修　士　論　文

カテゴリ共起を考慮した物体認識手法

Using the Co-occurrence of Multiple Categories
for Object Recognition

東京大学大学院
情報理工学系研究科
電子情報学専攻

66420　　近　藤　雄　飛

指導教員　　佐藤 洋一　准教授

平成20年2月

概要

実世界シーンの画像中には複数のオブジェクトカテゴリが含まれており，ある２つの物体は頻繁に共存するが別の２つの物体は共存しにくいなどカテゴリによって共起の仕方は異なっている．本研究ではカテゴリの共起を考慮することにより，アピアランスだけでなく他のオブジェクトカテゴリとの関連性から画像中におけるカテゴリの存在比率を求める手法を提案する．局所領域の特徴量からの識別を行う Bag of Features(BOF) モデルの枠組みの中で，テスト画像のヒストグラムが各カテゴリのヒストグラムの線形結合となっていると仮定し，カテゴリの共起を事前知識として学習させ，MAP 推定によって存在比率を表す結合係数を推定していく．

目 次

図 目 次

表 目 次

第1章　序論

1.1　はじめに

近年，デジタルカメラの普及やコンピュータの発達により，人々が大量の画像を扱うことが多くなった．画像をカメラからだけではなくWeb上からも容易に入手できるようになったため，画像を分類して整理したり，検索したりする技術のニーズが高まってきた．Web上においても検索エンジンのホームページでは画像検索のサービスが提供されているが，画像内容を理解した上での検索とは言い難く，さらなる精度の向上が求められている．また，画像検索はセマンティックWebなどコンピュータに自動的にコンテンツを理解させるという点からも注目されている。

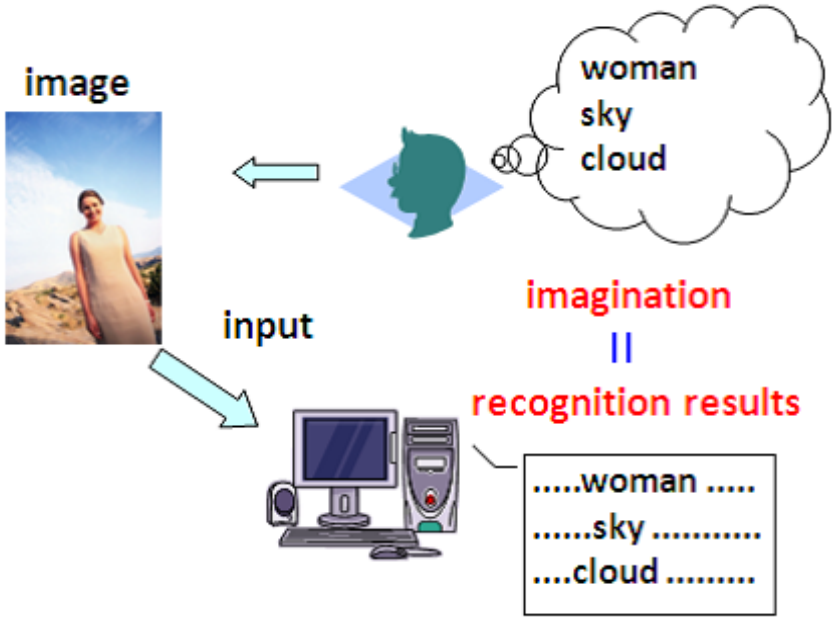


Fig. 1.1　Generic Object Recognition

一般物体認識はある画像が与えられたときにその画像内に映っている物体をコンピュータに自動的に認識させることであり，上記のようなアプリケーションに応用

が可能である．一般物体認識を最終的な目的としては，ある画像に存在する物体を認識することで人が画像から得る情報と同等の内容をコンピュータにも自動的に認識させることである．Fig. 1.1 に一般物体認識の基本的な概要を示す.

物体認識においては対象カテゴリにおける何らかの特徴を学習させ，未知画像内に存在する物体を識別する方法が一般的である．近年ではコンピュータの発達により大量のデータを学習させることが可能となり，コーナーやエッジなどの画像特徴を大量に学習して統計的な手法により，未知画像への認識を行うというアプローチが主流となってきた.

その中で Bag of Features (BOF)[7] モデルはテキスト処理におけるアプローチを画像に拡張したもので，画像中の局所特徴量のみを用いて物体認識を行う手法である．BOF モデルは現在その単純さと拡張性により様々な改良が行われている手法であり，シンプルな手法で高い識別性能を示している.

また，近年の研究では単純に画像からの特徴を用いて未知画像に存在する物体を認識するだけでなく，存在する物体間の関連性を考慮して認識を行うというアプローチも研究されるようになってきた．例えば，地面を認識することで人や車が地面に置いてあることを仮定して認識を行う [13] など画像内に写っている物体の周囲状況との関連性をアプローチに適用した手法も提案されている．しかし，物体間の関連性の扱いにおいては関連性を全て記述するのは困難である．その中で，一枚の画像中に複数カテゴリの物体が存在していることを利用して，物体カテゴリが一緒に存在しやすいかそうでないかという点に着目して既存の手法に取り入れた研究が注目されている．[24] 例えば，人と馬，人とバイクは同時に一枚の画像中に存在しやすいが，馬とバスはあまり同時に存在していないなどである．これらのカテゴリの共起の仕方を学習することで，複数物体が存在する画像の識別を行うというアイデアである．しかし，まだカテゴリの共起に関する研究はあまり行われておらず，共起性を上手く記述し取り入れることは物体認識において有用となるアプローチだと考えられる.

1.2　本研究の目的と着想

本研究では，物体カテゴリの共起性，例えばある 2 つの物体は同時に存在しやすいが他の 2 つの物体は同時に存在しにくいという情報をアプローチに取り入れることで物体認識を行うことを目的とする．前節で述べたように，単一物体が存在する画像ではなく，複数カテゴリの物体が存在する画像において物体認識を行う場合，物体カテゴリの共起性に関する情報は有用なものであると考えられる.

用いる物体カテゴリの共起情報としては，画像中におけるカテゴリの存在比率を学習させていく．存在比率は画像における全特徴点の数のうち，各カテゴリの物体から生じている特徴点の数の割合で表す．実験結果としては単一のカテゴリが画像中に存在するかどうかという結果ではなく，画像における対象カテゴリの存在比率を最終結果として求めていく．これにより，画像中に複数カテゴリの物体が存在している場合にも存在比率を求めることによりまとめて認識することが可能になる．

1.3　本論文の構成

以下，本論文では次のような構成となっている．第 2 章で，一般物体認識，Bag of Features ，カテゴリ間の関連性に関する関連研究を紹介し，第 3 章で既存のアプローチに提案したカテゴリの共起情報を組み入れる手法を述べる．続いて第 4 章で提案手法に対して実験を行い，実験結果に考察を加えていく．最後に第 5 章で本研究をまとめ，今後の課題を述べる．

第2章　関連研究

2.1　一般物体認識の概要

2.1.1　基本的な一般物体認識の説明

一般物体認識とは画像をコンピュータに自動的に理解させるシステムの一環として画像中に存在している物体を認識することである。日本国内では今まであまり研究が盛んに行われてはいなかった分野である。しかし，Web の発展やカメラの発達によって画像を個人単位においても画像を扱う機会の増加により，日本国内においても研究が盛んに行われるようになってきた．柳井 [31] は一般物体認識に関して国内研究者向けに，歴史や現状，今後の課題等について紹介しており，今後さらに研究が盛んとなる分野だと考えられる．

一般物体認識という分野においては一概に物体を認識するといっても何を目的とするかによって全くアプローチが異なってくる．例えば，画像内に富士山が写っているとする．すると,「富士山」そのものが存在しているかどうかを認識するのか,「山」というカテゴリ内のものが含まれているかを認識するのか,「富士山」の位置を認識するのか，それとも山中湖や川口湖などの関係から「富士山」や「山」のシーンという画像全体としての認識を行うのかなど様々な目的が考えられ，それによってアプローチも様々に考えられる．

しかし，一般的に物体認識という分野は困難であるとされていて，実用的なレベルにおいては人間の正面方向における顔認識ができるくらいである．人は数万種類の物体を認識することが可能だと言われている．しかし,「椅子」や「机」などでさえ機械に自動的に認識させるのは難しい．画像からの物体認識を困難としている理由はいくつか挙げられる．代表的なものを以下で述べる．

視点 画像において撮影された物体は静止している．よって，視点が異なると物体の見え方が変化するため同一物体と認識するのが困難となる．

照明 照明の当たり方によっては，物体に影ができたりと，見え方が変化してしまう．

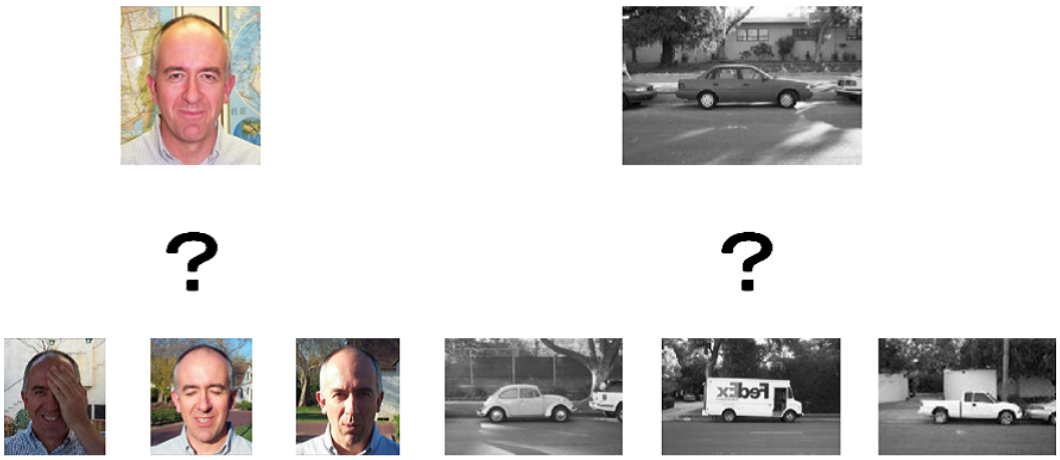


Fig. 2.1　物体認識における障害：左…オクルージョン，顔の変形，照明条件の変化，右…クラス内の変動

オクルージョン 人の顔を認識する際，画像中で髪などにより一部が隠されている場合はコンピュータが人の顔だと判断するのが難しくなる．

スケール 遠すぎてよく見分けられないなどの問題．しかし，人は周りの関係から認知できる場合もある．

物体の変形 動物などによくみられる．猫を例にとると寝転んだり立ったり毛の長さも様々だったりと変化が著しい場合もある．

背景のノイズ 画像中に背景が複雑すぎて，その前に存在している物体が見えにくい場合もある．

クラス内の変動 一概に「椅子」と言っても「パイプイス」や「ベンチ」,「ソファ」に近いものなど様々な形が存在する．

これらの問題が重なり合って存在しているため，現実的な研究においてはある程度認識する物体を限定して研究を行っている．一般的な認識手順としては，限定されたいくつかの物体の特徴を学習させ，未知画像中にどの物体の特徴が表れているかを判別することで物体を認識するという方法をとる．そのため，特に学習画像において上記の障害があまり存在せず，ラベル付けされたデータベースがいくつか国際標準的に用いられている．

一般物体認識の最終的な目標は，人間があるシーンを見たときにそのシーンに対して理解することをシステムに画像を通すことで計算機に自動的に理解させること

である．しかし，人間はシーンを理解する際，単に目に映ったものだけを理解するのではなく，シーンにおける目に映ったものの関連性をも考慮して理解を行う．よって，一般物体認識の研究は機械による人間認識機能の実現という点で，人工知能の分野にも応用が可能な研究であるといえる．

2.1.2　基本的なアプローチ

物体認識のアプローチは上記したように様々な種類が存在している．本研究ではその中で画像中に対象としたカテゴリの物体が写っているかどうかを認識するアプローチについて述べていく．上記したように物体認識は学習と認識の２つの段階から成り立っている．カテゴリ認識においては対象カテゴリ内に属するオブジェクトの特徴を学習させなければならない．近年において一般物体認識の分野における研究が盛んになってきた理由には，コンピュータの発達によって今まで扱えなかったデータ量が扱えるようになったことで統計や機械学習の分野における学習手法が画像へと適用できるようになってきたことが大きい．

ある画像が与えられた時に，何が写っているかをコンピュータに判断させたいとする．すぐに思いつくのは色情報であると思われる．木は茶や緑，空や海は青と色がある程度決まっているカテゴリがいくつかあるからである．色で画像をある程度切り分けて，ひとつひとつの領域に対して何が写っているかの認識を行う．このような考え方がベースとなって発展したのが領域に基づく方法である．また別のアプローチでは，ある位置に目みたいなものが写っていて，別の位置に鼻みたいなものが写っている．しかも目は２つあり中間に鼻があるのでこれは人の顔である，といったように画像内に写っている部分部分を組み合わせて，その特徴や位置関係から物体認識を行うといった考え方から発展したのが局所特徴量に基づく方法である．局所特徴量とは，あるせまい範囲（局所）における耳や目などのパーツの特徴を表すものである．この２つのアプローチが近年の一般物体認識における重要な手法である．

領域に基づく方法は基本的に領域分割（セグメンテーション）を行って，領域に対して自動的に物体の名称をつけていくアノテーションを行うための手法である．学習画像においては予め領域分割とラベリングがされているものを用いていく．このアプローチは画像を１枚クラス分類していくのではなく，大量の画像に対して複数のキーワードをつけるために提案された手法である．代表的な手法として，word-image translation model が挙げられ，Duygulu ら[9)] の研究が有名である．Translation model とは，日本語を英語に翻訳する際に，辞書を参照すると日本語１語１語に対応する

Fig. 2.2　Translation model[9)]：あらかじめセグメンテーションを行った領域に対して，その領域の物体を表すラベルがついている．

英語が存在するように，セグメンテーションを行った領域に対応するラベルを確率モデル化によって学習した辞書として，テスト画像の領域に尤もらしいラベルを対応させていく手法である．Translation model は最初は注目されたものの現在ではあまり用いられなくなっている．その理由としては初期の領域分割の結果にその後の処理が影響されてしまうためである．そのため，現在では領域に基づく方法ではなく，次で述べる局所特徴量を利用した方法の方が有効であるとされている．しかし，局所特徴量を用いた方法に比べ，領域にラベルがつくという結果はわかりやすいため，局所特徴量を利用した研究に応用されたりと様々な可能性が考えられる．（Fig. 2.2）

領域に基づく方法と比較して，局所特徴量を用いた方法は，一般的にセグメンテーションが不要であり，物体の変形や視点変動の問題にも対処できるという利点をもっている．局所特徴量を用いた手法において代表的な手法が Burl ら [6)] が提案した constellation model（星座モデル) である．Constellation model は星座が星の位置関係で形を表されるように，局所領域の特徴とその位置関係を確率モデルで表現したもので，物体の部分部分の特徴と相対的位置によって認識を行う手法である．この研究を発展させて，motorbike, airplane, face, car(side), car(rear), spotted cat など多くの種類のカテゴリに対応させたのが Fergus ら [10)] の研究である．

Fig. 2.3 が Fergus らが提案したモデルである．下の図におけるピンクの点は画像内における特徴点であり，色のついた楕円が画像内で仮説にもっともあてはまったものである．特徴と相対的位置を考慮することで，ここでは 6 つの局所特徴から顔であることを認識している．顔の向きとスケールはほぼ揃えられているものの，様々な人物を認識することができ，クラス内変動にも柔軟に対応している．

