論文 / 著書情報
Article / Book Information

題目(和文)	オブジェクト検出・認識のための画像特徴量に関する研究
Title(English)	
著者(和文)	三田雄志
Author(English)	Takeshi Mita
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第8818号, 授与年月日:2012年3月26日, 学位の種別:課程博士, 審査員:佐藤 誠
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第8818号, Conferred date:2012/3/26, Degree Type:Course doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

オブジェクト検出・認識のための
画像特徴量に関する研究

三田雄志

概要

本論文では，画像内のオブジェクトを検出あるいは認識するための特徴量に関する研究成果について述べる．オブジェクト検出とは，特定のカテゴリに属するオブジェクトの画像内での位置と大きさを求める処理であり，例えば，写真に写っている人物の顔を見つけ出す処理がそれにあたる．また，オブジェクト認識とは，画像内のオブジェクトがどのカテゴリに属するかを識別する処理で，例えば，写真に写っている鳥の種類を見分ける処理を指す．

オブジェクト検出・認識では，顔の個人差や鳥の姿勢の変化のようなカテゴリ内の変動を許容できること，照明条件の変動や撮像時のノイズのような外乱に左右されにくいこと，他のよく似たカテゴリと混同しないこと，短時間に実行できることが必要である．これらの技術課題に対処するアプローチとして，以下の 2 つを試みる．

第 1 に，カテゴリ内変動に対処するため，特徴量の確率分布を利用する．増分符号という照明変動に対して頑健な特徴量を複数の参照画像において観測し，その確率分布をテンプレートとする新しい照合法を提案する．従来，カテゴリ内変動を伴う対象を検出するには，平均画像をテンプレートとする照合法や部分空間法などの統計手法が一般的に用いられるが，1,000 枚を超える画像から顔を検出する実験によって，提案手法が識別精度と計算コストの両面で大幅に優れていることを示す．

第 2 に，類似カテゴリとの識別性能を高めるため，複数の特徴間の共起性を導入する．Sequential Forward Selection と Boosting を組み合わせた学習アルゴリズムを提案し，対象カテゴリと非対象カテゴリを識別するのに適した共起特徴を自動選択し，オブジェクト検出のための識別器を学習する．オブジェクト検出のフレームワークとしては，Viola と Jones の方法が最もよく知られている．顔や 3 種類の形状の手を検出する実験では，共起性を利用した提案手法は常に Viola と Jones の方法を上回る精度を同等の計算コストで実現できることを確認する．さらに，画像内の同じ位置から検出した複数の局所特徴量から識別に有効な共起関係を抽出する方法を提案し，サブカテゴリ識別への適用を試みる．Random Forest という木構造に基づく量子化アルゴリズムによって共起関係を記述するとともに，Multiple Kernel Learning によって共起特徴の重みをカテゴリごとに調節する．200 種類の鳥類を識別する実験で，一般物体認識において最も識別精度が高いとされる Spatial Pyramid と Multiple Kernel Learning を組み合わせた手法を上回る識別率を，より小さい計算量で実現できることを示す．

カテゴリ内変動に対する頑健性向上と類似カテゴリとの識別性能の改善は，オブジェクト検出およびオブジェクト認識の高精度化において重要である．提案手法は，このような高精度化と省処理化を両立させる点に特長がある．このような好ましい性質は，画像検索など様々なビジョンタスクに対して貢献できる．

目次

図目次

表目次

第 1 章

序論

1.1 目的

画像内のオブジェクトが属するカテゴリが何か (what) を認識したり，オブジェクトがどこにあるか (where) を検出する処理を自動化できれば，これまで人間しか行えなかった高度なタスクを機械によって代行する，あるいは人間の知的活動を支援することが可能となる．

本論文では，オブジェクト検出あるいはオブジェクト認識というコンピュータビジョンやパターン認識の分野における基本課題に対し，これを高精度かつ高速に実行する方法を確立することを目的とする．例えば，セキュリティシステム（図 1.1）やハンドジェスチャ認識システム（図 1.2）において画像内の顔や手を見落としなく高速に検出できるようにする．また，サブカテゴリ識別 (subordinate category classification) という最近取り組まれるようになった新しい課題の解決に取り組むことによって，専門家しか持っていない知識に一般の人々が「容易に」アクセスできるようにするという従来になかった価値を提供する．例えば，図 1.3 のようにカメラつきの携帯機器で撮影した鳥類の画像をアップロードすると，それに関連する Wikipedia の Web ページを提示するシステムの実現につながる．

1.1.1 技術課題

高精度かつ高速なオブジェクト検出・認識手法の確立や，サブカテゴリ識別の実現という目的を達成するためには，以下の技術的課題を解決する必要がある．

1. 照明条件の変動や撮影時に付加されるノイズなどの外乱に対する頑健性確保
2. 個体差や姿勢変化などのカテゴリ内の変動に対する許容性確保
3. 互いに類似したカテゴリを見分けるカテゴリ間の識別能力確保
4. 短時間に識別処理を実行できる高速性確保

図 1.4 に，照明条件の異なる環境で撮影された顔画像の例を示す．屋外や車内で発生するこのような外乱に左右されない頑健性が求められる．図 1.5 には，カテゴリ内変動の例として，顔の個人差や鳥の姿勢変化を示す．このようなカテゴリ内変動の影響を受けにくいアルゴリズ

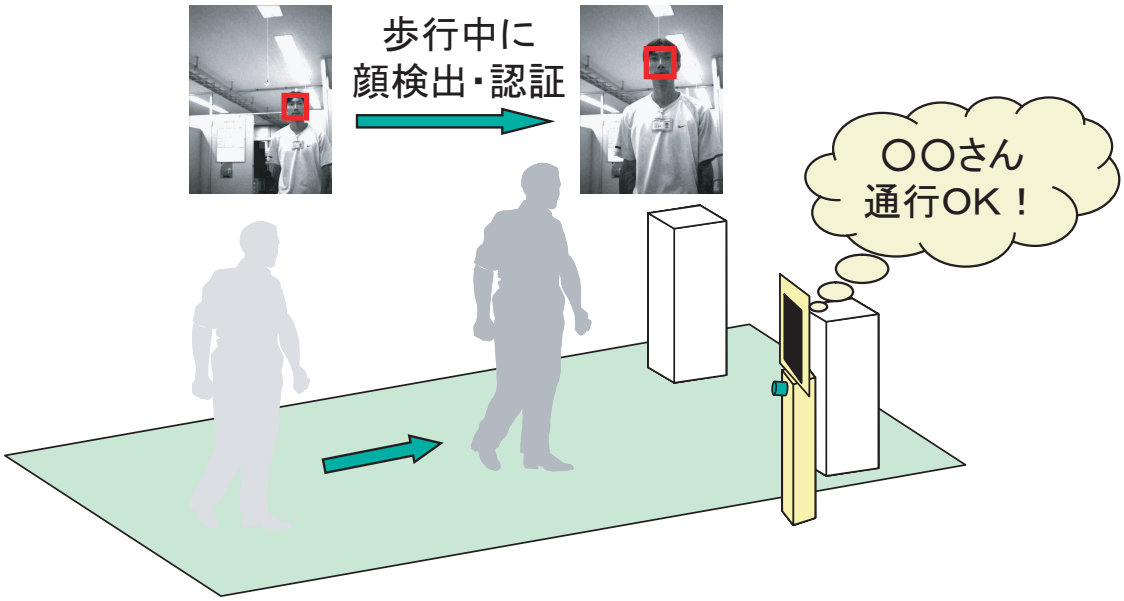

図 1.1. 歩行中に通行者の顔を検出，認識し，通行可能か否かを判定するセキュリティシステム．通行者は認証のために立ち止まったり，機器を操作する必要がない．

ムが必要である．図 1.6 に，異なるカテゴリに属するがよく似たパターンを有する画像あるいはカテゴリの例を示す．顔検出では，背景の一部に顔によく似た領域を誤って検出しないように識別性能を高める必要がある．サブカテゴリ識別では，一部の特徴のみが異なるようなパターンを区別できるようにすべきである．

1.1.2　アプローチ

外乱に対する頑健性確保のためには，画像の画素値そのものではなく隣接する画素値の差分のように空間的な勾配に基づく特徴を使用する．例えば，増分符号 [102] や rectangle feature [83, 56]，SIFT [43] を用いる．これらは，画素の絶対値でなく相対的な関係を利用しており，外乱に左右されにくい性質を有する．

カテゴリ内変動に対する許容性確保には，上記特徴量を評価するための統計的な枠組みを検討する．カテゴリ内変動を含む複数のサンプル画像から特徴量を観測し，その確率分布を考慮して他のカテゴリと識別する．さらに，AdaBoost [24] や Support Vector Machine [80] などの教師あり学習アルゴリズムを使用して，識別器を構築する．

カテゴリ間の識別能力確保には，類似したカテゴリ同士を見分けるのに適したわずかな差異を抽出する必要がある．これには，新たな特徴記述子を開発するなど特徴そのものを改良するアプローチが考えられるが，本論文では既存の特徴間の「共起性」を利用する方法を提案する．共起性とは，ある特徴を観測すると，別の特徴も同時に観測されるという性質である．例えば，顔を検出しようとする場合，顔はほぼ左右対称であるため，画像の左側で観測された特徴量が同時に右側でも観測されやすくなる．特徴記述子の改良は有効なアプローチであるが，

図 1.2. ハンドジェスチャ認識システムの例．ユーザの手形状や動作に応じて，リモコンを使わずに AV 機器などの操作ができる．

対象カテゴリにのみ最適化され，他のカテゴリにも適用できなくなる恐れがある．一方，共起性は多くのカテゴリに共通して観測されるため，これを利用した学習アルゴリズムは識別能力の面だけでなく汎用性の面でも優れたものとなる．

アルゴリズムの高速性確保には，抽出処理が簡便な特徴を使用すること，識別器の学習段階で識別に有効な特徴のみを選択すること，共起性の活用によって同じ計算量でより高い識別能力を獲得することを検討する．

これらの基本方針をふまえた以下 3 つのアプローチをとる．

- オブジェクト検出のための最も基本的な方法であるテンプレート照合において，技術課題 1〜4 を満たす新たな方法を検討する．テンプレート照合に関する先行研究では，特に 2 と 3 は十分に検討されてこなかったが，これらを評価できる新しい統計量を考案する．
- オブジェクト検出は，対象カテゴリとそれ以外（非対象カテゴリ）を識別する 2 カテゴリ識別問題ととらえることができる．特徴の共起性を導入することで上記 1〜4 を満たす 2 カテゴリ識別器を構築する．多数の学習サンプルから，識別に有効な共起特徴を自動選択するアルゴリズムを考案する．さらに，別の学習サンプルを与えれば，対象カテゴリを検出するための識別器が得られる汎用性に優れた手法を確立する．
- サブカテゴリの識別において特に重要となる技術課題 3 にフォーカスし，特徴の共起性を導入したサブカテゴリ識別を検討する．オブジェクト検出が 2 カテゴリの識別問題であるのに対し，サブカテゴリ識別ではより多数のカテゴリを対象とする．多数カテゴリ

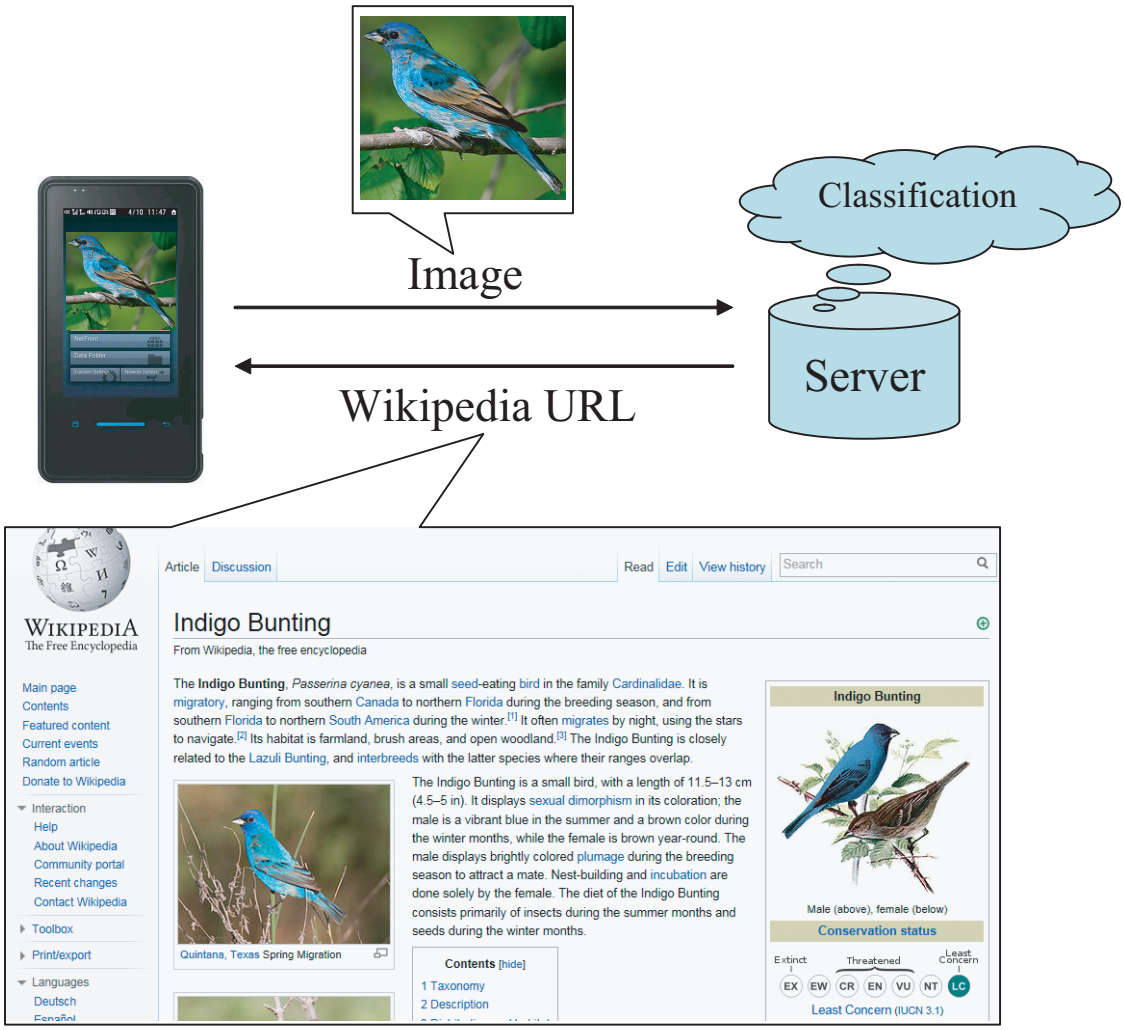

図 1.3. 鳥類の画像を入力すると，対応する Wikipedia ページを表示するシステム．そもそも名称を知らない対象の情報を知るにはキーワード検索では困難であるが，撮影した写真をクエリとすることで容易に専門知識にアクセスできるようになる．Perona [57] が Visipedia という同様のコンセプトを提案している．図中の Web ページの画像は，ウィキペディア (Wikipedia): フリー百科事典 [1]（更新日時 : 2011 年 12 月 13 日 21 時 56 分）より引用．

識別における共起性の活用方法を新たに考案する．

以上をまとめると，勾配に基づく特徴量や AdaBoost などの学習アルゴリズムは既存の方法を用いるが，特徴量の確率分布を利用する点や識別に有効な共起特徴を自動選択する点において新しく有効な方法を導く．さらに，特徴抽出から識別までを含めた処理全体で，計算コストを増加せずに識別精度を向上させる．言い換えれば，確率分布や共起性を活用することで，通常は二律背反となる高精度化と高速化の両立を図る点が本論文の重要な貢献となる．以下でより具体的に説明する．

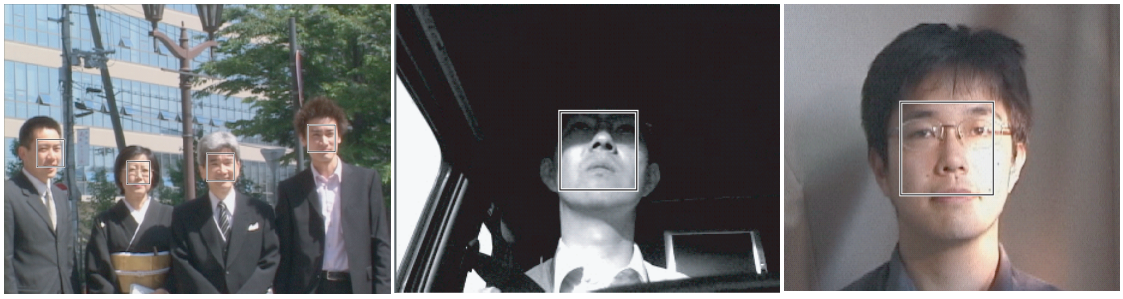

図 1.4. 照明条件の変動による顔画像の変化の例．屋外や車内で激しい照明変動が生じる．

図 1.5. カテゴリ内変動の例．人種，性別，年齢などの個体差や表情や姿勢の変化がある．[48] および CUB-200 データセット [87] より引用．

1.2　本論文の概要

ここでは，本論文の主要な貢献を示すとともに，本論文の構成を説明する．

1.2.1　貢献

本論文の主な貢献は，

- 照明変動などの外乱に対する頑健性と顔の個人差のようなカテゴリ内の変動に対する許容性を兼ね備えたテンプレート照合法
- 高精度かつ高速なオブジェクト検出のための，複数の特徴の共起性を利用した識別器学

