学習と共に境界線の最適化を行う場合，より一般化された条件付き確率場（Conditional Random Field: CRF）を使う方法も提案されている [26]．CRF の場合，隣接する画素もしくは隣接する Superpixel において現れる意味ラベルの出現確率が学習され，ラベル間の共起性に基づく最適化を行うことができる．

画素単位の識別

機械学習に基づく画像認識において CNN の活用が一般的となり，近年では画像領域抽出においても CNN を用いた手法が主流である．特に Fully Convolutional Network（FCN） [4] は 1 つのネットワークモデルを用いて入力画像を領域分割結果に変換する手法であり，入力画像と領域分割結果を直接的に学習することができるため CNN を用いた領域分割のスタンダードとなるモデルである．FCN はその名の通りネットワーク内の処理が全て畳み込み層で構成されていることを特徴とする．画像識別によく用いられる VGG19 [27] などの CNN は畳み込み層とプーリング層を繰り返すことにより画像全体から特徴抽出を行った後に，全結合層によって画像全体の特徴に基づくクラス識別を行う．FCN にはこの全結合層が存在せず，畳み込み層の出力として得られるある一定範囲の画像特徴に基づいて畳み込み層によるクラス識別を行い，画素単位の識別結果として物体領域抽出結果を出力する．すなわち FCN においてある画素毎の推定結果は周辺の一定領域から畳み込み層によって得られる特徴に基づく識別結果とみなすことができる．FCN を学習する際には各画素を 1 つの学習サンプルとしてパラメータ更新を行うため，画像の枚数に対して効率的に学習が進むことも特徴的である．FCN では畳み込み層とプーリング層を繰り返すことで画像特徴の解像度が劣化するため，精細な認識結果を得ることが難しかったが，この問題を解決するために多数の改良が提案されている．異なるスケールでの特徴量を結合した Hyper column を用いた PixelNet [28] では，全ての特徴を元画像の解像度にアップサンプリングすることで高精細な認識結果を得ている．ダウンサンプリングされた画像特徴を Deconvolution によりアップサンプリングし，異なるスケールでの特徴量が多層ネットワークの途中で失われないよう前方のレイヤーと後方のレイヤーをバイパスした U-Net [29] やプーリング層により画像特徴がなまされることを避けるための Atrous convolution を導入した DeepLab [30] も同様の問題を解決している．

不完全な教師情報による学習

十分な教師データが与えられている場合には，FCN によってどのような対象についても画素単位の意味づけを学習することができる．一方，部分的な教師データから画素単位のラベル付けを学習する問題も取り組まれている．全ての画素に対してラベルが付与され

図 1.3　同一物体領域抽出結果の例

ておらず，一部の画素にのみラベルが付与されている教師データから学習する問題を半教師学習，画像に対してどのような物体が含まれているかのラベル付けがされているもののどの画素が対応しているかが明記されていない教師データから学習する問題を弱教師学習と呼ぶ．半教師学習による物体領域抽出は，画像分割手法において述べたインタラクティブな画像分割手法と類似した問題となる．Lin ら [31] は一部のラベル付与された箇所をシードとして MRF 最適化等を適用することで教師データを拡張し，そのデータに基づいて FCN 等の識別器を学習することで半教師学習を行っている．また，限られた教師データから CNN を学習し，学習した CNN による推論結果と MRF による最適化を組み合わせて教師データを修正する手法も提案されている [32]．さらに，Tang ら [33] は Normalized Cut と同様の目的関数を CNN によって最大化することによって半教師情報から画素単位の識別を学習している．弱教師学習による物体領域抽出については，画像毎に付与されたラベル情報をもとに認識対象の局所的な画像特徴を学習することが課題となる．Multiple instance learning (MIL) と呼ばれる機械学習のアプローチはあるクラスに属するサンプルが少なくとも 1 つ含まれるサンプル集合と 1 つも含まれないサンプル集合を教師データとしてあるクラスに固有の特徴を抽出する．これは弱教師学習の問題設定と同様であるため，MIL によって特定のクラスを特徴づけるような局所的な画像特徴を学習することができる [34]．また，教師データに対して直接的に密なラベル付けを行い，ラベル付けされた教師データを使って密な学習を行うアプローチもよく用いられる [35, 36]．

1.2.3　画像集合からの同一物体領域抽出

画像に写る被写体に関するラベル付けがなくとも同様の物体が写る多数の画像が得られる場合が考えられる．例えば，自分で撮影した写真を大量に保存したストレージには家族や身の回りの人物・物品が繰り返し登場すると想定される．そうした場合に問題となるの

が，画像集合からの同一物体領域抽出である．多数の画像に写る同一物体をそれが何であるかの教師ラベルを与えられることなく，多数画像の共通領域として抽出することが目的となる．図 1.3 に物体領域抽出結果の例を示す．図 1.3(a) のように馬が写る多数の画像から共通する物体領域として馬の領域を抽出している．また，図 1.3(b) のように共通して写る物体が複数カテゴリ存在している場合にそれぞれの領域をカテゴリ分けと共に抽出する問題も取り組まれている．認識対象を単一のカテゴリとする問題の方がより容易であり，多くの取り組みが報告されているが，入力となる画像集合を適切にコントロールする必要があるため実用的とは言い難い問題設定である．多数の物体を被写体として想定する複数カテゴリの領域抽出については，前述の個人が持つ画像ストレージの整理など実用的な用途が考えられるが，高い精度で認識することは難しく，取り組みも少ない．また，複数カテゴリの領域抽出結果はその意味については言語化されていないため，どのような意味を持つカテゴリであるかは人が後から意味づけを行う必要がある．

単一カテゴリの抽出

同一の物体カテゴリが写る画像の集合から物体領域を抜き出す問題は Co-segmentation と呼ばれこれまで数多くの研究がなされている [37]．様々の認識対象に対して多数のアプローチが提案されているものの，標準的な手法が確立されていないことからも，画像集合からの物体領域抽出問題が今なお難しい問題であることがわかる．例えば，Joulin ら [38] は画像内の局所領域間の類似度によって類似度グラフを作成し，Spectral Clustering によって物体領域を抽出している．Rubinstein ら [39] はある画像のパーツを用いて別の画像に写る物体を再構成できることを仮定し，その再構成誤差によって物体領域抽出を行っている．Meng ら [40] は画像 1 枚から算出可能な顕著性マップに基づく前景らしさを導入して物体領域の抽出精度をあげており，Dong ら [41] は画像 1 枚の中でのラベルの連続性を MRF によって平滑化することで，抽出精度向上を狙っている．また，単一カテゴリ抽出においてもいくつかの異なる問題設定が存在する．最もシンプルな問題設定は見た目が全く同じ物体を異なる撮影条件で撮影した画像から物体領域を抽出する問題 [42, 40] である．これに対して見た目にある程度変化が生じる同一カテゴリの物体を写した多数の画像から物体領域を抽出する問題 [39] もある．全く同じ物体の領域抽出に比べて，同一カテゴリの物体領域抽出は，物体の色やテクスチャ，形状，姿勢などが変化する可能性があるため抽出が難しい場合が多い．同一カテゴリに含まれる複数の物体をそれぞれ別のクラスとして捉え，各物体の領域抽出を行う問題 [38] も提案されている．Rubinstein ら [39] は更に複雑な問題として画像集合にノイズ画像を含む場合を対象とした手法を提案している．抽出したい対象物体が支配的である必要があるため，ノイズ画像は相対的にわずかではあ

図 1.4　Wang らの手法 [1] による領域抽出結果

るが，わずかなノイズを許容することで，例えば検索エンジンによる画像検索結果から物体領域を抽出するといった現実的なシナリオでの活用が可能になったと主張している．

複数カテゴリの抽出

複数のカテゴリを画像集合から一括で学習する手法についてもいくつか既存の取り組みが存在する．Wang ら [1] は，図 1.4 に示すような多種の物体が写る画像集合から多数の画像にわたって共通して写っている子供やかぼちゃなどの抽出を実現している．Wang ら [1] は画像間の局所的な領域の類似性に基づいて対応関係を推定してグラフ関係を構築し，多数の画像について一貫性のある対応関係が取れる領域をグラフクラスタリングによって抽出している．図 1.4 のように抽出する同一物体の画像特徴が非常に類似している場合にはきれいな領域抽出ができているが，形状や姿勢などの変化によって画像特徴が異なる場合には局所領域の対応関係が得られないため適用が難しい．Cho ら [43] は Object Proposal によって多数の前景候補を抽出し，そこに含まれるパーツの類似から多種カテゴリが混在する複雑な画像集合からの物体位置検出を実現している．物体検出であるため，厳密には領域抽出までを行っていないが，MRF 等によって境界線の最適化を組み合わせることで領域抽出を行えることが想定される．大きく見えが変わるような多様なデータに対しても頑健に機能することが示されているが，各画像における前景物体の抽出精度が評価されており，どの物体が同一カテゴリであるかを精度良く認識することは将来課題としている．また，Niu ら [44] は文書解析で主に用いられるトピックモデルを活用し，画像集合に含まれる多種カテゴリをトピックとして扱うことで物体領域抽出を実現している．画像特徴の変化に対して，同一物体から抽出される画像特徴の傾向をモデリングする

ため，多様な物体の特徴を捉えられているものの局所的な画像特徴をその位置情報を考慮せず独立して扱うために，きれいな連続した領域を得ることが難しい．

1.2.4　物体領域抽出における課題

前節までは，画像理解のための物体領域抽出について従来の取り組みを述べた．近年，主に FCN に代表される多層ニューラルネットに基づく認識手法の台頭によって十分な教師データに基づく画素単位の意味づけは急速にその認識精度が高まっており，実用的なレベルに到達している．一方で，高精度な領域抽出器を学習するためには十分な教師データが必要となる．領域抽出器を学習するための教師データは画素単位のラベル付けが必要となるため，データ作成に非常に手間がかかる．現在の物体領域抽出における大きな研究課題として認識精度を保ったままいかに教師データを減らすことができるかが注目されており，半教師や弱教師による領域抽出器の学習が取り組まれている．半教師学習が可能になれば，画像データに対して部分的なラベルを付与するだけで，学習を進めることができ，教師データ作成のコストを大きく下げることができるだろう．また，更なる課題として教師なしでの物体領域抽出の学習が考えられる．これは前述の画像集合からの同一物体領域抽出に相当する課題であり，教師なし学習が実現されることにより，人が認識対象のラベルを予め決めることなく日々生成される画像データから新たな認識対象を抽出することができると考えられる．そのような学習の適用先を想定した場合，入力される画像集合は単一カテゴリが統一的に撮影された画像ではなく，多数のカテゴリが混在する画像集合を入力とする必要がある．画像集合からの複数カテゴリの物体領域抽出が現在の物体領域抽出技術において未解決の問題であり，爆発的に増え続ける画像データに対して人手を介さずに画像理解を深めるために鍵となる技術であろうと考えられる．

本研究では，複数カテゴリが混在する画像集合に対する同一物体領域抽出を主な課題としてその精度向上に取り組む．実利用上の観点では，いくつかの物体が教師データによって学習されていることが想定され，認識可能な物体と認識対象でない新規物体が混在するデータから新規物体の領域抽出を行うタスクが考えられるが，本研究においては未解決課題となっている教師データが付与されていないデータからの未知物体の学習を主な研究対象とするため，より単純な問題として複数の未知物体が共通して写る画像データセットから物体領域を抽出する問題に取り組む．

1.3 本論文の構成

本論文における次章以降の構成について以下に述べる．

第 2 章では，本研究において取り組む教師なし物体領域抽出問題について定め，提案する領域抽出手法のアプローチと枠組みについて概説する．提案手法では，領域へのラベル付与に加えて隣接する領域間の境界に対して空間的配置関係に相当するラベルを与えるアプローチをとる．

第 3 章では，教師なし物体領域抽出において異なる画像間での類似する物体領域を得るための領域特徴表現について述べる．意味レベルでの類似性を評価できる画像特徴を得るため，深層学習によって多様な教師データに基づいて学習されたモデルを活用する．多クラスの物体認識問題を学習済みの CNN の中間出力を深層特徴とよぶ．この深層特徴を球面クラスタリングすることによって同一物体のカテゴリ抽出が可能であることを示す．実験によって，事前に認識器を学習したカテゴリとは異なるカテゴリを含む多様な画像において教師情報を用いずに同一物体領域を抽出できていることを示す．

第 4 章では，教師なし物体領域抽出において空間的配置関係を得るための境界特徴表現について述べる．ここで，空間的配置関係とは境界線付近での上下左右及び前後の隣接関係によるものであるとして，前後関係を含む境界線形状を認識する問題を考える．多様なデータに対して境界線形状の教師データを用意し，それらに基づいた学習を行うことで，学習したモデルによって得られる認識のための中間表現を汎用的な境界特徴表現として得ることができる．実験では，この特徴を用いた境界線検出及び前後関係推定を評価し，従来手法と比較して高速かつ高性能に前後関係推定が行えることを示す．

第 5 章では，教師なし物体領域抽出問題に対して空間的配置関係を示す境界へのラベル付けを同時に最適化する手法としてトピックモデルをベースとした新たな確率モデルを提案する．従来の画像内の小領域に対するトピック推定に加えて，小領域間の境界に対する境界トピックを導入し，それらを同時に最適化することで物体領域抽出の精度が向上することを示す．実験においては，画像集合に対して 1 つの被写体カテゴリが設定されている場合と複数の被写体カテゴリが設定されている場合とで評価を行い，多数カテゴリへの適用性と従来手法に対する優位性を示す．

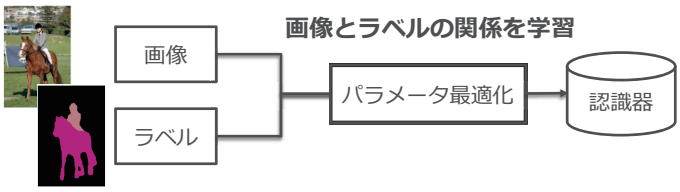
第 6 章では，第 3，4 章にて提案する画像特徴と第 5 章にて提案する確率モデルによって実現される教師なし物体領域抽出の評価を行う．各章における提案による貢献を確認し，またこれらを適切に組み合わせることで，高精度に物体領域抽出を行えることを示す．

最後に，第 7 章では，本研究の取り組み，成果，及び将来課題についてまとめる．

第 2 章

空間的配置情報を活用した教師なし物体領域抽出

第 1 章では，物体領域抽出技術における現状と課題について述べた．大量の教師データに基づく画素毎の意味ラベル付け問題については汎用技術としての研究課題は解決されつつあり，教師データが得られない場合での学習が残された問題となっている．



(a) 教師あり物体領域抽出

(b) 教師なし物体領域抽出

図 2.1　教師あり物体領域抽出処理と教師なし物体領域抽出処理の違い

2.1 本研究で取り組む問題

本研究では，以下のような設定において得られる画像データセットから物体領域抽出を行う問題に取り組む．

- 認識対象となる物体の有無について人手による教師情報が全く得られない．
- 画像データセットには同一カテゴリの物体が写る画像が多数含まれている．

認識する物体ラベルについて何らの教師データを得られないため，画像データに対して明示的なラベルを推定することは不可能であるが，複数の画像に共通して含まれている同一カテゴリの物体について同一の ID を付与するような領域抽出を目指す．データドリブンに抽出されたカテゴリに対する意味づけについては改めて人手によって行うことを想定する．教師データに基づく領域抽出処理と教師なし領域抽出処理との違いを図 2.1 に示す．また，画像に共通して写る同一カテゴリの物体が存在することを想定した領域抽出を行うため，全くの無作為な画像集合から推定を行うことは難しい．入力となる画像集合はある程度の偏りを持った画像集合であり，見た目によって同一性を判断することのできるような同一カテゴリに属する物体が十分な枚数の画像に出現しており，また各画像において十分な大きさで写っていることを想定する．

ここで，本研究で対象とする教師なし学習とは，領域抽出の対象となる画像集合に対して対象物体の有無に関する教師情報が付与されていないことを意味する．未知の物体に対する適切な領域抽出を達成するために，領域抽出の対象と直接的には関係しない一般的な事前知識を教師情報として予め学習しておくことは本研究の目的を損なうものではない．

2.2 空間的配置関係の活用

これまで，教師なし物体領域抽出の既存手法において主に用いられる手がかりは領域の切り替わりを示す境界線と抽出される領域内の類似性の 2 つであった．Wang らの手法 [1] や Niu らの手法 [44] では，境界線に基づく Superpixel 分割を前処理として，Superpixel 内の画像特徴の類似に基づいて同一物体の領域抽出を行っている．Wang らの手法では，Superpixel 間の類似性を直接評価するため，図 1.4 で示されるような画像間で見た目の類似性が非常に高い物体については，精度よく抽出できるものの同一物体カテゴリでありながらその色や形が変わるような多様性を持つ物体については抽出精度が低下してしまう．また，Niu らの手法では，カテゴリの持つ画像特徴を確率モデルによって表

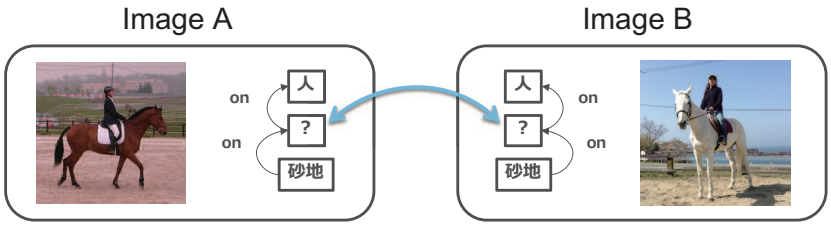


図 2.2　隣接する領域との空間的配置関係に基づく対応付け

現するため，見た目が多様なカテゴリの抽出をも実現しているが，同一カテゴリに含まれる領域の境界線を決める Superpixel 分割の影響が大きく，抽出されるべき領域が欠けるなど精度に問題がある．これらの従来手法は領域内の類似性の手がかりを重視しており，画像特徴の類似において一貫性のある領域抽出を実現するよう試みているが，領域の境界線については汎用的な Superpixel 分割の結果を用いており，Superpixel 単位で領域が欠損することによって領域抽出精度が低下している場合がある．

境界線に沿った分割と領域の類似性に基づく従来手法に対して本研究では，隣接する物体領域との空間的配置関係を新たな手がかりとして導入する．空間的配置関係とは，隣接する物体領域が上下左右前後にどのような配置関係にあるかを意味する．画像から未知の物体領域を得るにあたって，周囲の物体が何であるか，また周囲の物体とどのような配置関係にあるかを考慮することで，単純な Superpixel 内の画像特徴からは得られない類似性が評価できると考えられる．例えば，図 2.2 では馬の上にのる人の画像の例を示す．２つの画像において騎手の画像特徴と砂地の画像特徴は類似しており，対応関係が学習できると想定される．一方で，馬の画像特徴は類似性が小さく，対応関係を得ることが難しい．ここで，隣接領域との配置関係を考慮すると，馬の領域は砂地の上方・騎手の下方に位置している領域と見ることができる．この配置関係は２つの画像間で類似しており，この配置関係における類似性を手がかりとして馬の領域を抽出することができる．