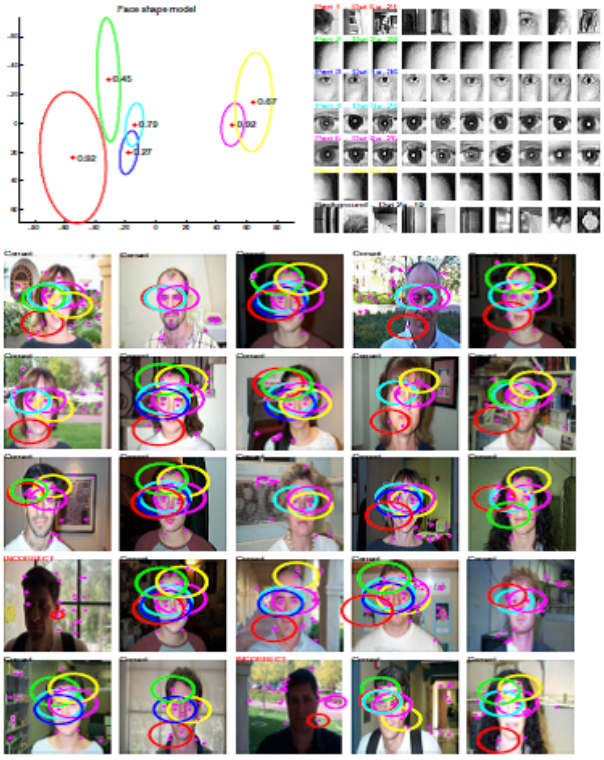


Fig. 2.3　Constellation model[10)]：左上の図が学習された部分の相対位置関係モデルである．右上は学習した部分の特徴であり，下の図では相対位置と特徴によって人であることを認識している．

2.2　Bag of Features

2.2.1　Bag-of-keypoints

Constellation model においては局所特徴の相対的位置の情報も確率モデル化することで用いていた．これに対して局所領域の特徴量のみを用いて，位置情報を用いず認識を行う方法が提案された．それが Csurka らが提案した Bag-of-keypoints [7] モデルである．Bag of Features (BOF) とも呼ばれるこの手法は統計的言語処理における Bag of Words model のアプローチを画像に適用したものである．Bag of Words model はテキスト分類の分野において語順を無視して，文書を単語の集合として考えてテキストを特徴づけて分類を行っていく手法である．

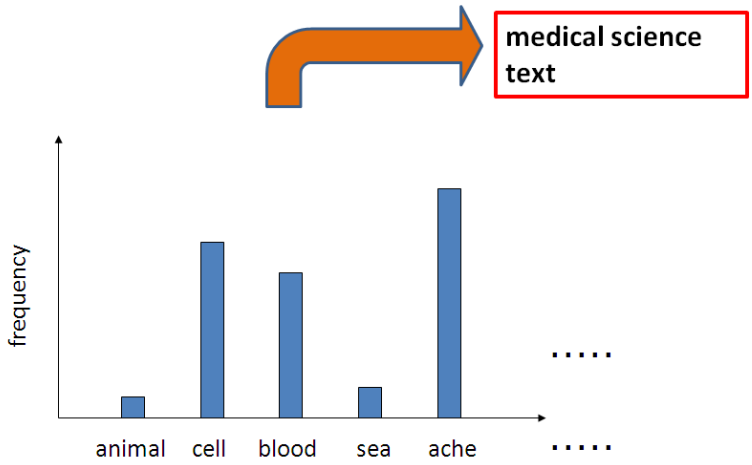


Fig. 2.4　Bag of Words model：単語の出現頻度によって，テキスト分類を行う．図の例では，医学用語の頻度が大きいため医学文書に分類することができる．

ある文書を単語の集合として，Fig. 2.4 のようにヒストグラム作成したときに図のように医学用語の頻度が高くそれ以外の語の頻度が低ければ，その文書は医学関連の文書であると分類することができる．同様に，Bag-of-keypoints モデルでは Fig. 2.5 のように Bag of Words モデルにおける単語を，画像における局所特徴 (keypoints) の集合として捉える考え方であり，位置情報は考慮しない．実際の処理においては，画像の局所特徴を単語として扱うために，クラスタリングによってある程度特徴が似ているものを同一のものとみなす必要がある．このクラスタリングによって，コードブック（ dictionary や vocabulary とも呼ばれる）を作成し，keypoint を word として扱えるようにする．このクラスタリングされた特徴を visual word と呼ぶこともあ

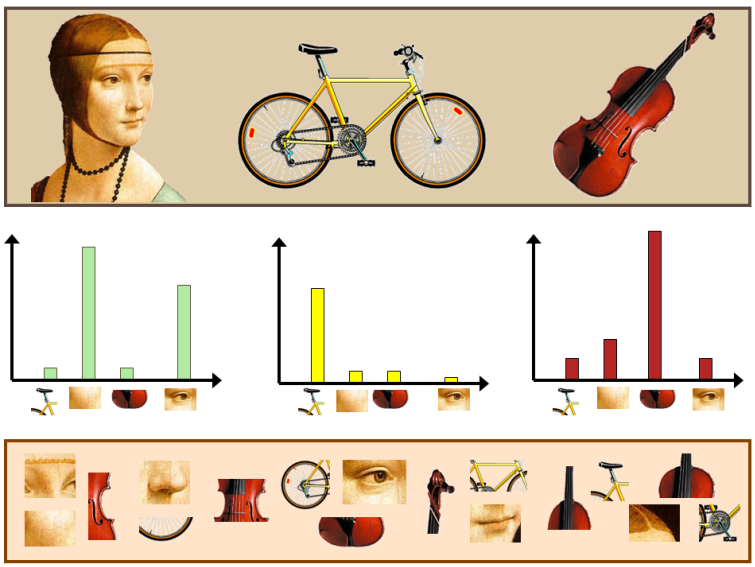


Fig. 2.5　Bag of Features model[5)]：Bag of Words model と同様に画像をカテゴリの特徴における出現頻度で表す.

る．画像はこの visual word の集合（bag）として，ヒストグラムとして表現される．こうして作成された対象カテゴリのヒストグラムを SVM などの識別器に学習させることでテスト画像のヒストグラムとの類似度を算出し，対象物体を認識していくアプローチである.

2.2.2　Bag of Features の関連研究

Bag of Features はアルゴリズムの単純なこともあり，様々な用途へと応用されている．Nilsback ら [21)] は一目見ただけでは区別できないような花の画像に対して，形・色・テクスチャ（模様）の 3 つの特徴からコードブックを作成し，その 3 つのコードブックを結合させることで花の種類の識別を行った．また，上東ら [29)] は Bag-of-keypoints のアプローチで Web 画像の分類を行い，湯ら [30)] は動画を対象とした映像認識を切り出されたフレーム画像を利用して行った.

他にも Bag of Features のアルゴリズムにおけるいくつかのフェーズに分かれた各構成要素の改良や他のモデルへの応用など様々なアプローチへと適用されている．特に各構成要素の改良においては，Bag of Features モデルの特徴的な部分といえる，コードブックを作成し，画像を visual word の集合（bag）を参照することで，ヒス

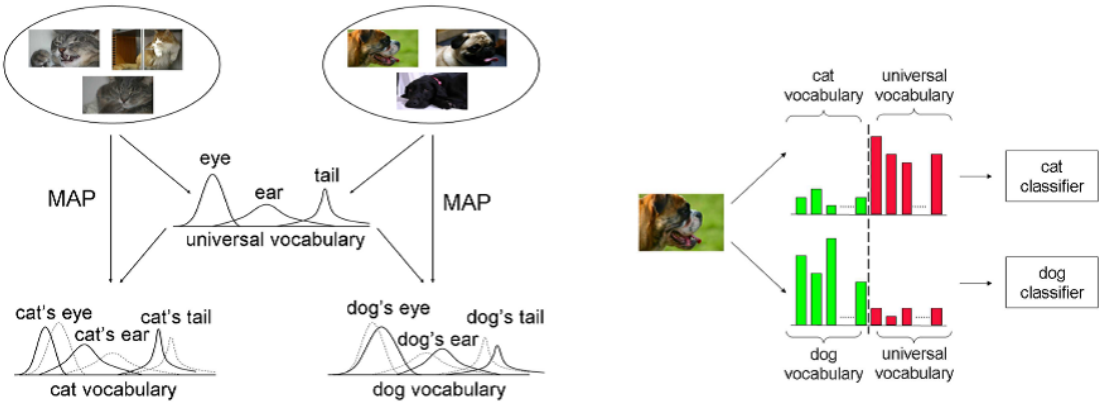


Fig. 2.6　Adpted vocabulary and bipartite histgram[22)]：全体の vocabulary では見分けがつかない特徴をオブジェクト毎の vocabulary によって識別する

トグラムとして表現するアルゴリズムに関して様々なアプローチがとられている．コードブック作成における特徴点のクラスタリングに関して，Moosmann ら[20)]は Randomized Clustering Forests と呼ばれるランダムに構築したクラスタのツリーの集合体を用いてクラスタリングを行った．また，他には Jurie ら[14)]は mean-sift を利用し，Mikolajczyk ら[19)]は階層的コードブックを作成し，効率的なクラスタリングを行っている．コードブックそのものを改良した研究では，Winn ら[26)]が識別精度を落とさず，コードブックのサイズを減らす手法を提案している．他には Bag-of-keypoints モデルを提案した Csurka らを含むグループにおいて，Perronnin ら[22)]がコードブックとヒストグラムの改良を行っている．彼らは universal vocabulary と呼ばれる対象カテゴリ全ての要素から最尤推定によって作成されるコードブックの他に，MAP 推定を適用することでカテゴリクラス毎の class vocabulary を作成し，２つのコードブックにより表現されるヒストグラムを結合させることで認識を行った．（Fig. 2.6）このアプローチによって，全カテゴリから作成したコードブックからでは見分けがつきにくい犬や猫などのカテゴリにおいても，カテゴリ毎のコードブックを作成することで認識ができている．

2.3　物体カテゴリの関連性

一般物体認識においては対象とするカテゴリをカテゴリ毎に認識していく手法が一般的である．PASCAL[3] などに代表されるようなコンテストにおいてもカテゴリ毎に精度を競い合うものである．しかし，一般的な物体認識の手法においてはカテゴリ同士の関連性などは暗黙的に考慮されてはいるが，複数の認識対象カテゴリが存在しているテスト画像においても画像全体というよりもカテゴリ毎に認識を行う手法が多数である．本節では認識対象とするオブジェクトカテゴリについてをカテゴリ間の関連性という観点とともに述べていく．

2.3.1　カテゴリの扱い

既存の研究においてカテゴリクラスの扱いを考慮した手法に Sivic ら [25] の研究がある．彼らは Bag-of-keypoints approach を用いて，大量の画像に対して probabilistic Latent Semantic Analysis (pLSA)[12] を適用することで，自動的にいくつかの分類クラスを決定するアプローチを提案した．

Fig. 2.7　Doublet による画像の表現 [25]：Doublet による記述により複数カテゴリの物体も認識できている．

教師無しの学習画像から自動的にクラスを探し出すというアイデアは，予め認識対象とするクラスを決めて，それに対応する学習画像を集める従来の一般的な方法とは全く異なるアプローチであった．また，提案した手法の中では認識対象とするカテゴリクラスだけではなく背景カテゴリについても言及しており，背景としていくつかのカテゴリクラスを組み入れた実験も行っている．また， doublet と呼ばれる特徴点を空間情報を考慮してペアにしたものを用いて，オブジェクトのレイアウトを表現しており，doublet を利用することで画像内で複数カテゴリの複数オブジェクトのレイアウトを表現し，識別している．(Fig. 2.7)

彼らの研究では大量の画像から自動的にクラスを探索し，doublet を利用することで複数カテゴリの物体を認識している．クラスを探索する際に学習画像として Caltech 101[1)] のデータを用いており，Caltech の画像においては１つの画像に１つのオブジェクトしか存在していないため，学習段階において次節から説明するようなカテゴリの共起は考慮されていない．しかし，画像中における複数カテゴリの物体を doublet を利用することまとめて認識するというアイデアは一般画像における複数物体の存在を明示的に扱ったという点で注目すべき研究だといえる．

また，Sivic らは背景カテゴリの追加についても言及しており，いくつかの背景カテゴリを考慮に入れるとカテゴリのクラス分けの精度が良くなるという実験を行っている．また，Zhang ら[27)] はカテゴリ識別において背景も有用な情報を持つことを示しており，背景カテゴリの導入によって対象物体カテゴリの認識精度左右されると考えられる．

2.3.2　物体カテゴリの関連性

一般物体認識においては 2.1.2 節で上記したように基本的には２つの手法に分けられる．領域に基づく方法では，画像内に複数のカテゴリ，複数のオブジェクトが存在していても領域分割を行うことでカテゴリ別，オブジェクト別に物体認識を行っている．また，局所特徴量に基づく方法においても基本的に単独の物体を認識するのに用いられる．しかし，実世界を撮影した画像中には複数の物体が含まれ，物体間にはそれぞれ何らかの関係をもって存在しているのが普通である．

例えば，Fig. 2.8 は交差点の画像であるが，奥の緑色のボックスと赤色のボックスはスケールが小さすぎるため画像特徴として表現することが難しく，そのものでは認識することが難しい．しかし，人間は道路上にあるという情報や交差点であるという情報を用いて，緑色のボックスは車であり，赤色のボックスは人であるとの認

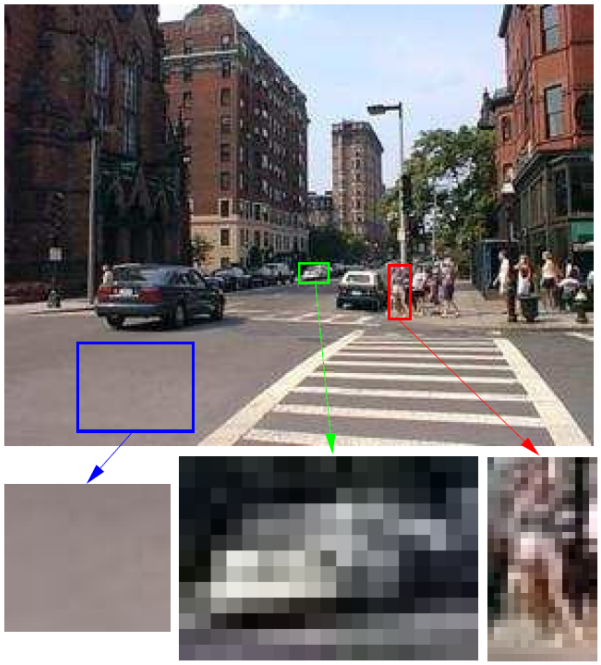
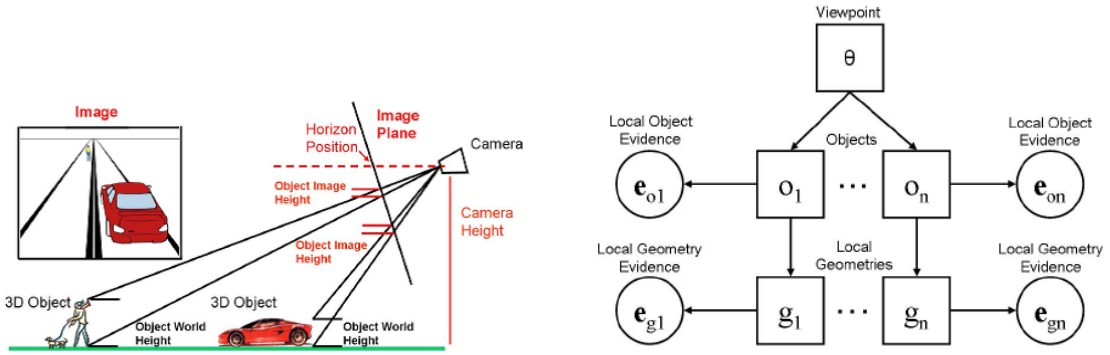


Fig. 2.8　物体同士の関連性 [13]：いくつかの認識可能な物体からシーンが理解できれば，小さくて特徴がとれない物体も識別可能となる．

知ができる．このように人がシーン理解に利用するように物体同士の関連性を物体認識に利用するというアプローチが考えられるようになった．物体間の関連性がわかっていれば，たとえ，外見だけでは何かわからないものでも認識が可能となる場合がある．図の画像においては，青色のボックスが道路であることが認識することで，緑色のボックスが車であることを認識するという手段である．こうした物体の関連性を利用した物体認識手法は context を用いた認識と呼ばれている．