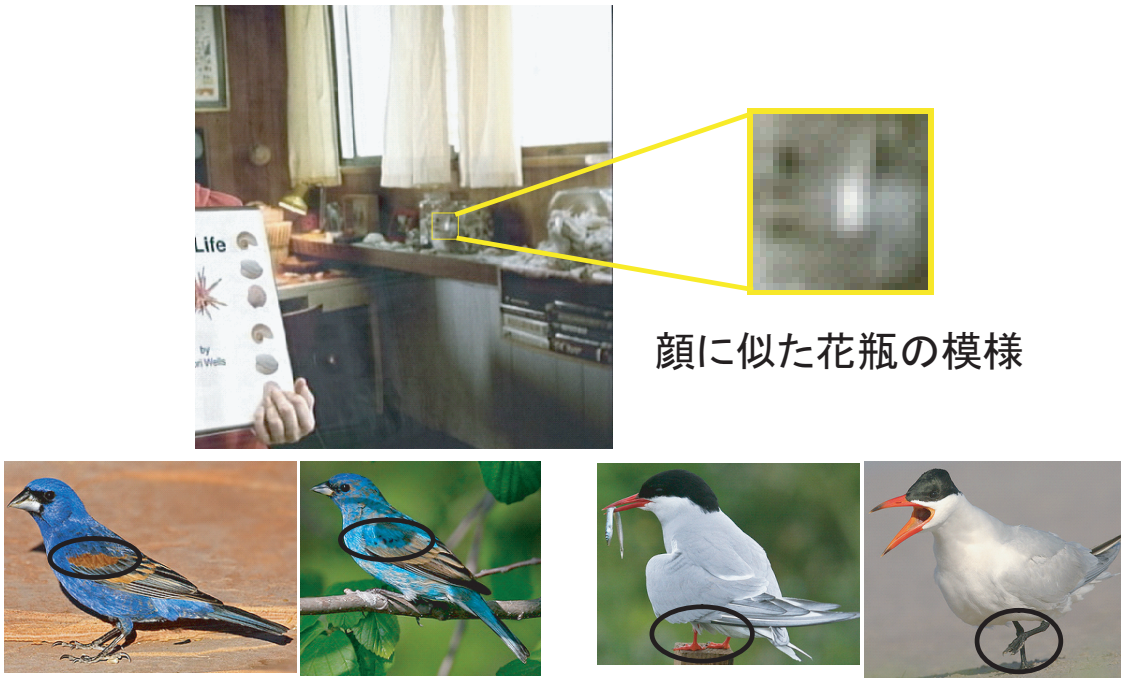

図 1.6. 類似カテゴリの例．顔検出では背景の一部に顔に似た領域が存在する．サブカテゴリ識別では，(a) と (b)，(c) と (d) のように一部の特徴のみが異なる別の種が存在する．このような類似カテゴリを混同しないように識別器を設計しなければならない．CUB-200 データセット [87] より引用．

習アルゴリズム

- 複数の特徴の共起関係を利用したサブカテゴリ識別手法

である．基盤となるアイデアは，特徴量の確率分布を利用することと共起性を評価する点であり，高精度化と省処理化を両立し，アルゴリズムの汎用性確保につながる．以降では，各手法を詳細に説明し，なぜ上述の技術課題を解決できるのかを示す．さらに，数千から数万枚の画像を使用した識別実験を通じ，最も近い技術課題を扱う先行研究に比べ，計算量を増やすことなく精度を高められることを示す．提案手法はいずれも学習サンプルを変更すれば異なるオブジェクトカテゴリを検出あるいは認識できる枠組みとなっており，汎用的であることを説明する．

1.2.2　構成

本論文の構成は以下のとおりである．

第 2 章　オブジェクト検出・認識に関する先行研究について述べ，1.1.2 で述べたアプローチの新規性について議論する．続く各章において，それぞれのアプローチの有効性を理論的，実験的に示していく．

第 3 章　確率的増分符号相関と呼ぶ新たなテンプレート照合法を提案する [97, 47, 95, 94]. 増分符号という照明変動に対して頑健な特徴量を複数の参照画像において観測し，その確率分布をテンプレートとする．従来，カテゴリ内変動を伴う対象を検出するには，平均画像をテンプレートとする照合法や部分空間法などの統計手法が一般に用いられるが，1,000 枚を超える画像から顔を検出する実験を通じて，提案手法が識別精度と計算コストの両面で大幅に優れていることを示す．さらに，2 カテゴリ識別，多数カテゴリ識別への拡張を試みる．

第 4 章　複数の局所特徴量の共起性を利用したオブジェクト検出手法を提案する [48, 98, 46, 96]．Sequential Forward Selection と Boosting を組み合わせた学習アルゴリズムにより，対象と非対象 2 つのカテゴリを識別するのに適した共起特徴を自動選択する．顔や 3 種類の形状の手を検出する実験で，計算コストを増加させることなく Viola と Jones の方法 [83] を上回る精度が得られることを示す．

第 5 章　画像内の同じ位置から検出した複数の特徴の共起性を利用したサブカテゴリ識別手法を提案する [99]．Random Forest という木構造の量子化アルゴリズムを導入して，特徴の共起関係を抽出することにより，互いによく似たサブカテゴリの識別精度を高める．200 種類の鳥類 [87] を識別する実験で，一般物体認識において最も識別精度が高いとされる Spatial Pyramid と Multiple Kernel Learning を組み合わせた手法を上回る識別率を，より小さい計算量で達成できることを示す．

第 6 章　前章までの議論を俯瞰して，本論文の主な貢献についてまとめる．提案手法で改善された課題とともに，残された課題について議論する．

第 2 章

関連研究

前章で指摘したように，オブジェクト検出あるいは認識の性能向上には，どのような特徴量と識別手段を用いるかが重要である．まず，最も簡便なオブジェクト検出手法としてテンプレート照合に関する先行研究をまとめる．そして，既存手法ではカテゴリ内の変動を扱う点において問題があることを指摘する．次に，局所特徴量について主要な先行研究を俯瞰する．局所特徴量は，対象物体の一部の領域から算出される特徴量である．これは，対象物体全体から特徴量を算出する多くのテンプレート照合法では，局所的な外乱への耐性が十分でないことや，対象の形状変化に対応し難いという問題があり，これを解決するのに有効とされている方法である．さらに，オブジェクト検出や認識に用いる識別器を学習する過程で，これらの局所特徴量がどのように用いられているかについて説明し，既存の枠組みでは識別性能の改善に限界があることを指摘する．最後に，指摘した 2 つの問題を解決するアイデアを示す．

2.1 テンプレート照合

事前に登録された参照画像（テンプレート）と対象画像の類似性を判定する処理は，テンプレート照合と呼ばれ，画像処理の基本課題として古くから研究されている [13]．通常，参照画像と対象画像では撮影条件が異なるため，照明変動などに影響されにくい照合法が必要となる．代表的な照合法である正規化相関 (Correlation Coefficient) [4] は，次式によって計算される．

$$CC = \frac{\sum \left(I(i) - \bar{I}\right)\left(I'(i) - \bar{I}'\right)}{\sqrt{\sum \left(I(i) - \bar{I}\right)^2}\sqrt{\sum \left(I'(i) - \bar{I}'\right)^2}} \tag{2.1}$$

ここで，I と I' はそれぞれ参照画像と対象画像である．$I(i)$ は参照画像 I の i 番目の画素値である．また，$\bar{I}$ は，I の画素値の平均である．正規化相関は，画像の分散で正規化するため，画像全体の一様な明度変化を許容できるが，局所的な明度変化や遮蔽といった外乱を扱うことができない．方向符号法 [77] は，コントラストの最大方向が照明変動に対して不変であることを利用するため，明度変化に対し頑健である．遮蔽に強い方法としては，ブロック分割した複数の部分テンプレートによる照合結果を統合する方法 [101] や，増分符号相関 [102]，増分符号画像によって遮蔽部分を除いて照合を行う選択的正規化相関 [93] がある．特に増分符号

相関は，照合に要する計算コストが小さく，その統計的性質に基づいてノイズや遮蔽に対する頑健性が示されている点で，実用性が高い．

増分符号相関 (Increment Sign Correlation; 以下 ISC) は，隣接画素間における明度の増減に着目した照合評価値である．注目画素と隣接画素の明度を比較し，その大小に応じて 0 もしくは 1 の 2 値の符号を与えることにより，画像から符号列を生成する．参照画像と対象画像それぞれの符号列を比較し，符号が一致する個数の割合として相関値を求める．ISC には以下の優れた特長がある．まず，照合する 2 枚の画像において相対的な明度の増減が保存されているかどうかのみを評価するため，一様な明度変化の影響を受けにくい．局所的な明度変化であっても，明度の大小関係が逆転しない大きさならば影響されない．また原理的に，正規化相関より計算コストが小さい．

しかし，これらの画像照合法は，基本的に参照画像と対象画像 1 対 1 の相関を評価する方法であるため，人物の顔のように個人差や表情変化などカテゴリ内変動を伴う対象を検出するには，参照画像の選択によって検出精度が左右されるという問題がある．複数の参照画像に対して照合を行うマルチテンプレート法を用いればカテゴリ内変動に対処できる可能性があるが，テンプレートの枚数に比例して計算量が増加するという問題がある．平均画像を参照画像として対象カテゴリを代表させる方法もよく用いられる [100] が，平均画像では参照画像間の変動の大きさ（ばらつき）を評価することができない．そのような変動を平均画像に加わったノイズとして吸収するのには限界があるため，変動を扱う新たな手段が必要である．

カテゴリに含まれる変動を扱う方法としては，部分空間法 [91] がよく用いられる．顔を対象とした方法では，固有顔（Eigenface） [76] という部分空間法と類似した考え方に基づく手法がある．これは，参照画像間の変動を表現し，それら変動の影響が吸収されるような低次元の特徴ベクトルを求める方法である．しかし，局所的な照明変動など参照画像のサンプルに含めることが困難な変動への対応が難しい上，画像間の内積計算を複数回行う必要があり，通常の相関に基づく照合法に比べて計算量が大きくなるという問題がある．

このように，既存のテンプレート照合法ではカテゴリ内変動に対処する能力に限界があり，カテゴリ内変動を扱う統計手法には計算コストが大きいという問題がある．

2.2　局所特徴量

上述したテンプレート照合法は，基本的に対象物体の全体から特徴量を算出するため，形状や姿勢の大きな変動や局所的に加わった外乱に対する頑健性が低い．そこで，対象物体の一部，すなわち局所領域から特徴量を算出し，局所特徴量の集合によって物体を表現するという方法がとられるようになっている [67]．以下に，よく知られた局所領域の特徴量を表現する特徴記述子について概説する．そして，それぞれの特徴記述子が，照明条件の変動や回転，並進，スケーリング，アフィン変換などオブジェクトの変形に対する頑健性を改善するため，どのように設計されているかについて述べる．

Rectangle feature: ウェーブレット変換における Haar 基底に類似した特徴であり，Haar wavelet 特徴あるいは Haar-like 特徴と呼ばれることもある．図 2.1 に示すように，数個の方

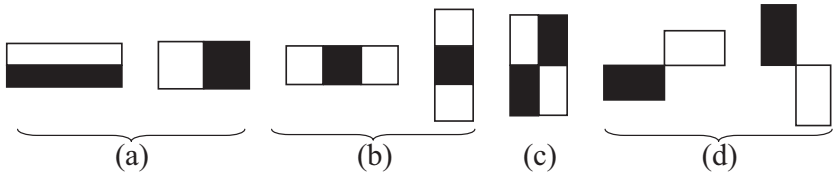

図 2.1. 様々な rectangle feature．2 つの方形を組み合わせた基本セット (a) に加え，方形の数を増やす，あるいは配置を工夫するなどの拡張セット (b)(c)(d) がある．

形を組み合わせた特徴である．白い方形と黒い方形それぞれに属する画素値の平均値の差を特徴量とする [56], [83]．すなわち，局所領域における特定の方向の勾配強度を求めるフィルタと言える．差をとるため，領域内に一定のオフセットが付加されるような変動に対して不変である．また，複数の画素の平均値を算出するため，撮影時などに付加されるランダムなノイズの影響が軽減される利点がある．積分画像（Integral image） [83] を用いると極めて小さいコストで特徴量を計算できるため，高速な処理に適している．

HOG(Histogram of Oriented Gradients): Dalal と Triggs [18] によって，人物検出における有効性が示されている．基本的には，注目領域内の各画素において勾配を求め，方向ごとに強度を投票した勾配ヒストグラムである．ヒストグラムはセルと呼ぶ 5×5 画素の小領域ごとに求め，さらにブロックと呼ぶ 3×3 セルの領域でヒストグラムの長さが 1 となるように正規化する．最終的に，各ブロックのヒストグラムを連結し，数千次元の特徴ベクトルを得る．これら一連の処理により，人物の姿勢変化や照明変動に対する頑健性が得られたと報告されている [18].

SIFT(Scale Invariant Feature Transform): Lowe [43] は局所領域における画素勾配のヒストグラムに基づく特徴記述子を提案している．Scale（特徴記述子を求める局所領域の範囲）と Orientation（局所領域の向き）に不変な特徴量を求めるために，Difference of Gaussian (DoG) を用いて interest point を検出する．勾配とそのヒストグラムを評価することによって照明変動と変形に対する頑健性を確保している．これらの好ましい性質から，一般物体認識の研究分野において広く用いられている．SIFT 特徴量を主成分分析によって次元圧縮するとともに，重要な特徴次元を強調した PCA-SIFT [36] が提案されている．また，SIFT がグレースケール画像における輝度勾配を求めていたのに対して，それを様々な色空間に適用した特徴記述子が提案されている．HSV-SIFT [10] は HSV の 3 チャンネルでそれぞれ 128 次元の SIFT 特徴量を求め，連結した 384 次元の特徴記述子を求める方法である．HueSIFT [79] は通常の SIFT 特徴量に Hue 成分のみから求めた特徴量を連結する方法である．その他，C-SIFT，rgSIFT，RGB-SIFT などがあり，それぞれ照明光の強度や色の変動に対する不変性が異なる．Sande ら [78] による詳細な比較実験によれば，OpponentSIFT が最も優れた識別性能を示している．OpponentSIFT は，照明光の強度変化に対して不変となるような変換 (2.2) を RGB 各成分に適用した後，変換後の成分 (O_1, O_2, O_3) それぞれについて SIFT 特徴量 [43] を算出し，384 次元の特徴記述子を取得する．

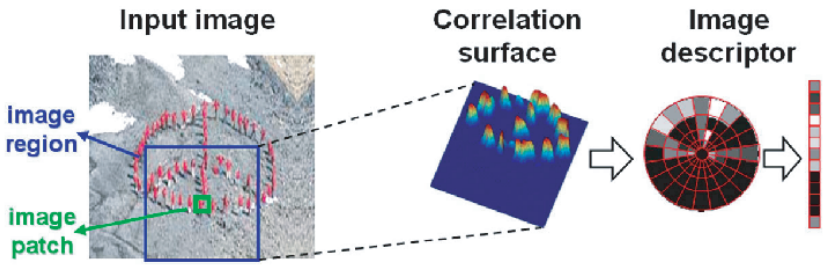

図 2.2. Self-similarity 特徴記述子の検出例．[70] より引用．

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix} \tag{2.2}$$

Harris-Affine: Mikolajczyk と Schmid [45] は Scale とアフィン変形に対して不変な interest points の検出方法として, Harris-Affine detector を提案している．Harris detector [31] は本来 Scale やアフィン変形に対する不変性を持たないが，スケール選択の仕組み [41] を取り入れ，さらにアフィン変形に対応させている．

Shape context: Belongie ら [5] はエッジあるいはオブジェクトの輪郭に沿って注目点を設定し，その周辺における他の点群の相対的な分布を表現する特徴記述子を提案している．すなわち，局所的なオブジェクトあるいはテクスチャの形状を表現する特徴量である．変形するオブジェクトの照合に対する有効性が示されている．

Self-similarity: Shechtman と Irani [70] は，図 2.2 に示すように注目する image patch とその周辺の patch 群との類似度を求め，類似度の空間的な分布を特徴記述子とする方法を提案している．輝度値そのものでも輝度勾配でもなく，注目領域と周辺との類似性のみに着目することで，対象の形状あるいはレイアウトを抽出することに成功している．また，Shape context と同様に log-polar bin を使用して特徴記述子を求めるため，変形に対する頑健性を備えている．

Geometric Blur: Berg と Malik [7] はアフィン変換などの幾何学的な変形を仮定し，変形後の画像における局所領域の積分値を特徴とする方法を提案している．オブジェクト認識においてもアフィン変形に対する頑健性が有効であることを示している [6].

Superpixels: Ren と Malik [61] は “superpixel” と呼ぶ色やテクスチャが似た小領域を Normalized cuts によって求める方法を提案している．図 2.3 に検出例を示す．

2.3　2 カテゴリ識別器を用いたオブジェクト検出

現在，最も有効とみなされているオブジェクト検出方法は，多くの学習サンプルから確率的な枠組みあるいは識別境界の探索によって 2 カテゴリ識別器を求める方法である．例えば，ニューラルネットワークに基づく顔検出器 [73, 63] や SVM による識別境界の学

図 2.3. Superpixels の検出例.