2.2.1　空間的配置関係の有効性

画像から物体領域を抽出するにあたって空間的な配置関係が有用な情報であることは間違いない．単純に，隣接する領域が前後に離れた配置関係であれば，異なるカテゴリを示す物体領域である可能性は高いと考えることができる．また，例えば，皿は机の上に並べて置かれることが多い，といった配置関係の偏りが存在していれば，机の上にある類似した物体として皿の領域抽出精度が高まることが考えられる．実際に深度情報の計測を用い

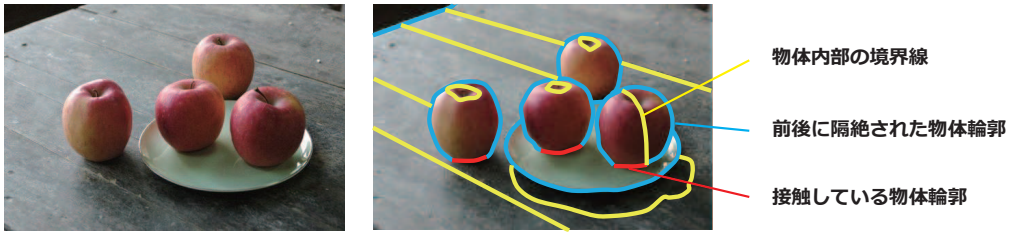


図 2.3　境界線上に定義される空間的配置関係の例

た物体領域抽出手法が提案されているが，画像データのみを用いる場合よりも大きな精度向上が認められている．

2.2.2　境界線への意味づけ

本研究では，画像集合から物体領域を抽出するにあたって抽出する領域と周辺の領域との空間的配置関係を手がかりとしたアプローチを提案する．提案手法では，隣接する領域間の配置関係を表現するため，境界線への意味づけを行う．隣接する領域間に発生する境界線に対して何らかのカテゴリ分類を行うことによって，領域間の配置関係を表現することができる．具体的な配置関係の表現として図 2.3 のような例を考える．画像はエッジの強さ等に応じて Superpixel に分割されており，隣接する Superpixel の間に境界線が定義されている．ここで，リンゴの領域内に含まれている境界線については，色の違いや影によって発生したエッジであるものの配置関係としては同一物体領域の内部となる．リンゴと背景との間の境界線については，深度の違いによって発生したエッジであり，配置関係として前後の隔絶した関係が発生する．更に，リンゴと皿との境界線については物体が異なるために色の違いや深度の違いなどによりエッジが発生しているが，配置関係としては一部が接触した前後関係と言える．このような配置関係を境界線に対して付与することができれば，抽出される物体領域やその物体のラベルは周辺の配置関係との関係性によってより良く抽出できることが期待される．

2.3　提案する教師なし物体領域抽出手法の枠組み

教師なし物体領域抽出の問題において空間的配置関係の手がかりを導入するため，隣接する領域間に発生する境界線に対してカテゴリ分類を行うアプローチを提案する．提案手法の枠組みについて概要を述べる．

2.3.1 領域と境界の画像特徴表現

多クラス教師なし物体領域抽出の従来手法と同様に，各画像を Superpixel に分割し，各 Superpixel が属する物体 ID を推定するアプローチを採用する．各 Superpixel についての画像特徴と Superpixel 間境界についての画像特徴を手がかりとして抽出し，各 Superpixel 間の関係性を境界線に紐づけてラベル付けすることを考える．本論文において，Superpixel から算出される画像特徴を「領域特徴」，境界線から算出される画像特徴を「境界特徴」と呼ぶ．教師なし学習において，画像特徴の類似性は非常に重要な手がかりとなるため，問題に応じて適切な画像特徴を用いることが重要となる．従来手法においては，領域特徴として SIFT [23] に代表される回転やスケール変化に対して不変性を持つ画像特徴が使われているが，抽出したい同一カテゴリにおいて物体の形状変化やテクスチャの変化を含む場合，SIFT 特徴に基づく類似性によって精度良く同一カテゴリを抽出できない問題がある．また，提案手法において導入される境界特徴についても，境界周辺の類似性評価に用いられている Geometiric Blur [45] などの画像特徴表現が提案されているものの，既存の特徴表現が必ずしも想定する空間的配置関係と相関しないという問題がある．本論文では，こうした問題を解決するために領域と境界の特徴表現について，同一物体抽出及び空間的配置関係抽出を想定した教師データに基づくデータドリブンな特徴表現を提案する．

提案する画像特徴表現は教師あり学習によって獲得する．画像特徴の算出によって抽出したい属性を予め設定し，それに基づいた教師データを用意することで教師あり学習を行うことができるが，この学習によって得られる認識器の中間出力を画像特徴として採用する．教師データが持つラベルが同じものに対して類似性が高くなるような中間特徴表現が学習されるため，抽出したい属性に適した類似性の評価を行うことができる．領域特徴表現については，大規模なカテゴリ数の物体認識を学習させ，その中間特徴表現を用いることで，撮影のスケールや照明条件による影響を抑えて，各物体のカテゴリに依存するような特徴表現を獲得する．また境界特徴表現については，空間的配置関係との関連を強くするために，物体輪郭であるか否か，前景背景の関係はどちら向きか，を含むような境界線形状を推定する問題を学習させることで特徴表現を獲得する．

2.3.2 領域と境界の同時カテゴリ分類

抽出された各種画像特徴に基づき，Superpixel に付与されるラベルと境界線に付与されるラベルが一貫性を持つように最適化を行う必要があるが，提案手法では，見た目の多様なカテゴリの抽出において有効性が示されている，トピックモデルに基づく物体領域抽出手法をベースとした最適化を行う．具体的には，領域へのラベル付けと境界へのラベル付けを画像特徴の類似性などから同様に最適化すると同時に両者の関係を定める確率分布を設定することによって双方のラベル付けに対して尤度が高まるような最適化を行う．このようにして領域と境界への同時カテゴリ分類を行うことで，物体領域のラベル推定と整合するような境界線における配置関係ラベルを得ることができ，またそれによってより良い物体領域抽出結果が得られる．

第 3 章

物体ラベルの学習に基づく領域特徴表現

本章では，教師なし物体領域抽出のための有効な画像特徴表現について述べる．予めラベル付けされた大量の画像を用いて学習した畳み込みニューラルネットワーク（CNN: Convolutional Neural Network）から得られる中間出力は深層特徴と呼ばれ，様々な画像認識問題において有用な特徴表現であることが知られている [46, 47, 48]．本章では，この深層特徴に対して球面クラスタリングを適用することで，教師なし画像集合から同一物体を示す特徴を抽出する手法を提案する [49]．画像集合から教師なしに同一物体位置を検出する問題（Co-localization）を例として，提案特徴表現によって従来手法を上回る高い精度が得られることを示す．

3.1　深層特徴による教師なし学習

深層学習が広く使われ始めてから，学習済みの CNN の中間出力値を画像特徴として取り出し，学習データとは異なる問題に適用するという手法が効果を示している．抽出された画像特徴は深層特徴と呼ばれ，これらを適用先の問題に対してさらに学習することで，物体識別や検出において有用な特徴となる．例えば，深層特徴を活用した Object Proposal の抽出 [46] や物体検出 [47] が提案されている．こうした事前学習された深層特徴は教師なし画像集合から共通して含まれている前景物体の位置を検出する Co-Localization 問題においても有用であり，深層特徴を活用することで教師なしでありながら各物体の負例画像が与えられる弱教師あり問題と同様の精度を実現できることが報告されている [48]．Wei らによる Deep Descriptor Transforming (DDT) [48] では，データセットから得ら

れる深層特徴に対して主成分分析（PCA: Principal Component Analysis）を行い，主成分が強い領域を抽出するというシンプルな手法によって高精度に Co-Localization が行えることが示されている．一般的に用いられる深層特徴は画像認識技術コンペティション ILSVRC において用いられている ImageNet [50] の 1000 物体を基に学習されたものであるが，こうした深層特徴の有効性は事前学習された 1000 クラスはもとより，ILSVRC に存在しない物体についても有用であることが示されており，様々な物体を対象として汎用的かつ強力な特徴表現になることが期待されている．

本章では，DDT [48] と同様に深層特徴のシンプルな解析による特徴表現を提案し，その有用性を Co-localization の実験によって示す．球面クラスタリング [51] に基づく共起特徴の抽出によって，更に高精度な Co-Localization を実現できることを示す．また，提案特徴表現は DDT では適用不可能であった多クラスを含む Co-Localization 問題についても同様の枠組みで適用可能であり，従来法である Cho らの手法 [43] を検出精度において大幅に超えることを示す．

3.2 提案する特徴表現

3.2.1 球面クラスタリングによる深層特徴モデリング

DDT [48] では，PCA によってデータセットの主物体らしさを抽出できることが示されている．この手法はデータセットから算出される第 1 主成分の値が一定値以上であれば，そのデータセットの主物体に帰属するとするシンプルなものである．抽出する深層特徴 x_i の集合 X から算出される第 1 固有ベクトルを ξ_1 とすると，各特徴の第 1 主成分は

$$p_i = \xi_1^T x_i \tag{3.1}$$

となり，p_i が閾値 τ を超えていれば主物体に属しているとみなせる．この処理は固有ベクトル ξ_1 の方向と x_i が同じ方向を向いていれば，主物体に属すとみなすことができる．すなわち，深層特徴 x_i はその向きによって物体クラスが表現されており，その大きさによってその物体に属する尤度が示されていると考えることができる．

我々はこの仮説に基づき深層特徴の方向に応じたクラスタリングを行う球面クラスタリングを導入する．球面クラスタリングによってデータセットに主として存在する物体クラスの持つ方向を抽出し，方向の類似する深層特徴をまとめることで，同一物体から得られる深層特徴を分類する手法を提案する．また，深層特徴の各クラスタ方向成分の強さによって信頼度を決定することもできる．

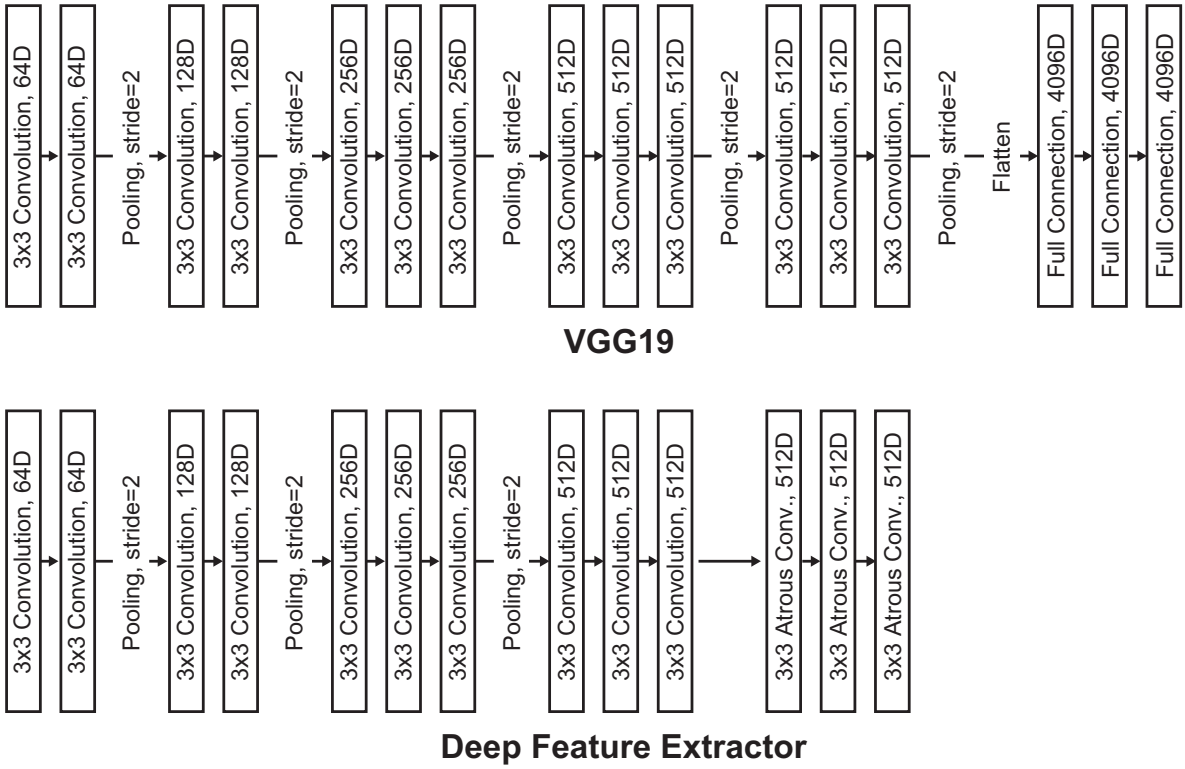


図 3.1　深層特徴抽出に用いる CNN の構成．上段は事前学習に用いられる CNN モデル，下段は学習済みパラメータを用いて深層特徴を抽出する CNN モデルを示す．

特徴抽出

まず，事前学習された CNN を用いて深層特徴を抽出する．事前学習モデルとして DDT [48] と同様に VGG19 [27] を用い，抽出する中間出力値として，畳み込み層の最終値である Conv5 層の出力（512 次元）を用いる．VGG19 ではプーリング層による縦横サイズの縮小を 5 回行っているため，そのままのモデルでは出力特徴の幅は入力画像の幅の 1/32 となってしまう．解像度を高め，より高精細な位置決めを行うために，VGG19 モデルに Atrous 畳み込み [52] を導入する．CNN モデルの構成図を図 3.1 に示す．プーリング層を除き，Atrous 畳み込みを用いることで，出力される深層特徴画像のサイズを入力画像の 1/16 としている．

図 3.2　Object Discovery Dataset に対する教師なし複数物体検出処理の流れ

球面クラスタリング

データセット全体を通して得られた深層特徴の集合を用いて球面クラスタリングを行う．球面クラスタリングは各データ点を半径 1 の超球面上に射影し，それらに基づいて k-Means クラスタリングを行う手法である．特徴ベクトルの持つ大きさよりも分布が意味を持つ場合に有効なクラスタリング手法であり，文書解析などに用いられている．本稿では，Dhillon らの提案する Spherical k-means clustering [51] に従い，超球面上に射影された各データ点を用いてユークリッド距離に基づく k-means クラスタリングを行う．クラスタ数 K は事前に与える必要があり，クラスタ中心の初期値は k-means++ [53] に基づいて設定する．クラスタリングを適用する前に入力特徴から全体の平均ベクトルを差し引くことで，超球面の中心はデータセット全体の平均値となるよう設定する．球面クラスタリングによって K 個のクラスタ中心 $c_k(k = 1 \cdots K)$ が与えられ，各データ点 x_i については近傍のクラスタ中心に応じたクラスタ番号 l_i が付与される．飛行機，馬，車が写る画像集合に対してクラスタリングを行った例を図 3.2 の 2 行目（Clusters）に示す．図のように同一物体から得られる深層特徴が同一クラスタにまとめられていることが分かる．

3.2.2　提案特徴を活用した教師なし物体検出

提案する特徴表現の有効性を確認するために，教師なし物体検出（Co-localization）問題に対しての適用性を評価する．提案特徴に基づいて教師なし物体検出を行う場合，抽出されるいくつかの物体クラスタのうち，どのクラスタが画像集合において共起している物体のクラスタであるかを推定する必要がある．このような主物体クラスタの選択において

も球面クラスタリングの結果を活用できるため以下に詳細を述べる．

主物体の位置検出手法

クラスタリングの結果，各画像から抽出される各特徴についてクラスタ番号が与えられている．データセットにおける主物体の位置を検出するためには対象物体を示す適切なクラスタ番号を選ぶ必要がある．ここで，深層特徴の持つ意味（物体クラス）が方向に表現され，その信頼度が大きさに表現されるという仮定を考慮すると，データセットにおいて最も信頼度が高くなるクラスタが主物体であると考えられる．抽出された各クラスタについて，クラスタに属する各特徴 x_i のクラスタ中心方向への大きさの平均値を算出し，各クラスタ k の強度 S_k とみなすことができる．

$$S_k = \frac{1}{N_k} \sum_{i \in k} c_k^T x_i, \tag{3.2}$$

ここで，N_k はクラスタ k に属する特徴ベクトルの数であり，クラスタ中心 c_k は絶対値が 1 となるように正規化されている．データセットについて最も強度 S_k が大きくなるクラスタを主物体クラスタとして選択する．

CNN によって抽出された画像特徴は入力画像に対してサイズが縮小しているため，入力画像サイズにアップサンプリングすることで，各データ点に対して付与されたクラスタ番号と画像内の領域とを対応させる．主物体クラスタに属する画像領域が抽出すべき物体領域となる．

最後に抽出された主物体領域のうち最も面積の大きいまとまりを選び，その領域の外接矩形を検出結果とする．

3.3 実験

実験によって提案手法が非常にシンプルでありながらも教師なし物体検出において強力な手法であることを示す．

3.3.1 データセットと評価指標

物体検出の評価手法として従来研究 [43, 48, 54] で用いられる CorLoc 指標を用いる．これは，推定された検出窓を B_D，正解枠を B_A とした時に $\frac{\mathrm{area}(B_D \cap B_A)}{\mathrm{area}(B_D \cup B_A)} > 0.5$ を満たす画像の割合を示す．$\mathrm{area}(x)$ は検出領域の面積を示す．画像に複数の物体が写っており，正解窓が複数ある場合には，最もスコアが高くなる窓を対象として評価する．

提案手法にはハイパーパラメータとしてクラスタ数を示すKが存在している．以下の定量評価においてはいくつかのKについて評価を行い，その精度変化を合わせて示す．性能評価のためのデータセットとして以下の3つを用いる．

Object Discovery Dataset

Object Discovery Dataset [39] はノイズ画像を含むデータを想定した際の Co-Localization 精度を評価するためのデータである．従来研究 [43, 48] における評価にならい，Airplane, Car, Horse の 3 クラスについてそれぞれ 100 枚のサブセットを評価に用いる．それぞれのカテゴリについて，18,11,7 枚のノイズ画像が含まれており，このノイズ画像は各物体を画像中に含まないものとなっている．ノイズ画像は学習の頑健性を確認するためのものであり，検出精度の評価においてはノイズ画像は用いない．