物体同士の関連性を視点位置と地面・空・ビルなどの表面形状から確率モデル化して認識に適用した研究として Hoiem ら [13] の研究がある．Hoiem らは，視点位置がオブジェクトのアピアランスに大きく影響すると考え，視点位置を求めるための変数として，水平線の位置とカメラの高さを定義した．(Fig. 2.9(a)) その後，Viewpoint, Objects，Surface Geometry の関係について，Fig. 2.9(b) のようにモデル化した．

つまり，水平線の位置とカメラの高さによって定義される視点位置はオブジェクトの位置やサイズに対して影響を与える．また，すべてのオブジェクトは地面に置いてあるという仮定のもとで，オブジェクトは表面形状に直接影響を与えるというものである．この研究においては自動車と歩行者しか扱っていないが，２つとも地面に存在しているとの仮定により，水平線の下側で上手く認識を行った．



(a) Scene projection　　(b) Graphical model

Fig. 2.9　Hoiem らの手法：視点，表面形状を考慮

Hoiem らの研究はある程度対象カテゴリを限定して，関連性について確率モデル化を行っている．しかし，対象カテゴリが多くなってくると大がかりなモデルが必要となってくるため，あまり現実的ではなかった．そこで，物体間の関連性を考える context を利用したアプローチにおいて，一般的な画像中に複数カテゴリの物体が写っていることを利用して，あるカテゴリ同士が一緒に存在しやすいまたは存在しにくいという画像中の物体の共起という情報を利用するという方法が考え出された．例えば，人と車，車と自転車，人と馬などは一緒に存在しやすいが，馬とバス，羊とバイクはあまり一緒には存在しないなどである．これらのカテゴリ共起など画像における特徴の共起について着目した認識手法について紹介する．

Guo-Jun ら [23] は Multiple Instance Learning (MIL)[18] と呼ばれる手法を使って領域特徴の共起を考慮したアプローチを提案した．MIL は 2 × 2 のピクセル間の色情報を領域特徴として 1 枚の画像中から抽出し，それをまとめて Bag としてその画像を特徴づける．その後，Bag 内の領域特徴を多次元空間に投影し，正例画像に共通して存在し，負例画像に存在しない特徴を手掛かりに認識を行うものである．(Fig. 2.10)

Guo-Jun らは MIL を改良したものを Concurrent MIL (ConMIL) と名づけ，Bag 内の領域特徴の共起をテンソル表現を利用することにより考慮した．これにより，領域特徴同士の関係性を記述し，シーン認識において有用となる手法を提案した．(Fig. 2.11)

カテゴリの共起を具体的に扱った手法に Rabinovich ら [24] の研究がある．彼らは Conditional Random Field (CRF)[16] を用いてオブジェクトカテゴリの共起を表現した．CRF はイメージラベリングの分野において，ラベル同士の関連を考慮する際に

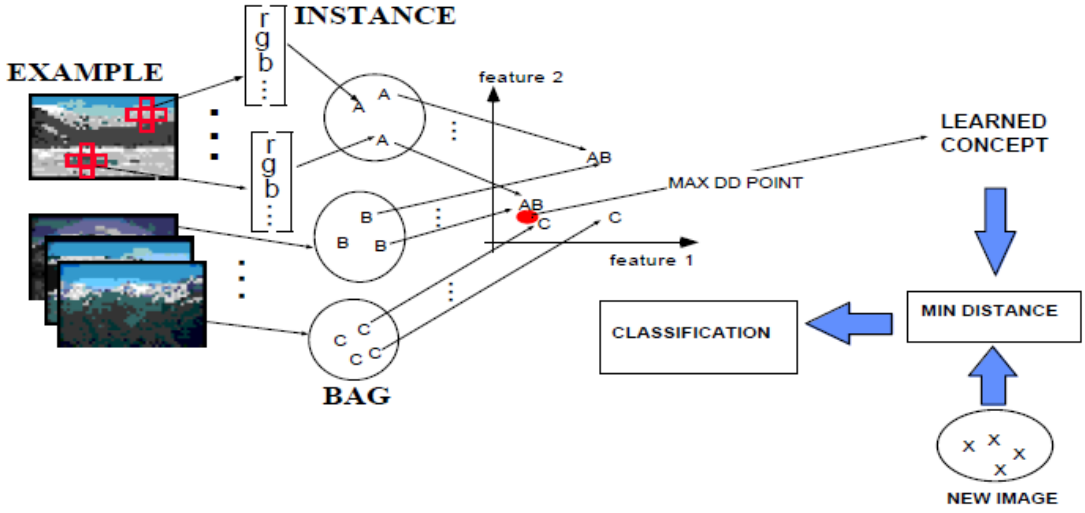


Fig. 2.10　Multiple Instance Learning[18)]

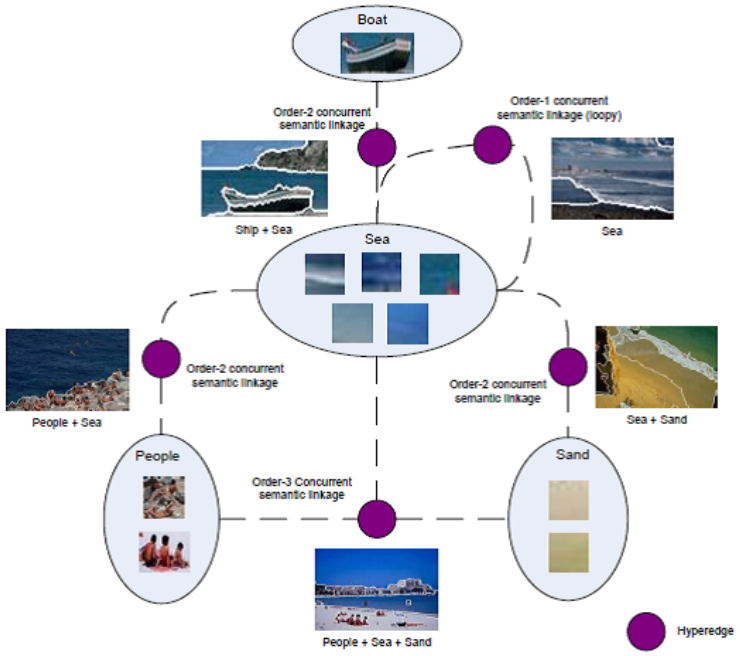


Fig. 2.11　Concurrent Multiple Instance Learning[23)]

用いられてきた．[11)15)] Rabinovich らは基本的な BOF の手法に対して事前処理にセグメンテーション，事後処理に context を CRF を用いて取り入れることで物体認識の精度を上げるモデルを提案した．（Fig. 2.12）

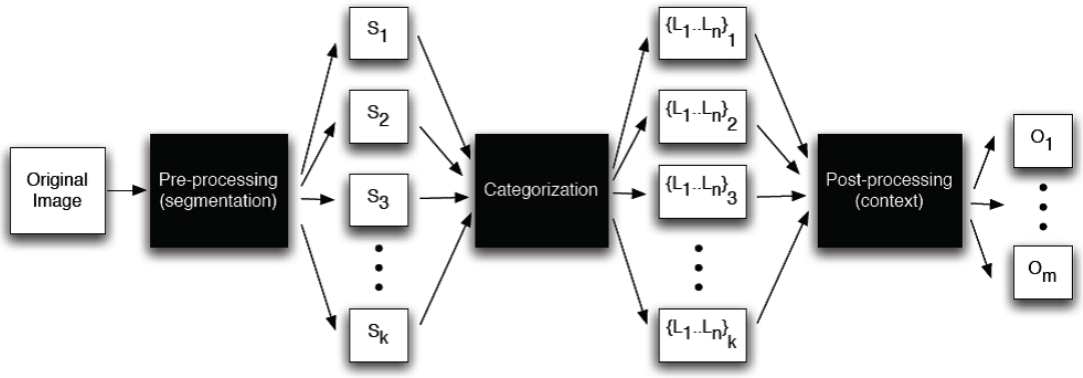


Fig. 2.12　Object Categorization using Semantic Context[24)]

Rabinovich らは context を学習するために Google Sets[2)] を用いた．Google Sets は入力語に関連のある言葉を出力してくれる．Small Sets と Large Sets があるが，あるカテゴリのラベルを入力した際に出力される語の中に別の対象カテゴリのラベルが含まれていれば共起しやすいとみなすことができる．この共起性を後処理として付け足すことで，カテゴリラベルの更新を行った．Rabinovich らの研究はオブジェクトレベルでの共起を物体認識に取り入れた点で，context によって物体認識を行うという研究に新しい概念を与えた．

Fig. 2.13 は Rabinovich らの実験結果である．context を用いていない時に誤認識していた結果もカテゴリの共起性を学習させて取り入れることで，正しくカテゴリラベルの更新が行われている．

2.4　本研究の特色

本研究では，BOF の枠組みの中でカテゴリ間の関連性として共起情報を考慮して物体認識を行っていく手法を提案する．具体的には BOF の出力として得られるテスト画像のヒストグラムが各カテゴリのヒストグラムの線形結合となっていると仮定し，MAP 推定によって結合係数を推定する．MAP（Maximum A Posteriori）推定は事後確率を最大にするようなパラメータを探索するアプローチであり，尤度だけで

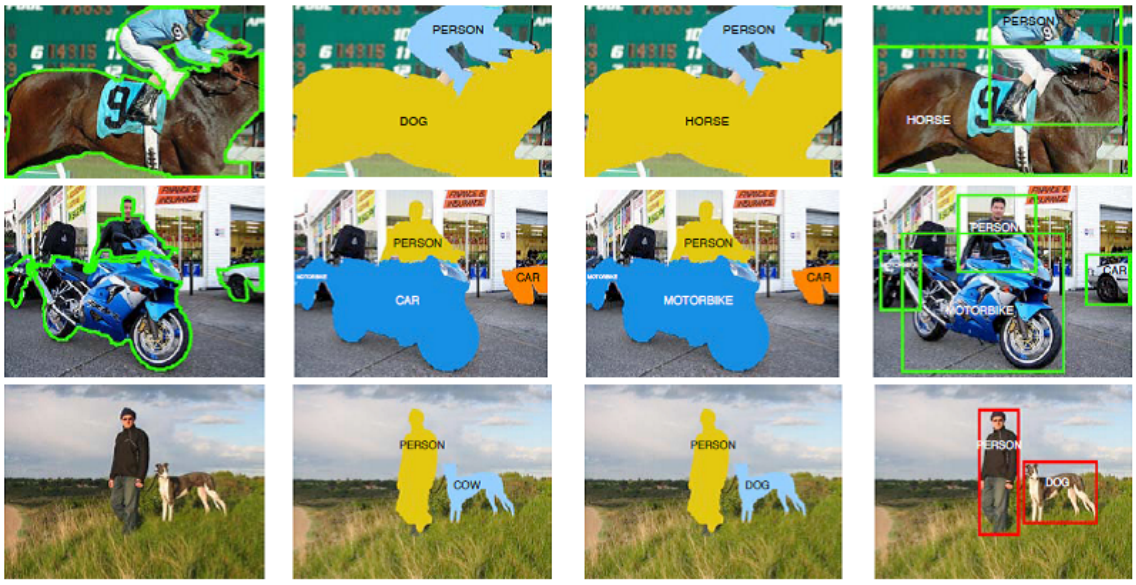


Fig. 2.13　カテゴリの共起性を用いた認識 [24]：(a)…元の画像にセグメンテーションを行ったもの　(b)…Context を用いていない認識　(c)…共起性を組み入れた認識　(d)…Ground truth

なく事前確率も考慮したものである．本手法ではカテゴリの共起を事前確率として学習させる．学習画像やテスト画像のデータとしては一般物体認識の標準的な評価用データセットとして知られている PASCAL 2006[3] のデータベースを用いる．

本研究では共起情報を考慮した物体認識を実現しているが、提案手法には次のような 3 つの特色がある。一つ目はセグメンテーションを行わないことである．Rabinovich らの手法も Guo-Jun らの手法もまずセグメンテーションを行い，カテゴリの共起や領域特徴の共起を考慮していた．しかし，複雑な画像においてはセグメンテーションが困難であり，またセグメンテーションの結果に後の処理が影響されてしまうため，本研究ではセグメンテーションを行わずに物体認識を行っていく．

二つ目は BOF の枠組みの中で結合係数をパラメータとして，テスト画像が与えられた時に事後確率が最大になるパラメータを推定していく．具体的には各カテゴリのヒストグラムの集合において要素毎に分布を考慮することで尤度を表し，さらに事前確率として共起情報を組み入れることで MAP 推定を行う．

三つ目は学習する共起情報としてはカテゴリの存在有無ではなく，各カテゴリの画像中における存在比率を利用することである．存在比率は画像中におけるカテゴリの bounding box 内の特徴点の数と画像全体の特徴点の数の比率で表し，面積の割

合として考えることができる．これにより最終結果をカテゴリの存在比率で表していく．

　これにより，最終的には画像中における各対象カテゴリの存在比率を求めることを目的とする．これにより，従来の手法が結果としていたあるカテゴリのオブジェクトが存在しているかどうかという画像理解よりも一歩進んだ形での物体認識が可能となる．また，物体の存在比率を求めることで，ある画像において何が主に映っているかということやシーンの理解にも利用することができ，高度な画像検索が可能になると考えられる．以下 Fig. 2.14 に本研究のフローチャートを簡単に示す．

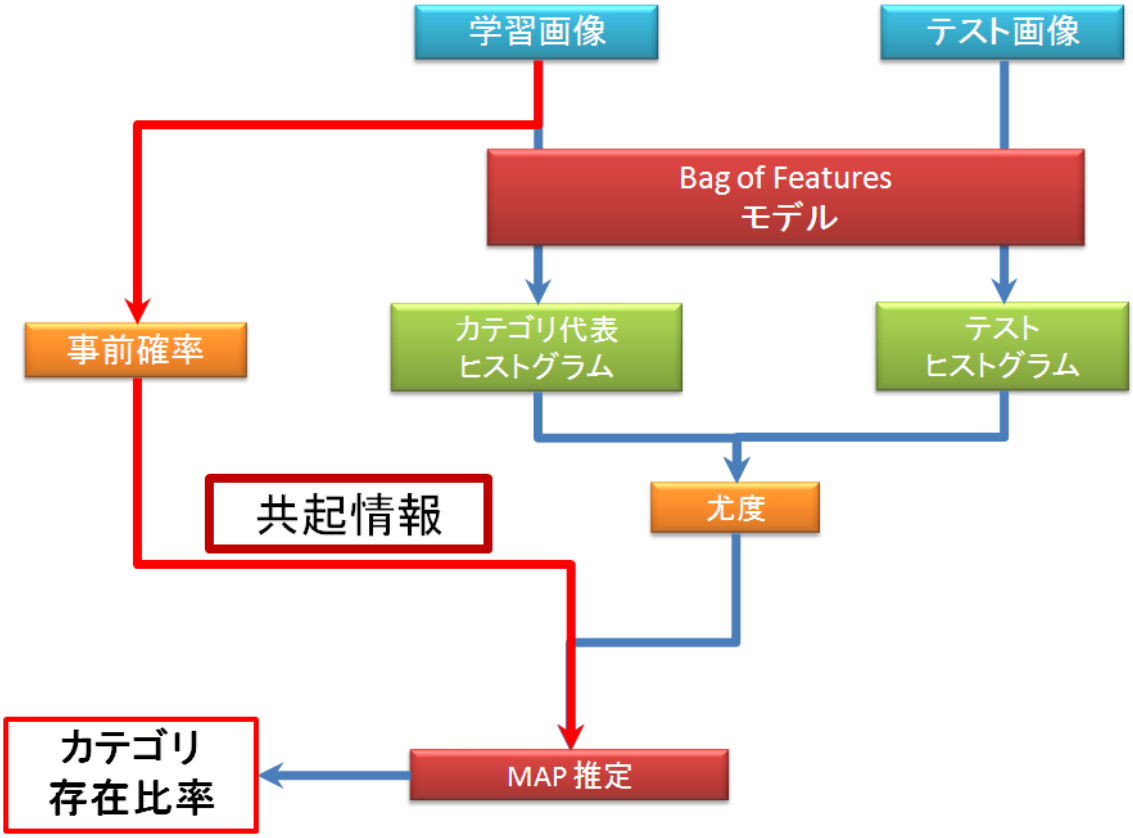


Fig. 2.14　Flow Chart

第3章　手法説明

3.1　Bag of Features

3.1.1　概要

Bag of Features (BOF) モデルは位置情報を用いないことで視点やオブジェクトが一部隠れていても認識ができる手法である．BOF は現在その単純さと拡張性により様々な改良が行われている手法であり，シンプルな手法で高い識別性能を示している．本研究ではカテゴリの共起情報を導入するために，拡張性が高い BOF の枠組みに沿って研究を進めていく．まず，一般的な BOF の手法を以下で説明する．

Bag of Features の手法は学習とテスト画像によるオブジェクトの認識という２つのフェーズに分けられる．以下でその２つのフェーズを説明する．

学習

1. 学習画像中から特徴点を探索し，記述する．
2. 抽出した特徴量に対してクラスタリングを行い，クラスタ中心を要素とするコードブックを構築する．
3. カテゴリ学習画像毎に特徴点をコードブックの要素に投票することでヒストグラムを作成する．
4. 識別器にカテゴリ毎にヒストグラムを学習データとして学習させる．

認識

1. 学習時と同様にテスト画像から特徴点を探索して記述する．
2. コードブックの要素に投票し，ヒストグラムを作成する．

3. 学習画像のヒストグラムをラーニング済みの識別器にテスト画像のヒストグラムをかけて認識を行う.