習 [55, 32, 56]，あるいはナイーブベイズ識別と呼ばれる統計学習手法 [68] がある．このうち，[73, 63, 55, 32] では画素値そのものを特徴量として用いているため，照明変動などの外乱に影響されやすいという問題がある．これに対し，局所領域における勾配を抽出する Haar ウェーブレット特徴 [56] や rectangle feature [83]，あるいは特定の方向や空間周波数にのみ反応する Gabor 特徴を用いれば，外乱への頑健性が増すことが知られている．

Viola と Jones [83] は，AdaBoost [24] を用いて識別に適した rectangle feature を選択し，識別器を構築する方法を提案している．選択された少数の特徴を使って小さい計算コストで高い識別精度が得られることが示されている．巨大な特徴セットから Boosting アルゴリズムによって識別に適した特徴だけを選択するというこのアプローチは，オブジェクト検出の基本的なフレームワークとなり，様々な拡張が試みられている．それらは大きく 2 つに分類される．まず，第 1 の拡張方法は Boosting アルゴリズムそのものの改善である．Real AdaBoost [66]，や KLBoosting [42]，あるいは FloatBoost [39] がある．FloatBoost は顔検出だけでなく，手の検出にも適用されている [54]．第 2 の拡張方法は，様々な画像パターンに対応できるよう特徴セットを追加することである．図 2.1 に示したように，基本的な特徴セット (a) に加え，(b)(c)(d) のように配置や数の異なる rectangle feature が使用されている [83, 88]．Lienhart ら [40] は 45° 回転した rectangle feature を効率よく計算する方法を提案している．Kölsch と Turk [37] は手検出用に独自の rectangle feature を追加している．Viola ら [84] は，空間方向の勾配だけでなく時間的な変化すなわち動きを取り入れた特徴を，歩行者検出に適用している．いずれの拡張方法も効果的ではあるが，さらにオブジェクト検出の精度を高めようとすると限界に直面する．なぜなら，これらの拡張方法は，特徴セットから特徴を 1 つずつ選択しながら，弱識別器を作っているからである．以下で詳しく説明する．

図 2.4 に Viola と Jones による識別器の学習方法を示す．それぞれの弱識別器 $h_t(x)$ は，数多くの特徴セットから最も識別に適した rectangle feature を 1 つ選択して学習される．計 T' 個の弱識別器の線形結合によって，入力画像 x が検出対象か否かを識別する．すなわち，識別時には T' 個の rectangle feature が観測され，その特徴量が評価される．しかしながら，単一の特徴のみによって形成された弱識別器では Boosting によって学習が進むにつれて識別性能

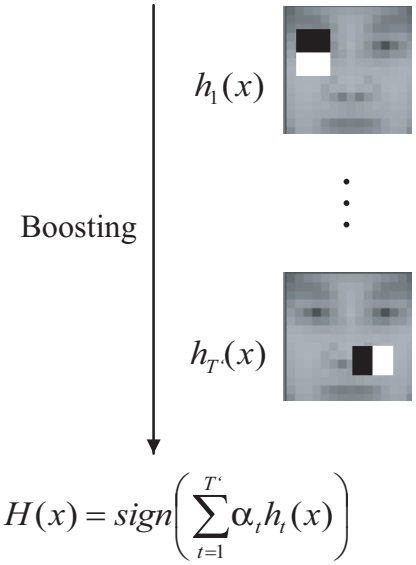

図 2.4. Viola と Jones による識別器の学習方法．T' 個の弱識別器 $h_t(x)$ の線形結合として識別器を学習し，入力画像 x を識別する．それぞれの弱識別器では，単一の rectangle feature を使用する．

が低下し，後段の弱識別器の識別能力は改善されなくなる．Viola と Jones も後段で選択された特徴を使用した識別誤り率は 0.4 から 0.5 の間であったことを報告している．2 カテゴリの識別において，誤り率の最悪値は 0.5 であることから，後段の弱識別器の識別能力は相当に弱い．このような “きわめて弱い” 識別器では，学習時の識別誤り率（訓練誤差）を低下することはできても，汎化誤差を小さくすることができない．これは Boosting アルゴリズム自体をいかに拡張しても，単一の特徴による弱識別器を使用する限り生じる問題である．単一の特徴であっても高い識別能力が得られるように，新たな特徴記述子を考案することは容易でなく，できたとしても現在の検出対象にのみ最適化された汎用性の低い方法となる問題がある．また，rectangle feature に加えて Gabor 特徴も使用するというように，別の特徴記述子を特徴セットに加えていく方法が簡単に思いつくが，特徴空間が巨大化し，学習のための計算コストが増大し過ぎるという問題がある．

まとめると，Viola と Jones のフレームワークは，高速計算可能で外乱に対して頑健な特徴を用い，識別に適した少数の特徴をあらかじめ選択しておくことによって，高速かつ高精度なオブジェクト検出を実現する優れた方法である．Boosting という学習アルゴリズムに特徴選択の機構を加えたという観点でも，重要な貢献である．しかし，すでに言及したとおり，さらなる識別性能の改善を図ろうとすると，学習に要する計算コストや汎用性の面で限界につきあたってしまう．

Dalal と Triggs [18] による HOG (Histogram of Oriented Gradients) は，人物全身像の検出に有効である．人物の学習サンプルとそれ以外のサンプルから HOG を計算し，線形 SVM によって識別を行う．数千次元というかなり高次元の特徴ベクトルとなる HOG に対し，SVM

によって識別に寄与する特徴次元とそうでない次元にそれぞれ異なる重みが付与される．特徴選択を行う Viola と Jones の方法とは異なり，特徴に重み付けをする方法と捉えることができる．やはり，さらなる識別性能の改善には，識別アルゴリズムの変更か新たな特徴を追加するという方策がとられる．Dalal ら [19] は，歩行者のオプティカルフローから動き情報を表現する特徴量を求め，HOG と組み合わせた．また，線形 SVM の代わりにガウシアンカーネルに基づく SVM を使用して識別精度が高まることも報告している．いずれも検出精度は向上するものの計算コストも増加している．

2.4　Bag of visual words によるオブジェクト認識

オブジェクト検出が特定のカテゴリに属するオブジェクトが画像内のどこにあるかを決定する問題であるのに対し，画像に写っている物体のカテゴリが何かを決定するのがオブジェクト認識である．オブジェクト検出では対象と非対象 2 つのカテゴリを識別するが，オブジェクト認識ではより多くのカテゴリを識別する．オブジェクト認識には，人物，車，飛行機といった一般的名称を認識する一般物体認識 (generic object recognition) と，鳥類のように一般物体より細分化された下位のカテゴリを扱うサブカテゴリ識別 (subordinate category classification) などがある．

オブジェクト認識は，古くから研究されており，線画解釈 [16] やルールベースの画像理解 [53]，円筒などの幾何学的なモデルに基づく方法 [59] などがある．扱える対象の種類や形状が限られていたり，認識ルールの記述が難しいという問題を抱えていた [105]．その後，大量の学習用画像から統計的な手法を用いて識別器を学習するアピアランスベースの手法が用いられるようになった．固有顔 [76] がその代表例であるが，対象物体全体から特徴量を求めており，大きな姿勢変化や遮蔽に対処できなかった．そこで，Schmid と Mohr [67] は局所特徴量の集合としてオブジェクトを記述する方法を提案した．Harris オペレータ [31] により抽出した interest point の画素値や周辺領域での勾配を特徴ベクトルとして，類似した特徴ベクトルを探し，投票を行う．さらに，Perona ら [14, 15, 85, 86, 23, 21] は，局所特徴量とそれらの相対的な位置関係を確率的に表現する constellation model を提案し，カテゴリ内の変動を扱う枠組みを与えた．

constellation model とは異なり，局所特徴量の位置関係は無視し，特徴量の観測頻度のみに基づいてオブジェクトを表現する方法が bag of visual words [17] である．それぞれの局所特徴量を k-means などの方法でベクトル量子化することにより，各オブジェクトもしくは画像をベクトル量子化された局所特徴量 (visual word) のヒストグラムで表現する．さらに，マルチカテゴリに対応した SVM を用いて識別関数を求める方法が典型的である．

Grauman と Darrell [27] は，SVM 識別器においてオブジェクト同士の類似度を定義するカーネルを改良している．Pyramid Match Kernel と呼ばれるカーネルは，ヒストグラムのビンサイズを徐々に細かくしながら，ヒストグラムインタセクションを算出する階層的な方法である．これにより，大局的な特徴の類似性からより詳細な類似性まで評価できるようになり，Caltech-101 [22] データセットの識別率を高めている．Lazebnik ら [38] は，Spatial Pyramid

Kernel により，局所特徴量の位置関係を階層的に評価することによって，さらに Caltech-101 の識別率を向上している．

このように一般物体認識は bag of visual words を駆使した各種の手法によって大きく進展しているが，サブカテゴリの識別についてはまだ取り組みが始まった段階であり，十分検討が進んでいない．サブカテゴリを高精度に識別するには，次のような課題がある．図 1.6 に，200 カテゴリという大規模な鳥類データセット CUB-200 [87] に含まれる 4 種類のカテゴリを示す．これらは，全体としてはよく似ているが，黒丸で囲われた特定の部分の色や形状が異なる．したがって，サブカテゴリの識別では，このようなわずかな差異を捉えることが重要となる．このような類似したカテゴリは，一般物体認識の研究で扱われてきた Caltech-101 [22], Caltech-256 [28]，あるいは VOC2010 [20] などのデータセットには見られない．そのため，従来の一般物体認識手法をそのまま使用しても十分な識別精度が得られない．

従来手法のうち最も高精度とされているのは，Bag of visual words [17] のフレームワークに基づき，Multiple Kernel Learning によって複数の特徴から算出されたカーネルを結合し [81, 82]，さらに特徴の空間的な配置を考慮した Spatial Pyramid Kernel [38] を用いた方法である．Branson ら [11] は，この方法による CUB-200 の識別精度は 19% と低かったことを報告している．これは，原理的に従来手法では特定部位の差異が重要となるカテゴリの識別が難しいからである．例えば，図 1.6 に例示したカテゴリを識別するには，黒丸で示す部位のように，特定の色がどのような形状で分布しているかという情報が重要となる．しかし，従来手法では特徴ごとに個別にヒストグラムを求める過程で，特徴同士の詳細な位置関係を無視してしまうため，細かな違いを識別する能力が劣化してしまう．Spatial Pyramid を用いたとしても，画像を 16 分割する程度 [38] のおおまかな解像度でしか特徴間の位置関係を評価できない．

従来手法を用いて識別精度を改善するには，新たな特徴を追加する方法しかないが，次に述べる理由から困難である．すでに有効と思われる特徴をすべて使ってしまっている場合は，対象としている物体の識別に有効な新たな特徴を考案しなければならないが，これは容易ではない．また，特徴を追加すると，その特徴抽出処理やヒストグラム計算など識別処理に要する時間が単調に増加してしまう．Nilsbak ら [51] や Joutou ら [35] は，それぞれ 102 種類の花や 50 種類の食物画像の識別を行っている．サブカテゴリ識別を課題とする先行研究と位置づけられるが，特徴の種類を変えた以外は従来手法の枠組みをそのまま適用しており，上述の理由から高精度化は困難である．

2.5　画像検索への応用

オブジェクト認識技術の有望なアプリケーションとして，図 1.3 に例示した画像検索がある．検索のための適切なキーワードが思い浮かない場合に，被写体の画像をクエリとして検索するシステムである．Google 社は 2011 年 6 月に，画像をアップロードすると類似画像やその画像に関する情報を検索できるサービスを開始している．図 2.5 は，有名なランドマークの検索結果である．米国イエローストーン国立公園の滝の名称とともに Wikipedia へのリンク

図 2.5. Google 画像検索の実行例．米国イエローストーン国立公園のロウアー滝のスナップ写真（ファイル名：IMG_4014.JPG）をクエリとして検索した結果，滝の名称とともに Wikipedia へのリンクが提示されている．2012 年 2 月 5 日実行．

が提示されている．しかしながら，図 2.6 に示すように，Yellow headed blackbird という鳥の画像をクエリとしたところ，全体として色合いや構図は類似しているものの異なる被写体の画像のみが提示され，目的の鳥の画像は検索結果に見当たらなかった．このことから，現状の画像検索サービスには，被写体を認識する機能は組み込まれておらず，類似画像を検索するのにとどまっていることが分かる．Web 上のあらゆる画像を対象とする同社のサービスにとっては，現状のオブジェクト認識技術で扱えるカテゴリ数が少ないことや，認識精度や計算量が十分でないことがその理由と考えられる．

2.6　まとめ：先行研究の問題と解決へのアイデア

オブジェクト検出や認識の方法は，オブジェクト全体から特徴量を算出するテンプレート照合から，局所特徴量の集合によるオブジェクト表現へと発展してきている．また，AdaBoost や SVM といった統計的な学習アルゴリズムを用いて，多くの事例（学習用画像）から識別器を導出するようになっている．識別器の学習過程では，識別に有効な特徴の選択あるいは重み付けがなされており，特徴量と学習アルゴリズムが密接に関連している．現時点で最も優れた

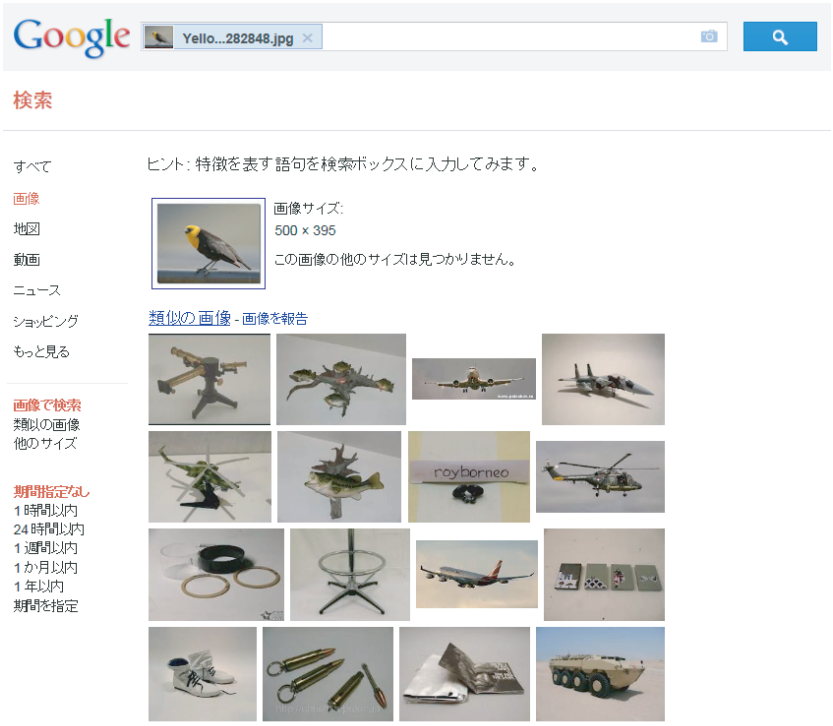

図 2.6. Google 画像検索の実行例．構図は類似しているものの被写体の鳥と類似した画像は検索されず，Yellow headed blackbird という鳥の名称は認識されていない．2012 年 2 月 5 日実行．

性能を示している，あるいは主流となっている手法は，テンプレート照合では増分符号相関，オブジェクト検出では Viola と Jones の方法，オブジェクト認識では bag of visual words である．すでに指摘したとおり，これらの方法には表 2.1 の問題がある．

本論文では，これらの問題に対して，2 つのアイデアに基づく新たな手法を提案し，以降の章で順に詳細を説明していく．まず，特徴量の確率分布をテンプレートとする照合法を提案する．このようなアイデアに基づくテンプレート照合法の例はなく新規であり，カテゴリ内変動に対処する方策としても有効と考えられる．次に，局所特徴量の共起性を利用したオブジェクト検出手法とサブカテゴリ識別手法を提案する．特徴の共起性に着目するアイデアそのものは，濃度共起行列 [106] などで古くから知られているが，識別に有効な共起関係を自動選択する，あるいは重み付けする新たな手法を示す．これにより，識別処理の計算コストを増加することなく，識別精度を向上させることが可能となり，極めて有効性の高い方法となる．

表 2.1. 先行研究の問題点と問題解決へのアイデア

問題	アイデア
既存のテンプレート照合法ではカテゴリ内変動に対処するのが難しい	特徴量の確率分布を利用
既存のオブジェクト検出・認識の手法ではさらなる識別性能改善が難しい	特徴間の共起性を評価

第 3 章

特徴量の確率分布を利用したテンプレート照合法

ここでは，個体差のある対象を，小さい計算コストで，かつ照明変動に左右されずに検出するための統計量を提案する．複数の参照画像から算出した増分符号の確率分布に基づく照合評価値であるため，確率的増分符号相関（Probabilistic Increment Sign Correlation; 以下 Probabilistic ISC）と呼ぶ．Probabilistic ISC は，画像内で特定の符号が観測される確率の高い位置，すなわち参照画像間で変動の小さい特徴に大きな重みを与えて照合を行うため，対象の個体差や形状変化などの影響を受けにくい．また，増分符号を用いるため，原理的に照明変動に対して頑健である．計算コストは増分符号相関と同程度に小さく，ハードウェアによる実現にも適している．

以下では，まず Probabilistic ISC の定義を示し，カテゴリ内変動に対する頑健性について考察する．統計的性質を明らかにし，解析的に照合しきい値を求める方法についても説明する．多数の実画像を用いた顔検出実験において，正規化相関などの相関に基づく照合法や代表的な統計手法である部分空間法との比較を行い，提案手法が従来手法よりも優れていることを示す．

3.1 確率的増分符号相関（Probabilistic ISC）

ここでは，まず Probabilistic ISC を定義する．定義に基づき，対象の変動に対して頑健であることを説明する．次に統計的性質を示し，それに基づいて解析的に照合しきい値を設定する方法を与える．

3.1.1 定義

簡単のため，画像を 1 次元明度列として説明する．実際の 2 次元画像の扱いは，3.3 において述べる．まず，参照画像が N 枚あるとする．n 枚目の参照画像 I_n における i 番目の画素の明度を $I_n(i)$ とする．

各参照画像において隣接する 2 画素の明度を比較し，符号列 B_n を生成する．i 番目の画素に対して割り当てられる符号を $B_n(i)$ とする．$B_n(i)$ を求める方法としては，式 (3.1) により 2 値の符号（増分符号）を得る方法 [102] と，式 (3.2) により隣接画素の明度が同値となる場合に特定の符号を与え，3 値の符号を得る方法 [100] の 2 通りを考える．なお，参照画像の明度列の長さが $M+1$ であるとき，符号列の長さは M となる．

$$B_n(i) = \begin{cases} 1 & I_n(i+1) \geq I_n(i) \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

$$B_n(i) = \begin{cases} 1 & I_n(i+1) > I_n(i) \\ -1 & I_n(i+1) < I_n(i) \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