PASCAL VOC 2007 Dataset

VOC データセット [55] については従来研究にならい，20 カテゴリの trainval セットに含まれる画像を対象とする．しかし，アノテーションされている物体全てに difficult もしくは truncated が付与されている画像については，データセットから除外し，対象外とする．VOC データセットには複数の対象物体が同時に含まれるような画像が多数含まれているため，画像毎に評価対象となる主物体が一意に定まらないことに注意したい．例えば，人と自転車が同時に写る画像がある場合に，人画像の Co-localization を評価する際には，人領域が主物体となり，自転車画像の Co-localization を評価する際には，自転車領域が主物体となる．

ImageNet Subsets Dataset

深層特徴は ImageNet [50] に含まれる 1000 クラスを事前学習することで得られる特徴表現であるため，この 1000 クラスに含まれる物体については有用な特徴表現が獲得されていると考えられる．一方で，1000 クラスに含まれない物体については，深層特徴で捉えるのが難しい可能性がある．認識対象の頑健性を評価するため，ILSVRC の 1000 クラスに含まれない ImageNet のカテゴリを 6 つ選び出し評価に用いる．従来研究 [48, 54] にならい，Chipmunk, Rhino, Stoat, Racoon, Rake, Wheelchair の 6 物体を対象としたデータセットを用いる．

表 3.1　Object Discovery Dataset に対する単一物体検出精度評価（CorLoc[%]）

Method	Airplane	Car	Horse	Avg.
Ours (K=2)	**95.1**	**97.8**	75.3	**89.4**
Ours (K=3)	92.7	34.8	76.3	68.0
Ours (K=4)	91.5	22.5	25.8	46.6
Ours (Best K)	95.1	97.8	76.3	**89.7**
Wei et al. [48]	91.5	95.5	**77.4**	88.1
Cho et al. [43]	82.9	94.4	75.3	84.2

3.3.2　単一物体検出

まず標準的な Co-Localization 問題への適用を試みた．各物体クラス毎にデータセットを分割し，各データセットから主物体の検出を行う問題設定である．提案特徴の有効性を確認するために，事前学習を一切用いておらず，人手によって設計された画像特徴を活用するアプローチである Cho らの手法 [43]，提案手法と同じく事前学習を活用しつつ，学習対象に対しては全くの教師情報を用いない Wei らの手法 [48]，Li らの手法 [54] との精度比較を行う．さらに，学習対象に関して物体を含まない負例画像データも合わせて学習する，いわゆる弱教師あり検出問題における State-of-the-art である Wang らの手法 [56] とも合わせて比較する．

Object Discovery Dataset における精度評価を表 3.1 に示す．このデータセットはノイズ画像が入っていることが特徴であるが，提案手法は全ての画像から共通する領域を探すようなアプローチではないため，ノイズの影響を大きく受けることなく高精度な検出を実現している．このデータセットは対象物体が画像中に目立って配置されており，背景もシンプルであるために K=2 の時に最も良い精度が記録されているが，K=2 の場合には DDT と同じような処理となるために，似たような精度が記録されている．馬の画像については，馬と人が写る画像がいくつかあり，背景の複雑さが増すために，K=3 で精度最大となっている．カテゴリ馬及び車，また全体の平均において従来手法の精度を超えた．

VOC Dataset における精度評価を表 3.2 に示す．このデータセットは複数クラスの物体が 1 枚の画像に写るなどしており，背景が多様な画像データであるため，K=3 や 4 において精度が最大となることが多く，全体の平均値においては K=3 で精度最大となっている．背景の複雑さに対して球面クラスタリングが功を奏しており，完全教師なしである

表 3.2　VOC Dataset に対する単一物体検出精度評価（CorLoc[%]）

Method	Cue	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow
Ours (K=2)	P+D	64.9	61.7	61.7	18.4	11.4	67.2	58.1	86.0	7.4	67.0
Ours (K=3)	P+D	69.0	**64.4**	66.4	38.3	11.9	72.1	70.7	**89.0**	0.4	**68.0**
Ours (K=4)	P+D	70.8	60.0	67.9	**42.6**	12.5	**78.7**	42.2	88.5	0.4	67.0
Ours (K=5)	P+D	71.3	59.4	**70.1**	36.9	13.6	9.8	40.2	20.5	0.4	22.0
Ours (K=6)	P+D	74.9	61.1	**70.1**	39.0	25.0	9.0	40.7	20.5	0.4	21.0
Ours (K=7)	P+D	75.4	46.7	67.9	38.3	**25.6**	2.5	19.8	20.0	0.4	23.0
Ours (Best K)	P+D	75.4	**64.4**	**70.1**	**42.6**	**25.6**	**78.7**	70.7	**89.0**	7.4	**68.0**
Wang et al. [56]	P+N+D	**80.1**	63.9	51.5	14.9	21.0	55.7	**74.2**	43.5	**26.2**	53.4
Cho et al. [43]	P	50.3	42.8	30.0	18.5	4.0	62.3	64.5	42.5	8.6	49.0
Wei et al. [48]	P+D	67.3	63.3	61.3	22.7	8.5	64.8	57.0	80.5	9.4	49.0
Li et al. [54]	P+D	73.1	45.0	43.4	27.7	6.8	53.3	58.3	45.0	6.2	48.0

Method	Cue	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Avg.
Ours (K=2)	P+D	20.4	78.3	66.0	74.9	18.6	20.6	54.4	38.7	49.1	15.7	47.0
Ours (K=3)	P+D	2.0	80.1	73.8	**78.1**	21.0	29.4	55.9	68.9	**81.3**	18.8	**53.0**
Ours (K=4)	P+D	0.0	83.0	73.8	15.0	5.7	**39.4**	55.9	**74.5**	**81.3**	54.8	50.7
Ours (K=5)	P+D	0.0	**83.4**	73.8	15.0	6.7	**39.4**	**58.8**	67.9	80.4	57.4	41.4
Ours (K=6)	P+D	0.0	18.1	**74.3**	14.4	6.6	38.9	55.9	66.0	29.5	58.4	36.2
Ours (K=7)	P+D	**42.9**	16.6	67.5	15.5	6.6	36.1	8.8	61.3	27.7	58.9	33.1
Ours (Best K)	P+D	**42.9**	**83.4**	**74.3**	**78.1**	21.0	**39.4**	**58.8**	**74.5**	**81.3**	58.9	**60.2**
Wang et al. [56]	P+N+D	16.3	56.7	58.3	69.5	14.1	38.3	**58.8**	47.2	49.1	**60.9**	48.5
Cho et al. [43]	P	12.2	44.0	64.1	57.2	15.3	9.4	30.9	34.0	61.6	31.5	36.6
Wei et al. [48]	P+D	22.5	72.6	73.8	69.0	7.2	15.0	35.3	54.7	75.0	29.4	46.9
Li et al. [54]	P+D	14.3	47.3	69.4	66.8	**24.3**	12.8	51.5	25.5	65.2	16.8	40.0

Cho らの手法 [43] はもとより，Co-localization の State-of-the-art である Wei らの手法 [48] や Li らの手法 [54] を大きく上回る精度を記録している．さらに，負例データを用いた学習に基づく手法である Wang らの手法 [56] をも上回る精度が得られた．従来手法と特に違いが顕著なクラスとして boat や dining table があげられるが，これらは特徴的なテクスチャを持たないために，局所特徴においても深層特徴においても画像間での共起を把握しにくかったからと思われる．提案手法では，超球面上に射影した特徴の方向のみを

表 3.3　ImageNet Subsets Dataset に対する単一物体検出精度評価（CorLoc[%]）

Method	Chipmunk	Rhino	Stoat	Racoon	Rake	Wheelchair	Avg.
Ours (K=2)	81.1	93.4	78.5	**87.9**	61.6	66.0	**78.1**
Ours (K=3)	81.8	**94.4**	79.3	**87.9**	**62.2**	61.7	77.9
Ours (K=4)	83.7	**94.4**	80.2	21.2	**62.2**	64.3	67.7
Ours (K=5)	**84.7**	83.6	**81.0**	20.8	35.4	31.9	56.2
Ours (Best K)	**84.7**	**94.4**	**81.0**	**87.9**	**62.2**	66.0	**79.4**
Wei et al. [48]	70.3	93.2	80.8	71.8	30.3	**68.2**	69.1
Li et al. [54]	44.9	81.8	67.3	41.8	14.5	39.3	48.3
Cho et al. [43]	26.6	81.8	44.2	30.1	8.3	35.3	37.7

表 3.4　Object Discovery Dataset に対する前景物体検出精度評価（CorLoc[%]）

Method	Airplane	Car	Horse	Avg.
Ours (K=7)	**87.8**	**96.6**	**83.9**	**89.4**
Cho et al. [43]	81.7	94.4	71.0	82.4

用いてクラスタリングを行うため，深層特徴の大きさに相当する反応の強さが乏しい場合においてもクラスタを抽出することができている．

ImageNet Subsets Dataset における精度評価を表 3.3 に示す．このデータセットではノイズ画像を含んでおらず，また対象物体のみが写っている比較的シンプルな画像が含まれているが，対象物体として事前学習に用いた ILSVRC1000 クラスに含まれていないものが選ばれており，事前学習したクラスと異なる物体に対しても深層特徴の恩恵が受けられるかを評価することができる．表に示された結果から，このデータセットにおいても他のデータと同様に従来手法を大きく上回る精度が得られた．特に Rake クラスにおいて大きな精度向上が見られる．K=2 の場合においても DDT [48] と大きな精度差が起きているが，これはレーキと同時に現れる特徴的な物体として人が存在しており，PCA によって得られる主成分方向は人特徴の影響を強く受けるのに対して，球面クラスタリングでは，人特徴の影響を受けることなく Rake に関する特徴をうまく分離できているものと考えられる．

表 3.5　VOC Dataset に対する前景物体検出精度評価（CorLoc[%]）

Method	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	
Ours (K=6)	**73.7**	**55.0**	**60.6**	**46.8**	**10.8**	**70.5**	**62.6**	**70.0**	**20.5**	**67.0**	
Cho et al. [43]	40.4	32.8	28.8	22.7	2.8	48.4	58.7	41	9.8	32	
Method	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Avg.
Ours (K=6)	**36.7**	**71.1**	**68.9**	**73.3**	4.6	7.2	**52.9**	**57.5**	**71.4**	8.1	**49.5**
Cho et al. [43]	10.2	41.9	51.9	43.3	**13.0**	**10.6**	32.4	30.2	52.7	**21.8**	31.3

表 3.6　ImageNet Subsets Dataset に対する前景物体検出精度評価（CorLoc[%]）

Method	Chipmunk	Rhino	Stoat	Racoon	Rake	Wheelchair	Avg.
Ours (K=11)	85.3	94.4	75.1	87.4	58.4	64.8	77.6

3.3.3　前景物体検出

提案手法はシンプルなクラスタリングのアプローチを採用しているため，簡単に多数クラスの Co-Localization へ対応させることができる．複数のクラスが混ざったデータセットに対して，教師情報を全く用いずに各画像の主物体位置を検出する問題に対して最高精度を示している Cho らの手法 [43] と比較する．データセット全体で主物体を示すクラスタ番号を決めることはできないため，画像毎に式 (3.2) のクラスタ強度を算出し，最も強度の高いクラスタを主物体として採用する．

表 3.4，3.5，3.6 に，それぞれ Object Discovery Dataset を用いた結果，VOC Dataset を用いた結果，ImageNet Subsets Dataset を用いた結果を示す．クラスタ数 K については，それぞれ 2-7，2-26，2-10 の範囲で最も平均精度が高くなった値を用いた結果を示す．表より，この問題設定においても Cho らの手法 [43] に比べて大幅な精度向上を実現していることがわかる．また，ImageNet Subsets Dataset についても単一カテゴリの問題設定とほぼ同様の認識精度が示されており，多数カテゴリにおいて非常に頑健に機能する手法であることが示された．さらに，Object Discovery Dataset の Horse クラスを見ると単一カテゴリにおける精度を上回る結果が得られた．Horse を含まない画像が追加されたことで，背景に含まれる人などの特徴をより良く捉えられるようになったと考えられる．また，VOC Dataset を用いた検出結果の例を図 3.4 に，ImageNet Subsets Dataset を用いた検出結果の例を 3.3 に示す．図 3.4 の上段の cat のように背景との違いが曖昧で捉え

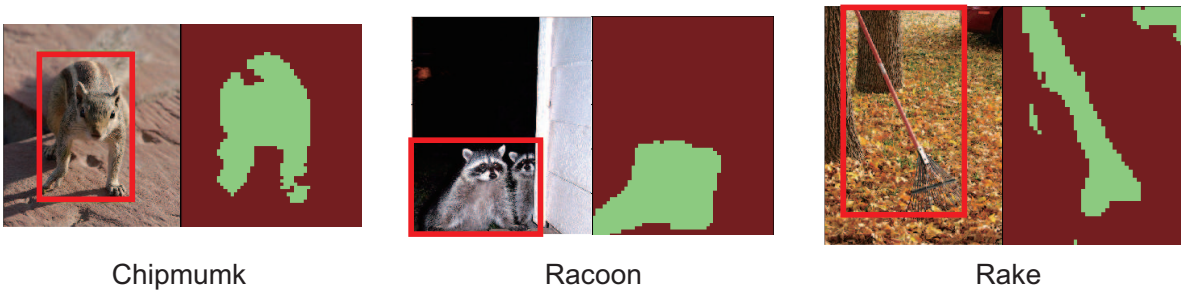


図 3.3　ImageNet Subsets Dataset において抽出された主物体領域（図右の緑領域）と対応する検出矩形（図左の赤枠）

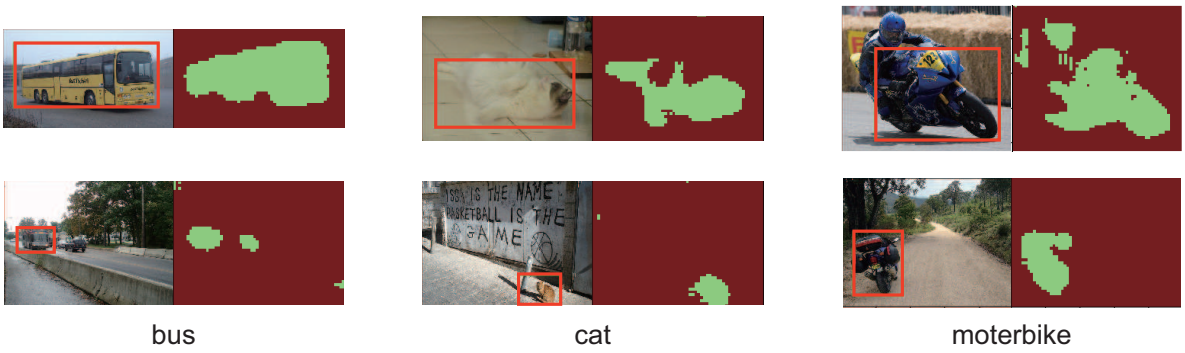


図 3.4　VOC Dataset において抽出された主物体領域（図右の緑領域）と対応する検出矩形（図左の赤枠）

表 3.7　Object Discovery Dataset に対する複数物体検出精度評価（CorLoc[%]）

Method	Airplane	Car	Horse	Avg
Ours (K=7)	89.0	95.5	83.9	89.5

にくい物体や，図 3.4 下段のように画像全体に対して面積の小さな物体についても精度良く検出できていることがわかる．

3.3.4　複数物体検出

提案する教師なし物体検出手法によって前景物体の検出ではなく，複数物体のクラス識別を同時に行いながら各物体を検出する実験も行った．これは本研究の最終目的としている複数物体の教師なし領域抽出に限りなく近い問題設定であるが，Co-localization においても複数物体のクラス毎の検出についてはこれまでに取り組みのない困難な問題である．

表 3.8　VOC Dataset に対する複数物体検出精度評価（CorLoc[%]）

Method	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	
Ours (K=39)	63.2	51.7	1.5	13.5	9.7	2.5	19.4	83.0	1.2	1.0	
Method	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Avg.
Ours (K=39)	38.8	0.4	22.3	38.5	29.1	35.6	1.5	58.5	58.9	54.3	29.2

表 3.9　ImageNet Subsets Dataset に対する複数物体検出精度評価（CorLoc[%]）

Method	Chipmunk	Rhino	Stoat	Racoon	Rake	Wheelchair	Avg
Ours (K=17)	2.3	92.5	11.8	66.7	57.1	42.6	45.5

3 つのデータセットに対して適用した結果をそれぞれ表 3.7,3.8,3.9 に示す．クラスタ数 K については，それぞれ 4-9, (20,25,30,35,40), (6,9,12,15,18,21) の中から最も平均精度が高くなった値を用いた結果を示す．評価指標はこれまでの実験と同様に各物体クラスについて正しく検出できている画像の割合を用いており，画像に写っていない物体クラスの検出枠が現れるような誤検出については評価しない．また，各クラスに対するクラスタ ID の割り当てについては，最も当てはまりの良いクラスから貪欲に割り当てることによって評価している．

表 3.7 より Object Discovery Dataset では前景物体検出設定とそこまで劣らない精度が出ているが，これはクラス数が 3 つと少ないことに加えて，各クラスの見た目が大きく異なっているためだと思われる．一方で VOC Dataset と ImagenNet Subsets Dataset では，前景物体検出設定に比べて大きく精度が低下している．これは例えば Chipmunk と Racoon は非常に似通った物体であるため，これらを分離することが難しく，精度劣化の要因となっていると考えられる．精度は劣化しているもののその傾向は VOC Dataset と ImageNet Subsets で共通しており，事前学習された 1000 カテゴリに含まれていない物体についても教師なしで発見できることが示唆されている．

VOC Dataset において精度良く認識できている物体としては aeroplane, bicycle, bus, cat などがあり，それぞれ boat, motorbike, car, dog などとの分離が難しいために片方の精度が著しく低下している．このような類似するクラスを見分けるためには，より細かな粒度で分類された特徴量を用いてクラスタ間での同一性を評価する手法が必要になると考えられる．