学習と認識の図は Fig. 3.1 で示す.

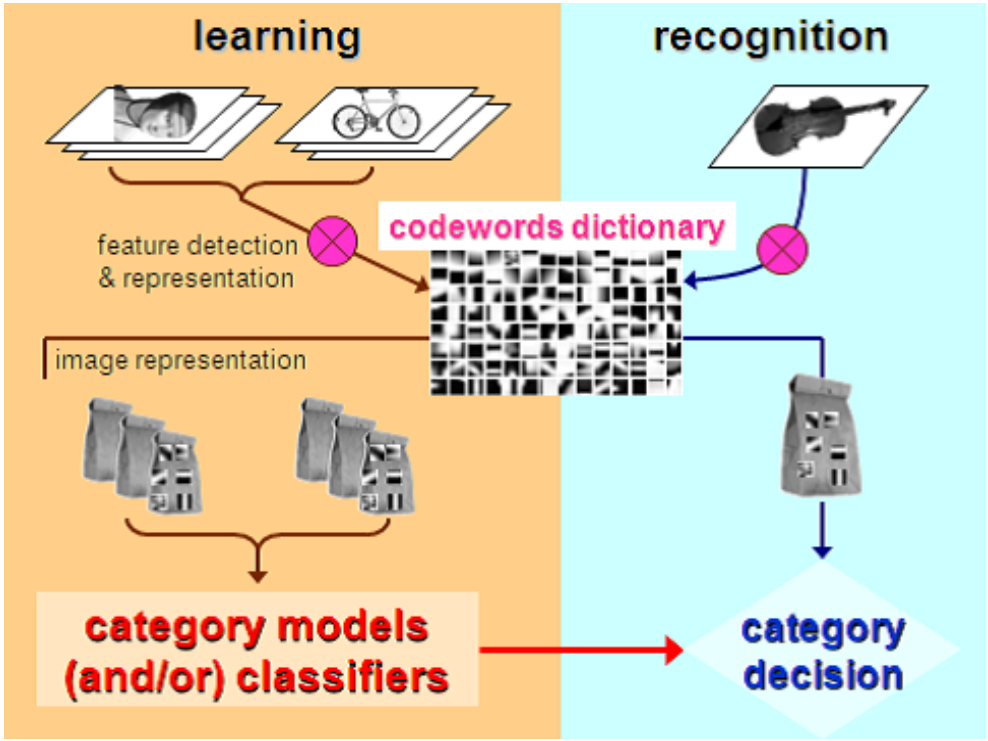


Fig. 3.1　Bag of Features model[5)]

Bag of Features モデルは近年，研究が盛んに行われており，学習や認識の各フェーズにおいて様々なアプローチが考えられている．Zhang ら [27)] は局所特徴と Supprot Vector Machine (SVM) に基づくテクスチャと物体カテゴリの認識に関する包括的な実験を行い，一般物体認識におけるモデルの構成要素に対して検証を行った．彼らはその中で特徴点の探索，記述のアプローチの比較，類似度を測るアプローチの比較，識別器のカーネルの比較などを行った．その中でシンプルな手法においても比較的高い認識性能が得られることを実験的に示した.

よって，本研究は Bag of Features モデルの各構成要素として，最も一般的と使用されているものを選択する．画像から局所特徴量を探索して記述する際には SIFT (Scale Invariant Feature Transform)[17)] を使用した．また特徴のクラスタリングやコードブック作成時には 2.2.2 節で上記したように様々なアプローチが存在するが，本研究ではクラスタリングにおいては k-means 法を用いて，ヒストグラムを構築する際

には一般的なユークリッド距離を用いる．また，識別段階においては SVM などの識別器を使わず，ヒストグラムを構築した後は共起情報を取り入れ，最終結果として画像中におけるカテゴリの存在比率を求める提案したアプローチを適用する．次節以降において各フェーズにおける本研究で行った BOF の手法の説明を詳細に行っていく．

3.1.2　SIFT 特徴量抽出

画像における特徴点は前述したように画像の拡大縮小，回転や視点のずれに対して，ロバストであるという特徴を持つ SIFT feature detector を用いて記述 (SIFT descriptor) していく．SIFT descriptor は，あるピクセルの代表輝度勾配方向を決定し，その方向を基準とした輝度勾配ヒストグラムを作成し，多次元ベクトルで特徴を記述したものである．

代表輝度勾配方向は全方向を 36 分割し，注目ピクセルを中心としたピクセルの輝度勾配にガウス窓関数をかけて作成した重み付け方向ヒストグラムの中で最大となる方向のものである．(Fig. 3.2)

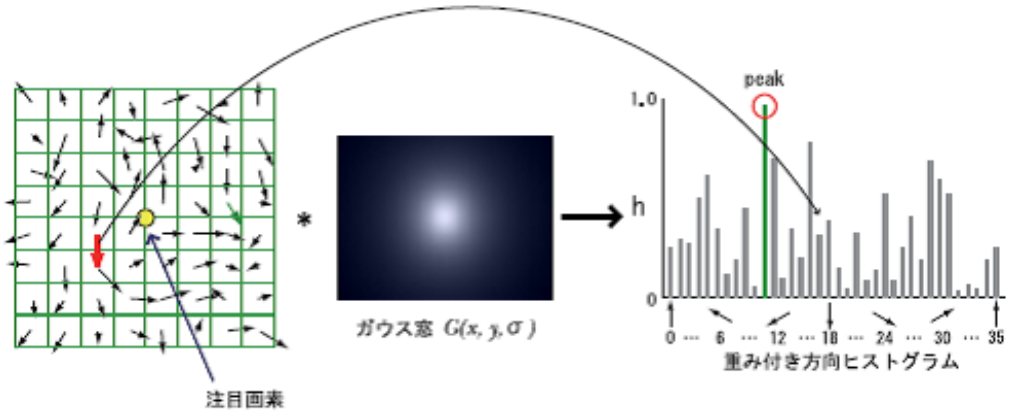


Fig. 3.2　重み付き方向ヒストグラム[28)]

この代表輝度勾配方向を基準とした周囲の輝度勾配ヒストグラムを作成する．注目ピクセルを中心とした 4x4 の領域に分割し，それぞれの位置で 8 方向の輝度勾配ヒストグラムを作成する．4x4 の領域にそれぞれ 8 方向ヒストグラムを作成するため，128 次元ベクトルの特徴量を持つことになる．この 128 次元の SIFT 特徴量を各ピクセルごとに抽出し，閾値によって特徴点として抽出する．(Fig. 3.3)

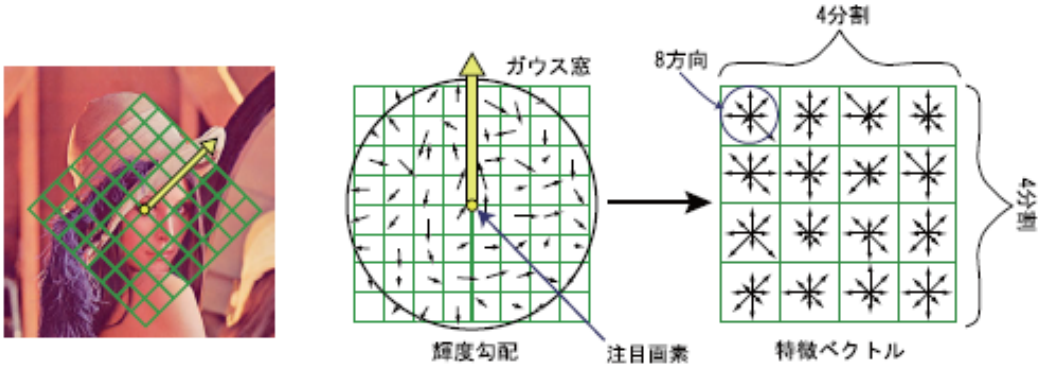


Fig. 3.3　SIFT 特徴量の記述[28)]

本研究では D. Lowe がホームページ上 [4)] で公開している SIFT Keypoint Detector のソフトウェアを使用して，特徴点抽出を行った．

3.1.3　コードブックの構築

コードブックとは学習する対象オブジェクトの全ての特徴点をひとつの辞書として表したものである．コードブック内の要素は 3.1.2 節で抽出した SIFT 特徴点の集合にクラスタリングを施したものである．理想的にはコードブックの各要素が各カテゴリオブジェクトにおける最も特徴的な部分を端的に表していて，それが他のカテゴリに見られない特徴であることが望ましい．しかし，現実的には人工的な要素が多数含まれるカテゴリと動物など自然的な要素が含まれるクラスタを区別することができたとしても，車とバスあるいは馬と牛など人工物同士あるいは動物同士などにおける特徴点は差別化を図ることが困難である．そのため，BOF では特徴点をクラスタリングしたものを用いて物体認識を行うのではなく，次節で述べるようにヒストグラムを作成することで認識精度を挙げている．

Bag of Features モデルにおけるクラスタリングの方法は 2.2.2 節で上記したように様々な方法が考えられている．本研究では最も基本的なクラスタリング法である k-measn 法を用いる．つまり，コードブックの要素数と特徴点は k-means 法によってクラスタリング中心を決定していく．コードブックの要素はクラスタリング中心であり，visual words と呼ばれる．一般的には対象とするオブジェクトの数が多ければ多いほどコードブックの要素数も多い方が望ましいが，演算効率や演算時間も膨大になっていくため適切な点を見つける必要がある．本論文では実装したシステムに

対し，適切な要素数を実験によって考察していく．

k-means 法は以下のようなアルゴリズムで実装される．データの数を n ，クラスタの数を K として説明を行う．

1. 各データ $x_i(i = 1 \cdots n)$ に対してランダムにクラスタを割り振る．
2. 割り振ったデータをもとに各クラスタの中心 $V_j(j = 1 \cdots K)$ を計算する．計算は通常割り当てられたデータの各要素の平均が使用される．
3. 各 x_i と各 V_j との距離を求め，x_i を最も近い中心のクラスタに割り当て直す．
4. 上記の処理で全ての x_i のクラスタの割り当てが変化しなかった場合は処理を終了する．それ以外の場合は新しく割り振られたクラスタから V_j を再計算して上記の処理を繰り返す．

結果は，最初のクラスタのランダムな割り振りに大きく依存することが知られており，1 回の結果で最良のものが得られるとは限らない．本研究では k-means 法の初期値依存の性質に対応するために一つの要素数で 10 回ほど k-means 法を試し，誤差が最小になるクラスタリング中心の集合をコードブックとして抽出している．

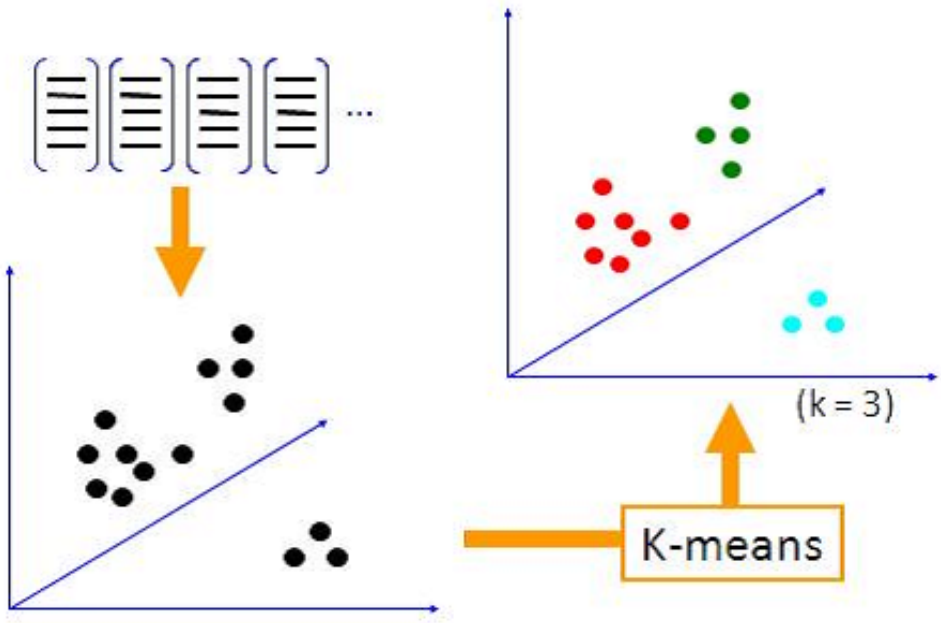


Fig. 3.4　k-means

3.1.4　ヒストグラムの作成

3.1.3 節で上述したようにコードブックを辞書として構築したが，今度はコードブックに記載されている要素から対象オブジェクトごとにヒストグラムを作成することでその画像を表現していく．

具体的には各カテゴリの学習画像から抽出した SIFT 特徴点がそれぞれコードブックの要素であるクラスタ中心との距離が最小となるものに投票して，ヒストグラムを作成する．出来上がったヒストグラムを画像において位置情報を用いずに特徴点の集合で表したものとして，各カテゴリ，各画像毎に画像を表現していく．(Fig. 3.5)

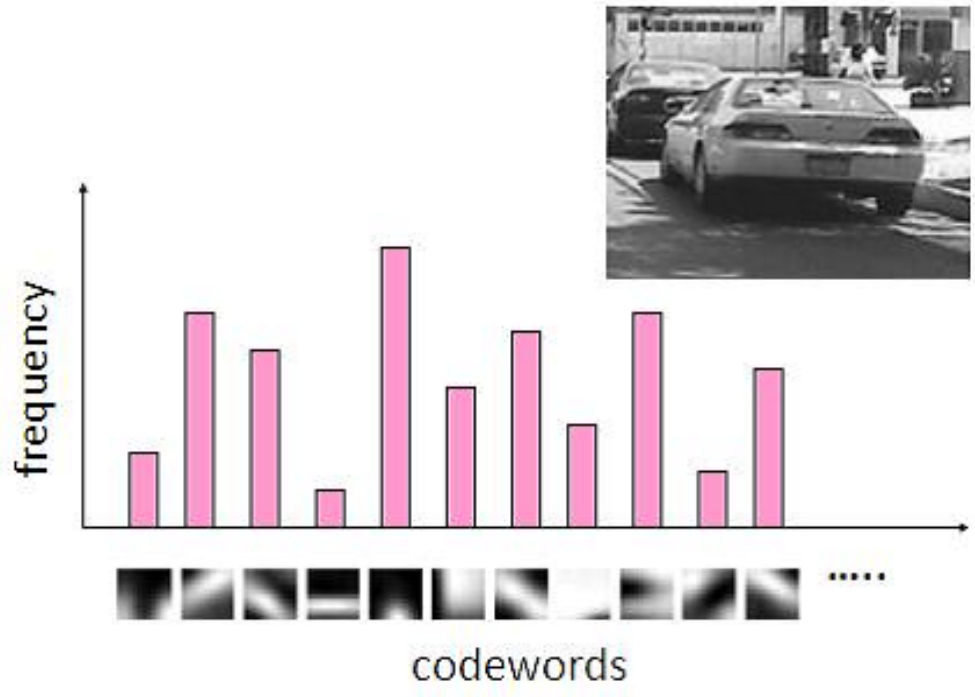


Fig. 3.5　Image representation

このコードブックのヒストグラムをそれぞれの対象オブジェクトにごとに学習画像の値として作成する．ヒストグラムは各カテゴリ毎に学習画像の枚数の数だけ作成される．ここでは，コードブックを構築するために用いた画像とヒストグラム作成するために用いた画像は同じものである．