次に，i 番目の画素において符号 s を観測する確率 $P(s,i)$ を計算する．$P(s,i)$ は N 枚の参照画像のうち符号 s が得られた画像の枚数の割合として，式 (3.3) により計算される．確率値 $P(s,i)$ の計算結果をルックアップテーブルに保持しておく．

$$P(s,i) = \frac{1}{N} \sum_{n=1}^{N} \delta(s, B_n(i)) \tag{3.3}$$

ただし，

$$\delta(s,z) = \begin{cases} 1 & z = s \\ 0 & z \neq s \end{cases} \tag{3.4}$$

とする．

照合時には，まず参照画像に対して行ったのと同様の手順で，対象画像 I' に対する符号列 B' を求める．次にテーブルから符号 $B'(i)$ に応じた確率値 $P(B'(i),i)$ を参照する．対象画像において，多くの参照画像から観測された符号と同じ符号を観測した場合，参照される確率値は大きくなる．各画素において符号が独立に観測されるとみなすと，参照された確率値の積 L によって相関を評価することができる（式 (3.5)）．L は，対象画像が参照画像カテゴリに属する尤度である．

$$L = \prod_{i=1}^{M} P(B'(i),i) \tag{3.5}$$

上式のかわりに，次式の対数尤度を用いれば，乗算でなく加算により相関を評価できる．これを Probabilistic ISC の相関値 C と定義する．なお，対数の計算は照合時に行う必要はなく，計算結果をルックアップテーブルに保持しておけばよい．

$$C = \sum_{i=1}^{M} \log P(B'(i),i) \tag{3.6}$$

3.1.2　カテゴリ内変動に対する頑健性

各画素において観測される符号の生起確率 $P(s,i)$ について考察する．$P(s,i)$ は，参照画像間において生じる増分符号の変動の大きさを示している．例えば，ある画素において特定の符

号が多くの参照画像から観測されるとき，それはカテゴリ内の変動が小さい特徴である．変動が小さいほど，その符号の生起確率は大きな値をとり，最大値は 1 となる．一方，各符号が観測される確率値に偏りがないとき，それは変動が大きい特徴である．最も変動が大きい場合，2 値符号列では $P(0,i) = P(1,i) = 0.5$ となる．このように，多くの参照画像から共通して観測される特徴であるほど，得られる確率値は大きな値となる．式 (3.6) より，Probabilistic ISC の相関値 C は，これら確率値の対数を加算することにより求められるので，変動が小さく照合に有効な特徴は C に対して大きく寄与することが分かる．つまり，Probabilistic ISC では，変動の小さい特徴に大きな重みを与えて相関を評価するため，変動の影響に左右されずに照合を行うことができる．これに対して，変動を評価しない他の手法では全ての画素が同じ重みとして扱われる．ある画素においていずれの符号を観測するかは，与えられた参照画像に依存して決定され，その符号が変動しやすいかどうかについては考慮されない．したがって，変動の影響を受けやすく，参照画像の選択方法によって照合精度が左右される．参照画像として平均画像を用いても，変動の大きさを評価することはできないため，高い照合性能は得られない．

3.1.3　統計的性質

ここでは，Probabilistic ISC の統計的性質について考察する．まず，相関値 C の分布について考える．式 (3.6) から，C は確率変数の和として定義される．よく知られているように各変数が独立であるとき，それらの和は正規分布で近似できるので，大きな画像では C は正規分布に従う統計量となる．この場合，次節に述べる方法により，損失最小の意味で最適な照合しきい値を解析的に求めることができる．また，対象画像が参照画像のカテゴリに属するかどうかを統計的に検定するなど，各種の統計手法を応用できる．

次に，C の分布の平均 $E(C)$ と分散 $V(C)$ を定式化する．これらの式を用いると，確率モデルを仮定できるカテゴリに対しては，多数の画像を収集して実際に照合処理を行うことなく相関値の分布を推定することができる．記述を簡単にするため，i 番目の画素において照合時にテーブルから参照される値を $p(i)$ とおき，

$$C = \sum_{i=1}^{M} \log P(B'(i), i) = \sum_{i=1}^{M} p(i) \tag{3.7}$$

とすると，平均と分散は，

$$E(C) = \sum_{i=1}^{M} E\left(p(i)\right) \tag{3.8}$$

$$V(C) = \sum_{i=1}^{M} V\left(p(i)\right) \tag{3.9}$$

と書ける．

参照画像と同様に，照合の対象となる画像の i 番目の画素において符号 s を観測する確率

$P'(s,i)$ が得られたとすると，

$$E\left(p(i)\right) = \sum_{s} \left\{P'(s,i) \log P(s,i)\right\} \tag{3.10}$$

$$V\left(p(i)\right) = \sum_{s} \left\{P'(s,i) \left(\log P(s,i)\right)^2\right\} - E\left(p(i)\right)^2 \tag{3.11}$$

と表せる．

対象画像が参照画像と同一カテゴリである場合，すなわち検出対象のカテゴリに対しては，

$$P'(s,i) = P(s,i) \tag{3.12}$$

として，平均と分散を見積もることができる．

また，対象画像が参照画像と異なるカテゴリである場合，すなわち検出対象でないカテゴリに対しては，検出対象以外のあらゆる画像に対する $P'(s,i)$ を推定する必要がある．しかし，検出対象を含まない画像同士は互いに相関がないとみなせるので，十分多くの画像を集められたとすると，2 値符号を用いる場合，

$$P'(0,i) = P'(1,i) = 0.5 \tag{3.13}$$

とすることができる．したがって，検出対象でないカテゴリに対する相関値は，以下の期待値をもつ．

$$E_N(C) = \sum_{i=1}^{M} \frac{\log P(0,i) + \log P(1,i)}{2} \tag{3.14}$$

これは，参照画像から得られる確率値のみによって決まる定数である．3 値符号では，

$$P'(1,i) = P'(-1,i) = \frac{1 - P'(0,i)}{2} \tag{3.15}$$

とすることにより，期待値を計算できる．ただし，2 値符号とは異なり，対象画像において隣接画素の明度が同値となる確率 $P'(0,i)$ が必要となる．$P'(0,i)$ は，対象画像の性質と 1 画素の明度値を表現する階調数によって決まり，例えば，1 画素 8 ビット（256 階調）の一様乱数画像に対する相関値の平均と分散は，$P'(0,i) = 1/256$ として見積もることができる．

2 値符号と 3 値符号のいずれを用いるのがよいかは，参照画像カテゴリおよびそれと区別したいカテゴリの画像の性質による．それぞれの符号を用いた場合で，カテゴリ間分散とカテゴリ内分散の比を評価するなどの予備実験を行うことで選択できる．ただし，同値符号を評価しない 2 値符号は，明度が均一な領域や低い階調の画像を扱う際に照合が不安定となる場合がある．

3.1.4　しきい値設定

画像から対象物を検出する問題では，テンプレートを走査させ，各位置において得られた相関値としきい値を比較することにより，検出対象かどうかを判定する．Probabilistic ISC による検出では，前節の統計的性質に基づき，損失最小の意味で最適なしきい値設定を行うこと

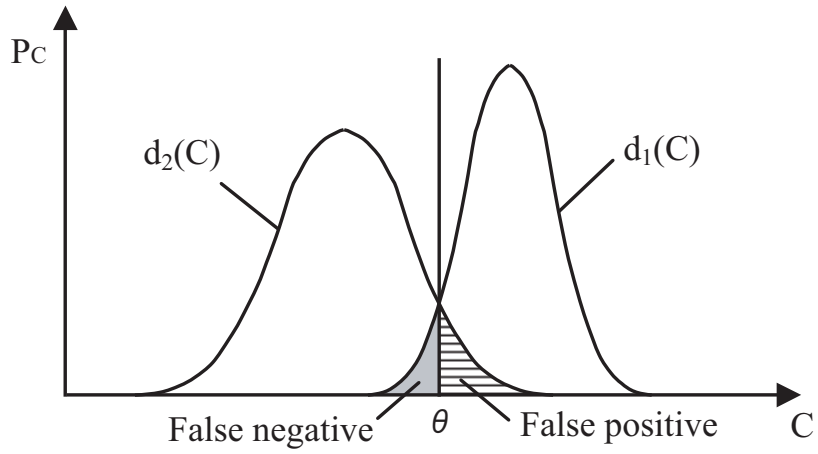

図 3.1. しきい値設定

が可能となる．一般に，検出対象のカテゴリおよびそれと区別したい非対象カテゴリそれぞれに属する画像に対して相関値を算出し，その確率密度を求めると，図 3.1 に示すような 2 つの分布を形成する．ほとんどの場合，分布同士は裾野の部分で重なり合うため，どのようにしきい値 θ を設定しても False negative（未検出:対象カテゴリの画像を検出しない誤り）と False positive（過検出:非対象カテゴリの画像を検出する誤り）を同時に 0 にすることはできない．このような場合，未検出と過検出に対する損失をそれぞれ c_1, c_2，対象カテゴリおよび非対象カテゴリの確率密度関数をそれぞれ $d_1(C)$, $d_2(C)$ とすると，次式を満たすしきい値 θ により期待損失を最小化できる [102].

$$\frac{d_1(\theta)}{d_2(\theta)} = \frac{c_2}{c_1} \tag{3.16}$$

相関値 C の分布が正規分布で近似可能な場合には，確率密度関数として式 (3.8) および (3.9) の平均と分散を持つ正規分布を用いることができる．損失 c_1, c_2 は用途に応じて設定するパラメータである．

3.1.5　ノイズに対する頑健性

Probabilistic ISC による照合では，隣接画素の明度の大小関係が保存されているかどうかのみを評価しているため，その関係が変化しない大きさのノイズには影響されない．画素間の明度差が大きいほど，また付加ノイズが小さいほど照合精度が保たれやすい．具体的には，全画素に一様なオフセットが加わるような明度変化には影響されない．局所的な明度変化でも，その領域内部では一様なオフセットが加わったとみなせるため影響されない．領域の輪郭部分においてのみ，明度の大小関係が保存されない場合があるが，このような画素は少数であるため影響が少ない．

3.1.6　計算コスト

照合時に必要となる計算は，隣接画素間の明度比較，明度比較結果に応じたテーブルの参照，参照された値の加算がそれぞれ M 回のみである．これは ISC と同等である．なお，3 値符号列を用いる場合には，明度比較の回数が各画素につき 1〜2 回必要となり，計算コストが増す．

3.2　多数カテゴリ識別への拡張

ここでは，Probabilistic ISC を拡張し，多数のカテゴリを識別する問題に適用する．K 個のカテゴリの識別では，K 個のルックアップテーブルを作成する．入力画像 I' から増分符号列 B' が得られたとすると，K 個の相関値 C_k が式 (3.6) により，計算される．以下では，まず 2 つのカテゴリを識別する問題について定式化し，次により多くのカテゴリを識別する問題を考える．

3.2.1　2 カテゴリの識別

オブジェクト検出は，検出対象のカテゴリと非検出対象のカテゴリの 2 つを識別する問題とみることができる．その識別関数は，

$$k^* = \begin{cases} +1 & C_{+1} - C_{-1} \geq \theta \\ -1 & \text{otherwise} \end{cases} \tag{3.17}$$

となる．ここで，C_{+1} と C_{-1} は，検出対象カテゴリと非検出対象カテゴリそれぞれに対する相関値であり，θ はしきい値である．式 (3.6) より，この識別関数は次に示す対数尤度の比で定義される．

$$C_{+1} - C_{-1} = \sum_{i=1}^{M} \log \frac{P(B'(i), i|k = +1)}{P(B'(i), i|k = -1)} \geq \theta \tag{3.18}$$

対数尤度比は，2 つのカテゴリを識別するのに有効な特徴 $B'(i)$ に，より大きな重みを与えていると解釈できる．

3.2.2　多数カテゴリの識別

カテゴリ数が 2 より大きい場合の識別関数は次のとおり容易に導ける．K 個の相関値のうち最大の値をとるカテゴリラベルを求めればよい．これは，ナイーブベイズ識別と呼ばれる方法と等価である．具体的には，上述の対数尤度比に基づく識別関数において $\theta = 0$ とし，K 個のうち 2 カテゴリずつ相関値を比較していき，最大の相関値をとるカテゴリラベルを求める．

$$k^* = \arg\max_k C_k \tag{3.19}$$

式 (3.6) が検出対象カテゴリのみに対する相関値を評価する “generative” な方法であるのに対し，2 以上のカテゴリを識別する式 (3.18) は “discriminative” な方法で識別関数を導いている．線形判別分析 (LDA) など他の discriminative な方法に対する提案手法の利点としては，識別対象のカテゴリが増減した場合に識別関数を更新する手間が小さいことである．具体的には，LDA は新しくカテゴリが追加されると識別関数を学習し直す必要があるが，提案手法では追加されたカテゴリからルックアップテーブルを作成すればよいだけであるので，計算コストが小さくて済む．

3.3　実験

顔には目鼻形状の違いやヒゲの有無など個人差があり，同一人物であっても表情の変化や経年変化が生じる．顔は，個体差のある対象の典型例であるといえる．本章では，多数の画像から顔を検出する実験を行うことによって，Probabilistic ISC の有効性を検証する．正規化相関などの相関に基づく従来の照合法や代表的な統計手法である部分空間法との比較を行う．さらに，参照画像（学習サンプル）の枚数を減らしていった場合の検出精度の劣化を調べ，Viola と Jones の方法 [83] などと比較する．最後に，多数カテゴリに対する識別能力を把握するため顔向きを推定する実験結果を示す．

3.3.1　顔検出アルゴリズム

参照画像より大きな入力画像内で，顔を検出する場合を考える．はじめに，顔の参照画像から照合時に参照するルックアップテーブルを作成しておく．次に，入力画像全体を符号列に変換する．本実験では，式 (3.20) と (3.21) に基づいて水平・垂直 2 方向に隣接する画素の明度を比較することにより，2 値符号列 BH および BV を得る．これにより，1 方向だけの明度比較では区別できないパターンに対する照合精度が増す．3 値符号列については，これらの式の簡単な修正により得られるので割愛する．

$$BH(x,y) = \begin{cases} 1 & I(x+1,y) \geq I(x,y) \\ 0 & \text{otherwise} \end{cases} \tag{3.20}$$

$$BV(x,y) = \begin{cases} 1 & I(x,y+1) \geq I(x,y) \\ 0 & \text{otherwise} \end{cases} \tag{3.21}$$

続いて，入力画像の原点から参照画像と同じ大きさの窓を走査させ，各窓領域に対する相関値を算出する．相関値は，2 方向についてルックアップテーブルから参照された値を加算して求め，その値が設定されたしきい値を上回っていれば「顔」，そうでなければ「非顔」と判定する．なお，各走査位置において部分画像を切り出し，部分画像の符号列を生成するのではなく，あらかじめ入力画像全体を符号化しておくことによって，明度の比較演算回数を大幅に削減することができる．

表 3.1. 比較手法

CC	正規化相関
ISC	増分符号相関 [102]
QTR	定性的 3 値表現 [100]
Subspace-1	部分空間法 [91]
Subspace-2	部分空間法：非顔サンプル使用
Probabilistic ISC-1	提案手法：2 値符号列
Probabilistic ISC-2	提案手法：3 値符号列

3.3.2　比較手法

表 3.1 に示す 7 種類の方法による顔検出精度および照合処理に要した時間を比較する．従来手法としては，相関に基づく照合法 3 種類 (CC, ISC, QTR) と代表的な統計手法である部分空間法 2 種類 (Subspace-1, Subspace-2) [91] を用いる．提案手法としては，2 値符号列に基づく場合 (Probabilistic ISC-1) と 3 値符号列に基づく場合 (Probabilistic ISC-2) の 2 種類を比較する．

CC は式 (2.1) により相関値を求める．$\bar{I}$ および $\bar{I}'$ は，それぞれ参照画像と各走査位置における部分画像の画素平均値である．CC は，一様な照明変動に対応可能であるが，局所的な照明変動を扱うことができない．

明度値自体に依存する CC に比べ，ISC と QTR は局所的な照明変動に対して頑健である．ISC は式 (3.20) と (3.21) に基づいて 2 値符号列を生成する．参照画像と入力画像中の窓領域内の符号列を比較し，符号が一致した総数を相関値とする．QTR は，2 値符号列ではなく 3 値符号列を用いて，相関値を求める．Probabilistic ISC-1 と Probabilistic ISC-2 は，それぞれ ISC と QTR を確率的に拡張したものに対応している．

Subspace-1 は，顔の参照画像から生成した部分空間に各走査位置から切り出した部分画像ベクトルを射影したときの射影成分の長さを相関値とする．部分空間を張る基底は，顔画像の明度値ベクトルの集合から自己相関行列を求め，それを KL 展開することによって得る．Subspace-2 は，顔の部分空間に加えて，顔以外（非顔）の参照画像から生成した部分空間にも射影し，両カテゴリへの射影成分の長さの差を相関値とする．顔か非顔いずれのカテゴリに似ているかを評価するため，Subspace-1 に比べて識別精度が向上する．

3.3.3　参照画像

参照画像として，HOIP 顔画像データベースの一部と独自に収集した画像の計 600 枚を用いた．両目の位置を基準として 19 × 19 画素のサイズとなるように縮小して顔を切り出した．画像の階調数は 256 である．図 3.2(a) は 600 枚の参照画像から生成した平均画像であり，CC,