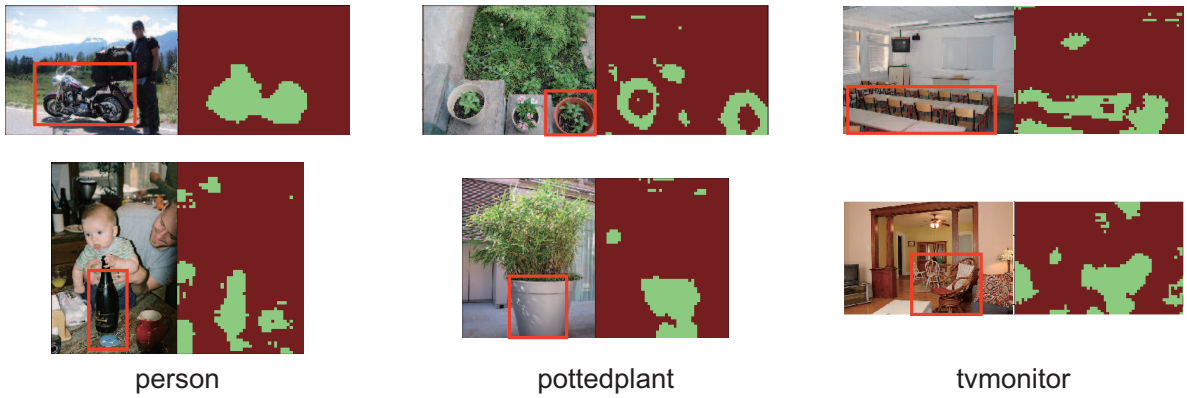


図 3.5　VOC Dataset における検出失敗の例

3.3.5　提案手法の限界

提案手法における検出失敗例を図 3.5 に示す．図は VOC dataset に対する前景物体検出実験において CorLoc 精度の低い 3 つのクラス（person, pottedplant, tvmonitor）を示している．person の例では，moterbike や bottle が誤って検出されており，画像に複数の物体が写る場合に主物体の選択において person が選ばれていないことが分かる．また，pottedplant の例では，植木鉢と植えられた植物が別のクラスタとなってしまっており，過分割によって potted plant 全体を捉えられていない．さらに，tvmonitor の例では，画像に写る monitor 領域を一定のクラスタとして捉えているものの，そのサイズが小さいために背景ノイズが検出枠として選ばれてしまい失敗となっている．

提案する特徴表現単体での教師なし物体検出における性能としては，前述のような課題が残されているものの，データから同一カテゴリの領域を抽出するための特徴表現として良好な性能が得られることが各種実験によって確認された．第 6 章にて空間的配置関係を考慮した境界特徴と組み合わせることで複数物体を含む画像集合に対してクラス毎の領域抽出を試みる．

第 4 章

空間的配置関係の学習に基づく境界特徴表現

本章では，空間的配置関係を適切に抽出するための画像特徴表現を提案する．本論文において想定する空間的配置関係として，上下・左右・前後の関係が含まれ，このうち上下左右については境界線の線形状から得られるものと考えられるが，前後関係については直感的な特徴表現の定義が難しい．本章において提案する境界特徴では，この境界線における配置関係を機械学習によって推定する推定器を構築し，その中間出力を画像特徴として抽出するアプローチをとる．この境界線の形状及び周辺の前後関係を推定する問題を「遮蔽境界線検出」と呼ぶこととする．遮蔽境界とはその境界線において奥行きの違いが発生している境界のことを指し，主に物体の輪郭を成す場合が多い．この遮蔽境界を検出し，奥行きの変化がどちら側にあるかを推定することが，本章における研究課題となる．多様な画像に対して遮蔽境界線を検出できるような推定器を学習することで，物体に依らずに適用可能な境界特徴を獲得することができる．

4.1　遮蔽境界線検出

画像を物体毎に領域分割する画像セグメンテーション技術は画像理解における基礎的な処理であるが，未知の物体に対してその物体領域を切り出すことは非常に困難な問題とされている．近年，こうした領域抽出の問題に対して領域の実空間での幾何的な関係性や深度情報を合わせて推定することで，物体に関する知識を用いることなく物体領域を抽出するアプローチが取り組まれている．画像から境界線を検出する問題に関しては，古くから様々な手法が提案されているが，物体領域の抽出に関しては，画像に存在する境界線のな

かでも物体輪郭線，すなわちある物体によって背後の物体の遮蔽が発生している境界線を検出しなければならない．こうした時に境界の遮蔽関係（境界のどちら側が手前に存在する物体で，どちら側が遮蔽された物体であるか）を考慮することは高精度な領域抽出のために有用であると考えられる．なお，この検出すべき境界を遮蔽境界と呼ぶ．本研究では，画像から遮蔽境界を検出し，検出された境界での遮蔽関係を推定することを目的とする．

境界線検出の分野では Berkeley Segmentation Dataset (BSDS) [12] をベンチマークとして様々な手法が競われている．なかでも Arbelaez らの手法 (gPb)[12] が良好な結果を示しており，多くの画像理解研究において特徴抽出や事前処理として用いられている．しかし，BSDS では遮蔽境界でない物体内部のテクスチャとして存在している境界線も正解に含まれている場合があり，必ずしも物体輪郭抽出の精度を評価しているとは言い難い．そのため，BSDS で高精度を示している手法であっても遮蔽境界線検出においては精度が大幅に低下することが想定される．

一方で，最近では境界線検出と同時に境界での遮蔽関係を推定することで遮蔽境界線検出を行う研究が取り組まれている [2, 57]．遮蔽境界検出のアプローチは大きく分けて 2 つあり，事前に gPb [12] などの境界線検出を適用し，得られた境界線に対してその遮蔽関係を推定するアプローチ [58, 59] と遮蔽関係の推定と同時に遮蔽境界の検出を行うアプローチ [2, 57] がある．前者のアプローチでは，人手によって与えられた境界線に対して遮蔽関係を精度良く推定できているが，その推定精度が与えられる境界検出の精度に大きく影響されるという問題がある．Ren らの手法 [58] では，人が正しい境界線を与えた場合の遮蔽関係推定精度が 82.8 %であったのに対して，gPb による境界検出結果を用いた場合推定精度が 68.9 %まで落ちることが報告されている．それに対して，後者のアプローチでは遮蔽関係の推定と境界線の検出を同時に行うことで，遮蔽の発生している物体輪郭線を精度良く抽出することができ，かつ遮蔽境界線検出が改善することで遮蔽関係の推定精度も向上することが期待できる．本研究においても同様に，遮蔽境界の検出と遮蔽関係の推定を同時に行うことで両者の精度向上を狙う．しかし，現在良好な結果を示している Maire の手法 [2] は，局所的な遮蔽関係の推定結果を Spectral Partitioning によって大域的な領域分割に用いることで境界線検出を行っているため，計算コストが非常に高いという問題がある．また，Hoiem ら [57] は画像の領域に対して地面や空，直立物体などの幾何的な意味付けをすることでその関係によって遮蔽境界を検出する手法を提案しているが，推定対象として風景画像に限定した手法となっており多様な画像に対しての適用が難しい．

本研究では，こうした課題を解決するため Lim らの提案する輪郭形状の中間特徴表現

[13] を用いて高速かつ高精度に遮蔽境界検出を行う手法を提案する．

4.2 提案手法

4.2.1 Sketch Tokens

従来，gPb [12] などの高精度な境界線検出手法に関して計算コストが高いという課題があったが，Lim らの手法 [13] では，Sketch Tokens と呼ばれる境界線の部分形状を表す中間特徴表現を導入し，境界線形状識別器を学習することで境界検出の精度を保ちつつ大幅な高速化を実現している．提案手法においても同様の中間特徴表現を用いるため，Sketch Tokens の詳細について以下に解説する．

Lim らは境界線を検出するにあたり，従来のような輝度値の勾配強度を評価するのではなく，物体検出のような枠組みにおいて境界線の部分的な形状（Sketch Tokens）を検出するアプローチをとっている．検出すべき境界線形状は境界線画像のデータセットから代表的な線形状を抽出する．部分形状の識別を画像全体の各画素に対して行うことで，各部分形状に関する境界らしさを得ることができ，それらを統合することで高精度な境界検出を実現している．また，識別器として Random Forest [60] を用いることで高速に境界線形状の識別を行っている．

Lim らは Maire の手法 [2] や Hoiem らの手法 [57] のように画像全体に関して境界らしさを評価する大域的な処理を行っていないが，一方でパッチ画像に基づく特徴表現と境界線形状を中間表現とした分割統治のアプローチによって従来手法と同程度の検出精度を実現している．さらに，処理コストの大きい大域的な処理を行わないことで大幅な高速化を実現している．

しかし，Lim らの手法は遮蔽境界検出に特化したアプローチではないため，物体輪郭を精度良く捉えることはできない．また，境界での遮蔽関係の推定を実現するためには，得られた境界線に対して改めて遮蔽関係推定を行う必要がある．以下に，Sketch Tokens と同様の中間特徴表現を活用し，物体輪郭を捉えるための遮蔽境界検出と境界上での遮蔽関係の推定を同時に行う提案手法 [61] について解説する．

4.2.2 遮蔽関係を考慮した境界線特徴

提案手法では，高精度な遮蔽境界検出を実現するため遮蔽関係を考慮した輪郭形状毎に画像の見えを学習する．具体的には，境界線が付与された訓練データから図 4.1(c) のような様々な輪郭形状クラスタを抽出する．ここで，訓練データには遮蔽境界線に加えて線

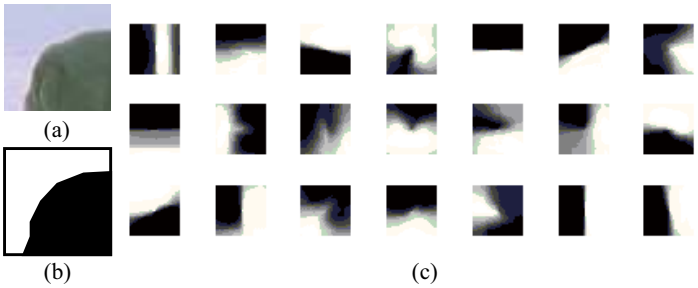


図 4.1　遮蔽関係を表現するパッチ画像と抽出されるクラスタの例．(a) 一定サイズに切り出された画像，(b) 画像に対応する遮蔽関係を表す 2 値画像 (c) クラスタリングされた遮蔽関係画像の平均画像

上の遮蔽関係（どちら側が手前でどちら側が奥か）が付与されているが，その遮蔽関係を輝度値によって表現した 2 値画像を輪郭形状パッチとする．図 4.1(a) の画像に対応する遮蔽関係の表現として，図 4.1(b) 中の黒い領域が手前側，白い領域が奥側を示している．また，輪郭形状によってはパッチ内に 3 つ以上の領域を含む場合もある．領域が 3 つであった場合，その遮蔽関係に応じて 3 値画像を輪郭形状パッチとし，4 つ以上の場合は訓練データから除外する．輪郭形状パッチのクラスタリングは Lim らの手法と同様にパッチ画像から DAISY 特徴 [62] を抽出し，kMeans クラスタリングを用いて行う．抽出するクラスタ数は N_t 個とする．図 4.1(c) はクラスタリングによって抽出される各クラスタの平均的な輪郭形状を示している．

次に，得られた輪郭形状クラスタ及び非境界クラスを識別クラスとする識別器を学習する．訓練データに含まれる境界線上の画素はクラスタリングによっていずれかの輪郭形状クラスタに属しているため，各画素を中心とした一定サイズのパッチ画像を対応するクラスのサンプルとして抽出する．非境界クラスのサンプルについては，境界線上でない画素からランダムに抽出する．パッチ画像から得られる画像特徴については，Lim らの手法 [13] と同様に Dollar らの提案する画像特徴 [63] に基づき Self-similarity 特徴を導入したものを用いる．この画像特徴は勾配特徴やテクスチャ特徴だけでなく，パッチを区切ったセル間の類似性を表現しており，この Self-similarity 特徴によって遮蔽境界と関係の強い影やぼけといったテクスチャの変化を抽出できると考えられる．識別器の学習は Random Forest 識別器 [60] を用いる．Random Forest はデータのばらつきに頑健な識別器とされており，また多クラス識別を高速に行うことができるという性質を持つ．

境界の検出を行う際には，入力画像の各画素に対して画素周辺のパッチ画像を用いて学習された輪郭形状識別器による識別を行う．識別器の出力は，式 (4.1) に従い，各画素 i

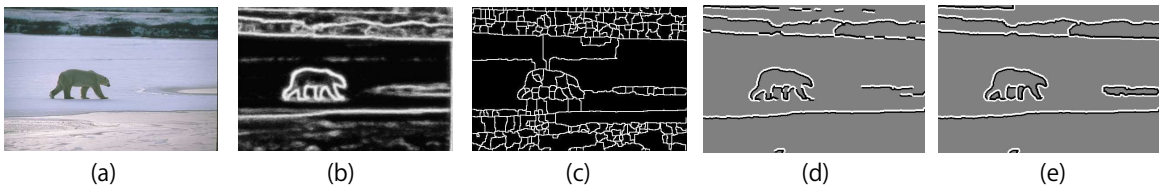


図 4.2　提案する遮蔽境界線検出手法の処理フロー : (a) 入力画像，(b) 各画素について境界線確率の算出，(c)Superpixel 分割による境界線候補の算出，(d) 各境界線候補について遮蔽関係の識別，(e) 画像全体全体での遮蔽関係の最適化

のパッチ画像 $\boldsymbol{x}_i$ が各輪郭形状クラスタ j に属する確率値 t_{ij} を要素とするベクトル $\boldsymbol{t}_i$ を得ることができる．

$$\boldsymbol{t}_i = F_s(\boldsymbol{x}_i), \tag{4.1}$$

ここで，F_s は学習された Random Forest 識別器による出力を示す．このようにして得られる輪郭形状を示す中間特徴表現 $\boldsymbol{t}_i$ を Occlusion Tokens と呼ぶ．

さらに，式 (4.2) に従って非境界クラス $j = 0$ を除いた全輪郭形状クラスについて足し合わせることで各画素の境界線確率 p_i を得ることができる．

$$p_i = \sum_{j=1}^{N_t} t_{ij} \tag{4.2}$$

各画素の境界線確率 p_i を算出した結果の例を図 4.2(b) に示す．

4.2.3　境界線特徴を活用した遮蔽境界線検出

前述の輪郭形状識別器による出力である Occlusion Tokens を用いて遮蔽境界線の検出及び遮蔽関係の推定を行う．輪郭形状識別器によってすでに図 4.2(b) に示されるような境界線らしさは得られているが，この結果から直接遮蔽境界線の遮蔽関係を得ることはできない．また，全ての輪郭形状についての境界らしさを足し合わせた出力となっているため，遮蔽関係がはっきりしないような物体内部の境界線についても高い境界確率が出力されていると考えられる．提案手法では，輪郭形状識別器による中間特徴表現 Occlusion Tokens を用いてさらに識別を行うことで遮蔽境界の遮蔽関係を推定する．まず，遮蔽関係の推定にあたって画像中の遮蔽境界線となり得る境界線候補の線を獲得する．境界線候補の抽出は，セグメンテーション手法の一種である Watershed 変換を用いる．図 4.2(b) のような輪郭形状識別器の出力（式 (4.2)）による境界確率画像に対して Watershed 変換

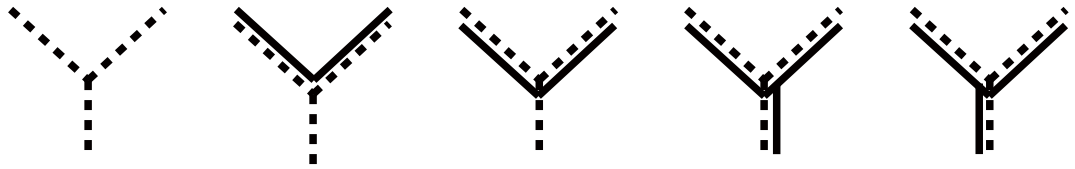


図 4.3　交点で取りうる遮蔽関係の組み合わせ

を適用することで，確率値が極大値を示す部分を抽出し，図 4.2(c) のような細線化された境界線画像を得ることができる．

次に，ここで得られた境界線候補に対して，遮蔽境界の有無及び遮蔽関係の識別を行う．この識別は境界線候補上の各画素に対して行い，前述の Occlusion Tokens$\boldsymbol{t}_i$ を入力とする．識別器の学習は境界線情報付きの訓練データを用いるが，訓練画像に対して学習済みの輪郭形状識別器を適用し，得られた境界線候補のうち遮蔽境界から一定距離以内であった画素を遮蔽境界のサンプルとして用いる．さらに，抽出されたサンプルをその境界線の向きによって N_a 個の方向に分ける．境界線画素は N_a の方向と 2 つの遮蔽関係の組み合わせで $N_a \times 2$ のクラスに分けられる．また，境界線候補でありながら遮蔽境界でなかったものを非境界線のサンプルとして用い，これらもまた境界線の向きに応じて N_a 方向のクラスに分けられる．こうして得られた訓練サンプルを用いて，各方向について N_a 個の識別器を学習する．識別手法として輪郭形状識別と同様に Random Forest を用いる．

得られた識別器を用いて境界検出を行う際には，境界線候補の各画素 i に対して，境界線の向き θ_i を算出し，対応する識別器を適用する．式 (4.3) に従って，Occlusion Tokens$\boldsymbol{t}_i$ から遮蔽関係の確率値を要素とするベクトル $\boldsymbol{o}_i$ が得られる．

$$\boldsymbol{o}_i = F_o^{(\theta_i)}(\boldsymbol{t}_i) \tag{4.3}$$

ここで，$F_o^{(\theta_i)}$ は方向クラス θ_i に属する遮蔽関係識別器の出力を示す．

得られた 3 つのクラスの確率値 $\boldsymbol{o}_i$ を同一境界線の全画素に関して掛け合わせることで，該当する境界線の遮蔽関係確率ベクトル $\boldsymbol{q}_l$ とする．

ここで，l は境界線候補，B_l は境界線候補 l に含まれる画素集合を示す．また，k は遮蔽関係を示す 3 つのクラスを表す．各境界線候補 l に関して q_{lk} が最大となる遮蔽関係 k を選ぶことで，図 4.2(d) のような出力を得ることができる．

最後に，各境界線候補の遮蔽関係確率ベクトルを用いて画像全体での最適化処理を行う．画像に存在する遮蔽境界の性質を考えた時，一続きの境界線において遮蔽関係が入れ替わるなど通常発生し難い状況がいくつか考えられる．図 4.2(c) のような境界線候補に

対して遮蔽関係のラベル付けを行う場合，境界線候補の交点において取りうる遮蔽関係の組み合わせは図 4.3 に示す 5 つのパターンとこれらを回転させてものに限られる．図は破線が非境界，実線と破線の二重線が遮蔽境界を表しており，二重線の実線側が手前にある関係を示している．そこで，提案手法ではこれらの組み合わせのみを用いて，遮蔽関係の同時確率が最も高くなるようなラベル付けを求める．すなわち，図 4.3 に示される制約の中で次式を最大化するラベルの組み合わせ $\hat{\boldsymbol{K}}$ を求める．

$$\hat{\boldsymbol{K}} = \arg\max_{\boldsymbol{K}} \prod_{l} q_{lK_l} \tag{4.4}$$