Bag of Features モデルの既存の手法においては，このヒストグラムを SVM などの識別器に学習させることでテスト画像のヒストグラムを認識していく．位置情報を用いていないため，今までの手法と比較して演算効率が高く，動画や Web 上への適応や実時間での物体認識など様々な用途が考えられる．

本研究ではヒストグラムをそのまま識別器に学習させるのではなく，共起情報を利用するために異なるアプローチをとる．一般的に識別器にかける分類法においては，１枚の画像に対してそれぞれのカテゴリが含まれるかどうかの試行を行っていく．しかし，実際の画像には複数カテゴリの物体が含まれている可能性も考えられる．仮にあるカテゴリの物体が存在するとみなされたときに，そのカテゴリの物体と共存しやすいのはどのカテゴリの物体かという情報を学習することで，間違ったカテゴリ識別も改良できると考えられる．共起情報を組み入れた提案手法の詳細を次節から述べていく．

3.2　共起確率の導入

3.2.1　線形結合による表記

本研究では，ある画像において対象物体カテゴリがどれくらいの比率で写っているかを求め，物体認識を行うことを目的とする．本節では，3.1 節で説明した Bag of Features モデルから得られた結果を物体カテゴリの存在比率を導出するために適用した手法を説明する．

ある一つの画像が 3.1.4 節で述べたコードブックから作成されたヒストグラムで表される際に，そのヒストグラムが対象とするいくつかの物体カテゴリにおける代表（平均）ヒストグラムの線形結合となっていると仮定する．

具体的には式 (3.1) のように仮定する．

$$\vec{b} = \sum_{k=1}^{K} c_k \vec{h}_k \tag{3.1}$$

$\vec{b}$ はコードブックから作成されたテスト画像のヒストグラムである．また，$\vec{h}_k$ は同じコードブックから作成された各対象カテゴリの学習画像における複数のヒストグラムの代表（平均）ヒストグラムであり，K は対象カテゴリの数である．

式 (3.1) における結合係数 $\vec{c} = (c_1, c_2, ..., c_K)^T$ を推定することで，画像における対象カテゴリの存在比率を求めることができる．

まず，各カテゴリの平均ヒストグラムのみを用いると式 (3.1) は単純な線形問題とすることができる．求めるテスト画像のヒストグラムの次元も各カテゴリのヒストグラムの次元もコードブックの要素数として与えられる．一般的にこのヒストグラムの次元が対象とする物体カテゴリの数よりも大きければ結合係数 $\vec{c}$ は一意に決定する．つまり，

$$\vec{b} = H\vec{c} = [\vec{h}_1 \cdots\cdots \vec{h}_K]\vec{c}$$

$$\begin{pmatrix} b_1 \\ \vdots \\ \vdots \\ b_N \end{pmatrix} = \begin{pmatrix} h_{1,1} & \cdots & \cdots & h_{K,1} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ h_{1,N} & \cdots & \cdots & h_{K,N} \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ \vdots \\ c_K \end{pmatrix} \tag{3.2}$$

となる結合係数 $\vec{c}$ を推定することができる．ここで，N はヒストグラムの次元，つまりコードブックの要素数である．また，H は対象カテゴリの各代表（平均）ヒストグラム $\vec{h}_k$ を結合させたものである．
よって，

$$\vec{c}_{\mathrm{LS}} = \underset{\vec{c}}{\operatorname{argmin}} ||H\vec{c} - \vec{b}|| \tag{3.3}$$

を最小にする結合係数 $\vec{c}_{\mathrm{LS}}$ を探索することになり，線形最小二乗問題の最適化へと帰着することができる．本研究ではこの線形最小二乗問題の最適化を様々な制約条件の下で行うために，MATLAB の Optimization Toolbox 内で実装されている「lsqlin」関数を用いて実験を行っていく．

本研究では上記の式（3.3）で求めた結合係数 $\vec{c}_{\mathrm{LS}}$ を次節以降説明する MAP 推定最適化における初期値として用いていく．

また，画像内におけるオブジェクトの比率を求めるために結合係数 $\vec{c}$ の推定において制約を課していく．まず一つの目の制約は $\vec{c}$ の値は 0 以上であることである．これは，結合係数が単純にどのカテゴリのオブジェクトがどれくらいの割合で含まれているかということ表しているためである．またもう一つの制約として $\vec{c}$ の足した値が 1.0 となるようにする．これは学習させるヒストグラムもテスト画像のヒストグラムも特徴点の数をそろえるために正規化しているためである．この制約によって，必然的に結合係数 $\vec{c}$ の値は 0 ～ 1 範囲の値をとることになる．この制約条件を初期値として式 (3.3) を推定する際と後述する MAP 推定における最適化を行う際に取り入れる．

3.2.2　統計的要素の導入

上記で説明したアプローチでは，学習した対象カテゴリの各平均ヒストグラムしか用いておらず，各カテゴリ内におけるヒストグラムの分布を全く考えていない．よって，本節では統計的な要素を加味したアプローチとして最尤推定を適用した手法を説明する．

まず，最尤推定の一般的原理を説明する．[8)]

ω_j を求めたい未知変数とする．標本の集合をクラスによって分割し，その結果，D_i 内の各標本が確率 $p(\mathbf{x}|\omega_j)$ に従って独立に抽出されるような，a 個のデータ集合 $D_1, \ldots D_a$ が得られるとする．$p(\mathbf{x}|\omega_j)$ は既知のパラメータ形式を持ち，そのため，

パラメータベクトル θ_j の値によって一意的に決定されると仮定する．問題は，訓練標本から得られた情報を使って，各カテゴリに対応した未知のパラメータベクトル $\theta_1, \ldots, \theta_a$ の良好な推定結果を得ることである．

問題を簡単にするために，D_i 内の標本は $i \neq j$ のとき θ_j についての情報を何も持っていない，すなわち，異なるクラス間のパラメータは他のクラスの分布関数からは独立であると仮定する．これによって，各クラスを別々に考慮することが可能になり，また，クラスの別を表すインデックスを省くことで表記が簡単になる．この仮定の結果，次のような形式の a 個の別々の課題が得られる．すなわち，確率密度 $p(\mathbf{x}|\omega_j)$ のもとで互いに独立に抽出された訓練標本からなる集合 D を用いて，パラメータベクトル θ を推定せよという問題である．

D は n 個の標本，$\mathbf{x}_1, \ldots, \mathbf{x}_n$ からなると仮定する．すると，標本は独立に抽出されるので，式 (3.4) が得られる．

$$p(D|\theta) = \prod_{k=1}^{n} p(\mathbf{x}_k|\theta) \tag{3.4}$$

$p(D|\theta)$ を θ の関数だと見なせば，それは標本集合に関して θ の尤度であると定義される．θ の最尤推定値は，定義によれば，$p(D|\theta)$ を最大にする $\hat{\theta}$ である．この推定値は，実際に観測にされた訓練標本に最も適合する，または，観測値を最も支持するような θ に対応しているといえる．

本研究では，結合係数 $\vec{c}$ のときに $\vec{b}$ が得られる確率 $p(\vec{b}|\,\vec{c})$ を最大にする結合係数を推定するアプローチとして最尤推定を行っていく．まず，本手法では $p(\vec{b}|\,\vec{c})$ を計算するために，複数枚の学習した対象カテゴリのヒストグラムが要素ごとに正規分布すると仮定する．つまり，あるカテゴリのヒストグラム $\vec{h}_k$ において，n 番目の要素が $N(\mu_{k,n}, \sigma^2_{k,n})$ の正規分布に従うとする．また，各要素が独立しているとすると，$p(\vec{b}|\,\vec{c})$ は下の式 (3.5) のように表される．

$$p(\vec{b}|\,\vec{c}) = \prod_{n=1}^{N} p(b_n|\vec{c}) \tag{3.5}$$

よって式 (3.2) より，$b_n = \sum_k c_k h_{k,n}$ となるので，正規分布の再生性より，b_n は $N(\sum_k c_k \mu_{k,n}, \sum_k c_k^2 \sigma^2_{k,n})$ の正規分布に従うと仮定できる．

一般的な正規分布の式は式 (3.6) のように表される．

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{3.6}$$

このとき，本研究における $p(\vec{b}|\,\vec{c})$ は式 (3.7) のように表されるので，

$$p(\vec{b}|\,\vec{c}) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi \sum_k c_k^2 \sigma_{k,n}^2}} \exp\left[-\frac{1}{2} \frac{(b_n - \sum_k c_k \mu_{k,n})^2}{\sum_k c_k^2 \sigma_{k,n}^2} \right] \tag{3.7}$$

対数をとって式 (3.8) のようになる．

$$\ln p(\vec{b}|\,\vec{c}) = \sum_{n=1}^{N} \left[-\frac{1}{2} \ln 2\pi \sum_k c_k^2 \sigma_{k,n}^2 - \frac{1}{2} \frac{(b_n - \sum_k c_k \mu_{k,n})^2}{\sum_k c_k^2 \sigma_{k,n}^2} \right] \tag{3.8}$$

よって，最尤推定として事前確率を学習させていない段階では，対数尤度の符号を反転させ，定数部分を除去した以下の式によって $p(\vec{b}|\,\vec{c})$ を最大にする結合係数 $\vec{c}_{\mathrm{ML}}$ を推定する．

$$L_{\mathrm{ML}} = \sum_{n=1}^{N} \left\{ \frac{(b_n - \sum_k c_k \mu_{k,n})^2}{\sum_k c_k^2 \sigma_{k,n}^2} + \ln \sum_k c_k^2 \sigma_{k,n}^2 \right\}$$

$$\vec{c}_{\mathrm{ML}} = \underset{\vec{c}}{\operatorname{argmin}}\, L_{\mathrm{ML}} \tag{3.9}$$

これによりコードブックの分布という要素を既存のシステムに加味することができる．

3.2.3　共起情報を導入した MAP 推定

3.2.2 で説明した最尤推定のアプローチにおいて，各学習画像のヒストグラムを正規分布と仮定することにより統計的要素を考慮する．本節では，統計的要素に加え，学習画像におけるカテゴリの共起情報を事前確率として組み入れることで，最大事後確率を推定する MAP 推定を適用したアプローチを説明する．

事後確率はテスト画像のヒストグラム $\vec{b}$ が与えられた時に結合係数 $\vec{c}$ が得られる確率 $p(\vec{c}|\vec{b})$ として表され，この事後確率を最大にする結合係数 $\vec{c}$ を推定する．

$p(\vec{c}|\vec{b})$ はベイズ則により，次のように表される．

$$\begin{aligned} p(\vec{c}|\vec{b}) &= \frac{p(\vec{b}|\,\vec{c})p(\vec{c})}{p(\vec{b})} \\ &\propto p(\vec{b}|\,\vec{c})p(\vec{c}) \end{aligned} \tag{3.10}$$

$p(\vec{b}|\,\vec{c})$ は式 (3.7) で与えられるため，事前確率 $p(\vec{c})$ にカテゴリの共起情報を学習させる．本研究では事前確率 $p(\vec{c})$ を学習するために，$p(\vec{c})$ が多変量正規分布に従うと仮定する．一般的に多変量正規分布は次のように表される．

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right]$$

よって事前確率 $p(\vec{c})$ は

$$p(\vec{c}) \propto \exp\{-\frac{1}{2}(\vec{c}-\vec{\nu})^T\mathbf{\Sigma}^{-1}(\vec{c}-\vec{\nu})\} \tag{3.11}$$

となり，式 (3.8) と式 (3.10) より，

$$L_{\mathrm{MAP}} = L_{\mathrm{ML}} + \lambda(\vec{c}-\vec{\nu})^T\mathbf{\Sigma}^{-1}(\vec{c}-\vec{\nu})$$

$$\vec{c}_{\mathrm{MAP}} = \underset{\vec{c}}{\mathrm{argmin}}\, L_{\mathrm{MAP}} \tag{3.12}$$

となる結合係数 $\vec{c}$ を推定する．ここで λ の値は事前確率としての共起情報の重みづけのための値であり，適切な値を探索する必要がある．

本研究ではカテゴリの共起情報を考慮するために上記のような式 (3.12) を求めたが，この式においては平均 $\vec{\nu}$ と分散共分散行列 Σ における共起情報の導入が重要となる．本研究では，学習画像において対象物体カテゴリがどのような比率で含まれているかという共起情報を用いていく．よって，平均 $\vec{\nu}$ と分散共分散行列 Σ はカテゴリの存在比率を画像毎に学習させることで計算する．

PASCAL の学習画像にオブェクトの bounding box が与えられることを利用し，画像中の各オブジェクトにおいて SIFT 特徴点の数の比を共起情報として学習させる．具体的には，Fig. 3.6 で示すように元の画像の特徴点の合計数と画像内の各 bounding box 内における特徴点の数を用いて比率を求めることで，共起情報として平均 $\vec{\nu}_1$ と

分散共分散行列 Σ_I を導出する．なお，オブジェクトの bounding box が重なっていることによりに特徴点が重複している場合もあるため，各オブジェクトの比を足した時に 1.0 を超える場合もある．

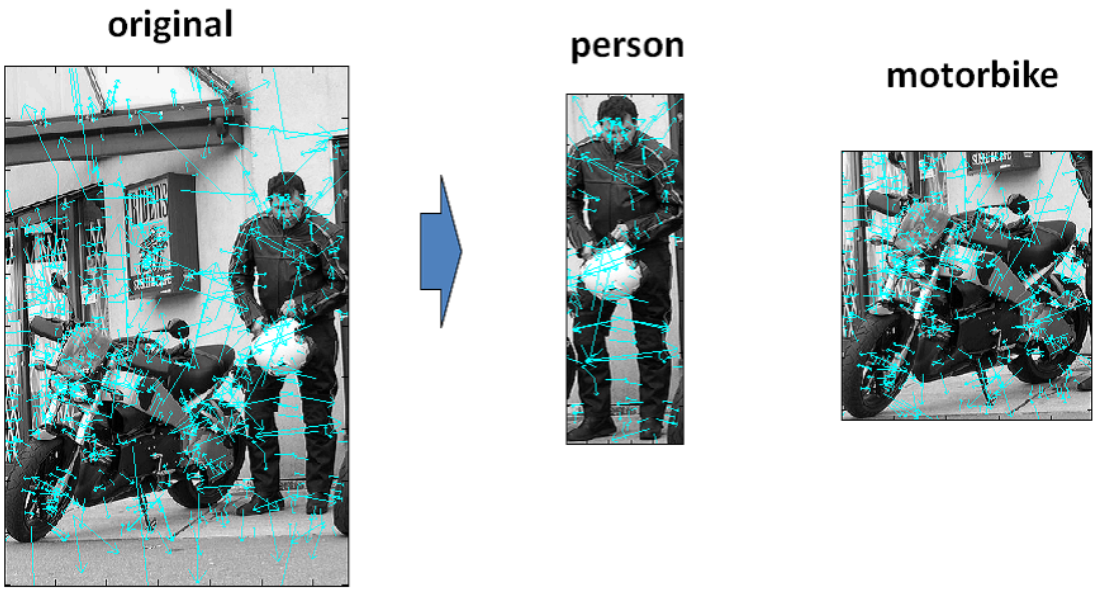


Fig. 3.6　共起情報の学習

第4章　実験と考察

4.1　学習

4.1.1　データベース

本研究では実験のデータベースとして The PASCAL Visual Object Classes Challenge 2006[3] においてコンテストの対象として提供されている学習画像とテスト画像を用いて実験を行う．PASCAL の提供画像の特徴としては学習画像とテスト画像共に Fig. 4.1 の様に画像中に存在する各対象物体に対して， bounding box が与えられていることである．