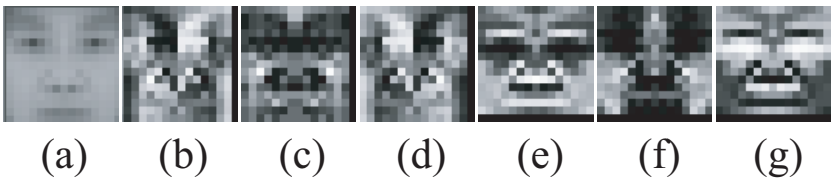

図 3.2. 確率値テーブルの例．(a) 平均画像，(b)〜(g) 確率値テーブル．

ISC および QTR による照合のためのテンプレートとして用いた．

Subspace-1 および Subspace-2 は，600 枚の参照画像を画像平面上で ±10 度ずつ回転させた画像も加え，1,800 枚の画像から顔辞書を作成した．これら顔画像の明度値ベクトル（361 次元）の集合から固有ベクトルを求めた．固有値の大きい方から 30 個の固有ベクトルを選び，それらによって張られる部分空間を顔辞書とした．Subspace-2 については，顔を含まない画像 22,328 枚から同様の方法で非顔辞書を作成した．なお，辞書作成および照合対象となる部分画像に対する前処理として，2 乗誤差が最小となる平面をあてはめて照明勾配を補正した上で，画像内の明度値の平均 μ と分散 σ^2 が $\mu = 128$，$3\sigma = 128$ となるように正規化した．

図 3.2(b)〜(g) は，600 枚の参照画像から作成した Probabilistic ISC-2 の確率値テーブルを画像として表現したものである．確率値が大きい画素ほど明るく表示している．(b),(c),(d) は水平方向の明度比較を行うことにより得られた確率値を表しており，それぞれ符号-1,0,1 の生起確率に対応している．(e),(f),(g) は垂直方向の明度比較により得られた確率値のテーブルである．これらのテーブルは顔の特徴をよく表現していることが分かる．例えば，空間的な明度変化が大きい目や鼻の付近では多くの顔画像から-1 もしくは 1 の符号が観測されるため，(b),(d),(e),(g) において高い確率値が得られている．また，明度が均一な頬のあたりでは符号 0 が観測されるため，(c),(f) において高い確率値が得られている．実際の照合では，この確率値の対数を格納したテーブルを参照した．

さらに詳細に，参照画像から得られた確率値について調べる．図 3.3 には，参照画像内の 2 つの画素 A と B に注目し，右隣の画素との明度差分値のヒストグラム (a)(c) と 3 値符号の生起確率 (b)(d) を示している．まず，グラフ (a)(b) を与えている画素 A に注目する．画素 A は，まゆげのやや上に位置している．この地点は，前髪の長さが異なるなど個人差が大きい．そのため，明度差分値や符号の生起確率に大きな偏りが生じない．明度差分値のヒストグラム (a) は，平均値がほぼ 0 であり，正負の値が均等に観測されている．3 値符号の生起確率を示す棒グラフ (b) も，それを示している．一方，グラフ (c)(d) を与える画素 B は，左目と眉間の境界あたりに位置している．画素 B の右隣の画素は，ほとんどの顔画像において B より明度値が小さい．そのため，明度差分値のヒストグラム (c) は負の値の方向に大きく偏っており，符号-1 を観測する確率は 0.9 を超えている．画素 A ではいずれの符号を観測したとしてもテーブルから参照される値は高々 0.5 ほどであるが，画素 B では符号-1 を観測したとき約 0.9 という値が参照される．すなわち，画素 A よりも B の方が照合に適した特徴であるとして，より大きな重みが与えられており，これが対象に含まれる変動に対して頑健な照合につながる．

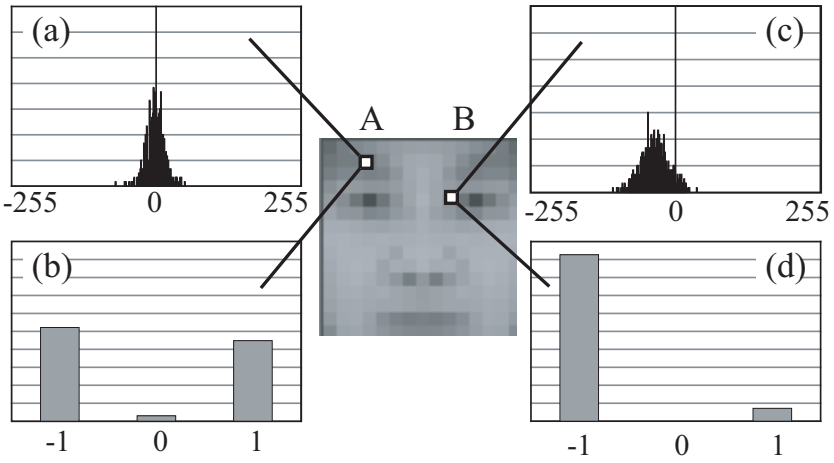

図 3.3. 確率による参照画像間の変動の評価

3.3.4　テストセット

検出精度を測定するため，2 種類のテストセットを準備した．テストセット 1 は，様々な映像からほぼ正面を向いた顔が含まれる場面を集めて作成した．屋外で撮影されたシーンなど複雑な背景を伴う 1,244 枚の画像からなる．テストセット 2 は，様々な年代の男女 295 名の顔画像からなる XM2VTS データベース [44] のうち左右から照明をあてて撮影された 1,176 枚の画像を用いた（図 3.4）．テスト画像はすべて 256 階調の濃淡画像であり，画像内に顔を 1 つだけ含んでいる．これらの画像に対して，手作業で左右の瞳と鼻孔の位置を入力した．さらに，入力位置を基準として，画像内の顔領域の大きさが参照画像と同じ大きさになるように，あらかじめ画像を縮小した．参照画像と異なる大きさの顔を検出するには，様々な尺度で拡大・縮小した複数の画像に対し検出処理を行うのが一般的であるが，これらの画像間でオーバーラップして検出された領域を統合するなどの後処理が必要となる [63]．今回の実験では，事前にテスト画像を縮小することによって，後処理のアルゴリズムに依存することなく照合精度のみを計測した．

3.3.5　検出精度

7 種類の方法によってテストセットから顔を検出し，その検出精度を比較する．しきい値を変化させながら，正検出率（顔総数に対して，正しく検出された顔の数の割合）と過検出率（総照合回数に対して，誤って背景を顔と判定した回数の割合）をプロットしていくことにより図 3.5 および図 3.6 に示す ROC 曲線を得た．照合によって「顔」と判定された領域に，あらかじめ手作業によって入力した瞳と鼻孔が含まれる場合を「正検出」，そうでない場合を「過検出」として計数した．しきい値は，各照合法の相関値がとり得る最小値と最大値の間を 1,000 段階に区切って変化させた．例えば，ISC の相関値は 0 から 1 の間の値をとるので，し

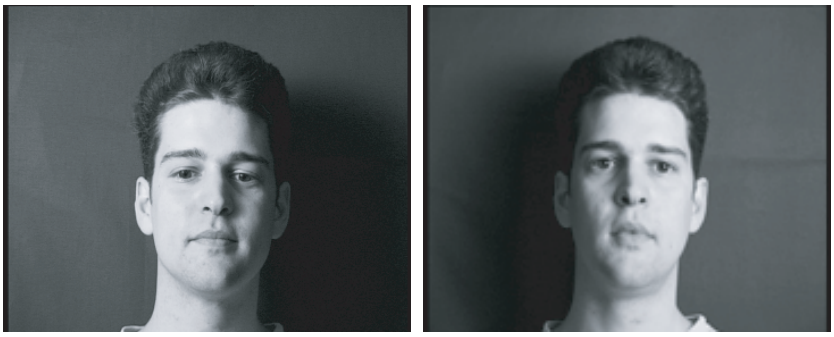

図 3.4. テストセット 2 の画像．XM2VTS データベース [44] のうち左右から照明をあてて撮影した画像.

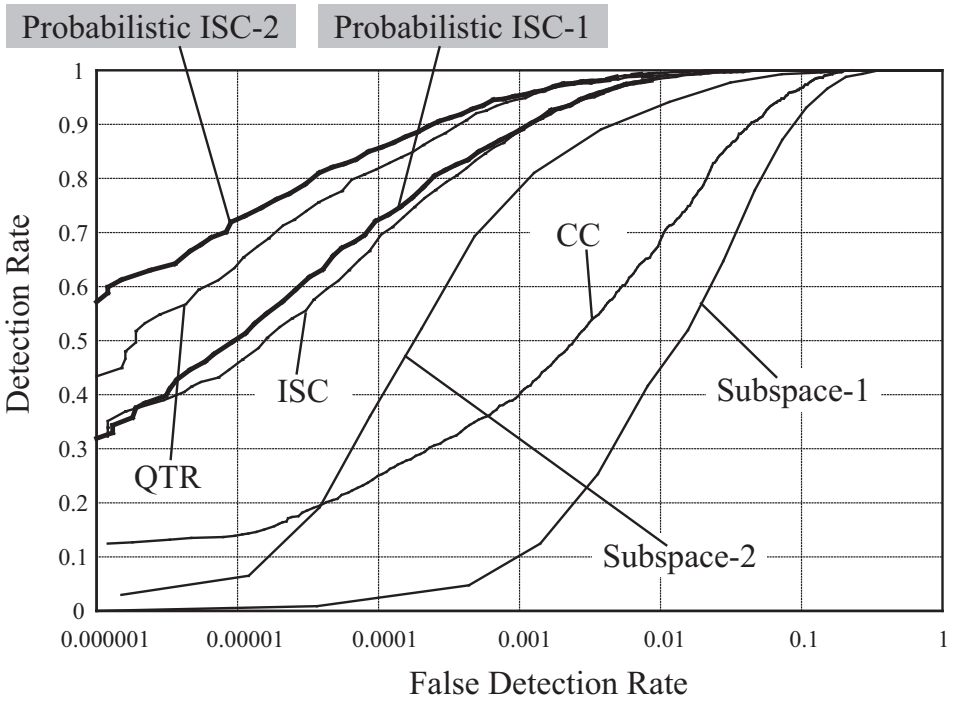

図 3.5. テストセット 1 に対する ROC 曲線

きい値を 0.001 ずつ増加させた．提案手法の相関値は 0 以下の値をとり，相関のないパターンに対しては理論的には無限に小さくなるが，実際には相関値が無限小となることはほとんどないため，十分小さな値から 0 の間でしきい値を変化させた．図 3.5 および図 3.6 では，検出精度が高い方法ほど曲線が左上に描かれている．

まず，図 3.5 に注目する．CC の検出精度は低く，ISC と同じ正検出率を得るようにしきい値を調整したとすると，最大で 100 倍の過検出が発生している．これは，ISC が照明変動やノイズに対して頑健であることを示している．QTR は ISC よりさらに過検出を削減している．これは，3 値符号により隣接画素の明度が等しい場合を評価することの効果を示している．一

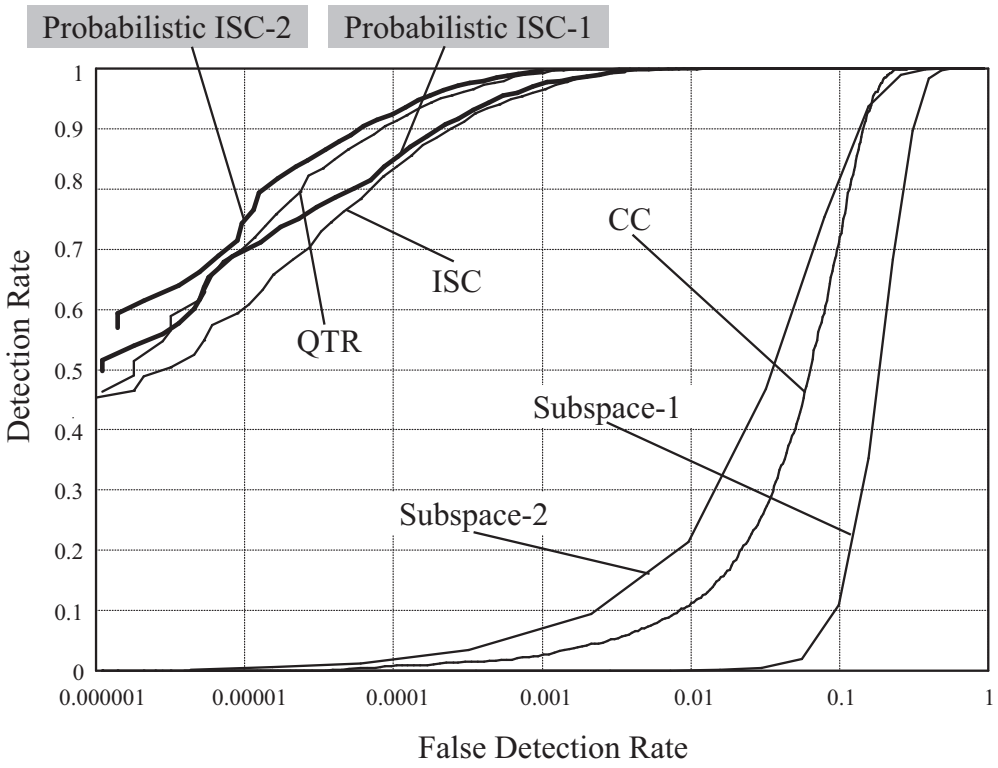

図 3.6. テストセット 2 に対する ROC 曲線

般に，画像には空間的に明度が均一な領域が含まれる．例えば，空や壁といった背景領域であり，テスト画像にもそのような領域が多く存在する．逆に，顔領域では明度が均一な領域は少ない．そのため，同値符号を評価することによって両者の違いが強調され，検出精度が向上する．提案手法は，ISC と QTR に比べてそれぞれ検出精度が向上している．これは，顔の個人差や表情変化に対して頑健であることを示している．また提案手法は，部分空間法に比べても高い検出精度が得られている．部分空間法は，前処理により照明変動の影響を軽減しているとはいえ，明度値自体を特徴量として用いているため，照明変動の影響を大きく受け，高い検出精度が得られなかった．次に，テストセット 2 に対する結果を示す図 3.6 に注目する．各手法の優劣の順位は図 3.5 と同様である．しかし，照明変動に対して頑健な方法と明度値自体を特徴量として用いている方法の間に，より大きな差が現れていることが分かる．提案手法が，照明変動に対しても頑健であることが分かる．

3.3.6　処理時間

各照合法の処理時間を実測した結果を表 3.2 に示す．Intel® Xeon™ 3GHz を搭載した PC 上で，テストセット 1 に対して顔検出処理を行い，すべての画像に対する照合処理の総時間数を計測した．ISC，QTR および提案手法については，入力画像から符号列を得る処理に要する時間も含めて計測している．マルチメディア命令を用いた並列化などは行っていない．提案

表 3.2. 処理時間の比較

手法	照合に要した処理時間 [秒]
CC	24.3
ISC	13.5
QTR	14.3
Subspace-1	1020.0
Subspace-2	2173.3
Probabilistic ISC-1	13.3
Probabilistic ISC-2	14.4

手法は，CC に比べて高速であり，ISC や QTR と同等の処理時間であった．部分空間法と比較すると 1/70〜1/150 の時間で照合を行うことができた．以上のように，実測値からも提案手法の計算コストが小さいことを確認した．

3.3.7　顔の検出例

屋外あるいは強い照明変動下で撮影された画像に対して，相関値が最大となった領域を方形で表示した結果を図 3.7 に示す．顔の一部に強い照明をあてて撮影した画像も含まれている．明度値自体を特徴量としている CC と Subspace-1 は，複雑な明度変化に対応できずに顔以外の領域において相関値が最大となる例が多い．一方，ISC，QTR，Probabilistic ISC-2 については良好な結果が得られており，これらの方法が照明変動に対して頑健であることを確認できる．最下段の画像においては，Probabilistic ISC-2 のみが正しい位置を検出している．

3.3.8　参照画像枚数による識別性能の変動

図 3.8 は，12 種類の異なる顔検出器によって得られた ROC 曲線である．これらの検出器は，参照画像の枚数を変えながら 3 つの方法で学習されたものである．“Subspace” は，部分空間法 [91] に基づく顔検出器である．部分空間を張る固有ベクトルは累積固有値が 0.99 より大きくなるように選択されており，その数は 12 から 25 の間である．“AB” は，Viola と Jones [83] の方法により学習された顔検出器を指し，AdaBoost で rectangle feature を選択することによって構築されたものである．rectangle feature の数は 760 個とした．これは，Probabilistic ISC で評価される特徴（増分符号）の数と同数である．“AB1000” は，それぞれ 1,000 枚の顔画像と非顔画像を用いて学習した顔検出器の ROC 曲線である．参照画像の数を 1,000 から 100 に減少させるにつれて，“AB” と “Subspace” の識別性能が顕著に低下する様子が分かる．一方，Probabilistic ISC は参照画像枚数が 100 枚以上であれば，識別性能をほぼ維持している．すなわち，提案手法は参照画像（学習サンプル）が少数の場合では先行研究より優れた識別性能が得られ，多くの学習サンプルを収集する手間を小さくできる利点があ