この最適化問題は整数計画法によって解くことができる．整数計画ソルバーとして GLPK [64] を用いた．

この大域的最適化処理によって得られる推定結果の例を図 4.2(e) に示す．図 4.2(d) に示される最適化前の推定結果に比べて，大域的最適化によって同一線上の遮蔽関係の矛盾や遮蔽境界が途切れてしまう問題が解決されていることがわかる．

4.3 遮蔽境界線検出実験

提案する遮蔽境界検出手法を評価するため実験を行った．評価には BSDS [12] を用いたが，BSDS500 は境界線に遮蔽関係は付与されておらず，また遮蔽境界以外の物体領域内部の境界線も正解として含まれているため，遮蔽境界の学習・評価には BSDS の訓練セット 200 枚に遮蔽関係を付与したデータセット [65] を用いた．各パラメータについては，パッチ画像のサイズは一辺 35 ピクセル，輪郭形状のクラスタ数 N_t は 150，境界方向の分割数 N_a は 4 とした．

実験結果を表 4.1 に示す．遮蔽境界が付与された画像 200 枚のうち，100 枚を訓練データ，100 枚をテストデータとして用いた．Maire の手法 [2] については公開されているプログラムを用いて実験を行った．表に示す検出精度は，BSDS で用いられる境界画素の検出に関する F 値 (ODS) によって評価している．境界検出精度は，遮蔽関係を考慮せず遮蔽境界を示す画素が検出される精度を示しており，遮蔽関係検出精度は遮蔽関係を含めて正解データと合致する精度を示している．また，遮蔽関係正解率は正しく検出された遮蔽境界において正しく推定された遮蔽関係の正解率を示している．計算時間は画像 1 枚（321×481）あたりの推定にかかる平均時間を示している．

表 4.1 より，遮蔽関係と同時に遮蔽境界検出を行う従来手法である Maire の手法に対して，同等程度の遮蔽境界検出精度を実現しており，遮蔽関係の検出精度に関しては精度

表 4.1　遮蔽境界検出精度，遮蔽関係推定精度及び計算時間の評価

	Ours	Maire [2]	Ren [58]	Palou [59]
F	55.0%	54.8%	-	-
F (with f/g)	40.5%	36.3%	-	-
f/g only	73.5%	66.4%	68.9%	71.3%
time	14s	180s	-	-

が 4.2% 向上している．検出された遮蔽境界に対するまた，遮蔽関係推定の正解率では，7.1% もの精度向上を示している．Ren らの手法 [58]，Palou らの手法 [59] については，境界線を先に推定し，検出された境界線に基づいてその遮蔽関係を推定する手法であるが，推定精度の文献値を記載した．こうしたアプローチとの比較においても，従来の最高精度を示す Palou らの手法を 2.2% 上回った．また，画像 1 枚あたりの計算時間では，高速な輪郭形状識別器と簡易な大域的処理によって，Maire の手法に比べて約 10 倍の高速化を実現している．

図 4.4 に提案手法によって境界検出を行った結果の例を示す．遮蔽関係付きデータの 200 枚全てを使って学習し，BSDS のテストセットに対して手法を適用したものである．図の奇数行に提案手法による推定結果，偶数行に Maire の手法による推定結果を示す．全体として，従来手法に比べて提案手法ではいくらか検出漏れが見られるものの誤検出が抑えられ精度良い遮蔽関係推定が行えていることが見て取れる．これは，提案手法がパッチ画像内部の Self-similarity を特徴とした識別を行っているためであり，虎や蝶の柄のような細かい繰り返しをテクスチャとして捉えることで誤検出が抑えられている．また，輪郭形状識別結果を中間特徴とする遮蔽関係の識別を行うことによって遮蔽関係が明確でない境界線が遮蔽境界として検出されにくく，同様に物体内部や背景に含まれるエッジの誤検出が抑えられている．

図 4.5 に提案手法による遮蔽境界検出が難しい例を示す．(a) は物体輪郭が見えの変化として顕著に表れていない例であり，局所的には物体内部のテクスチャと境界部分とを識別することができない．また，(b) は物体内部のテクスチャに強いエッジが多く発生している例であり，逆に物体内部に遮蔽境界が誤検出されてしまう．(c) は物体輪郭が複雑な形状をしている例であり，その複雑な輪郭形状を捉えきれないため，境界として識別することができない．(a) や (c) の例を解決する指針としては，境界部分の識別と合わせて物体内部のテクスチャを用いた識別を行うアプローチが考えられる．前景物体領域と背景領域のテクスチャの違いを捉えられれば，境界線として明確でない物体輪郭も検出すること

図 4.4　遮蔽境界検出と遮蔽関係推定結果の比較．奇数行に提案手法による結果，偶数行に Maire [2] による結果を示す．赤線が奥側，緑線が手前側として遮蔽関係を表現している．

図 4.5　提案手法による遮蔽境界検出が難しい例

ができるだろう．また，(b) の例については，大域的なテクスチャを捉えることで解決できると考えられる．提案手法では，画像パッチによって境界の識別を行っているため，図 4.4 における虎や蝶の模様のような細かなテクスチャを捉えることができるが，同様にしてより大きなテクスチャを捉えることができれば物体内部テクスチャの誤検出を抑えることができるだろう．

4.4　提案する境界特徴

本章では遮蔽境界線検出を課題として設定し，精度良く境界線を見つけ，またその前後関係を推定するための新たな特徴表現として Occlusion Tokens を提案した．局所的な境界線の形状情報とその境界線における前後関係に基づく 150 のクラスタを設定し，そのクラスタに属する識別スコアを特徴量とすることで，輪郭形状と遮蔽関係に特化した特徴が得られている．また，実験によって提案特徴が遮蔽境界検出に有効であることが示された．

第 6 章では，提案する境界特徴を第 3 章にて提案した領域特徴と組み合わせて教師なし物体領域抽出を試みる．

第 5 章

空間的配置関係を考慮したトピックモデルによる領域抽出

本章では，教師なし物体領域抽出の具体的なアプローチについて述べる．提案手法は，変化を伴う多種の物体カテゴリを抽出するため，確率的な生成モデルをベースとする．観測される画像特徴が潜在的な物体カテゴリから生成されることを表現したモデルとして，トピックモデル [66] を用いた物体領域抽出手法 [3] が提案されているが，本章ではこれに基づいて物体領域間の空間的な配置関係を表す確率変数として境界トピックを導入し，領域の類似性と空間的な配置関係を同時に考慮するモデルを提案する．

5.1　トピックモデルによる物体領域抽出

トピックモデルは，自然言語処理において何らかの潜在的なトピックに基づいて文書が生成されることを仮定した生成モデルであり，Latent Dirichlet Allocation (LDA) [66] が広く知られている．画像認識の分野においても，画像特徴を文書内の単語の集合とみなすことで，トピックモデルの適用が可能であり，物体領域抽出に適用することができる．

5.1.1　LDA を用いたトピック抽出

トピックモデルの代表的な手法である LDA を用いて画像内局所特徴のトピックを抽出する枠組みを紹介する．トピックモデルでは，各画像を Visual Word と呼ばれる量子化された局所特徴の集合として扱い，局所特徴集合として表された複数の画像から，各特徴の出現頻度に従い画像特徴と関連づいているトピックを自動で抽出する．このトピックを

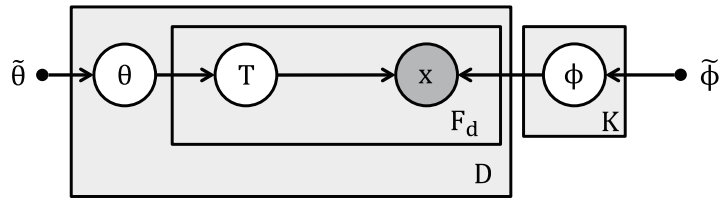


図 5.1　LDA のグラフィカルモデル．画像集合を D，画像内の局所特徴集合を F_d で表す．x は Visual Word，T は Word 毎に設定されるトピックを示しており，ϕ はトピックごとの Word 分布，θ は画像内のトピック分布，$\tilde{\phi}, \tilde{\theta}$ はそれぞれ ϕ, θ の事前分布を支配するパラメータである．

‘人物’や‘車’などの物体ラベルに対応させることで，一般物体認識手法として応用することができる．具体的には，LDA は図 5.1 のようなグラフィカルモデルによって表すことができる．グラフィカルモデルとは，変数同士の確率的な関係性をグラフによって表したものであり，ある変数に応じて生成される変数が有向グラフによって示されている．また，図中の白い丸が確率変数，灰色の丸が観測された値，黒い点があらかじめ設定されたパラメータを示す．矩形で囲まれた領域はその中にある変数が同様に複数存在していることを示しており，その個数が右下に記されている．このモデルに従うと各変数は次のように生成される．まず，画像毎にトピックの出現頻度を示す $\boldsymbol{\theta}$ が生成される．この $\boldsymbol{\theta}$ はパラメータ $\tilde{\boldsymbol{\theta}}$ によって定義されるディリクレ分布に従うとする．画像毎のトピックの出現頻度が決まると，その $\boldsymbol{\theta}$ に従って各 Visual Word のトピック T が生成される．最後にトピック T が与えられると，トピックに対応した単語出現分布 $\boldsymbol{\phi}$ に従って Visual Word x が生成される．各トピック毎の単語出現頻度を示す $\boldsymbol{\phi}$ は，パラメータ $\tilde{\boldsymbol{\phi}}$ を持つディリクレ分布によって生成される．LDA では，このモデルに従って画像内の Visual Word が生成されていることを仮定し，観測情報として多くの Visual Word が与えられたときに各トピックの事後確率を最大化することでトピックの推定を行う．しかし，ある変数について事後確率を求めるためには，他の確率変数をそれぞれ積分消去する必要があり，大変煩雑な式になる．そのため，サンプリング法や変分ベイズ法などの近似的な手法によって変数の事後分布を取得するアプローチが一般的である．大量の画像からトピックを抽出する流れとして，まず対象画像集合から Visual Words を抽出し，モデルにしたがって $\boldsymbol{\phi}$ の事後分布を計算する．得られた事後分布を最大化する $\boldsymbol{\phi}$ を求めることで，Visual Word $\boldsymbol{x}$ とトピック T との関係性を得ることができる．得られたパラメータを用いて，画像のトピックを推定する際には，学習した $\boldsymbol{\phi}$ の事後分布を $\boldsymbol{\phi}$ の事前分布として用いて，トピック T の事後確率を最大化すれば良い．

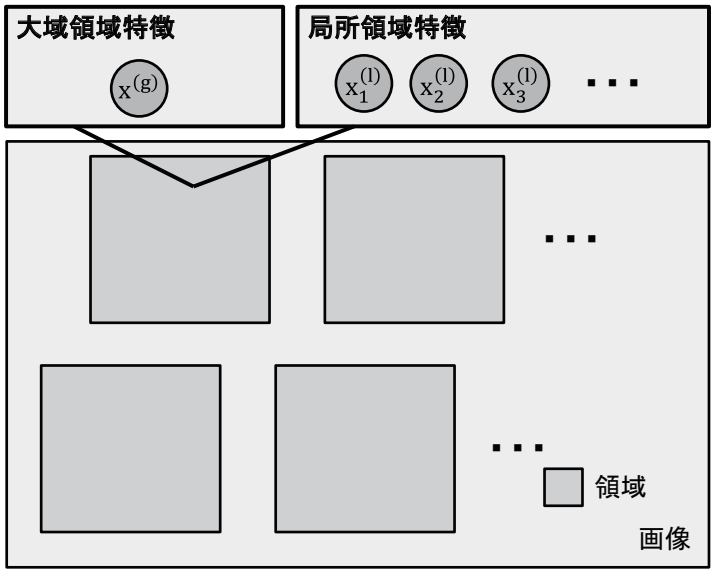


図 5.2　Spatial-LTM モデルにおける画像内小領域と観測される特徴の関係.

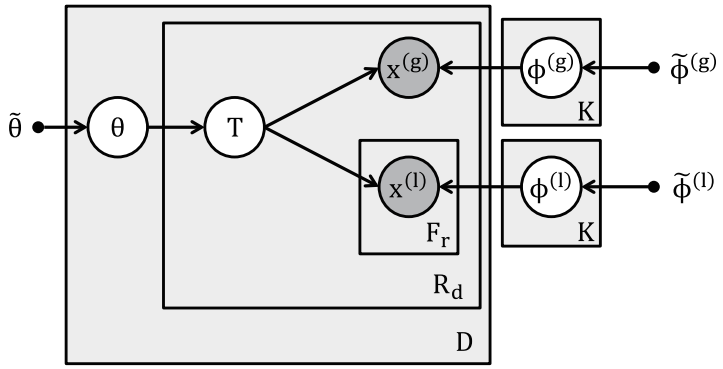


図 5.3　Spatial-LTM [3] のグラフィカルモデル．R_d は画像中に含まれる領域集合を示す．トピック T は領域毎に設定され，パラメータ $\boldsymbol{\phi}^{(g)}$ に従って大域領域特徴 $\boldsymbol{x}^{(g)}$ を，パラメータ $\boldsymbol{\phi}^{(l)}$ に従って局所領域特徴 $\boldsymbol{x}^{(l)}$ を生成する．

5.1.2　領域ベースのトピックモデル

LDA による画像解析では，全ての画像特徴に対して独立にトピックを推定しており，画像内での Visual Word の共起性のみを用いてトピック抽出を行っている．Bag of Visual Words と呼ばれるこうしたアプローチは，画像特徴の局所的なマッチングを用いず，画像全体としての傾向を用いることで，物体向きの変化や模様の変化といった局所的な見え

の変化に対して頑健に画像全体のトピック推定を行うことができる．その一方で，同一物体の画像特徴が画像内で連続して存在していることは，画像認識において強力な情報であり，画像特徴の配置関係を用いないことは大きな欠点にもなりうる．

こうした問題に対して，領域ベースのトピックモデルによる画像セグメンテーション手法が提案されている [3, 67]．Cao らの提案する Spatial-LTM [3] では，図 5.2 のように各画像を superpixel に分割することで，superpixel 内の局所特徴は同一のトピックを持つという情報を組み込んでいる．superpixel 内には局所特徴 $\boldsymbol{x}^{(l)}$ が複数存在し，また領域全体から得られる特徴 $\boldsymbol{x}^{(g)}$ も観測することができる．同一領域内の $\boldsymbol{x}^{(l)}$ については，同一トピックとなるように設定し空間的に連続する局所特徴が同一トピックとして抽出されるよう学習する．グラフィカルモデルは図 5.3 のようになり，トピックに対応した観測特徴の分布としてパラメータ $\boldsymbol{\phi}^{(l)}$ と $\boldsymbol{\phi}^{(g)}$ を導入している．このモデルでは，superpixel にトピックを付与することで，その superpixel を内包している物体を同一のトピックとして得やすくなり，結果として物体と関連したトピックを抽出することができる．また，superpixel 単位でトピックを与えることで，トピック推定結果をセグメンテーション結果として見ることもできる．こうした複数の画像特徴を持つ superpixel に対してトピックを付与するアプローチはトピックモデルを用いた画像解析において様々に応用されており，最近では標準的な技術となっている [68, 69].

5.1.3 隣接する領域関係の利用

Spatial-LTM では，superpixel 毎にトピックを設定することで，領域内で一貫性を持ったトピック抽出を行うことができる．しかし，各 superpixel は独立に扱われており，領域間の関係性は特に考慮していない．このことから，Spatial-LTM は隣接する superpixel に対する一貫性はなく，superpixel 毎にまばらな推定結果を得やすい欠点を持つ．Zhao らは，Spatial-LTM に MRF を組み込み，隣接領域間のトピック変化にコストを与えるモデルを提案しているが，目立った改良は示されていない [67]．一方で，実際の画像中には物体領域と物体領域の間には特徴的な表現が現れることが多く，領域間の画像変化は画像解析において重要な情報となりうる．例えば，Martin らの手法 [70] では，領域境界の特徴を学習して画像から物体領域を抽出している．そこで，本論文では領域境界の持つ潜在的なトピックとして境界トピックを設定し，隣接する superpixel 間の配置関係をトピックとして抽出，推定するモデルを提案する [71, 72, 73]．提案するトピックモデルを RABIT-Model: Region And Boundary Integrated Topic Model と呼ぶ．推定する境界トピックは隣接する領域トピックの組み合わせと境界線の画像特徴に応じた性質を表して

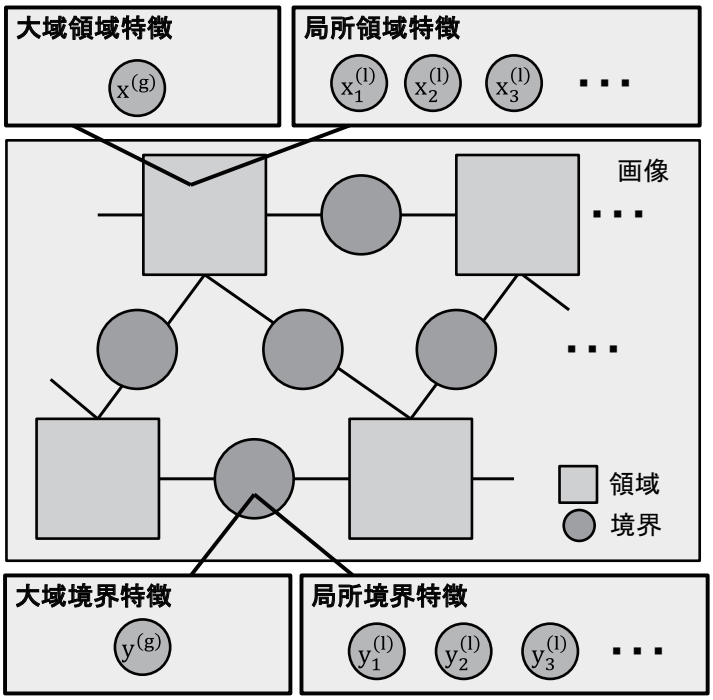


図 5.4　提案するトピックモデルにおける superpixel とその境界，観測される特徴の関係.