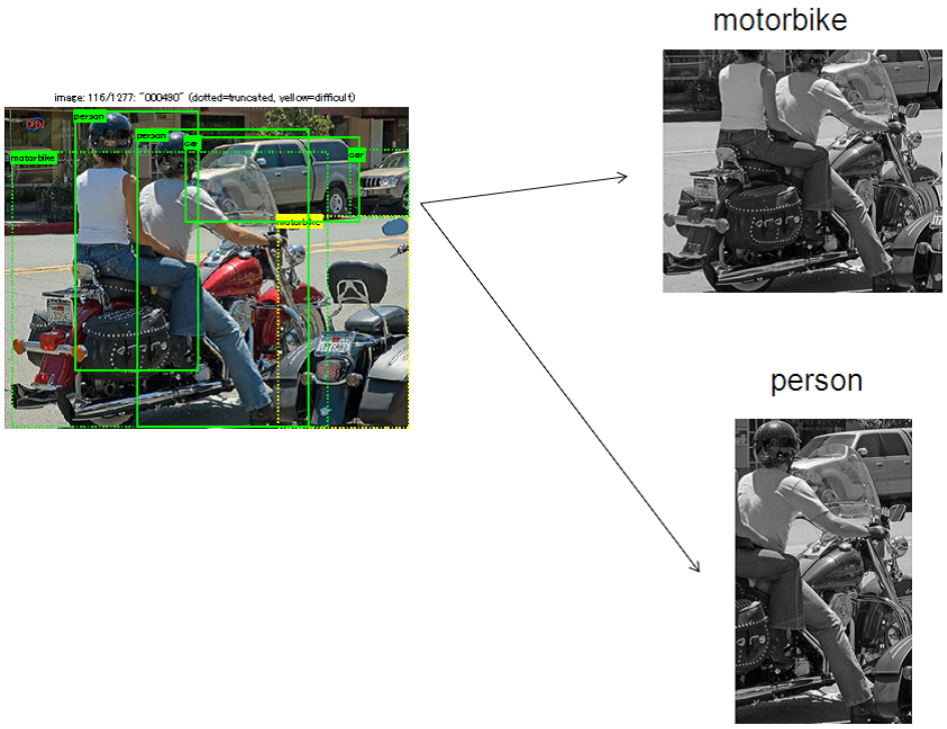


Fig. 4.1　Bounding Box

本研究では SIFT 特徴量を抽出する画像として， bounding box から各対象カテゴリの物体を切り出して，特徴点を抽出した．その際，小さすぎる画像においては特徴点の数が少なすぎるのである程度の大きさ以上の切り出した画像を用いて特徴点を抽出した．

本研究では，The PASCAL Visual Object Classes Challenge 2006 に合わせて次の 10 カテゴリを対象オブジェクトとして学習・認識を行うことにする．

- bicycle, bus, car, motorbike
- cat, cow, dog, horse, sheep
- person

対象カテゴリから切り出した各カテゴリの画像に SIFT descriptor を適用した様子を図に記す．（Fig. 4.2）

Fig. 4.2　SIFT 特徴点

本研究では上記の 10 カテゴリに対して，各々50 枚の切り出し画像の SIFT 特徴量を学習させた．

4.1.2　コードブックとヒストグラム

4.1.1 節で記したように学習画像から SIFT 特徴点を抽出したが，それぞれの特徴点の数はカテゴリの毎に多いものと少ないものが見受けられる．（Fig. 4.2）しかし，本研究では切り出し画像を使っているとはいえ，背景の要素を含む画像が数多いため，背景要素の影響も考慮し，またコードブック作成時においては学習画像の枚数をそろえているため特徴点の個数の違いはそのままで k-means 法によって，いくつかの要素数にクラスタリングを行っていった．

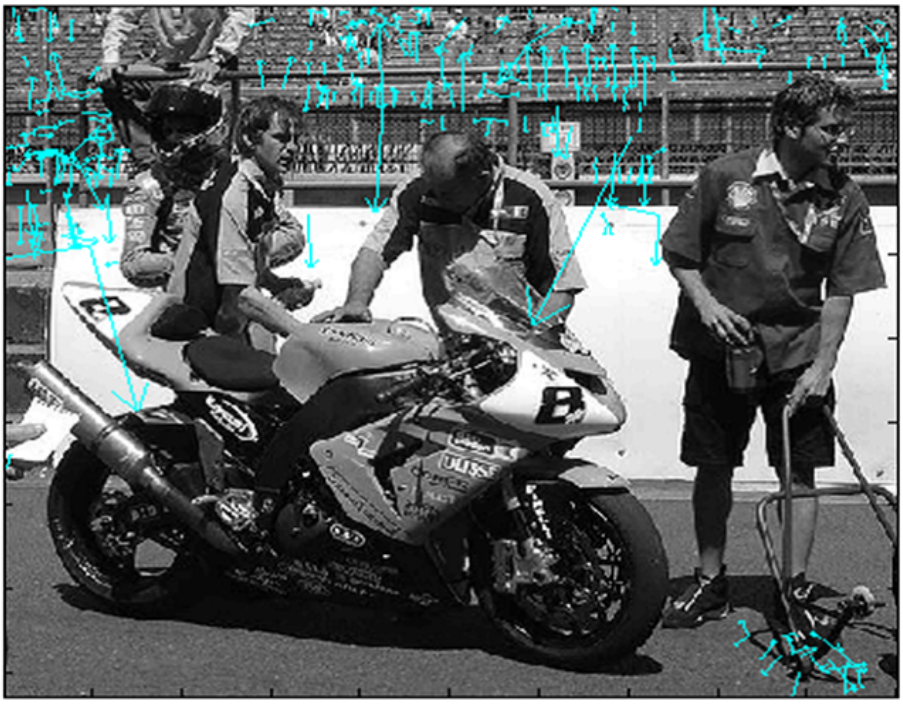
具体的には $100, 300, 500, 800, 1000, 1500, 2000$ 個の要素数をもつ７つのコードブックを作成した．これは，Bag of Features の先行研究において [7] は計算時間と精度のトレードオフから７つのオブジェクトに対して 1000 個の要素数をもつコードブックが適当であると結論づけていたため，1000 周辺の値からいくつか選んだものである．なお，k-means 法は初期値に依存するので各要素数につき 10 回ほど試行を行い，入力した特徴点の集合と誤差が最少となった値をその要素数におけるコードブックとした．

コードブック作成にあたり，上記したように 10 カテゴリから特徴点を抽出し，まとめて k-means 法によってクラスタリングを行ったものとは別に，本研究では新たに背景カテゴリを加えたコードブックを比較のために作成した．

背景の例として，本研究では２つの背景カテゴリを加えた．一つはオフィスやビルなどのコーナーやエッジによる人工的なものを多く含む背景で二つ目は草や木など自然のものを多く含む背景である．上記の 10 カテゴリに加え，背景の 2 カテゴリを加え特徴点を抽出して，まとめて k-means 法によってクラスタリングを行った．なお，MATLAB のメモリの制限から合わせて 12 カテゴリ毎に 40 枚の学習画像を用いた．要素数は $100, 300, 500, 800, 1000, 1500, 2000$ 個のコードブックを同様に作成した．学習する背景カテゴリの特徴点は Fig. 4.3 のように画像からオブジェクトの bounding box 内の特徴点を除いたものを使用した．

ヒストグラムは学習画像毎に抽出した特徴点をひとつひとつ作成したコードブックの中で最も距離が近いものに投票する形で表現する．学習画像はコードブックを作成する際に用いたものと同じものを使用する．

各カテゴリ毎に 50 個（40 個）の Bag of Features となるヒストグラム作成し，特徴点の数をそろえるために正規化を行う．その後，それぞれの要素毎に平均と分散を求め各カテゴリの代表ヒストグラムとする．



(a) 背景カテゴリ I（人工物などの背景）

(b) 背景カテゴリ II（自然における背景など）

Fig. 4.3　背景の SIFT 特徴点：背景カテゴリ 2 種類は人手で無作為に選択

4.1.3　共起情報の学習

3.2.3 節で説明したように，共起情報を利用するにはにはカテゴリの共起情報を含むように事前確率 $p(\vec{c})$ を学習させる必要がある．

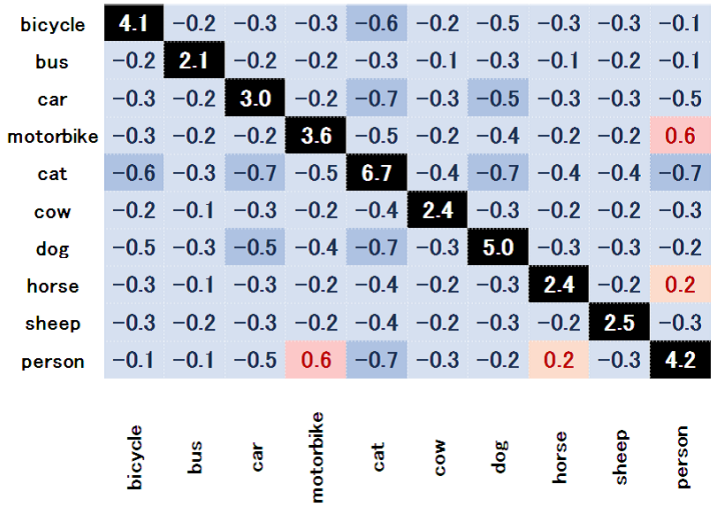
本研究では共起情報を学習するために PASCL のデータベースにおいて bounding box が与えられることを利用し，Fig. 3.6 のように特徴点の数を用いて，各カテゴリの特徴点の比率を求めることで共起情報としての，平均 $\vec{\iota}_{\mathrm{I}}$ と分散共分散行列 Σ_{I} を導出した．

また，共起情報の学習にあたり，PASCAL 2006 のデータベースを用いることで，共起情報としてカテゴリの存在比率を定義できる．しかし，学習画像におけるカテゴリの有無のみを考慮した Rabinovich らの手法との比較のため，対象とする 10 カテゴリのオブジェクトが存在すれば 1 とし，存在しなければ 0 として平均 $\vec{\iota}_{\mathrm{II}}$ と分散共分散行列 Σ_{II} を導出した．どちらの事前確率においても PASCAL 2006 から学習画像として与えられている 2618 枚全てを用いて学習させた。

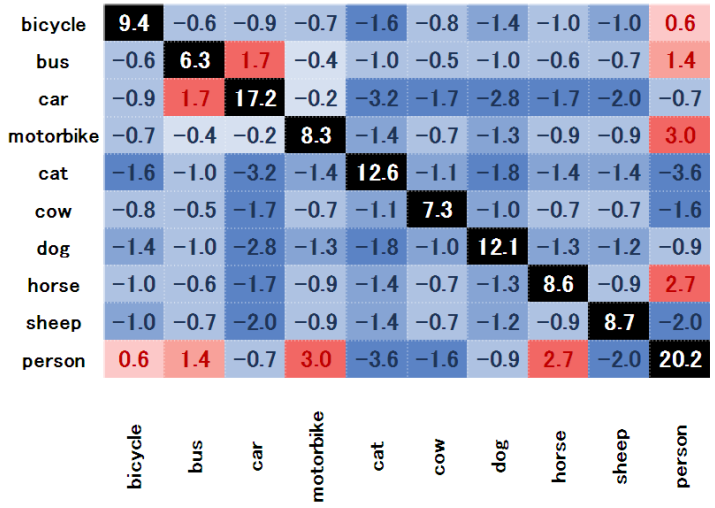
それぞれの分散共分散行列の値を Fig. 4.4 に記す．分散共分散行列は対角成分を中心として対称である．

Fig. 4.4 は PASCAL の bounding box を利用して存在比率を学習させた Σ_{I} と存在を 1，非存在を 0 として学習させた Σ_{II} である．まず，単純な存在，非存在だけを考慮したため，全体的に Σ_{II} の方が値の絶対値が大きい．また，学習画像においては単一のカテゴリオブジェクトのみが存在している場合が多いため，対角成分は大きい値となっている．

カテゴリの共起情報としては，いくつか赤の部分において正の相関つまり共起しやすいものが表されている．例えば，馬と人，モーターバイクと人などである．濃い青の部分はマイナスの値が大きいものであるが，負の相関つまり共起しにくいという情報も有用なものであると考えられる．また，いくつかのカテゴリ間において，Σ_{I} の方では負であるにも関わらず Σ_{II} の方で正の値となっている部分が見られる．これは，Σ_{II} において大きさに関わらず存在を 1 としたため，共分散の値を計算する際に 2 つのカテゴリが共起している画像の場合必ず正になってしまうためだと考えられる．一方，Σ_{I} においては共起していても片方の値が平均を超えていなければその画像における値は負になってしまう．



(a) 分散共分散行列 $\Sigma_{\mathrm{I}}(\times 10^{-2})$



(b) 分散共分散行列 $\Sigma_{\mathrm{II}}(\times 10^{-2})$

Fig. 4.4　分散共分散行列：正の相関は赤，負の相関は青で記述されている．赤は共起しやすく，青は共起しにくいことを示している．

4.2　認識結果

本研究では式 (3.12) において非線形多変数関数の最小値の探索を様々な制約条件の下で行うために，MATLAB の Optimization Toolbox 内で実装されている「fmincon」関数を用いて実験を行っていった．

4.2.1　正解データと評価方法

本研究ではテスト画像として，各カテゴリを含むテスト画像をそれぞれ 50 枚合計 500 枚を無作為に選択した．ただし，他のカテゴリ用に選択した画像内にも共起により対象カテゴリのオブジェクトが含まれる場合があるため，カテゴリ毎に出現回数は異なる．結合係数 $\vec{c}$ と画像毎に正解データを照らし合わせて，カテゴリ毎に ROC 曲線を描き，AUC で評価していく．

ROC 曲線とは，Receiver Operating Characteristic curve（受信者動作特性曲線）の略語であり，物体認識においては手法の精度を表す指標として用いられる．テスト画像において，カテゴリの正解ラベルが与えられている際に，認識結果における数値を評価することができる．横軸は false positive と呼ばれる不正解を正解としてしまった確率で，縦軸は true positive と呼ばれる正解を正しく正解とした確率である．これを認識結果における各カテゴリの存在確率において，正解とする閾値を変えながら描いたものである．グラフの曲線が左上に位置すればするほど精度が高いといえる．

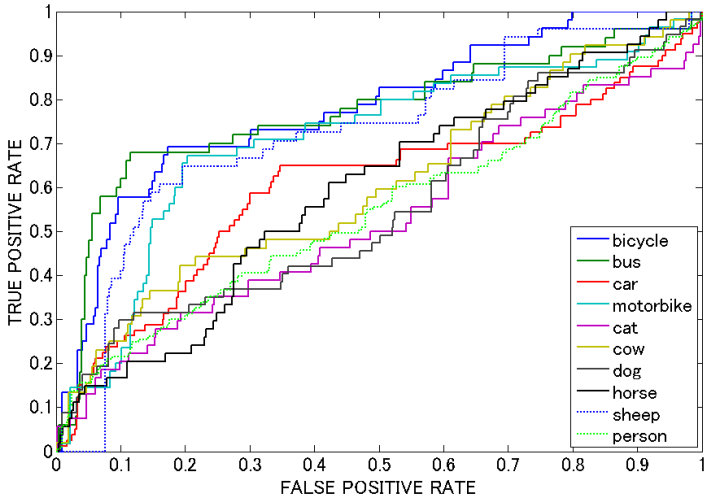
本研究では推定した結合係数 $\vec{c}$ における存在比率の値に関して，閾値を上回ったものを存在として，閾値を変えながらカテゴリ毎に正解ラベルと照らし合わせた．結合係数は制約条件により 0〜1 の値をとるため，一般的な物体認識におけるカテゴリ毎の存在確率と同様に扱うことで，ROC 曲線を描くことができる．また，曲線のみの評価が難しいため，一般的に評価数値としては AUC が用いられる．AUC は Area Under the Curve の略であり，ROC 曲線より下の面積を表している．この AUC の値が 1 に近いほど精度が高い手法であるといえる．