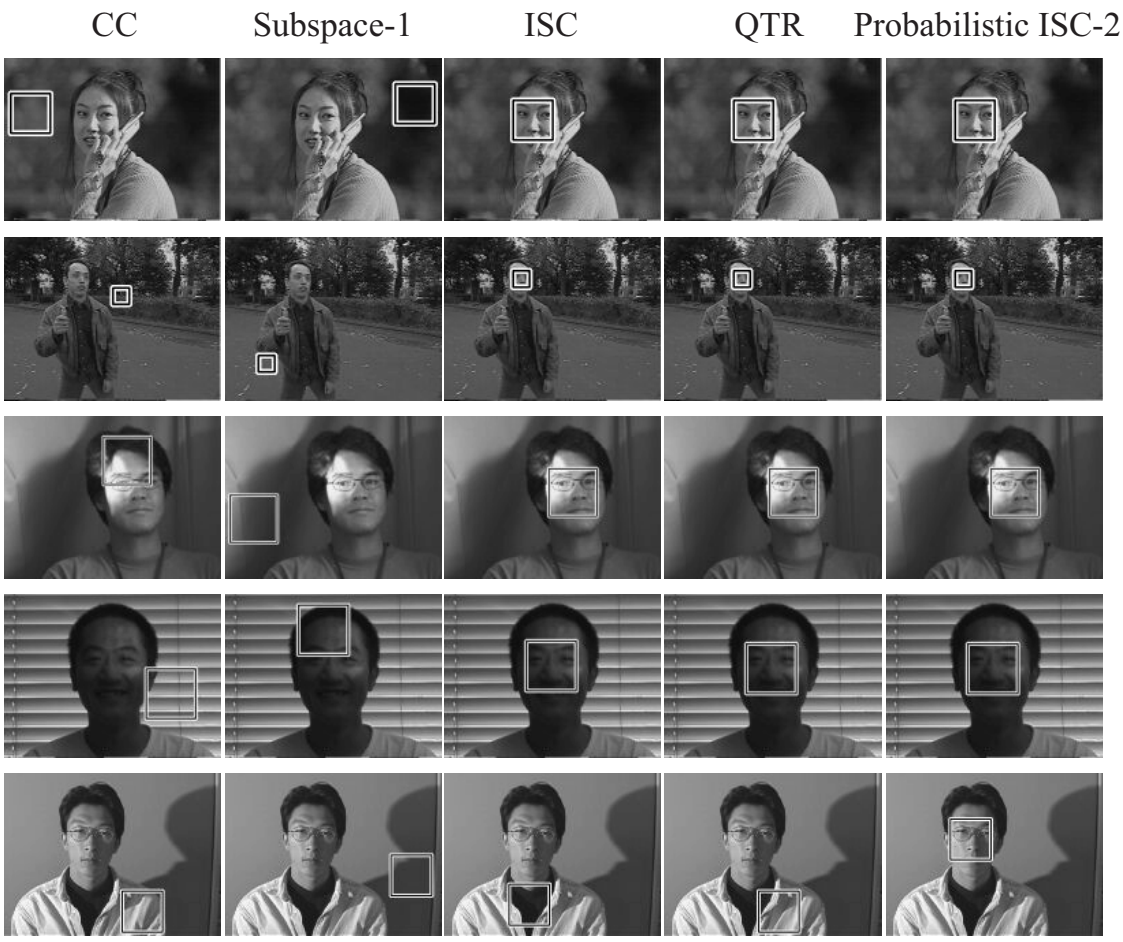

図 3.7. 顔の検出例

る．また，提案手法は学習に要する計算時間が短いことも特長である．例えば，100 枚の参照画像を使って Probabilistic ISC のルックアップテーブルを作成する計算時間は，0.0036 秒と極めて短かったが，“Subspace100” や “AB100” は，それぞれ 0.078 秒と 8 分を要した．

3.3.9　顔向き推定への応用

ここでは，Probabilistic ISC の顔向き推定への応用を検討する．図 3.9 に，照明変動と部分的な遮蔽を伴うテスト画像を使った実験の概要を示す．実験では，9 つのカテゴリを識別する．具体的には，7 種類の面内回転と 2 種類の面外回転を加えた顔画像を用意した．面内回転画像は 0° と表記された正面向きの顔画像を 30° ずつ回転させることにより作成した．テスト画像は，ランダムパターンによって部分的に遮蔽されている．遮蔽領域の面積比率は 20% である．テスト画像は合計 9,401 枚である．表 3.3 に，誤識別数と識別率を示す．提案手法では ISC と同等の計算時間で 97.5% のテスト画像を正しく識別でき，最も識別率が高かった．Subspace はその 13 倍の計算時間を要したが，86.9% の識別率であった．なお，標準的な AdaBoost アルゴリズムは 9 カテゴリの識別に適用できない．

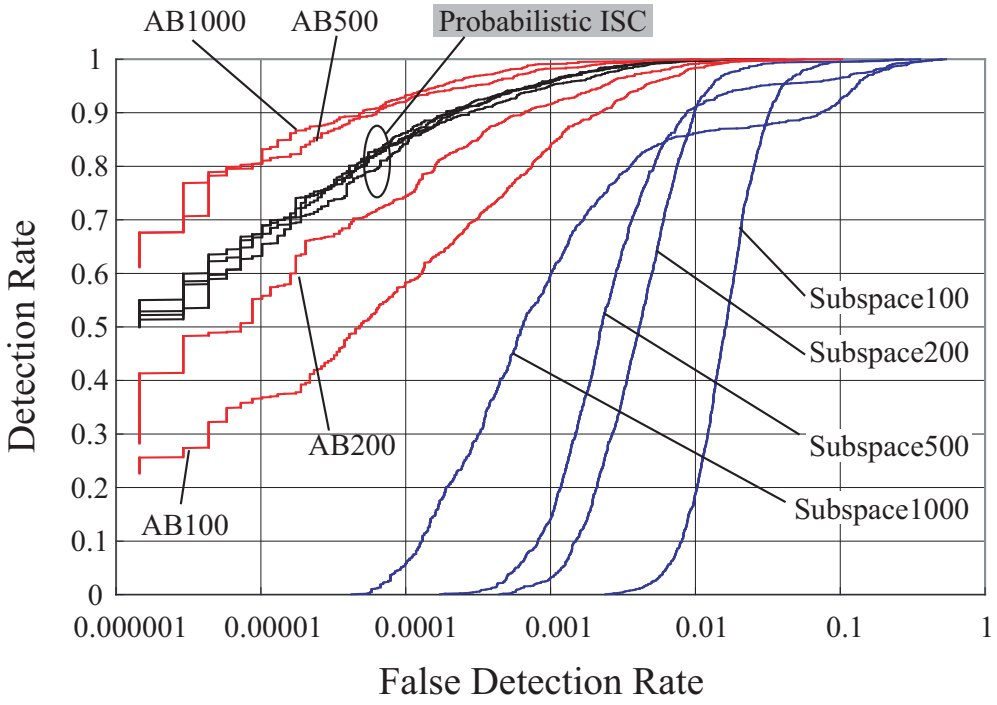

図 3.8. 参照画像枚数による ROC 曲線の変動

表 3.3. 顔向き推定の精度

	誤識別数	識別率 (%)
CC	3,531	62.4
Subspace	1,230	86.9
ISC	459	95.1
Probabilistic ISC	237	97.5

3.3.10　考察

実験結果から，提案手法が既存の相関に基づく手法や部分空間法に比べ，高精度に顔を検出できることが分かった．また，これを相関に基づく手法と同等，部分空間法に比べて格段に小さい計算コストで実現できた．

Probabilistic ISC は顔検出を直接の目的とした手法ではないが，代表的な顔検出手法との違いについても説明する．画像から顔を検出する方法としては，古くは固有顔と呼ばれる上記 Subspace-1 と同様の原理に基づく方法 [76] が有名である．しかし，顔によく似た背景領域を誤って検出しやすいという問題があるため，2 段階の処理によって検出精度を向上させる方法が提案されている [103, 104]．この方法では，固有顔によって顔が存在する候補領域を限定した後，固有顔では区別できない背景画像と顔を判別分析によって識別する．また，

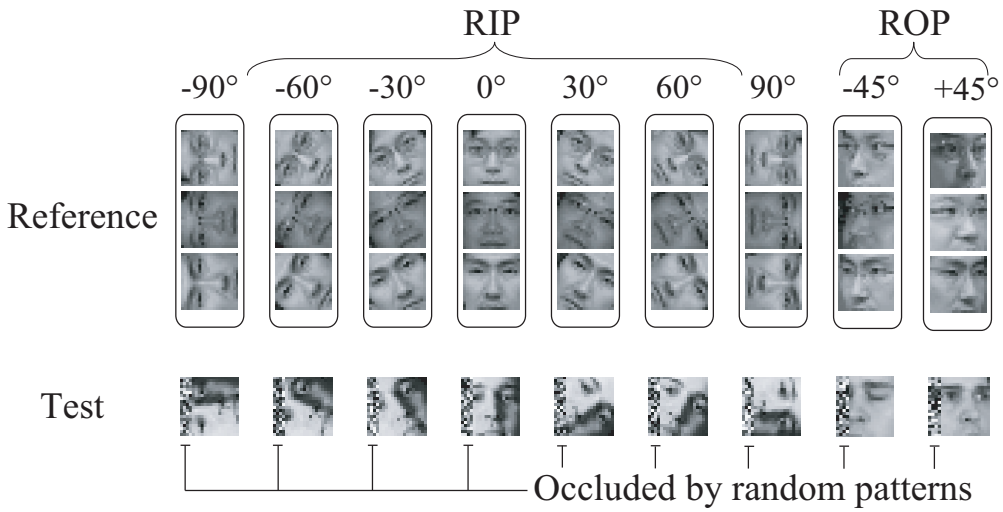

図 3.9. Probabilistic ISC による顔向き推定実験．テスト画像は，照明変動と部分的な遮蔽を含む．

ニューラルネットワークや SVM，AdaBoost といった学習方式を用いたものが提案されている [63, 55, 83]．これらは，顔と顔以外（非顔）の 2 カテゴリのサンプル画像から両者を識別する境界を学習しておき，入力画像がどちらのカテゴリに近いかを比較して識別を行う方法である．一方，Probabilistic ISC による顔検出では，固有顔と同様に入力画像がどれぐらい顔に似ているかのみに基づいて顔か否かを判定する．つまり，代表的な顔検出手法と提案手法とでは，検出対象と区別すべき対象を与えるかどうかにおいて基本的な条件設定が異なる [92]．非顔のサンプル画像を多数収集できる場合は，2 カテゴリ識別に基づく方法が有利である．ただし，これらは数千から数万枚のサンプル画像を用い，学習された識別器が正しく識別できなかった非顔画像を収集して再学習する枠組みであるため，一定の検出精度が得られるまでには手間と時間を要する．実験で示したとおり，提案手法は数百枚程度と比較的少数の顔のサンプル画像のみを収集すれば，数秒程度で確率テーブルを得ることができるため，実現が容易である．例えば，顔画像照合によって入退室を管理するシステムなど顔が最大で 1 つだけ含まれるような画像を対象とする場合など，複数個の顔を検出するのに比べて問題設定を簡単化できる用途に適している．このような用途では，画像中で相関値が最大となる位置を求め，照合しきい値と比較すればよい．このとき，3.1.4 の方法によって，過検出に対する損失を高く設定して照合しきい値を求めておけば，画像中に顔が写っていない場合にも誤って背景を検出することがなくなる．

また，顔向き推定の実験結果から，多数カテゴリの識別に拡張しても優れた識別性能を示すことを確認できた．

3.4　まとめ

増分符号の確率分布をテンプレートとするアイデアによって，外乱への頑健性，カテゴリ内変動への対応，省処理化という 3 つの課題を大きく改善できた．また，Probabilistic ISC を拡張し，他のカテゴリとの識別に有効な特徴を重み付けする方法を与えた．これにより，類似カテゴリとの識別性能改善につながる．

Probabilistic ISC はこれらの技術課題をバランスよく改善する優れた性質を持つため，顔以外に目や鼻といった顔部品，あるいは臓器等の医用画像へも適用可能である．応用としては，監視システムや映像中の人物認識によるインデクシング [65] などがある．しかしながら，特定カテゴリに属するオブジェクトを検出するという目的においては，実験で示したように上述の方法 [63, 55, 83] の識別性能に及ばない．また，一般物体認識やサブカテゴリ識別のような大きなカテゴリ内変動とさらに類似したカテゴリを含むような課題には対処し難い．以降では，「共起性」の導入というアイデアによって，新たなオブジェクト検出の枠組みを導く．

第 4 章

局所特徴量の共起性を利用したオブジェクト検出手法

ここでは，局所特徴量の共起性を利用したオブジェクト検出手法について述べる．具体的には，Boosting による弱識別器の学習において，複数の特徴の共起性を利用することで，計算コストを増加させずに高い識別精度が得られるオブジェクト検出のフレームワークを提案する．このフレームワークでは，Sequential Forward Selection という貪欲な探索方法によって識別に適した共起特徴が自動選択され，その共起特徴に基づく弱識別器の線形結合によって最終的な識別器を求める．

図 4.1 において，rectangle feature を用いた顔検出器を学習するための 3 種類の特徴選択方法を比較する．(a) は Viola と Jones による Boosting のみによる特徴選択方法，(b) は Boosting は用いず共起関係のみを探索する特徴選択方法，(c) は提案手法である．(c) では，最終的に得られる強識別器 $H(x)$ は，弱識別器 $h_1(x)$ から $h_T(x)$ の線形結合である．(a) とは異なり，それぞれの弱識別器は複数の特徴を使用する．例えば，$h_1(x)$ は F 個の特徴を同時に観測し，それぞれの特徴の同時統計量 (joint statistics) を評価する．単一の特徴では捉えることができない顔の構造的な類似性，例えば，目は周辺より明度が低い ($z_{1,1}$)，鼻孔は周辺より明度が低い ($z_{1,f}$)，さらに眉間は目より明度が高いといった情報を統計的に評価することが可能となる．F 個の特徴の組み合わせは，それぞれの弱識別器の学習段階において自動的に求められ，図 2.1 に例示した拡張した特徴セットを事前に用意する方法とは異なる．そのため，図 4.1 のように空間的に離れた特徴の組み合わせを得ることができる．別の視点で考えると，(c) の提案手法は，従来手法 (a) と (b) の一般化と位置づけることができる．すなわち，(a) や (b) は提案手法の特殊な場合と解釈できる．提案するフレームワークは，識別に使用する特徴の総数を変えずに，すなわち計算コストを増やさずに，より高精度な識別器を構築するには，どのような特徴を選択すればよいかという問いに対する 1 つの解である．

以降では，まず複数の特徴の共起関係を表現する方法を示す．次に，Boosting による弱識別器の学習時に Sequential Forward Selection によって共起特徴を選択するアルゴリズムを示す．最後に，顔と 3 種類の形状の手を検出する実験において提案手法の有効性を証明する．10-fold クロスバリデーションによる詳細な性能評価を実施する．また，通常の AdaBoost [24]

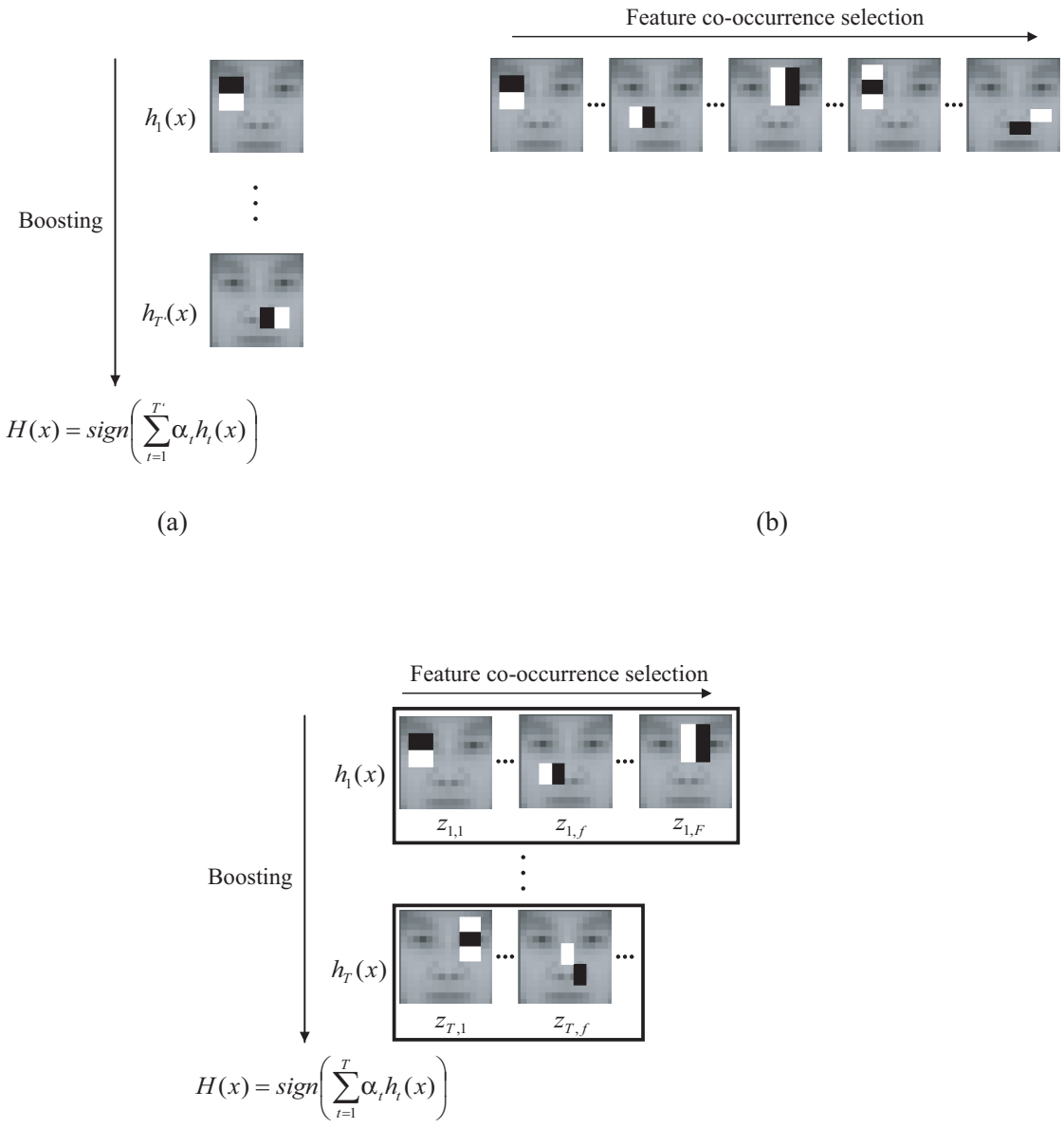

図 4.1. オブジェクト検出器を学習するための 3 種類の特徴選択方法：(a)Viola と Jones の方法，(b)Boosting を用いずに共起関係のみを探索する方法，(c) 提案手法．(c) は (a) と (b) の一般化であり，(a) と (b) は提案手法 (c) の特殊な場合と解釈できる．