おり，領域・境界トピックを同時に推定することで，境界線の性質に応じた領域トピックの傾向が得られ，また，妥当な境界線による領域の切り替わりを実現する．提案する生成モデルとその解き方については，次節にて説明する．

5.2　提案手法

5.2.1　境界属性を考慮したトピックモデル

提案するトピックモデルは図 5.4 のような概念図によって表される．各画像から複数の superpixel が得られており，隣接する superpixel に対して境界を設定する．各 superpixel からは 1 つの大域領域特徴と複数の局所領域特徴が観測され，1 つの境界から 1 つの大域境界特徴と複数の局所境界特徴が観測される．各 superpixel，各境界にそれぞれ領域トピック，境界トピックが設定され領域トピックは画像全体のトピック分布と領域画像特徴から，境界トピックは隣接する領域トピックと境界画像特徴から決定される．グラフィカルモデルは，図 5.5 のように設定する．superpixel に対して領域トピック T を設定し，領域境界に対して境界トピックとして潜在変数 U を設定する．superpixel から得られる

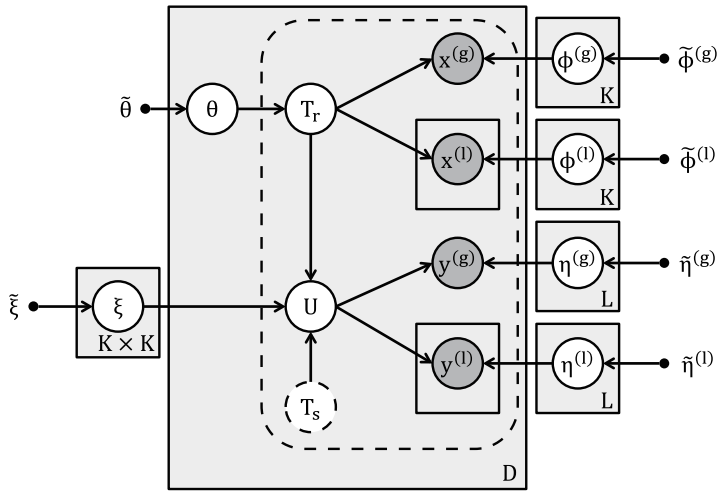


図 5.5　提案するトピックモデルのグラフィカルモデル．パラメータ $\boldsymbol{\xi}$ は隣り合う領域トピック T から境界トピック U を生成するパラメータ，$\boldsymbol{\eta}^{(g)}, \boldsymbol{\eta}^{(l)}$ は U から境界特徴 $\boldsymbol{y}^{(g)}, \boldsymbol{y}^{(l)}$ を生成するパラメータである．

大域領域特徴 $\boldsymbol{x}^{(g)}$ と局所領域特徴 $\boldsymbol{x}^{(l)}$ に加えて，境界からも大域境界特徴 $\boldsymbol{y}^{(g)}$ と局所境界特徴 $\boldsymbol{y}^{(l)}$ を示す Visual Word を得る．抽出する特徴ベクトルについては，次節にて詳述する．境界は全ての隣接領域について定義されるため，各 superpixel を次々とつなぐようにして境界トピックが設定されるが，図 5.5 のグラフィカルモデルでは，簡単のため 1 対の隣接領域についての関係性のみを抜き出して表記した．領域トピック T は画像全体のトピック分布を示す $\boldsymbol{\theta}$ から生成され，そのトピックに応じて superpixel 内の各特徴 $\boldsymbol{x}^{(g)}, \boldsymbol{x}^{(l)}$ が生成される．境界トピック U は隣接する 2 つの領域トピックの組み合わせに応じて，パラメータ $\boldsymbol{\xi}$ に従って生成される．そして，境界トピックにから境界内の各特徴 $\boldsymbol{y}^{(g)}, \boldsymbol{y}^{(l)}$ が生成される．各 Visual Word のトピックに対する分布を示すパラメータは $\boldsymbol{\phi}^{(g)}, \boldsymbol{\phi}^{(l)}, \boldsymbol{\eta}^{(g)}, \boldsymbol{\eta}^{(l)}$ で示され，それぞれパラメータ $\tilde{\boldsymbol{\phi}}^{(g)}, \tilde{\boldsymbol{\phi}}^{(l)}, \tilde{\boldsymbol{\eta}}^{(g)}, \tilde{\boldsymbol{\eta}}^{(l)}$ を持つディリクレ分布によって生成される．境界トピックが隣接する領域トピックから生成されていることから，境界トピックは隣接する領域トピックの組み合わせに対して偏りを持つ．領域トピックが物体ラベルを示しているとすると，隣接する superpixel が同一物体である時の境界特徴の偏りやある物体の輪郭での境界特徴の偏りを境界トピックとして得ることができる．

5.2.2　トピック抽出処理の流れ

提案手法において領域トピック，境界トピックを抽出する流れを解説する．

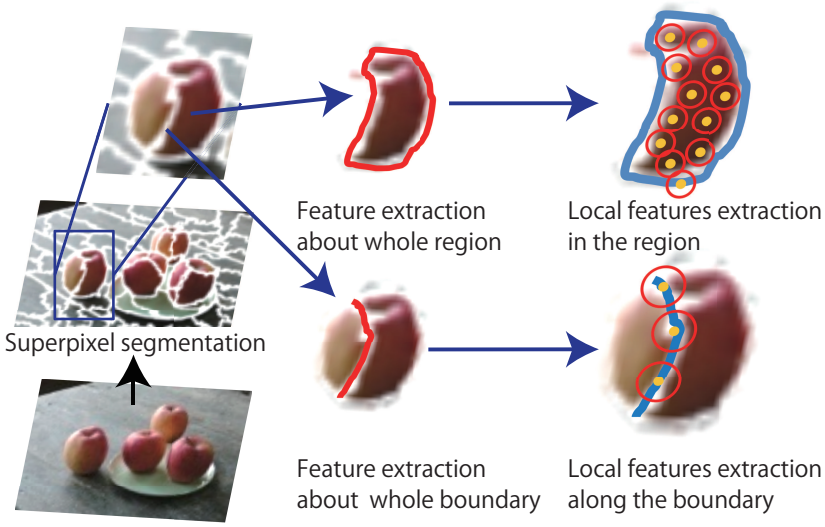


図 5.6　superpixel 分割と特徴抽出の流れ．各 superpixel から 1 つの大域領域特徴と複数の局所領域特徴，各境界から 1 つの大域境界特徴と複数の局所境界特徴が抽出される．

領域・境界特徴抽出

初めに，入力画像に対して Visual Word 生成のための特徴抽出処理を行うが，その流れを図 5.6 に示す．まず，トピックを付与するための superpixel・境界を得るため，画像を superpixel に分割する．この superpixel は複数の物体が同じ領域となることのないように十分に細かく分割する必要がある．本研究では，Liu らの提案する画像分割手法 [20] を用いて画像を superpixel に分割している．次に，画像特徴を抽出する．画像特徴は，各 superpixel とその境界線から 2 種類の画像特徴を得る．superpixel 全体の見えを表す大域領域特徴として，領域内のカラーヒストグラムとテクスチャヒストグラム [68] を抽出し，superpixel 内の局所的な勾配を表す局所領域特徴として，SIFT 特徴 [23] を用いる．SIFT 特徴は領域面積に応じて superpixel 内のランダムな特徴点から複数の特徴ベクトルを抽出する．さらに，境界線の勾配を表す大域境界特徴として，境界線上で角度毎に勾配強さを加算したヒストグラムを用いる．このヒストグラムは以下のような式で表される．

$$H_a = \sum_i (\Delta_{xi} \cos a_i + \Delta_{yi} \sin a_i) \delta(a_i, a) \tag{5.1}$$

ここで，i は境界線上の画素，a_i は画素 i における境界線の法線方向，Δ_{xi}, Δ_{yi} は画素 i での横方向及び縦方向の勾配強度を表す．$\delta(a_i, a)$ はディラックのデルタ関数を示し，

$a_i = a$ の時に 1 となる．法線方向は 16 段階で表現し，全画素における境界線に沿った勾配強度を境界線方向毎にヒストグラム化している．勾配強度にはソーベルフィルタの出力を用いている．また，境界線上の局所的な勾配を表す局所境界特徴として，Geometric Blur 特徴 [45] を用いる．Geometric Blur 特徴は物体境界上に設定される特徴点の記述子として，その頑健性が知られており物体ラベルに関連した境界線を抽出する提案モデルに適した特徴量と言える．得られた各種特徴ベクトルはベクトル量子化辞書によって離散的な値に量子化される．量子化辞書は画像集合全体の特徴ベクトルに対して，k-means クラスタリングを適用することで得られるいくつかのクラスタ中心を代表ベクトルとすることで構成する．各特徴ベクトルから最も近い代表ベクトルを検索し，そのインデックスを量子化された Visual Word として用いる．こうして得られた分割画像と各種 Visual Word は図 5.4 のような構造で対応している．境界トピックは隣接する 2 領域に付き 1 つずつ存在する．

領域・境界トピック推定

次に，得られた観測変数に従って，画像のトピック推定を行う．画像のトピック推定は，データ集合を用いたモデルパラメータ推定とモデルパラメータを用いた画像毎のトピック推定に分けることができる．

まず，画像データ全体の分布を表現するモデルパラメータ $\boldsymbol{\phi}^{(g)}, \boldsymbol{\phi}^{(l)}, \boldsymbol{\eta}^{(g)}, \boldsymbol{\eta}^{(l)}, \boldsymbol{\xi}$ を推定する．モデルパラメータの推定は図 5.5 のような生成モデルの関係性から各パラメータの事後分布を求めることにより行う．パラメータ $\boldsymbol{\phi}^{(g)}$ の場合，その事後分布はベイズの定理によって以下のように表すことができる．

$$P(\boldsymbol{\phi}^{(g)}|\boldsymbol{x}^{(g)}, \tilde{\boldsymbol{\phi}}^{(g)}) \propto P(\boldsymbol{x}^{(g)}|\boldsymbol{\phi}^{(g)})P(\boldsymbol{\phi}^{(g)}|\tilde{\boldsymbol{\phi}}^{(g)}) \tag{5.2}$$

このように，事後分布 $P(\boldsymbol{\phi}^{(g)}|\boldsymbol{x}^{(g)}, \tilde{\boldsymbol{\phi}}^{(g)})$ は，事前分布 $P(\boldsymbol{\phi}^{(g)}|\tilde{\boldsymbol{\phi}}^{(g)})$ と尤度 $P(\boldsymbol{x}^{(g)}|\boldsymbol{\phi}^{(g)})$ の積で表現することができるが，尤度の計算には，その他の潜在変数が複雑に関係しており解析的な計算が難しい．そこで，近似的な事後分布を求めるため変分ベイズ法の一種である Variational Message Passing (VMP) [74] を用いる．VMP では，各変数の事後分布を逐次的な更新によって近似的に求めることができる．

データ集合から各パラメータの事後分布が得られると，それらの事後分布を用いて画像のトピック推定を行うことができる．トピック推定もパラメータ推定と同様にして，図 5.5 のモデルにおいて変数 T と U の事後分布を求めることで行う．この時，各パラメータの事前分布 $\tilde{\boldsymbol{\phi}}^{(g)}, \tilde{\boldsymbol{\phi}}^{(l)}, \tilde{\boldsymbol{\eta}}^{(g)}, \tilde{\boldsymbol{\eta}}^{(l)}$ として，モデルパラメータ推定において得られた各パラメータの事後分布を用いる．VMP を用いて領域・境界トピックの事後分布を求め，次

式 (5.3),(5.4) のように事後確率が最大となるトピック $\hat{T}, \hat{U}$ を推定値とする．

$$\hat{T} = \arg\max_{T} P(T|\boldsymbol{x}^{(g)}, \boldsymbol{x}^{(l)}, \boldsymbol{y}^{(g)}, \boldsymbol{y}^{(l)}) \tag{5.3}$$

$$\hat{U} = \arg\max_{U} P(U|\boldsymbol{x}^{(g)}, \boldsymbol{x}^{(l)}, \boldsymbol{y}^{(g)}, \boldsymbol{y}^{(l)}) \tag{5.4}$$

5.3 実験

5.3.1 単一物体の領域抽出

提案手法を用いて多数の画像データから物体領域を抽出できることを確認するため，Weizmann Horse Dataset [75] を用いてトピック抽出を行った．Weizmann Horse Dataset には，馬の画像が 327 枚含まれており，領域抽出精度評価用に人の手によって馬の領域を示した正解データが用意されている．トピック抽出精度の定量的な評価を得るため，これらの馬画像からトピック抽出を行い，同一トピックとしての馬の領域抽出精度を評価する．なお，実験に用いた画像は全て横幅が 400 ピクセルになるようにスケールしてあり，また，実験に用いた各種パラメータとして，各特徴の量子化辞書サイズ V は 500，画像中の superpixel の数 R_d は 50，ハイパーパラメータ $\tilde{\boldsymbol{\phi}}^{(g)}, \tilde{\boldsymbol{\phi}}^{(l)}, \tilde{\boldsymbol{\eta}}^{(g)}, \tilde{\boldsymbol{\eta}}^{(l)}, \tilde{\boldsymbol{\xi}}$ はそれぞれ全要素 1 のベクトルとした．経験的に境界トピック数 L は 2，領域トピック数 K は 5 と設定した．

図 5.7 上段に領域トピック抽出の一例を示す．特に違いが顕著であった例を示しているが，従来手法では各 superpixel が独立に扱われるためバラバラになりがちであった領域が，提案手法ではまとまった領域として推定されていることがわかる．

定量的な評価として表 5.1 に提案手法による領域抽出精度と従来手法との比較を示す．ここで領域抽出精度とは，全画素のうち正解データと同じラベルが推定されている画素の

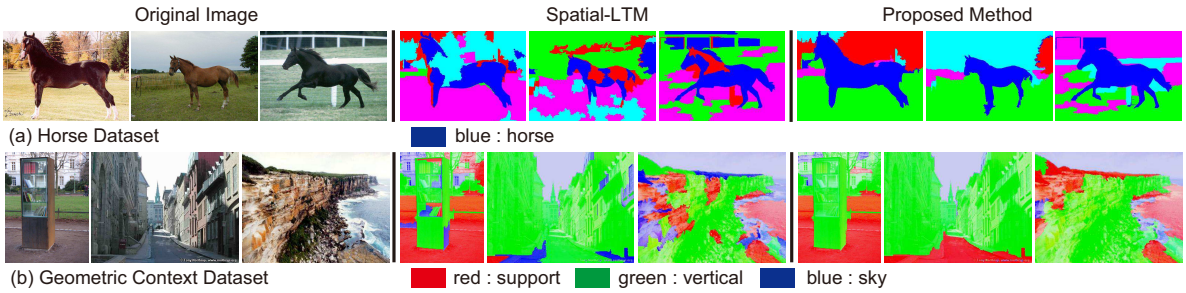


図 5.7　Horse Dataset 及び Geometric Context Dataset に対する教師なし領域抽出結果の例

表 5.1　Horse Dataset に対する領域抽出精度

	RABIT-Model	Spatial-LTM	TRF
Accuracy	**87.9%**	79.6%	75.4%

比率を示している．ただし，抽出された領域トピックのうち最も馬の領域に含まれていたものを馬ラベルとして算出した．Spatial-LTM [3] 及び Topic Random Field [67] については，文献を参考に実装し直したもので，提案手法と同様の画像分割や画像特徴を用いている．提案手法によって領域抽出精度が向上していることがわかる．

5.3.2　複数物体の領域抽出

提案手法を用いて境界属性に基づいて複数カテゴリの物体領域を抽出できることを確認するため，Geometric Context Dataset [76] を用いて教師なし領域抽出を行った．Geometric Context Dataset には，300 枚の風景画像が含まれており各画像の画素毎に垂直面や地面，空などの立体的なレイアウト情報が付与されている．レイアウト情報は空間的配置情報の偏りが顕著であるため，提案手法によって空間的な配置情報の偏りを加味できているかを確認することができる．

実験に用いた各種パラメータとして，各特徴の量子化辞書サイズ V は 1000，画像中の superpixel の数 R_d は 100 とした．学習に用いる画像は全ての画素が‘地面’，‘垂直面’，‘空’のいずれかに含まれるため，抽出する領域トピック数 K は 3，境界トピック数 L は 2 とした．その他のパラメータについては前節の実験と同様である．

表 5.2　Geometric Context Dataset に対する領域抽出精度

	RABIT-Model	Spatial-LTM	TRF
Accuracy	**80.1%**	70.4%	50.2%

実験結果を図 5.7 下段と表 5.2 に示す．図 5.7 下段より従来手法に比べて精度良く 3 種の領域抽出ができている．例えば空のようなテクスチャのない領域は，空だけでなく道路上にも発生しており，従来手法においては地面領域に空のラベルが付与される誤りが発生しているが，提案手法においては隣接領域の配置関係を暗に考慮しているために連続する地面領域を正しく抽出できていることがわかる．表 5.2 に各手法による領域抽出精度を示す．従来手法に比べ大幅に精度向上していることがわかる．

5.4　更なる精度向上に向けて

領域トピックと同時に境界トピックを推定する提案モデルによって境界によって示唆される配置関係を考慮した物体領域抽出が実現した．しかし，生成モデルにおいてはsuperpixel 内の特徴と境界での特徴が明確に分かれており異なる意味を持つのに対して，実際に観測される画像特徴には重複した成分を含むものとなってしまっている．これによって，提案モデルが意図する領域トピックと境界トピックとの補完関係が十分に機能していないことが想定される．次章にて本章にて提案した RABIT-Model による領域抽出を更に高精度化・頑健化するため第 3 章にて提案した領域特徴及び第 4 章にて提案した境界特徴の活用し，領域抽出の性能向上を確認する．

第 6 章

事前学習された領域・境界特徴に基づく教師なし物体領域抽出

第 5 章では Superpixel 間の境界線に潜在的な意味づけを行う境界トピックを導入した新たなトピックモデルを提案し，教師なし物体領域抽出を行えることを示した．各 Superpixel 内から抽出される領域特徴及び Superpixel 間から抽出される境界特徴として従来の物体認識等に一般的に用いられていた SIFT 特徴や Geometric Blur 特徴などを用いて評価を行ったが，これらの特徴を事前学習によって獲得された特徴表現に置き換えることで更に高精度な領域抽出が期待される．本章では，第 3 章にて提案した球面クラスタリングによって量子化された深層特徴と，第 4 章にて提案した前後関係を加味した輪郭形状の分類を行う Occlusion Tokens 特徴とをそれぞれ領域特徴，境界特徴として適用し，提案するトピックモデルを用いた教師なし物体領域抽出の評価を行う．

6.1 実験設定

6.1.1 画像特徴表現

教師なし物体領域抽出に用いる画像特徴として第 3 章で提案した領域特徴，及び第 4 章で提案した境界特徴を用いる．各特徴の抽出方法について述べる．

領域特徴

第 5 章における実験と同様に，入力画像をまず Superpixel に分割する．Superpixel 分割には Liu らの手法 [20] を用いる．入力画像を R_d 個の過分割された Superpixel に分割