4.2.2　認識結果

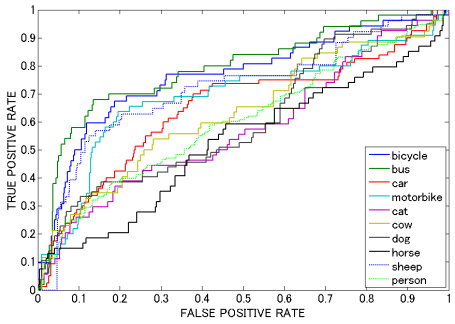
コードブックは 4.1.2 節で上記したように基本的に 100，300，500，800，1000，1500，2000 の 7 つの要素数のものを作成したため，要素数別，カテゴリ別に実験結

果を記していく．また，式 (3.12) において λ の値も 0，1，10，100，200 といくつか変化させて実験を行った．なお $\lambda = 0$ の時は共起情報を全く用いていない時である．まずはカテゴリ毎に ROC 曲線を描いたものを Fig. 4.5 示す．ここでは要素数が 100 のコードブックを用いて λ の値を 0，1，10，100，200 と変化させた時の ROC 曲線である．

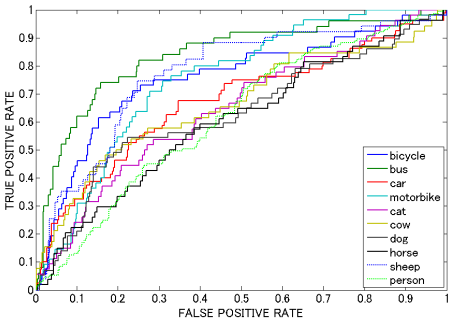
ROC 曲線を見ると，AUC が高く出やすいカテゴリと低く出てきてしまうカテゴリが存在する．例として，cat や horse などに比べ bicycle や bus などの人工物において AUC が高い値となった．しかし，これら ROC 曲線を見ても共起情報が有用であるかどうかの判断が難しい．また，PASCAL のコンテストにおいてはカテゴリ毎の ROC 曲線による AUC によって評価を行い，従来の手法では各カテゴリについて存在の有無を示していたが，本研究ではカテゴリの共起に着目して存在比率を結果として出力する．よって，ひとつひとつのカテゴリの AUC を評価するのではなく，全カテゴリの AUC の平均を手法の比較として用いていく．



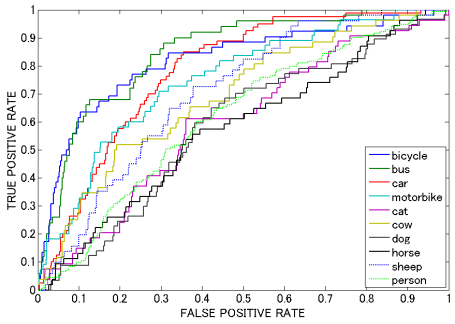
(a) $\lambda = 0$（共起情報を用いていないもの）



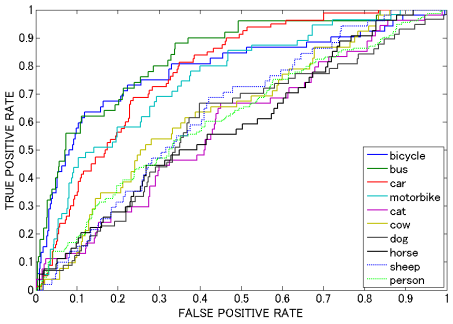
(b) $\lambda = 1$



(c) $\lambda = 10$



(d) $\lambda = 100$



(e) $\lambda = 200$

Fig. 4.5　ROC 曲線の比較

Table 4.1　共起情報の適用

λ	要素数						
	100	300	500	800	1000	1500	2000
0	0.64	0.63	0.64	0.66	0.63	0.64	0.64
1	0.66	0.64	0.63	0.64	0.64	0.65	0.63
10	0.69	0.69	0.65	0.67	0.66	0.65	0.63
100	0.70	0.68	0.68	0.67	0.67	0.65	0.65
200	0.68	0.70	0.68	0.66	0.68	0.65	0.64

Table 4.1 が λ の値を変えて全カテゴリの AUC の平均を算出したものである．共起情報の学習は PASCAL の bounding box を利用して存在比率を学習させた Σ_{I} を用いている．$\lambda = 0$ の共起情報を用いていないときよりも，要素数によっても多少変化するが結果としては良くなった．

次に背景カテゴリ 2 つも加えて，対象の 10 カテゴリを認識した結果を記す．なお，共起情報としては存在比率を学習させた．

bg_map は Table 4.1 と同様に背景カテゴリを加えた 12 カテゴリにおいて結合係数 $\vec{c}$ を求め，対象となる 10 カテゴリについて AUC を求めたものである．bg_map と map を比較すると，背景カテゴリを加えたことにより全体的に AUC が下がっている．これは，背景カテゴリ 2 つを加えたことで制約条件によって対象カテゴリの値が全体的に下がったためだと思われる．

最後に共起情報の学習において PASCAL の bounding box を利用して存在比率を学習させた Σ_{I} と，単純に存在を 1，非存在を 0 として学習させた Σ_{II} との比較を行った．

Table 4.3 において，単純に存在の有無を学習させた Σ_{II} においても Table 4.1 における共起情報を学習させないものよりも精度が良く出ている．よって，共起情報の有効性は示されたが，PASCAL の bounding box を利用して存在比率を学習させたもの Σ_{I} においてはさらに AUC が高く出ており，存在比率を学習させたことによる共起情報の有効性を示す結果となった．

Table 4.2　背景カテゴリの適用

	λ	要素数						
		100	300	500	800	1000	1500	2000
map	0	0.64	0.63	0.64	0.66	0.63	0.64	0.64
	1	0.66	0.64	0.63	0.64	0.64	0.65	0.63
	10	0.69	0.69	0.65	0.67	0.66	0.65	0.63
	100	0.70	0.68	0.68	0.67	0.67	0.65	0.65
	200	0.68	0.70	0.68	0.66	0.68	0.65	0.64
bg_map	0	0.64	0.66	0.66	0.63	0.62	0.63	0.59
	1	0.65	0.66	0.64	0.64	0.64	0.64	0.61
	10	0.69	0.69	0.66	0.66	0.66	0.65	0.61
	100	0.69	0.67	0.66	0.67	0.64	0.62	0.63
	200	0.68	0.66	0.65	0.65	0.61	0.64	0.61

Table 4.3　共起情報の比較

	λ	要素数						
		100	300	500	800	1000	1500	2000
map Σ_{I}	1	0.66	0.64	0.63	0.64	0.64	0.65	0.63
	10	0.69	0.69	0.65	0.67	0.66	0.65	0.63
	100	0.70	0.68	0.68	0.67	0.67	0.65	0.65
	200	0.68	0.70	0.68	0.66	0.68	0.65	0.64
map Σ_{II}	1	0.63	0.63	0.63	0.65	0.63	0.64	0.63
	10	0.67	0.64	0.64	0.64	0.63	0.65	0.62
	100	0.68	0.68	0.68	0.66	0.65	0.65	0.63
	200	0.68	0.69	0.68	0.67	0.65	0.66	0.64

4.3　実験考察

本節では共起情報を用いて MAP 推定を行った結果に対して，考察を加えていく．

4.3.1　適切なパラメータの選択

本研究ではコードブックの要素数と式 (3.12) における λ の値をいくつか変えながら実験を行っていった．式 (3.12) においては 2 つの項が存在する．片方は要素数を足し合わせる項で，もう片方が共起情報の重みづけとして λ をかけたものである．本研究ではこの 2 つの項の和が最小となる結合係数 $\vec{c}$ を推定することを目的としていて，ROC カーブにおける AUC の値は要素数が 100 で λ が 100 もしくは要素数が 300 で λ が 200 となるところで最大の値をとった．これは要素数が少ないコードブックで十分な認識できる可能性を示唆している．実際，各カテゴリにおいてヒストグラムを作成した際に他のヒストグラムとの違いがはっきりと表れるような visual words の数はあまり多くはなかった．λ も小さすぎると共起情報が考慮されず，また大きすぎると $\vec{v}$ の値に近づいていくだけになってしまう．要素数と λ の値は式 (3.12) においてどの要素数においてはどの λ の値が適切であるかについては他の要素の改良によっても影響されると考えられるため，さらに詳細な実験を行う必要がある．

4.3.2　評価方法

本研究ではデータベースとして PASCAL 2006 の学習画像とテスト画像を用いて実験を行った．実際の PASCAL のコンテストにおいては各カテゴリ毎に ROC 曲線を描き，その AUC において手法としての評価を競っていく．そのため，本研究においても同様に各カテゴリ毎の AUC において比較をした．しかし，既存の手法が対象カテゴリ毎に存在の有無のみを識別するのに対し，本研究では対象カテゴリの存在比率をまとめて算出する手法である．また，そのため結合係数 $\vec{c}$ には制約条件として存在比率の和が 1 になるという条件の下で最適化を行った．よって，カテゴリ毎の値が相対的に小さく出てきてしまう．ゆえに，存在の有無で評価を行うのではなく，正解データもカテゴリの存在比率で表したものを用いて評価を行う必要がある．正解データを Fig. 3.6 と同様の方法でカテゴリの存在比率によって表し，認識結果と比較したものを Fig. 4.6 で記す．

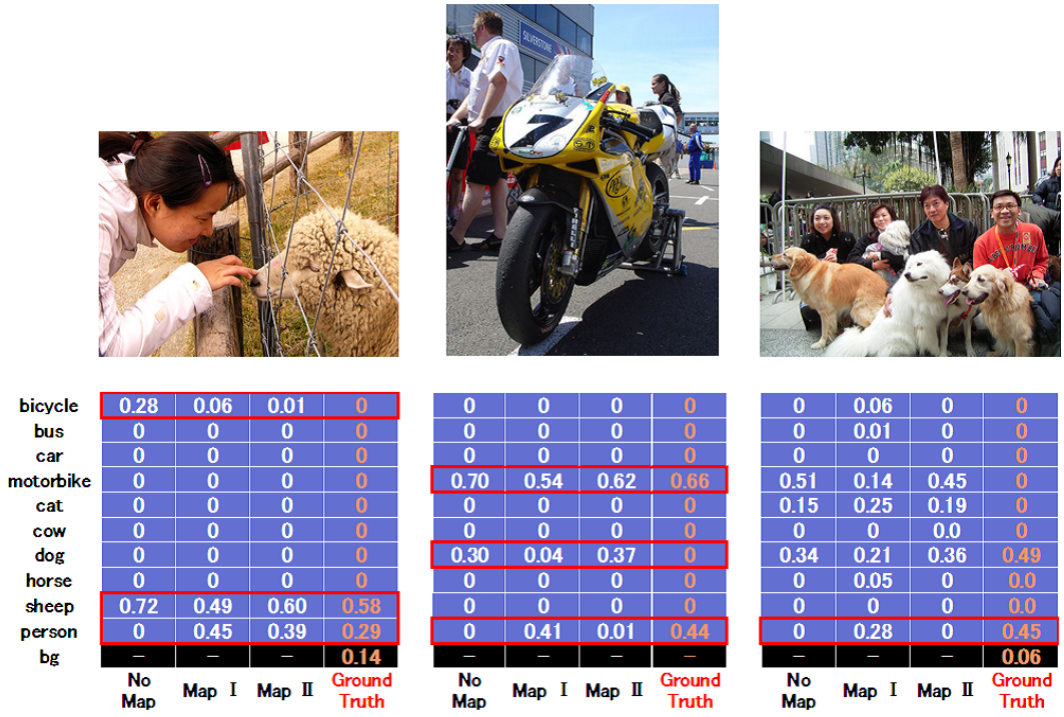


Fig. 4.6　共起情報組み入れた認識：No_Map は共起情報を用いていないもの，Map I は存在比率を Map II は存在有無を共起情報として学習させたものである．値は各カテゴリの存在比率であり，Ground Truth は Fig. 3.6 と同様の方法で算出した．

１枚目の画像においては，共起情報を用いていない場合には bicycle と sheep が認識されて person が認識されていないが，共起情報を用いたことにより人が認識されており，bicycle の存在比率も下がっている．２枚目においても，同様に dog の誤検出が減り，person が新しく認識されている．３枚目では，背景の影響から人工物における存在比率が多少出ているが，person が新しく認識されており，共起情報を組み入れたことで上手く認識ができるようになっている．

4.3.3　背景の扱い

上記で説明したように正解データをカテゴリの存在比率として表現した場合，背景の扱いが重要となってくる．一般の画像においては物体が存在していない領域（＝背景）が必ずあり，本研究の目的であるカテゴリの存在比率を求めるのであれば，背景の比率も求める必要がある．背景カテゴリを導入しない場合，背景の特徴点の集合を対象カテゴリのヒストグラムを組み合わせて表現しており，誤検出につながると考えられる．よって，いくつかの背景カテゴリを加えた上で，結合係数の和が 1.0 になるような制約条件を課す必要があると考えられる．背景カテゴリを加えたことで誤検出がなくなるようないくつかの例を Fig. 4.7 で記す．

テスト画像において，対象カテゴリのオブジェクトが小さい場合，画像のほとんどは背景であるが，背景カテゴリを加えた MAP 推定においては背景の要素が認識されていて，Ground Truth に近い認識結果となった．２枚目と３枚目においては，背景カテゴリを追加したことで認識できていなかったカテゴリも認識されている．しかし，背景カテゴリをどうやって定義するかという問題が残っている．本手法においては背景カテゴリの影響で対象カテゴリの存在比率が相対的に小さくなっている場合が多々見られた．また，馬や羊の周りには牧草が生えていることが多いなど，背景要素も対象カテゴリの特徴要素に含めることもできるため，背景の扱いについては慎重なアプローチをとる必要がある．

4.3.4　アプローチの改良

本研究では BOF から共起情報の導入まで，基本的なアプローチをとっている．しかし，各構成要素においてさらに効率が良く，精度が高いアプローチが考えられている．BOF のアプローチにおいては特徴点抽出，クラスタリング，ヒストグラムの表現の仕方などである．

	Map	Bg Map	Ground Truth
bicycle	0.11	0	0
bus	0.01	0	0
car	0	0	0
motorbike	0	0	0
cat	0.79	0	0
cow	0	0	0
dog	0	0	0
horse	0.02	0	0
sheep	0.03	0.03	0.08
person	0.05	0	0
bg	–	0.97	0.92

	Map	Bg Map	Ground Truth
	0.09	0.14	0
	0	0.10	0
	0	0.07	0
	0	0.03	0
	0	0	0
	0	0	0
	0	0	0
	0	0.08	0.09
	0.17	0.02	0
	0.44	0.17	0
	–	0.40	0.91

	Map	Bg Map	Ground Truth
	0.12	0.01	0
	0.06	0.05	0
	0	0.06	0.12
	0.16	0	0
	0	0	0
	0	0	0
	0.21	0.13	0
	0	0.04	0.15
	0	0	0
	0.46	0.02	0.05
	–	0.71	0.68

Fig. 4.7　背景カテゴリを導入した認識：背景カテゴリを導入した認識：Map は対象 10 カテゴリ，Bg_Map は 2 つの背景カテゴリを加えた 12 カテゴリにおいて認識を行ったものである．なお，図中のカテゴリ「bg」は 2 つの背景カテゴリの比率を足したものである．

本研究では比較実験としてコードブックを改良したものについての認識も行った．対象となる 10 カテゴリまたは背景の 2 カテゴリを加えた 12 カテゴリの学習画像から抽出した特徴点をまとめて k-means 法によってクラスタリングを行うのではなく，カテゴリ毎に k-means 法によって 100 個のクラスタに分類し，それを結合させることでコードブックを作成した．要素数は 1000 または 1200 のコードブックとなり，MAP 推定において共起情報の重みづけの値である λ の値を変化させていった．実験結果は以下の Table 4.4 のようになった．