の代わりに，Real AdaBoost [66] を導入することによって，さらなる性能向上が可能となることも示す．

4.1　特徴の共起表現

ここでは，複数の rectangle feature の共起関係を表現する方法を説明する．rectangle feature を用いる理由は，特徴量を解像度や場所に依存せず一定の時間で計算できる利点があるからである．

4.1.1　Rectangle feature

rectangle feature の特徴量は，近接する方形領域の平均明度の差分値として求められるスカラー量である．絶対的な明度には依存せず，局所的な明度勾配を抽出する．方形領域には複数の画素が含まれるため，それらの平均値を用いることによって，撮像や伝送の段階でランダムに付加されるノイズの影響を低減できる．

識別器の基準解像度，すなわち学習サンプルの画像サイズに基づいて，方形の位置や大きさ，方形同士の配置を網羅的に変化させることによって，異なった位置や空間周波数，あるいは方向に対応した明度勾配を捉える特徴を生成できる．例えば，画像サイズが 25×25 画素であるとき，図 2.1 の特徴セット (a)〜(d) からは，合計 239,408 通りの特徴が生成される．それぞれの特徴に対して，すべての学習サンプルから特徴量を算出すると，特徴ごとに一次元の確率密度分布が得られる．図 4.2 に示す例は，対象カテゴリと非対象カテゴリから得られた特徴量の分布である．[83] では，これら二つの分布を最も小さい誤り率で分離するしきい値を選択し，弱識別器を構成している．これを，AdaBoost アルゴリズムによって重み更新されたサンプル集合に対して繰り返し，識別に有効な少数の特徴を選択する．しかし，上述したように，学習の後半では異なるカテゴリに属しながらも互いによく似たサンプルの重みが増加するため，弱識別器の誤り率は大きくなってしまう．たとえ 239,408 通りの候補のうち最良の特徴であっても，良好な識別性能が得られなくなる．図 4.3 に，筆者が実装した Viola と Jones の方法に基づく顔検出器の識別性能を示す．弱識別器の数，すなわち識別に使用する特徴の数に対して，訓練誤差（学習サンプルに対する識別器の識別誤り率）と汎化誤差（テストサンプルに対する識別器の誤り率）をプロットしている．訓練誤差は，特徴の数が 500 程度でゼロに収束している．訓練誤差がゼロとなった後も汎化誤差が徐々に減少しているが，これは AdaBoost が識別境界とのマージンを最大化するアルゴリズムだからである．しかし，特徴数が 1,000 を超えた後は 3,000 個まで追加しても，汎化誤差はほとんど減少していない．これは，識別に有効な特徴を使い果たしたことを意味しており，今後学習を継続しても大幅に識別精度を向上させる可能性は期待できない．Wu ら [88] は，特徴量を複数のしきい値で区切り，64 の区間ごとに識別を行う方法を提案しているが，分布の重なりが大きい場合は，しきい値処理と同様の問題に直面する．

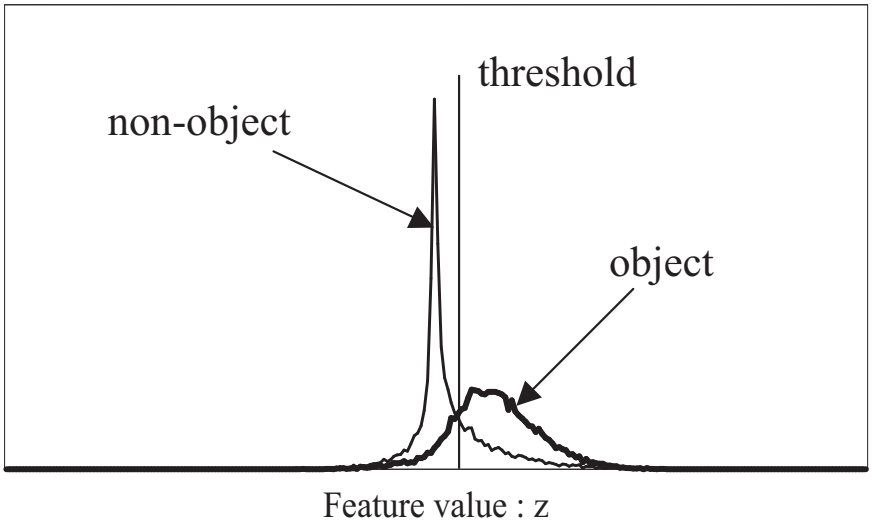

図 4.2. rectangle feature の特徴量の確率密度分布．[83], では，対象カテゴリと非対象カテゴリの分布を最も小さい誤り率で分離するしきい値を選ぶことによって，弱識別器を学習する．

4.1.2　特徴量の量子化

提案手法では，汎化誤差を減少させるために，各弱識別器において複数の特徴を同時に観測する．すなわち，特徴の共起性を利用して，単一の特徴では識別が困難なサンプルに対する識別精度を向上させる．特徴間の共起性は，それらの同時確率によって表現する．ここでは，それぞれの特徴量 z を 2 値化し，対象カテゴリと非対象カテゴリに対応する 1 もしくは 0 の 2 進符号によって表現し，同時確率を求める．符号 s はサンプル x から，

$$s(x) = \begin{cases} 1 & p \cdot z(x) > p \cdot \theta \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

によって算出される．ここで，θ はしきい値，p は特徴量 $z(x)$ としきい値 θ を比較する不等号の向きを決定する変数で，$+1$ もしくは -1 のいずれかの値をとる．θ と p は，学習サンプルに対する識別誤り率が最小となるように決定する．この 2 値化方法は，Viola と Jones の弱識別器と同じである．本論文では，特徴間の共起性を活用することの有効性を示すため，複数の特徴を組み合わせること以外は Viola と Jones の方法と同一の処理としている．2 値化ではなく多段階に量子化すると，より複雑な確率分布に適応する可能性はあり，何段階に量子化するべきかという新たな問題を検討する必要がある．提案手法は 2 値化に限定されるものではないが，ここでは多値化は扱わない．

なお，2 値化の利点は，ノイズや照明変動に対して一定の頑健性が得られることである．例えば，符号 s は式 (4.1) の不等号が逆転しない大きさの明度変化には影響されず，不変である．

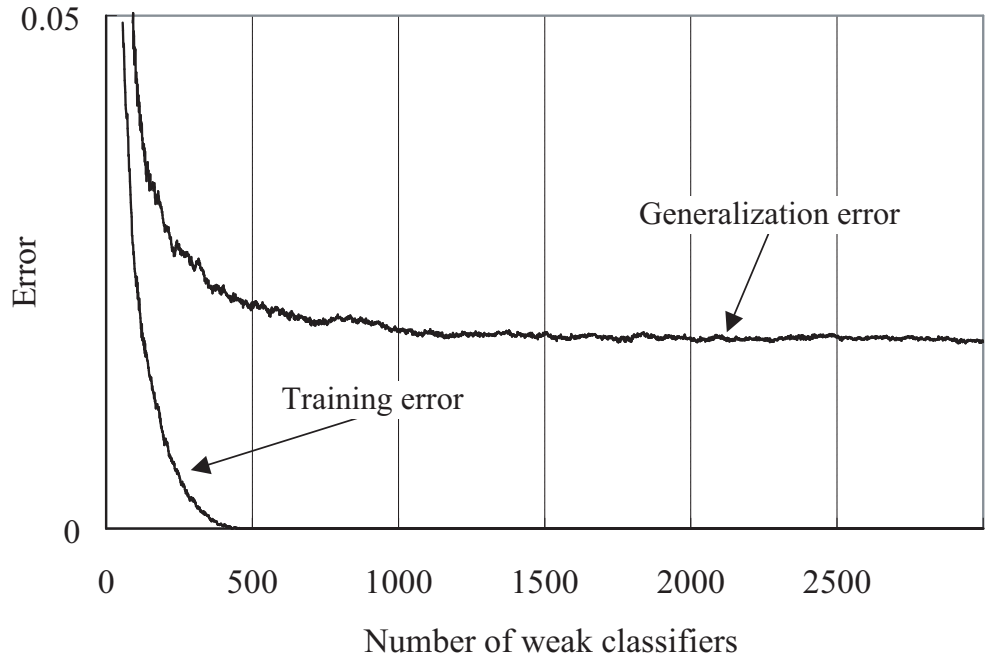

図 4.3. Viola と Jones の方法により学習された顔検出器の識別性能．Training error（訓練誤差）がゼロに収束するのに対して，Generalization error（汎化誤差）は特徴数が 1,000 を超えた後はほとんど減少していない．これは，識別に有効な特徴を使い果たしたことを意味しており，今後学習を継続しても大幅な識別精度の向上は期待できない．弱識別器が単一の特徴を使用することを前提とした方法は，いかに優秀な学習アルゴリズムを使っても同様の問題に直面する．

4.1.3　複数の特徴の共起表現

特徴間の共起性は，複数の特徴から算出した 2 値符号の組み合わせによって表現する．図 4.4 に，3 つの rectangle feature の共起性を表す具体例を示す．あるサンプル x に対して，3 つの rectangle feature からそれぞれ 1，0，1 という符号が観測されたとき，

$$J(x) = (101)_2 = 5 \tag{4.2}$$

とする．$J(x)$ は 2 進表現された特徴の組み合わせのインデックス番号であり，組み合わせる特徴の数を F とすると，2^F 通りの値をとり得る．

あるカテゴリに属するそれぞれのサンプル x_i について $J(x_i)$ を観測することによって，特徴間の統計的な依存関係を知ることができる．ある入力パターン x から得られた特徴の組み合わせ $J(x)$ が，対象カテゴリと非対象カテゴリのいずれから観測されやすいかを評価することにより，x が検出対象か否かを識別することができる．複数の特徴の組み合わせは，非対象カテゴリからは観測されないような対象カテゴリの構造的な類似性を捉えるように選択するのが望ましい．以降では，識別に有効な特徴の配置や組み合わせを自動選択する方法を示す．

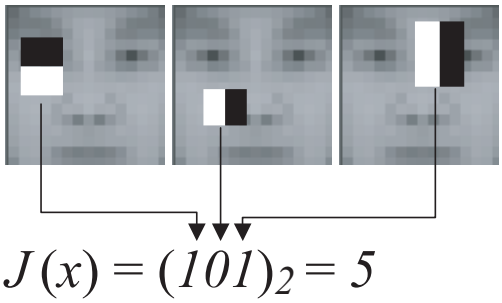

図 4.4. 特徴の共起表現．3 つの rectangle feature から観測された 2 進符号の組み合わせによって表現される．

4.2　Sequential Forward Selection と Boosting を用いた共起特徴の自動選択アルゴリズム

ここでは，共起特徴を自動選択することによってオブジェクト検出のための識別器を学習するアルゴリズムについて説明する．まず，複数の特徴の共起に基づく弱識別器を定義する．次に，Boosting によって弱識別器を逐次的に学習する学習手順を示す．その過程において，識別に有効な共起特徴がいかに選択されるかを説明する．以下では，標準的な AdaBoostt [24] と Real AdaBoost [66] の 2 種類の Boosting アルゴリズムを導入した場合の定式化を与えるが，提案手法は LogitBoost [25] など他の Boosting アルゴリズムも利用可能な汎用的なフレームワークとなっている．以下では，標準的な AdaBoost は RAB (Real AdaBoost) と区別するために，DAB (Discrete AdaBoost) と表記する．

4.2.1　弱識別器

共起特徴に基づく弱識別器を定義する．

まず，DAB (Discrete AdaBoost) に向けた弱識別器を定式化する．弱識別器 $h_t(x)$ において，サンプル画像 x から共起特徴を観測する処理を関数 $J_t(x)$ で表す．x から特徴量 $J_t(x) = j$ を観測したとき，$h_t(x)$ を以下の条件付き確率に基づくベイズルールによって，

$$h_t(x) = \begin{cases} +1 & P_t(y=+1|j) > P_t(y=-1|j) \\ -1 & \text{otherwise} \end{cases} \tag{4.3}$$

と表す．ここで，$y \in \{+1, -1\}$ はカテゴリラベル，$P_t(y=+1|j)$ と $P_t(y=-1|j)$ は共起特徴量 j を対象（正）カテゴリと非対象（負）カテゴリから観測する条件付き確率である．これらは，複数の特徴の組み合わせを観測する同時確率であり，次式のように学習サンプルの重み分布 D_t に基づいて算出される．

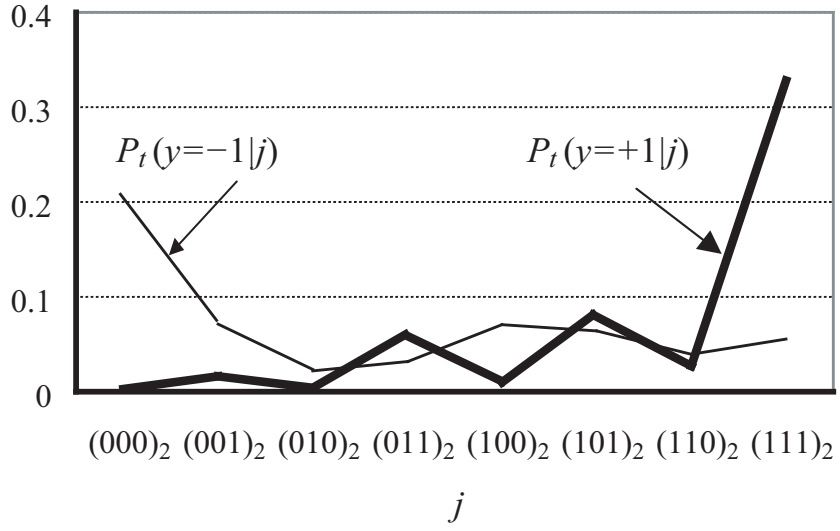

図 4.5. 条件付き確率に基づく弱識別器．$P_t(y=+1|j)$ と $P_t(y=-1|j)$ は，3 つの rectangle feature から得られる．3 つの特徴から算出された 2 進符号は 8 通りの値をとる．$j=(011)_2$ や $(101)_2$ あるいは $(111)_2$ のとき，入力画像は検出対象であると識別される．

$$P_t(y=+1|j)=\sum_{i:J_t(x_i)=j\wedge y_i=+1} D_t(i) \tag{4.4}$$

$$P_t(y=-1|j)=\sum_{i:J_t(x_i)=j\wedge y_i=-1} D_t(i) \tag{4.5}$$

ここで，x_i および $y_i\in\{+1,-1\}$ は，i 番目の学習サンプルとそのカテゴリラベルである．また，$D_t(i)$ はサンプル x_i の重みである．これらの変数の詳細は次節で説明する．

図 4.5 に，3 つの特徴を用いた場合に実測した $P_t(y=+1|j)$ と $P_t(y=-1|j)$ の分布を示す．3 つの特徴の組み合わせなので，$(000)_2$ から $(111)_2$ の 8 通りの値をとる．例えば，入力画像から $j=(011)_2=3$ や $(101)_2=5$，あるいは $(111)_2=7$ が観測されたとすると，入力画像は検出対象であると識別される．その他の特徴量では，逆に検出対象ではないと判定される．

RAB (Real AdaBoost) に向けた弱識別器の定義は以下のとおりである．

$$h_t(x)=\frac{1}{2}\log\frac{P(y=+1|j)}{P(y=-1|j)} \tag{4.6}$$

しかし，この定義式では，分母の $P_t(y=-1|j)$ が極めて小さい値もしくはゼロとなる場合が考えられ，そのとき $h_t(x)$ は無限大となる．これを避けるため，[66] に示されている平滑化手法を用い，以下のように定義式をあらためる．

$$h_t(x)=\frac{1}{2}\log\frac{P(y=+1|j)+\nu}{P(y=-1|j)+\nu} \tag{4.7}$$

ここで，ν は小さな正の数である．以降では，$\nu = 1/N$ とする．N は学習サンプルの画像枚数である．

RAB における弱識別器は，式 (4.3) に示した DAB の弱識別器とは異なり，カテゴリラベルではなく識別の信頼度を返す．信頼度は 2 つの確率値の比によって算出され，比の値が 1 に近いとき，すなわち両カテゴリから j が同様に観測されるとき，信頼度はゼロに近い値をとる．逆に，確率値が大きく異なるとき，すなわち共起特徴が識別に有効である場合，信頼度はゼロからかけ離れた正か負の値をとる．DAB の弱識別器は，2 つの確率値のいずれが大きいかのみを評価するのに対し，RAB は確率値の違いがどの程度かも評価するため，より高い識別性能が得られる．4.3.5 節で実施する実験によって，これを確かめる．

4.2.2　DAB (Discrete AdaBoost) の学習手順

DAB によって特徴選択を行う手順を図 4.6 に示す．N 枚の教示された学習サンプル $(x_1, y_1), \ldots, (x_N, y_N)$ が用意されているとする．ここで $y_i \in \{+1, -1\}$ は，サンプル x_i のカテゴリラベルである．$D_t(i)$ は，x_i の重みである．重みは，$D_1(i) = 1/N$ と初期化する．最終的に得られる強識別器 $H(x)$ は，T 個の弱識別器 $h_t(x)$ の線形結合として，

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{4.8}$$

により構成される．ここで，α_t は重みつきサンプル分布 D_t における弱識別器 $h_t(x)$ の誤り率から算出される弱識別器の信頼度である．Boosting 学習の各段階では，図 4.6 のステップ (A)〜(E) によって，最も識別に適した共起特徴が選択される．

4.2.3　RAB (Real AdaBoost) の学習手順

図 4.7 に RAB によって共起特徴を選択する手順を示す．上述の DAB との違いは，ステップ (D) において弱識別器を選択する際の規準である．DAB が弱識別器の識別誤り率最小化を規準としていたのに対し，RAB で用いられる規準は，重みつきサンプル分布 D_t における Bhattacharyya 限界 Z_t である．もう一つの違いは，式 (4.3) および式 (4.7) に示す弱識別器の定義によるものである．RAB における弱識別器は，識別の信頼度を求めるため，DAB で用いられた α_t は最終識別器の定義式から省略されている．

4.2.4　特徴の組み合わせ探索

弱識別器を構築するには，識別に有効な特徴の組み合わせを探索する必要がある．基本的には，あらゆる特徴の組み合わせの候補を全探索することによって，最良の共起特徴を選択すればよい．しかし，組み合わせる特徴の数を増やすと，共起特徴の候補の数は指数的に増加するため，全探索には多大な計算コストがかかる．通常は，現実的な計算時間内に実行できなくなってしまう．M 個の特徴から F 個を選んで組み合わせる場合，探索の計算コスト

1. Prepare a set of N labeled samples as $(x_1, y_1), \cdots, (x_N, y_N)$.
   $y_i \in \{+1, -1\}$ is the class label associated with the sample image x_i.
2. Initialize weights $D_1(i) = \dfrac{1}{N}$.
3. For $t = 1, ..., T$:
   (A) For each feature, calculate a feature value.
   (B) Binarize each feature value and assign a binary variable
   according to Eq.(4.1).
   (C) Train a weak classifier based on a
   combination of features.
   (D) Choose $h_t(x)$ with the lowest error ϵ_t.
   The error is evaluated with respect to the sample weight $D_t(i)$,
   $\epsilon_t = \sum_{i:y_i \neq h_t(x_i)} D_t(i)$.
   (E) Update the weights:
   $D_{t+1}(i) = \dfrac{D_t(i) \exp\left(-y_i \alpha_t h_t(x_i)\right)}{\sum_i D_t(i) \exp\left(-y_i \alpha_t h_t(x_i)\right)}$,
   where $\alpha_t = \dfrac{1}{2} \log\left(\dfrac{1 - \epsilon_t}{\epsilon_t}\right)$.
   End For
4. Output the final strong classifier:
   $H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.