し，各 Superpixel について領域特徴を抽出する．

領域特徴には第 3 章と同様にして事前に学習された CNN を活用した深層特徴を用いる．各 Superpixel からその面積に応じて特徴点数を決定し，特徴点数分の座標を Superpixel 内部からランダムに取り出す．選ばれた座標値に該当する深層特徴を抽出し，これらを各 Superpixel に対応する領域特徴ベクトルとする．トピックモデルへの適用にあたってこれらの特徴ベクトルを量子化して扱う必要があるため，クラスタリングによって一定数の離散的な数値への変換を行う．領域特徴のクラスタリングは推定対象とするデータセットに応じて抽出された特徴ベクトルを用いて適宜行うが，ここで第 3 章において効果が確認されている球面クラスタリングを導入する．抽出された領域特徴ベクトルから全ベクトルの平均値を引き，これをコサイン類似度に基づく球面クラスタリングによって V_r 個のクラスタへと分類する．算出された各クラスタの中心ベクトルを用いて，各領域特徴ベクトルとのコサイン類似度が最も大きいクラスタを決定し，そのクラスタインデックスを領域特徴 $\boldsymbol{x}^{(l)}$ とする．また，各種パラメータについて，本実験においては R_d を 100，V_r を 1000 としている．

境界特徴

境界特徴は Superpixel 間の境界線上から画像特徴を算出する．境界における隣接関係を表す特徴として第 4 章にて提案した Occlusion Tokens を用いる．第 4 章の実験にて用いた Berkeley Segmentation Dataset によって学習された Random Forest 識別器を用いて，150 クラスの Occlusion Tokens への識別スコアを得る．識別はパッチ画像を入力として行われるが，パッチの中心画素を境界線上に設定することで，その境界線から得られた境界特徴とみなすことができる．境界線の長さに応じてランダムな点を複数選択し，各点に対応する Occlusion Tokens の識別スコアを各境界線から得られる境界特徴ベクトルとする．トピックモデルによる領域抽出の際には，データセットから得られる境界特徴ベクトルを kMeans クラスタリングによって量子化し，割り当てられたクラスタのインデックスをもって境界特徴 $\boldsymbol{x}^{(l)}$ とする．また，設定パラメータについてはクラスタ数 V_b を 100 としている．

6.1.2 評価データセット

評価には，背景がシンプルで前景として単一カテゴリの物体を対象とする Weizmann Horse Dataset [75] と背景が非常に複雑であり，また前景物体のカテゴリも豊富な Pascal VOC Dataset [55] の 2 種類を用いる．

Weizmann Horse Dataset はデータセット中の全ての画像に共通して馬が写っており，また背景には空や草原などが多く他物体の写り込みが少ないために前景抽出が比較的容易なデータセットである．第 5 章においてもこのデータセットを用いたが，特徴表現の改良による認識精度の向上を簡単なデータセットにおいて確認する．

一方で Pascal VOC Dataset は多様な状況で撮影された多数の前景物体を含む画像データセットである．多数の前景物体に対して背景に写る物体についても多様であり，前景抽出は比較的難しいデータセットといえる．第 3 章においては単一物体の検出問題において VOC Dataset への適用性を評価したが，本章では多数物体の領域抽出問題への適用性を評価する．前景物体のカテゴリ数によって抽出精度が変わることが想定されるため，データセットには含まれる 20 カテゴリの前景物体のうち 2 種のカテゴリを選択的に用いた評価と 20 種全てを用いた評価を行う．2 種のカテゴリに基づく評価は Wang ら [1] が用いている PASCAL-multi Dataset と同様の設定を用いる．事前に選択された 2 つのカテゴリについて，それらのカテゴリのみを含み，他のカテゴリを含まない画像を収集し処理対象とする．また，前景領域の面積が画像全体の 0.5% に満たない画像は用いない．

6.1.3 評価指標

評価指標として，Weizmann Horse Dataset については，第 5 と同様に画素単位の正解率を用い，Pascal VOC Dataset については，領域抽出精度評価によく用いられる Intersection over Union (IoU) を用いる．IoU は以下の式で求められる．

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{6.1}$$

ここで，TP は True Positive に該当する画素数，FP は False Positive に該当する画素数，FN は False Negative に該当する画素数を意味する．各表にはこれらを 100 倍し百分率として表した値を表記する．

6.2 単一物体抽出

各種手法を用いて Weizmann Horse Dataset に対して領域抽出を行った結果を表 6.1 に示す．表の最左列に，評価に用いるモデルと特徴量の組み合わせを簡略化して表記している．RABIT は第 5 章にて提案したトピックモデルを示すが，括弧内は領域特徴と境界特徴として採用された特徴表現を示す．S,G,D,O がそれぞれ SIFT 特徴，Geometric Blur 特徴，深層特徴，Occlusion Tokens 特徴を意味する．表に記載されている通り，

表 6.1　Weizmann Horse Dataset に対する単一物体領域抽出における画像特徴の違いによる抽出精度変化．表中の S,G,D,O はそれぞれ SIFT 特徴，Geometric Blur 特徴，深層特徴，Occlustion Tokens 特徴を示す．

Method	Region feature	Boundary feature	K	L	Accuracy(%)
SLTM [3]	SIFT Color Hist.	-	5	-	79.6
RABIT(S,G) [73]	SIFT Color Hist.	Geometric Blur Line shape	5	2	86.2
RABIT(S,O)	SIFT Color Hist.	Occlusion Tokens	5	5	79.8
RABIT(D,G)	Deep Feature	Geometric Blur Line shape	3	2	91.9
RABIT(D,O)	Deep Feature	Occlustion Tokens	3	5	**92.2**

表 6.2　Weizmann Horse Dataset に対する単一物体領域抽出におけるトピック数の違いによる抽出精度変化

		L			
		3	5	7	9
K	2	59.1	55.7	76.6	68.5
	3	90.2	**92.2**	92.1	91.9
	4	91.5	91.1	91.7	92.0
	5	90	90.8	90.8	89.6

SIFT 特徴は色ヒストグラムと共に，Geometric Blur 特徴は線形状のヒストグラムと共に用いられる．また，K, L はトピックモデルにおけるトピック数を予め定めるパラメータであり，それぞれ領域トピック数，境界トピック数を示す．表 6.1 における K, L は実験的に定めた値である．第 5 章では，従来のトピックモデルに対して提案するトピックモデルを導入することで領域抽出精度が向上することを示したが，表 6.1 では，それに加えて，特徴量を変更することで性能向上していることが見て取れる．特に領域特徴としての深層特徴の導入が大きな精度向上につながっていることが見て取れる．境界特徴として Occlustion Tokens を導入することで大きな精度の変化は見られないが，最も精度の良い特徴量の組み合わせは深層特徴と Occlusion Tokens となっている．一方で領域特徴を

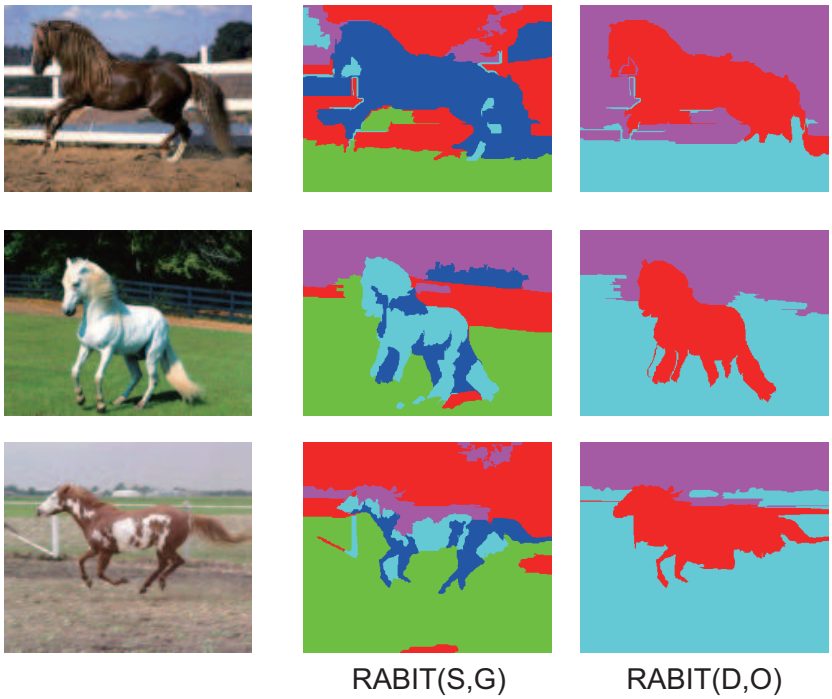


図 6.1　Horse Dataset に対する教師なし領域抽出結果の例

SIFT に設定した際には境界特徴が Occlusion Tokens よりも Geometric Blur の方が良い結果が得られている．これは，周辺の勾配情報を抽出する Geometric Blur から馬らしさのテクスチャが得られ，SIFT 特徴を補うように働いたためと考えられる．Occlusion Tokens は境界線の性質のみを表すため，個別の対象に対して偏った特徴量は得られない．深層特徴と合わせて活用する場合には，深層特徴が SIFT 特徴に比べてより良く物体固有の特徴を捉えられていたために，Geometric Blur 特徴によるテクスチャ表現は補助的に機能せず，境界線情報に特化した Occlusion Tokens の方が精度向上につながったと考えられる．

また，深層特徴と Occlusion Tokens を用いた最高性能の手法において領域トピック数及び境界トピック数を変えた場合の領域抽出精度を表 6.2 に示す．表から領域トピック数が 2 の場合に精度が急激に悪化しているものの，3-5 については大きな精度劣化がないことがわかる．領域トピックの数は入力画像をいくつの領域に分割するかの分割数を表しているため，数字が小さいほどパラメータの変更に対して過敏に反応すると考えらえる．この結果からもデータセットが持つと想定されるカテゴリの数に対して少し高めのトピック数を設定する方が高い精度が得られやすいといえる．境界トピックの数については 3-7 においてあまり大きな精度の変化は見られなかった．データセットに含まれるカテゴリの種

表 6.3　Pascal-multi Dataset の一部を用いた複数物体抽出精度の詳細な評価

	Bottle + Dining table			Chair + Potted plant		
Method	K	L	IoU(%)	K	L	IoU(%)
SLTM	3	-	17.9	5	-	6.6
RABIT(S,G)	3	2	21.3	5	2	7.0
RABIT(S,O)	3	5	19.7	5	5	6.3
RABIT(D,G)	5	2	30.9	10	2	25.8
RABIT(D,O)	5	5	**35.4**	10	5	**36.8**

表 6.4　Pascal-multi Dataset に対する複数物体領域抽出精度評価

Categories	Wang et al. [1]	RABIT(S,G)	RABIT(D,O)
Bike + Person	**40.1**	15.5	33.6
Boat + Person	**44.6**	17.0	44.0
Bottle + Dining table	**47.6**	21.3	35.4
Bus + Car	**49.2**	19.5	48.3
Bus + Person	55.5	25.9	**60.9**
Chair + Dining table	**40.3**	19.3	32.4
Chair + Potted plant	22.3	7.0	**36.8**
Cow + Person	45.0	18.7	**60.8**
Dog + Sofa	49.6	17.8	**55.3**
Horse + Person	42.1	16.5	**53.0**
Potted plant + Sofa	40.7	19.5	**43.6**

類によらず，一般的な実写画像に現れる境界線の分類をしていると考えられるため，以降の検証実験においては Occlusion Tokens 特徴の境界トピック数として 5 を設定する．

提案手法によって推定された領域抽出結果の例を図 6.1 に示す．図には，特に SIFT 特徴と色ヒストグラムによって抽出が難しくなる白い馬の例を示している．SIFT 特徴を用いた手法では，馬の領域が複数に分かれてしまっているが，深層特徴を用いた手法では一貫して馬の領域を抽出できていることがわかる．

6.3 複数物体抽出

複数物体抽出への適用性を評価するために，上述の PASCAL-multi Dataset を用いて評価を行った．事前に選択された 2 つのカテゴリと抽出精度を表 6.3，6.4 に示す．まず，単一物体と同様に各特徴による効果を確認するため，“Bottle + Dining table” と “Chair + Potted plant” を用いて性能評価を行った結果を表 6.3 に示す．単一物体抽出の場合と同様に深層特徴と Occlusion Tokens 特徴の導入によって抽出精度が大幅に向上していることが見て取れる．また，Occlusion Tokens 特徴は深層特徴と組み合わせた際に良い効果を発揮していることも改めて確認された．境界トピック数 L については Horse Dataset での結果に基づき，Geometric Blur 特徴に対しては 2, Occlusion Tokens 特徴に対しては 5 を設定している．領域トピック数 K については $3, 5, 10$ の値を評価し，最も精度の良いものを設定している．

他の組み合わせについての領域抽出結果を表 6.4 に示す．各特徴量の相性が明らかになったため，第 5 章にて用いた SIFT 特徴＋ Geometric Blur 特徴に基づく手法と深層特徴＋ Occlusion Tokens 特徴に基づく手法の 2 手法について結果を示している．また，同様の実験を行っている Wang らの手法 [1] について論文に掲載されている評価値を記載する．表より，全ての組み合わせにおいて提案する画像特徴によって抽出精度が向上していることがわかる．また，教師なし複数物体領域抽出において最高精度を示している Wang らの手法に対してもいくつかのカテゴリにおいて精度向上していることがわかる．領域トピック数 K については，RABIT(S,G) の評価については常に $K = 5$ とし，RABIT(D,O) の評価については $K = 5, 10$ を評価し，良い方を採用している．

評価結果が特に良好であった “Cow + Person” と “Dog + Sofa” の領域抽出結果の例を図 6.2 に示す．従来の SIFT 特徴に基づく手法では，きれいな物体領域を抽出できていないが，提案する画像特徴に基づく手法では，牛や犬の領域が精度よく抽出できていることがわかる．図左の “Cow + Person” では，牛の領域が赤，人の領域が茶色で抽出されているだけでなく，林の領域が黒，地面の領域が水色で示されており，精度の良い領域分割ができていることがうかがえる．また図右の “Dog + Sofa” では，犬の領域が青，ソファの領域が緑，床の領域が紫で塗り分けられており，こちらも良好な結果が得られている．

また，表 6.5 に Pascal VOC の全カテゴリについて同時に領域抽出を行った評価結果を示す．“Best assignment” 欄には，各カテゴリについて最もあてはまりの良いクラスタを選択した際の IoU を示しており，“Exclusive assignment” 欄には各カテゴリのクラスタ選択を重複しないように選んだ場合の結果を示している．抽出するトピック数として

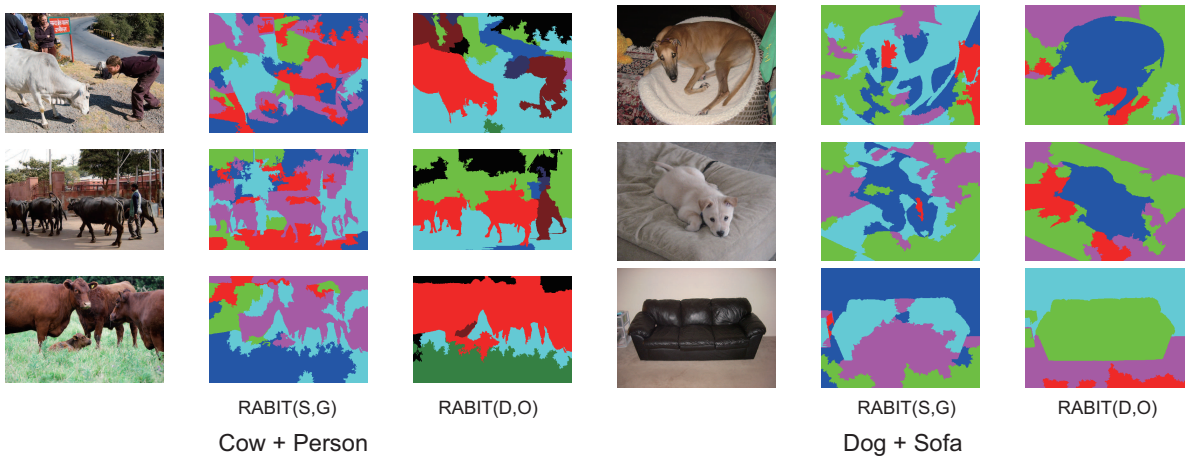


図 6.2　Pascal-multi Dataset に対する教師なし領域抽出結果の例

表 6.5　VOC Dataset 20 クラスに対する複数物体領域抽出精度評価

Best assignment										
Method	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow
RABIT(S,G)	3.5	1.4	2.9	2.6	2.4	8.8	4.8	5.8	2.3	3.4
RABIT(D,O)	27.7	14.3	22.8	8.5	17.0	27.5	23.1	29.1	9.0	14.1
Method	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv
RABIT(S,G)	2.7	6.7	3.9	5.3	11.9	2.6	2.9	2.1	6.6	2.1
RABIT(D,O)	12.3	23.1	13.9	19.9	31.5	13.4	18.8	13.0	16.3	24.6
Exclusive assignment										
Method	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow
RABIT(S,G)	3.5	1.4	2.9	1.6	2.4	8.8	3.4	5.8	2.0	1.6
RABIT(D,O)	27.7	14.3	22.8	8.5	17.0	27.5	23.1	29.1	9.0	14.1
Method	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv
RABIT(S,G)	2.7	3.5	1.5	0.8	1.7	0.6	1.3	1.9	2.0	1.8
RABIT(D,O)	12.3	23.1	2.3	19.9	31.5	13.4	18.8	13.0	16.3	24.6

は $K = 50, L = 30$ を設定した．評価結果から，RABIT(D,O) が一定の抽出精度を示しているものの Pascal-multi の結果よりも全体的に低い抽出精度となっていることがわかる．最も大きな要因としてデータセットにおける対象カテゴリの出現頻度の違いが考えられる．Pascal VOC Dataset では，特に Person が登場する画像が他のカテゴリに比べ多

表 6.6　ノイズ画像に対する頑健性評価

Method	K	L	0%	10%	30%	50%	70%	90%
RABIT(D,O)	5	5	90.8	90.5	91.4	90.6	87.0	75.7

くなっているが，表 6.5 の結果においても Person クラスが最も高い精度を示している．Person 以外のカテゴリについては出現頻度が約 10% 程度であり，50% 以上の画像に出現するよう設計された Pascal-multi に比べて精度低下する要因となっているといえる．