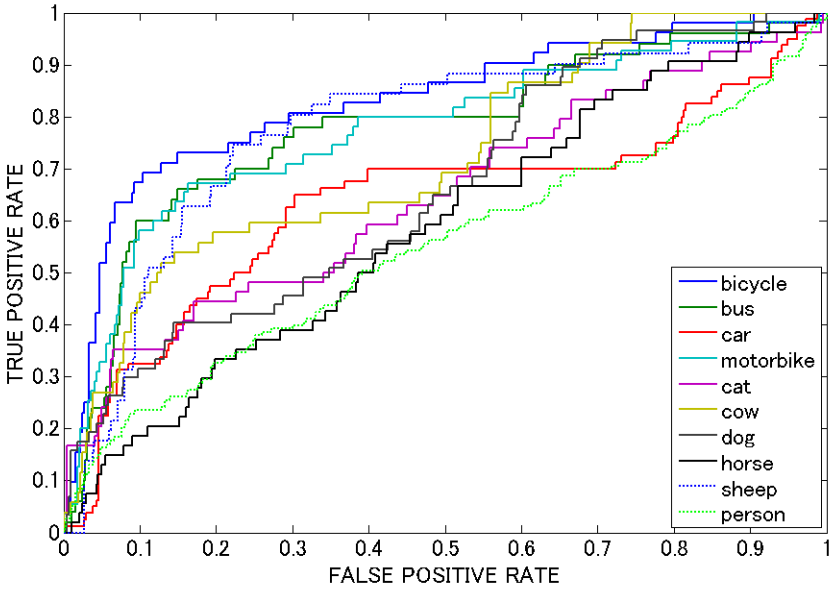
Table 4.4　コードブックの改良

λ	1	10	50	100	200
no-bg (1000)	0.65	0.70	0.68	0.67	0.69
bg (1200)	0.65	0.67	0.69	0.67	0.65

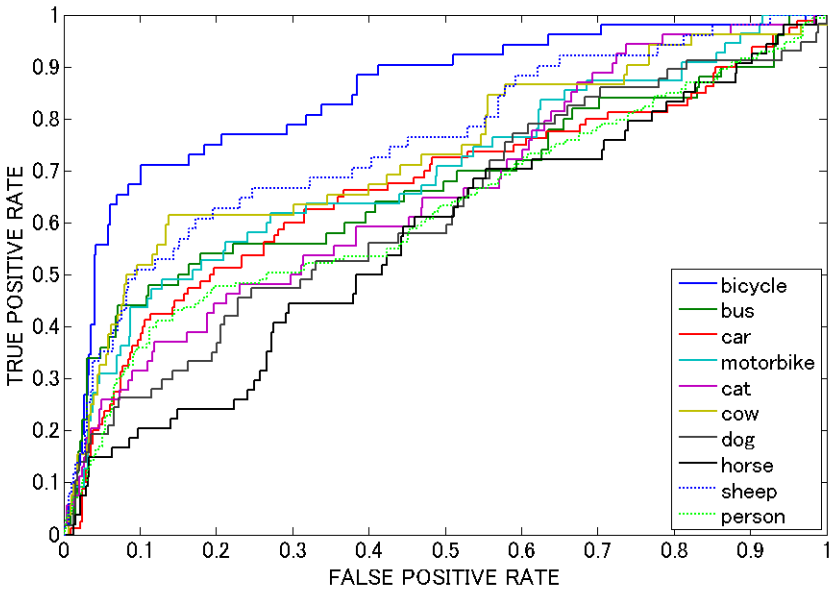
またそれぞれのコードブックで最も AUC の平均が良かった ROC 曲線を Fig. 4.8 に描く．

Fig. 4.5 に比べると Fig. 4.8(a) は cat や dog などの AUC が若干高くなっている．これは，Fig. 4.2 の特徴点抽出において，bicycle や bus と比べ cat や dog の特徴点の数は相対的に少なくなる．よって，特徴点をクラスタリングする際に bicycle や bus の visual words が多く出てきてしまったため，相対的に cat や dog の特徴を表す visual words が少なくなってしまうためだと考えられる．Fig. 4.8(a) では各対象カテゴリ毎に visual words の数が等しくなるために cat や dog の特徴を表す visual words もコードブックに反映されたためだと考えられる．また，Fig. 4.8(b) では自転車以外の ROC 曲線は同じような軌道に収束している．これは背景カテゴリを導入したことで，背景カテゴリに存在比率が割り振られるため，誤検出が減ったためだと考えられる．背景カテゴリを導入しない場合，背景の特徴の集合もカテゴリのの存在比率によって表されるため，カテゴリ毎に背景の要素を表しやすいカテゴリと表しにくいカテゴリができてしまう．これらの BOF の構成要素の研究ではクラスタリングだけにとどまらず，現在多数のアプローチがあるため，本研究に最も適したアプローチを適切に選択する必要がある．

共起情報の導入において本研究では，画像の bounding box から特徴点の比を学習させた．しかし，複数カテゴリ，複数物体の bounding box が重複している場合については特徴点をどう分配するかなどについてさらに試行を重ねる必要がある．また，



(a) no-bg ：$\lambda = 10$



(b) bg ：$\lambda = 50$

Fig. 4.8　コードブックを改良した認識

共起情報は用いるデータベースに依存されるため，データベースとして何を用いるかという問題や比率を学習させるのか，存在の有無を学習させるのかという問題についてもさらに実験を重ねる必要がある．

また，実用的な時間内に識別を可能にするためには共起情報の導入において式 (3.12) の最適化におけるさらに効率が良いアルゴリズムを考えなければならない．本研究においては，上記したように MATLAB の Optimization Toolbox 内で実装されている「fmincon 」関数を用いて最適化を行ったが，誤差や反復回数と時間についても実験を重ねて適切なパラメータを与える必要がある．

第5章　結論

5.1　まとめ

本研究では，一般物体認識をテーマとして画像中に存在している物体のカテゴリを認識することを目的としてきた．様々なアプローチが考えられている中で，その単純さと拡張性により様々な研究へと応用されている Bag of Features（BOF）のアプローチを手法の基盤として用いた．BOF は位置情報を用いず画像を局所特徴量の集合で表すモデルであるが，その枠組みの中で結果として出力されたテスト画像のヒストグラムが各対象カテゴリの代表ヒストグラムの線形結合となっていると仮定した．

本研究では，線形結合における結合係数が画像中におけるカテゴリの存在比率となっているとみなしてその結合係数を推定するアプローチを提案した．まず，各カテゴリのヒストグラムにおける要素が学習した複数枚の画像において正規分布し，お互い独立したものと仮定した．その後，正規分布の再生性により，テスト画像のヒストグラムが正規分布するため，線形結合によって求められた平均と分散の値から最尤推定によって結合係数を推定した．

また，本研究ではカテゴリの共起に着目し，物体認識にカテゴリの共起関係を情報として組み入れる手法の提案も行った．具体的には，上記の最尤推定において事前確率として共起情報を学習させることで事後確率を最大にする MAP 推定を用いて，結合係数を推定した．共起情報の学習においては，共起確率が多変量正規分布に従うと仮定し，各カテゴリの存在比率をデータベースから学習させて用いた．

実験においては，コードブックの要素数や共起情報の重みづけなどいくつかパラメータを変化させながら行った．また，対照実験として２つの背景カテゴリを加えたコードブックや共起情報として単純に存在の有無を学習させたものを用いて実験を進めた．

その結果，共起情報を取り入れていないときには認識できていなかったカテゴリが認識されたりと共起情報を取り入れたことで評価の指標である AUC の精度が良くなった．また，単純にカテゴリの存在有無を学習させただけでも共起情報を用いて

いない手法よりも精度が良くなったが，カテゴリの存在比率を学習させたものではさらに精度が良くなった．これにより，カテゴリ間の共起情報を学習させることの有効性が示され，学習方法も存在比率を学習させる方が有効であることが示された．しかし，単純な存在有無を学習させる方法でも精度が良くなるため，様々なデータベースにも応用できることと考えられる．

また，対照実験として行った背景カテゴリを加えたことによる実験では，背景カテゴリを導入した手法において AUC の値が低くなった．しかし，背景部分が多い画像を認識する際に，背景カテゴリにも結合係数を割り振ることで，物体カテゴリの誤検出が減るなど，背景カテゴリが有効となる例もいくつかみられた．

5.2　今後の課題

5.2.1　存在比率の評価方法

本研究では物体の画像中におけるオブジェクトカテゴリの存在比率を求めることを目的とした．4.3 節でも述べたが現在の物体認識の評価方法では各カテゴリに ROC 曲線を書き，AUC の値で評価を行うことが一般的である．しかし，本研究では一つのカテゴリの存在有無を個別に導出するのではなく，対象全カテゴリの存在比率を算出することができる．よって，既存の手法と同様の評価方法ではなく，本手法に沿った形での評価を行う必要がある．そのためにはまず正解とする画像をカテゴリの存在有無ではなく，全カテゴリの存在比率で表す必要がある．その後，正解の存在比率と認識結果の存在比率の類似度を評価できるような方法が必要なる．しかし，正解には背景が含まれたりするため，評価方法に関しては様々な面からの考慮が必要になる．今後は，正解データの作成方法や背景の要素も含めて適切な評価方法を吟味していく必要がある．

5.2.2　背景の扱い

画像中におけるカテゴリの存在比率を求めるにあたり，背景の要素について考えなければならない．本研究においては，結合係数の制約条件として，和が 1.0 となるようにしているが，これでは背景を表すことができなかった．よって，対照実験として背景カテゴリを含む実験も行ったが，AUC の結果としては精度が落ちる結果となった．しかし，Fig. 4.7 で示された様に背景カテゴリを加えたことで画像自体の

存在比率としては正解に近づくという例もみられた．また一方で，背景の要素が前景のカテゴリの認識において有効になるという例もある．[27] よって，背景の要素をどう扱うかはとても難しい問題である．そのために結合係数の制約条件をどう扱うかや背景カテゴリをどのようにモデル化するかなど慎重に実験を重ねていく必要がある．

5.2.3　カテゴリ間におけるさらなる情報の利用

本研究では共起情報として，存在比率を学習させた．また対照実験としてカテゴリの存在有無を学習させたものについても実験を行った．どちらにおいても共起情報を学習させない手法よりも精度が良くなった．単純なカテゴリの存在有無を学習させた理由としては，存在比率を学習させるにはデータベースにおいて物体の領域が与えられる必要があるためである．共起情報の学習においてはデータベースの選択によって何らかのバイアスがかかる必要があるため，より一般的なデータベースを用いる必要がある．しかし，一般的なデータベースから存在比率を学習させることは困難である．よって，本研究では単純な存在有無を学習させることでの有効性が示されたため，Web 上のデータベースなど様々なデータベースに応用が可能であると考えられる．

また一方で，単純な存在有無より存在比率を学習させた方が精度が良く出たため，存在比率以上の情報を学習することができればさらに精度が上がる可能性があると考えられる．例えば，物体同士の相対的な位置関係などである．バイク・自転車や馬の上には人が存在しやすいなどの情報をカテゴリ間の関連性の情報として用いればさらに精度が高い認識が可能となり，物体の位置を探索する認識方法においても利用が可能となると考えられる．しかし，位置情報を用いるためには学習方法や認識手法の改良が必要となるため，さらなる検討が必要である．

5.2.4　アプローチの各要素の改良

本研究では Bag of Features の枠組みの中で共起情報を取り入れた手法を提案した．しかし，BOF における構成要素においては様々に改良されたアプローチが提案されている．今後は特徴点探索・記述，クラスタリング，ヒストグラムの作成等の各構成要素に対しても様々な試行を重ねる必要がある．また，本研究では式 (3.12) において結合係数を推定するために，非線形多変数関数の最小値の探索を行う，MATLAB

の Optimization Toolbox 内で実装されている「fmincon 」関数を用いて実験を行っていった．しかし，実用的な時間で精度の高い認識を行うにはさらなるアルゴリズムの利用や適切なパラメータの設定を行う必要があると考えられる．

謝辞

本研究を遂行するにあたり，多大なる御指導そして御協力を頂きました，佐藤洋一准教授に心より御礼申し上げます.

本研究を進めるにあたって様々な有益なる御助言と御教示をして下さいました，佐藤研究室助教の岡部孝弘氏に深く御礼申し上げます.

研究のことのみならずに様々な御指摘，御助言をして下さいました木谷クリス真実氏に心より感謝いたします.

また，同じ研究室の仲間として研究だけでなく，様々な面でお世話になりました佐藤研究室の皆様に深く感謝いたします.

最後に，学生生活を支えて下さった全ての方に深く感謝いたします.

平成 20 年 2 月 4 日

近藤 雄飛

参考文献

1) Caltech101. http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html.

2) Google Sets. http://labs.google.com/sets.

3) The pascal visual object classes challenge 2006. http://www.pascal-network.org/challenges/VOC/voc2006/index.html.

4) Sift keypoint detector. http://www.cs.ubc.ca/ lowe/keypoints/.

5) Two bag-of-words classifiers. http://people.csail.mit.edu/fergus/iccv2005/bagwords.html.

6) M. Burl and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. *Proc. of European Conference on Computer Vision*, pp. 628–641, 1998.

7) G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *Proc. of ECCV Workshop on Statistical Learning in Computer Vision,*, pp. 1–22, 2004.

8) R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley-Interscience, 2000.

9) P. Duygulu, K. Barnard, Nd. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *Proc. of Europe-an Conference on Computer Vision*, pp. 97–112, 2002.

10) R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale invariant learning. *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 264–271, 2003.

11) X He, R. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labeling. *Proc. of IEEE Computer Vision and Pattern Recognition*, Vol. 2, pp. 695–702, 2004.

12) T Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, Vol. 42, pp. 177–196, 2001.

13) D. Hoiem, AA. Efros, and M. Hebert. Putting objects in perspective. *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 2137–2144, 2006.

14) F. Jurie and B. Triggs. Creating efficient codebook for visual recognition. *Proc. of EEE International Conference on Computer Vision*, Vol. 1, pp. 604–610, 2005.

15) S. Kumar and M. Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. *Proc. of IEEE International Conference on Computer Vision*, Vol. 2, pp. 1150–1157, 2003.

16) J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. of International Conference on Machine Learning*, pp. 282–289, 2001.

17) D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, pp. 91–110, 2004.

18) O. Maron and A. Ratan. Multiple-instance learning for natural scene classification. *International Conference on Machine Learning*, pp. 341–349, 1998.

19) K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. *Proc. of IEEE Computer Vision and Pattern Recognition*, Vol. 1, pp. 26–36, 2006.

20) F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. *Advances in Neural Information Processing Systems*, pp. 985–992, 2006.

21) M-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. *Proc. of IEEE Computer Vision and Pattern Recognition*, Vol. 2, pp. 1447–1454, 2006.

22) F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. *Proc. of European Conference on Computer Vision*, pp. 464–475, 2006.

23) Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, Tao Mei, Jinhui Tang, and Hong-Jiang Zhang. Concurrent multiple instance learning for image categorization. *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

24) A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. *Proc. of IEEE International Conference on Computer Vision*, pp. 1–8, 2007.

25) J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman. Discovering objects and their location in images. *Proc. of IEEE International Conference on Computer Vision*, pp. 370–377, 2005.

26) J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. *Proc. of IEEE International Conference on Computer Vision*, Vol. 2, pp. 1800–1807, 2005.

27) J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories A comprehensive study. *International Journal of Computer Vision*, Vol. 73, No2, pp. 213–238, 2007.

28) 永橋知行, 藤吉弘亘. 領域分割に基づく sift 特徴を用いた物体識別. 電気学会 システム・制御研究会, pp. 39–44, 2007.

29) 上東太一, 柳井啓司. Bag-of-keypoints 表現を用いた web 画像分類. 情報処理学会研究報告. CVIM, pp. 201–208, 2007.

30) 湯志遠, 柳井啓司. Bags-of-keypoints による trecvid データに対する映像認識. 情報処理学会研究報告. CVIM, pp. 209–216, 2007.

31) 柳井啓司. 一般物体認識の現状と今後. 情報処理学会研究報告. CVIM, pp. 121–134, 2006.

発表文献

近藤 雄飛, 岡部 孝弘, 木谷 クリス 真実, 佐藤 洋一 ”カテゴリ共起を考慮した物体認識手法”, 情報処理学会コンピュータビジョンとイメージメディア研究発表会, March 2008　（発表予定）