6.4　ノイズ画像に対する頑健性

本論文では同一カテゴリの物体が写る画像集合に対して教師情報を用いることなく物体領域抽出が行えることを目的としているが，入力となる画像集合は見た目に同一性が判断できる同一カテゴリの物体を多数含むことを想定している．一方で，提案手法の枠組みでは必ずしも全画像に対象物体を含む必要はなく，ある程度無関係な画像が混じっていても領域抽出を行うことができる．対象物体の写る画像量と抽出精度との関係性を評価するため，単一物体を含むデータセットに対して関連のないノイズ画像を加えることで精度の変化を確認した．ベースとする評価画像を Weizmann Horse Dataset とし，これに含まれる画像を一定の割合で Pascal VOC Dataset の Person ラベルが付与された画像に置き換えて領域抽出を行った．ノイズ入りの画像集合を用いてモデルのパラメータ推定を行い，その後，推定されたパラメータによる領域抽出を Weizmann Horse Dataset の全画像に対して行い領域抽出精度を算出した．評価結果を表 6.6 に示す．ノイズ画像の存在により抽出されるカテゴリ数が増えると予想されるため，領域トピック数 K は 5 とした．実験結果よりデータセットの 70% 程度が他のカテゴリの画像であっても，高精度に馬の領域抽出が行えることが示された．既存の Co-segmentation 手法と比較すると圧倒的な頑健性を持つ手法であると言える．提案手法は生成モデルをベースとしており，画像特徴を分布として捉えるため，データセット内に対象外の画像が多く含まれていても大きく影響を受けずに領域抽出することができる．一方で，ノイズ画像が 90% になると大きく精度が劣化している．提案手法によって領域抽出を行うためには，同一カテゴリの物体が大きく写っている画像を少なくとも 30% 程度含むような画像集合が望ましいといえる．

第 7 章

結論

本論文では，多様な画像に対して頑健な教師なし物体領域抽出を目的として，3 つの提案を行った．1 つ目に局所領域の類似性を頑健に抽出可能な領域特徴表現手法として，球面クラスタリングされた深層特徴を提案し，提案する特徴表現によって多様な見た目の変化を伴う画像データセットに対して頑健に同一物体の検出が行えることを示した．2 つ目に空間的な配置関係を適切に抽出可能な境界特徴表現手法として遮蔽関係を加味した輪郭線形状を分類する Occlusion Tokens 特徴を提案し，この特徴表現によって高速に遮蔽境界を検出し，高精度にその前後関係を推定できることを示した．3 つ目に局所領域の類似性と空間的な配置関係性を同時に考慮する統合手法として，トピックモデルをベースとして，領域とその境界に同時にトピックを付与する RABIT モデルを提案し，提案する RABIT モデルが空間的な配置関係を考慮した領域抽出の枠組みとなり得ることを示した．3 つの提案を組み合わせた物体領域抽出手法によって多様で整理されていない画像集合に対しても頑健かつ高精度に物体領域抽出を行えることを実験によって示した．本研究によって，世界中に爆発的に増加する画像メディアを構造化して管理することを考えた場合に，教師情報を用意するというわずらわしい作業を経ることなく，画像データのみから自然な分類や対応付け，検索などが可能な技術の実現に向けて貢献しうる技術を創出した．

しかしながら，実際の商用サービスとして活用可能なレベルまでにはまだいくつかの解決すべき将来課題が残る．十分な教師データを用いて学習を行った場合の領域抽出結果に比べ，精度面で大きく劣ることが直近の課題となるが，実用的には，教師なしで得られた認識結果を人手によって修正を加えるなど必要に応じて人手による補助を与えることで大きく精度が改善することが期待される．より実際に即した課題として，多数のカテゴリについて予め十分な教師データによる学習が済んでいる状態で新奇な物体を学習することが

考えられる．そうした場合には学習済みのカテゴリとの関連性に基づいて更に有効な手がかりを導入することができるかもしれない．また，特に大きな将来課題として，画像集合から抽出すべき物体カテゴリの数と粒度の決定方法の問題があげられる．例えば，馬の画像を対象とした際に，抽出すべきカテゴリとして馬全体を抽出すべきなのか白い馬と茶色い馬を分けて抽出すべきなのかはデータのみから推定することは難しい．同様にして画像集合からいくつのカテゴリを抽出すべきかも決めがたい問題である．階層ディリクレ過程 [77] のようなノンパラメトリックな確率モデルが提案されてはいるが，いずれにしてもクラスタ数をコントロールするハイパーパラメータの設定は必要であり，ニーズに応じた適切なパラメータ設定の手段が課題となる．更に，抽出すべき対象が必ずしも排他的でない場合も想定される．例えば，馬カテゴリと馬車カテゴリを同時に抽出するような場合である．このような問題に対しては，粒度の異なるカテゴリを階層的に表現し，階層構造の推定と共に抽出する技術が必要となるであろう．

謝辞

本研究の遂行及び本論文の執筆に際して，東京理科大学工学部谷口行信教授には NTT 研究所での元上司として，また在学中の指導教員として，終始あたたかいご指導を賜りました．心より感謝申し上げます．谷口先生が私の所属するグループのグループリーダになられた 2012 年以来，実に 7 年間にわたって様々な形でご支援をいただきました．社会人学生として東京理科大の博士課程に入学する決断をできたのも谷口先生にお声がけいただけたからこそであり，谷口先生のご指導の下こうして学位論文を執筆することができたことに改めて深く感謝いたします．

東邦大学理学部数藤恭子教授には，主に NTT 時代のテーマ設定及び研究遂行において直接指導者としてご指導賜りました．心より感謝申し上げます．実利を重視すべき企業研究所において，本論文執筆につながる学術的なテーマ設定を実現できたのはひとえに数藤先生のご指導のおかげです．改めて深く感謝いたします．

NTT サービスイノベーション総合研究所杵渕哲也主幹研究員には東京理科大在学中の職場上司として，博士学生としての研究と会社業務との両立を快く認めていただきました．深く感謝いたします．

埼玉大学工学部島田裕助教には，共著での研究発表こそできませんでしたが，進捗報告の度に的確なご指摘を多数いただき研究をご指導していただきました．深く感謝いたします．

柳瀬直人さん，高木基希さん，佐藤大己さん，沖谷卓哉さん，宮尾恵さん，鈴木智裕さんをはじめ谷口研究室の皆様には，ほとんど研究室に顔を出さない私にも快く接していただきました．研究室で過ごした時間はわずかでしたが，楽しいキャンパスライフを送る皆さんの様子を見ているだけでも楽しい時間を過ごすことができました．感謝いたします．

最後に，社会人学生として仕事と博士研究に追われる日々の生活をサポートし，学位取得に向けて応援してくれた妻と娘たちに心から感謝いたします．

参考文献

[1] F. Wang, Q. Huang, M. Ovsjanikov, and L. Guibas. Unsupervised Multi-Class Joint Image Segmentation. In *Proc. CVPR*, pp. 3142–3149, 2014.

[2] M. Maire. Simultaneous Segmentation and Figure/Ground Organization using Angular Embedding. In *Proc. ECCV*, pp. 450–464, 2010.

[3] L. Cao and L. Fei-Fei. Spatially Coherent Latent Topic Model for Concurrent Object Segmentation and Classification. In *Proc. ICCV*, pp. 1–8, 2007.

[4] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In *Proc. CVPR*, pp. 3431–3440, 2015.

[5] D. Comaniciu and P. Meer. Mean Shift: A Robust Approach toward Feature Space Analysis. *IEEE Trans. PAMI*, Vol. 24, No. 5, pp. 603–619, 2002.

[6] P. Felzenszwalb and D. Huttenlocher. Efficient Graph-Based Image Segmentation. *Int. J. Comput. Vision*, Vol. 59, pp. 167–181, 2004.

[7] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From Contours to Regions: An Empirical Evaluation. In *Proc. CVPR*, pp. 2294–2301, 2009.

[8] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Trans. PAMI*, Vol. 22, No. 8, pp. 888–905, 2000.

[9] V. Kolmogorov and R. Zabih. What Energy Functions Can Be Minimized via Graph Cuts? *IEEE Trans. PAMI*, No. 2, pp. 147–159, 2004.

[10] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, et al. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. PAMI*, Vol. 34, No. 11, pp. 2274–2282, 2012.

[11] J. Canny. A Computational Approach to Edge Detection. *IEEE Trans. PAMI*, No. 6, pp. 679–698, 1986.

[12] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour Detection and Hier-

archical Image Segmentation. *IEEE Trans. PAMI*, Vol. 33, pp. 898–916, 2011.

[13] J. Lim, C. Zitnick, and P. Dollar. Sketch Tokens: A Learned Mid-Level Representation for Contour and Object Detection. In *Proc. CVPR*, pp. 3158–3165, 2013.

[14] P. Dollár and L. Zitnick. Structured Forests for Fast Edge Detection. In *Proc. ICCV*, pp. 1841–1848, 2013.

[15] S. Xie and Z. Tu. Holistically-Nested Edge Detection. In *Proc. ICCV*, pp. 1395–1403, 2015.

[16] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive Foreground Extraction using Iterated Graph Cuts. *ACM Trans. Graphics*, Vol. 23, No. 3, pp. 309–314, 2004.

[17] Y. Boykov, O. Veksler, and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts. *IEEE Trans. PAMI*, Vol. 23, No. 11, pp. 1222–1239, 2001.

[18] M. Unger, T. Pock, W. Trobin, D. Cremers, and H. Bischof. TVSeg-Interactive Total Variation Based Image Segmentation. In *Proc. BMVC*, Vol. 31, pp. 44–46, 2008.

[19] F. Chung. *Spectral Graph Theory*. American Mathematical Soc., 1997.

[20] M. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa. Entropy Rate Superpixel Segmentation. In *Proc. CVPR*, pp. 2097–2104, 2011.

[21] A. Levinshtein, A. Stere, K. Kutulakos, D. Fleet, S. Dickinson, and K. Siddiqi. Turbopixels: Fast Superpixels using Geometric Flows. *IEEE Trans. PAMI*, Vol. 31, No. 12, pp. 2290–2297, 2009.

[22] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic Segmentation with Second-Order Pooling. In *Proc. ECCV*, pp. 430–443, 2012.

[23] D. Lowe. Object Recognition from Local Scale-Invariant Features. In *Proc. ICCV*, pp. 1150–1157, 1999.

[24] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Proc. CVPR*, Vol. 1, pp. 886–893, 2005.

[25] J. Tighe and S. Lazebnik. Finding Things: Image Parsing with Regions and Per-Exemplar Detectors. In *Proc. CVPR*, pp. 3001–3008, 2013.

[26] X. He, R. Zemel, and M. Carreira-Perpiñán. Multiscale Conditional Random Fields for Image Labeling. In *Proc. CVPR*, Vol. 2, 2004.

[27] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-

Scale Image Recognition. In *Proc. ICLR*, 2015.

[28] A. Bansal, B. Russell, and A. Gupta. Marr Revisited: 2D-3D Alignment via Surface Normal Prediction. In *Proc. CVPR*, pp. 5965–5974, 2016.

[29] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proc. MICCAI*, pp. 234–241, 2015.

[30] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. PAMI*, Vol. 40, No. 4, pp. 834–848, 2018.

[31] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-Supervised Convolutional Networks for Semantic Segmentation. In *Proc. CVPR*, pp. 3159–3167, 2016.

[32] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple Does It: Weakly Supervised Instance and Semantic Segmentation. In *Proc. CVPR*, pp. 1665–1674, 2017.

[33] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers. Normalized Cut Loss for Weakly-Supervised CNN Segmentation. In *Proc. CVPR*, 2018.

[34] A. Vezhnevets and J. Buhmann. Towards Weakly Supervised Semantic Segmentation by Means of Multiple Instance and Multitask Learning. In *Proc. CVPR*, pp. 3249–3256, 2010.

[35] A. Kolesnikov and C. Lampert. Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation. In *Proc. ECCV*, pp. 695–711, 2016.

[36] Y. Wei, J. Feng, X. Liang, M. Cheng, Y. Zhao, and S. Yan. Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach. In *Proc. CVPR*, pp. 6488–6496, 2017.

[37] T. Daryanto, S. Arif, and S. Yang. Survey: Recent Trends and Techniques in Image Co-Segmentation Challenges, Issues and Its Applications. *Int. J. Computer Science and Software Engineering*, Vol. 6, No. 5, p. 99, 2017.

[38] A. Joulin, F. Bach, and J. Ponce. Multi-Class Co-Segmentation. In *Proc. CVPR*, pp. 542–549, 2012.

[39] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised Joint Object Discovery and Segmentation in Internet Images. In *Proc. CVPR*, pp. 1939–1946, 2013.

[40] F. Meng, H. Li, G. Liu, and K. Ngan. Object Co-Segmentation Based on Shortest Path Algorithm and Saliency Model. *IEEE Trans. Multimedia*, Vol. 14, No. 5, pp. 1429–1441, 2012.

[41] X. Dong, J. Shen, L. Shao, and M. Yang. Interactive Co-Segmentation using Global and Local Energy Optimization. *IEEE Trans. Image Processing*, Vol. 24, No. 11, pp. 3966–3977, 2015.

[42] F. Wang, Q. Huang, and L. Guibas. Image Co-Segmentation via Consistent Functional Maps. In *Proc. ICCV*, pp. 849–856, 2013.

[43] M. Cho, S. Kwak, C. Schmid, and J. Ponce. Unsupervised Object Discovery and Localization in the Wild: Part-Based Matching with Bottom-Up Region Proposals. In *Proc. CVPR*, pp. 1201–1210, 2015.

[44] Z. Niu, G. Hua, L. Wang, and X. Gao. Knowledge-Based Topic Model for Unsupervised Object Discovery and Localization. *IEEE Trans. Image Processing*, Vol. 27, No. 1, pp. 50–63, 2018.

[45] A. Berg and J. Malik. Geometric Blur for Template Matching. In *Proc. CVPR*, Vol. 1, pp. 607–614, 2001.

[46] A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, and L. Van Gool. DeepProposal: Hunting Objects by Cascading Deep Convolutional Layers. In *Proc. ICCV*, pp. 2578–2586, 2015.

[47] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proc. CVPR*, pp. 580–587, 2014.

[48] X. Wei, C. Zhang, Y. Li, C. Xie, J. Wu, C. Shen, and Z. Zhou. Deep Descriptor Transforming for Image Co-Localization. In *Proc. IJCAI*, pp. 3048–3054, 2017.

[49] K. Murasaki, Y. Taniguchi, and T. Kinebuchi. Unsupervised Multi-Class Object Discovery by Spherical Clustering of Deep Features. *ITE Trans. MTA*, Vol. 7, No. 1, pp. 2–10, 2019.

[50] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet Large Scale Visual Recognition Challenge. *Int. J. Computer Vision*, Vol. 115, No. 3, pp. 211–252, 2015.

[51] I. Dhillon and D. Modha. Concept Decompositions for Large Sparse Text Data using Clustering. *Machine learning*, Vol. 42, No. 1-2, pp. 143–175, 2001.

[52] G. Papandreou, I. Kokkinos, and P. Savalle. Untangling Local and Global De-

formations in Deep Convolutional Networks for Image Classification and Sliding Window Detection. arXiv:1412.0296, 2014.

[53] D. Arthur and S. Vassilvitskii. k-means++: The Advantages of Careful Seeding. In *Proc. ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035, 2007.

[54] Y. Li, L. Liu, C. Shen, and A. Hengel. Image Co-Localization by Mimicking a Good Detector's Confidence Score Distribution. In *Proc. ECCV*, pp. 19–34, 2016.

[55] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *Int. J. Computer Vision*, Vol. 111, No. 1, pp. 98–136, January 2015.

[56] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly Supervised Object Localization with Latent Category Learning. In *Proc. ECCV*, pp. 431–445, 2014.

[57] D. Hoiem, A. Efros, and M. Hebert. Recovering Occlusion Boundaries from an Image. *Int. J. Comput. Vision*, Vol. 91, pp. 328–346, 2011.

[58] X. Ren, C. Fowlkes, and J. Malik. Figure/Ground Assignment in Natural Images. In *Proc. ECCV*, pp. 614–627, 2006.

[59] G. Palou and P. Salembier. Monocular Depth Ordering using T-Junctions and Convexity Occlusion Cues. *IEEE Trans. Image Processing*, Vol. 22, pp. 1926–1939, 2013.

[60] L. Breiman. Random Forests. *Machine Learning*, Vol. 45, No. 1, pp. 5–32, 2001.

[61] K. Murasaki, K. Sudo, and Y. Taniguchi. Occlusion Boundary Detection Based on Mid-Level Figure/Ground Assignment Features. In *Proc. ICIP*, pp. 4707–4711, 2014.

[62] S. Winder, G. Hua, and M. Brown. Picking the Best DAISY. In *Proc. CVPR*, pp. 178–185, 2009.

[63] P. Dollar, Z. Tu, P. Perona, and S. Belongie. Integral Channel Features. In *Proc. BMVC*, pp. 1–5, 2009.

[64] GNU Linear Programming Kit. http://www.gnu.org/software/glpk/.

[65] C. Fowlkes, D. Martin, and J. Malik. Local Figure/Ground Cues are Valid for Natural Images. *J. Vision*, Vol. 7, No. 8, pp. 1–9, 2007.

[66] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *J. Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[67] B. Zhao, L. FeiFei, and E. Xing. Image Segmentation with Topic Random Field.

In *Proc. ECCV*, pp. 785–798, 2010.

[68] L. Li, R. Socher, and L. FeiFei. Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework. In *Proc. CVPR*, pp. 2036–2043, 2009.

[69] Z. Niu, G. Hua, X. Gao, and Q. Tian. Context Aware Topic Model for Scene Recognition. In *Proc. CVPR*, pp. 2743–2750, 2012.

[70] D. Martin, C. Fowlkes, and J. Malik. Learning to Detect Natural Image Boundaries using Local Brightness, Color, and Texture Cues. *IEEE Trans. PAMI*, Vol. 26, pp. 530–549, 2004.

[71] 村崎和彦, 数藤恭子, 島村潤, 森本正志. 物体領域とその隣接関係獲得のための境界情報を考慮したトピックモデル. 画像の認識・理解シンポジウム予稿集, pp. OS8–01, 2012.

[72] 村崎和彦, 数藤恭子, 谷口行信. 空間的連続性を考慮した物体領域とその境界線を同時表現するトピックモデル. 電子情報通信学会論文誌, Vol. J96(D), No. 8, pp. 1844–1853, 2013.

[73] K. Murasaki, K. Sudo, and Y. Taniguchi. RABIT-Model : Unsupervised Scene Understanding via Region and Boundary Integrated Topic Model. In *Proc. Scene Understanding Workshop, CVPR*, 2013.

[74] J. Winn, C. M. Bishop, and T. Jaakkola. Variational Message Passing. *J. Machine Learning Research*, Vol. 6, pp. 661–694, 2005.

[75] E. Borenstein and S. Ullman. Learning to Segment. In *Proc. ECCV*, pp. 315–328, 2004.

[76] D. Hoiem, A. Efros, and M. Hebert. Recovering Surface Layout from an Image. *Int. J. Comput. Vision*, Vol. 75, pp. 151–172, 2007.

[77] Y. Teh, M. Jordan, M. Beal, and D. Blei. Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. In *Proc. NIPS*, pp. 1385–1392, 2005.