図 4.6. DAB による学習手順

は $O(M^F)$ である．Branch-and-bound アルゴリズム [50] は，最適解を効率よく探索できるアルゴリズムとして知られている．しかし，候補数が大きい場合，ワーストケースでは全探索と同じになる．さらに，特徴選択の規準関数が単調であることが必要であるが，多くの場合これは成り立たない条件である．最適解は保障できないが，より効率的な特徴選択アルゴリズムがいくつかある [34]．最もよく知られているのは，Sequential Forward Selection (SFS) あるいは Sequential Backward Selection (SBS) である．SFS は，貪欲な探索方法の一種で，最も良い単一の特徴を探索するところから始め，その特徴に他の特徴を順次組み合わせていくやり方である．反対に，SBS はすべての特徴を組み合わせておき，順に 1 つずつ特徴を除いていく方法である．いずれも事前に定義された規準を評価しながら，特徴の追加あるいは削除を行う．Plus-l-Minus-r 法 [72] は，SFS と SBS を組み合わせたアルゴリズムで，l 個の特徴を SFS を使って加え，その後 r 個の特徴を SBS で取り除くという操作を繰り返す．Sequential Forward Floating Selection (SFFS) や Sequential Backward Floating Selection (SBFS) [60] は，Plus-l-Minus-r 法の一般化であり，l や r の値を自動決定する仕組みを持つ

1. Prepare a set of N labeled samples as $(x_1, y_1), \cdots, (x_N, y_N)$.
   $y_i \in \{+1, -1\}$ is the class label associated with the sample image x_i.
2. Initialize weights $D_1(i) = \dfrac{1}{N}$.
3. For $t = 1, ..., T$:
   (A) For each feature, calculate a feature value.
   (B) Binarize each feature value and assign a binary variable
   according to Eq.(4.1).
   (C) Train a weak classifier based on a
   combination of features.
   (D) Choose $h_t(x)$ with the lowest Bhattacharyya bound Z_t:
   $Z_t = 2\sum_j \sqrt{P(y=+1|j)P(y=-1|j)}$.
   $P(y=+1|j)$ and $P(y=-1|j)$ are calculated by Eq.(4.4) and (4.5),
   respectively.
   (E) Update the weights:
   $D_{t+1}(i) = \dfrac{D_t(i)\exp\left(-y_i h_t(x_i)\right)}{\sum_i D_t(i)\exp\left(-y_i h_t(x_i)\right)}$.
   End For
4. Output the final strong classifier:
   $H(x) = sign\left(\sum_{t=1}^{T} h_t(x)\right)$.

図 4.7. RAB による学習手順

探索法である．Pudil ら [60] は，逐次的な探索法を比較し，SFFS が最も良い性能を示したことを報告している．本論文では，実装が容易なことから SFS を使用する．具体的には，SFS によって特徴を 1 つずつ追加しながら，より識別精度が高まる組み合わせを選択する．識別精度は，DAB では ϵ_t，RAB では Z_t により評価する．探索のための計算コストは $O(FM)$ と大幅に小さくなる．図 4.8 に，SFS による弱識別器の学習手順を示す．

組み合わせる特徴の数 F をどのように決定するかも重要な問題である．多くの特徴を組み合わせるほど複雑な確率分布を表現できるが，学習サンプルに過剰適合する可能性が高まる．また，特徴を 1 個追加すると共起特徴量 j がとり得る値の範囲が 2 倍ずつ増加するため，その増加に見合った十分な数の学習サンプルがなければ確率密度の推定における統計的な信頼性が低下してしまう．これを避けるため，F の上限値 F^{max} を，

$$2^{F^{max}} \times 10 < N \tag{4.9}$$

と設定し，すべての j が均等に観測されたとしても，少なくとも 10 個のサンプルが割り当てられるようにする．ここで，N は学習サンプルの数である．

1. Initialize a feature subset: $S_0 = \emptyset$.
2. For $i = 1, ..., F$:
   For $j = 1, ..., NumOfFeatureCandidates$:
   (A) Generate a weak classifier by combining S_{i-1} with one feature f_j.
   (B) Evaluate accuracy of the weak classifier by a criterion function $G(S_{i-1} \cup f_j)$. G is evaluated by ϵ_t for DAB or Z_t for RAB.
   End For
   Select the best feature f^*:
   $f^* = \arg\min_j G(S_{i-1} \cup f_j)$.
   Add it to the subset:
   $S_i = S_{i-1} \cup f^*$.
   End For
3. Output the weak classifier with the subset S_F.

図 4.8. SFS による弱識別器の学習手順．F 個の特徴が選択される．

F の決定方法として，以下の 2 通りを考える．
(1) F の値を数種類設定して複数の強識別器を学習し，識別精度が最も高いものを選択する．すべての弱識別器で同じ F 個の特徴を組み合わせる．それぞれの強識別器の学習に必要な計算コストは Viola と Jones の方法と同等となるため，複数の強識別器を学習するコストの合計は，何種類の F を調べるかによって，線形に増加する．
(2) それぞれの弱識別器 h_t において F_t を自動決定する．これには，Leave-one-out 法や Bootstrap 法などを使用することができるが，学習とテストを多数回繰り返す必要があり学習時間が増大するので，ここでは学習サンプルとは別の検証用サンプルを用いる hold-out 法を用いる．各弱識別器の学習段階で，F_t を 1 から F^{max} まで増やしながら，次式で示す損失 $L_{T'}$ を検証用サンプルを用いて算出する．$L_{T'}$ が最小となるような F_t を選択する．

$$F_t^* = \arg\min_{F_t} L_{T'} \tag{4.10}$$

最終的な強識別器の学習コストは，Viola と Jones の方法の F^{max} 倍必要となる．

学習に要する計算コストを低減する方策として，F^{max} まで特徴の数を増やす過程において，損失 $L_{T'}$ が減少から増加に転じた段階で学習をやめる方法がある．この場合，$F_t + 1$ 個の特徴を評価する必要がある．学習のための計算コストは Viola と Jones の方法の高々 2 倍である．ワーストケースは，すべての弱識別器において，2 個の特徴の組み合わせまで評価するが，結果として 1 個の特徴が選択される場合である．

$L_{T'}$ は，検証用サンプルにおける指数損失として定義する．識別誤り率ではなく指数損失を用いる理由は，識別境界とのマージンが考慮されるからである．N' 個のカテゴリラベルつき

検証用サンプル (x'_i, y'_i) を用いて，

$$L_{T'} = \frac{1}{N'} \sum_{i=1}^{N'} \exp\left(-y'_i H_{T'}(x'_i)\right) \tag{4.11}$$

と定義する．ここで，$H_{T'}(x)$ は $t = T'$ における強識別器であり，次式で求められる．

$$H_{T'}(x) = sign\left(\sum_{t=1}^{T'} \alpha_t h_t(x)\right) \tag{4.12}$$

ここで，$h_{T'}$ は学習された直後の弱識別器である．

以上の手続きによって求められた特徴の組み合わせにおいて，最初に選択された特徴は Viola と Jones の方法で選ばれる特徴と同一である．上述のワーストケース，すなわちすべての弱識別器が単一の特徴を評価する場合となる以外，その他の特徴は異なる．

hold-out 法によって特徴の数を決定する方法は，[83] のアルゴリズムをわずかに修正すれば，カスケード構造の識別器の学習にも適用可能である．[83] では，特徴の追加と検証用サンプル上での評価を繰り返しながら，段階的に非対象カテゴリのパターンを棄却するためのしきい値を調整する．さらに，検出率の下限値と過検出率の上限値のように，事前に設定された性能目標に識別器が到達すると，現在の識別器では対象カテゴリとして検出されてしまう非対象カテゴリのサンプルを収集し，次の識別器を学習する．Boosting のみによって特徴を追加する代わりに，SFS と Boosting のいずれかを用いて特徴を追加するようにアルゴリズムを修正すればよい．検証用サンプル上での指数損失が減少している間は，SFS によって特徴を追加し，指数損失が増加した場合は Boosting による特徴選択に移行する．検証用サンプルは，しきい値の調整と組み合わせる特徴の数の決定の両方に利用できる．

4.2.5　共起性を利用した先行研究の違い

これまでに示した方法によって，識別に有効な特徴の共起関係を選択することが可能となる．ここでは，特徴の共起性を利用した先行研究と提案手法との違いについて説明する．

Hadid ら. [30] により提案された Local Binary Pattern (LBP 特徴) は，注目画素と周辺画素の画素値の大小関係によって，エッジやコーナー点のようなプリミティブな特徴を表現する方法である．近傍の画素対の共起性に着目した方法と考えることができるが，共起関係を評価する特徴の組み合わせは 256 種類のみと少数であり，事前に組み合わせを与えている点が提案手法とは異なる．このような組み合わせが識別に有効であるかどうかは，検出対象に依存し，事前には分からない．また，これ以外に有効な組み合わせがあったとしても表現できない．提案手法では，Boosting による学習過程において，特徴の組み合わせ探索を自動的に行い，LBP 特徴では表現できない空間的に離れた特徴同士の共起性を利用できる点で優れている．

Schneiderman と Kanade [68] は，画像をウェーブレット変換することにより求められた係数間の依存関係を利用する．彼らの顔検出器は，最も高精度はオブジェクト検出方法の 1 つである [89]．しかし，依存関係を評価する係数の組み合わせ方法を事前に定めている．LBP 特

徴に比べれば表現可能な共起関係の種類は多いが，必ずしも識別に有効な組み合わせのみを選択する枠組みではないので，計算コストが大きくリアルタイム処理に適さないという問題がある．提案手法は，識別に適した特徴の組み合わせのみを少数選択するため，処理コストの面で有利である．

rectangle feature のような局所特徴量ではなく，テンプレートマッチングや主成分分析 [75] を用いて広域の画像情報に基づく弱識別器を構成すれば，画素同士の共起性をある程度評価することができる．特に，主成分分析を用いる場合には，特徴に自動的に重み付けできる．Zhang ら [90] は，主成分分析に基づく弱識別器を Viola と Jones の枠組みに導入した方法を提案している．具体的には，Boosting による学習の前段では rectangle feature を用い，後段では識別精度を高めるために固有ベクトルを選択する．しかし，主成分分析に基づく弱識別器は，rectangle feature に基づく弱識別器に比べて計算コストが大きいため，識別器全体の計算コストの増加と識別精度の向上が釣り合うように，弱識別器の切り替えタイミングを実験的に調整する必要がある．提案手法は，Viola と Jones の方法に比べて識別時の計算コストを増加させずに識別精度を向上させるアプローチであるため，そのような調整が不要である．さらに，[90] の主成分分析に基づく弱識別器は，画素値を要素とする特徴ベクトルを入力とするため，rectangle feature に比べて照明変動の影響を受けやすいという欠点がある．

4.3　実験

4.3.1　データ収集

図 4.9 に，実験に使用したサンプルの一部を示す．顔および 3 種類の形状の手を検出するための計 4 つの識別器を学習する．

まず，対象カテゴリに属するサンプル（正事例）をどのように収集したかを説明する．顔のサンプルは，AT&T [64]，FERET [58]，CMU-PIE [71]，XM2VTS [44]，Yale [26] の顔画像データベースからランダムに 10,000 枚を抽出して作成した．カメラに対して真正面を向いた顔のみを選んだ．横を向くなどの姿勢変化は含まないが，多様な照明変動を伴う画像が多い．瞳と鼻孔の位置を手作業で入力し，これらの点を基準に 25×25 画素となるように顔領域を正規化して切り出した．手のサンプルは，人が 3 種類の手形状を作る様子をビデオ撮影することによって作成した．手を握った状態でカメラに手のひら側を向けるジェスチャ，手を開いて手のひらをカメラにむけるジェスチャ，さらに人差し指でカメラを指差すジェスチャが含まれる．ここでは，それぞれを “Fist”，“Open”，“Point” と表記する．ビデオ映像は異なる照明環境で撮影し，手の上部と下部を人手で入力した．25×25 画素となるように正規化して，それぞれ 5,000 枚の画像を得た．各ジェスチャには左右両方の手形状が含まれる．

次に，非対象カテゴリに属するサンプル（負事例）の収集方法について説明する．Web 上の画像のうち，顔や手を含まない画像のみ 8,000 枚をダウンロードした．各画像は，25×25 画素のタイル状に区切り，ランダムに切り出した．負事例の枚数は，正事例の枚数の 40% の枚数となるようにした．すなわち，顔検出器には 4,000 枚，手検出器には 2,000 枚ずつの負事

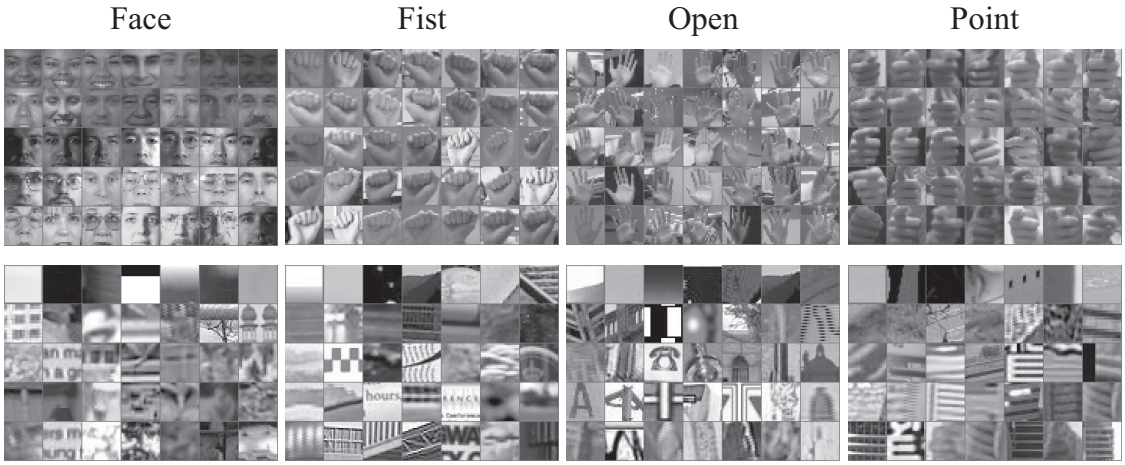

図 4.9. 実験サンプルの一部．上段：正事例，下段：負事例．bootstrapping によって正事例に似たパターンを負事例として収集した．

例を用意した．このデータを使って，まず Viola と Jones の方法により検出器を学習した．この検出器で誤って正事例として検出された非顔画像を 2,000 枚と非手画像を 1,000 枚ずつ追加し，再度 Viola と Jones の検出器を学習した．これを正事例と負事例が同数となるまで繰り返した．この bootstrapping [73] 操作は，合計で 3 度行われ，結果として正事例によく似たパターンの負事例を含む識別が困難なデータを作成した．収集したデータを図 4.9 に示す．それぞれ正事例に似た負事例が含まれていることが分かる．

4.3.2　交差検証による精度評価

10-fold クロスバリデーション（交差検証）により，4 つの識別器それぞれの識別精度を精密に評価する．まず，学習サンプルを 10 のグループに分ける．9 つのグループは学習に使用され，学習された識別器で残りの 1 グループをテストする．順にテストグループを変えながら 10 回試行し，識別誤り率の 10 回の平均値を算出する．誤り率は，誤って識別したサンプルの数の割合であり，正事例を負と誤った場合と負事例を正と誤った場合の両方を含む．

4.3.3　実験 (1)：DAB による識別器の精度比較

検出対象（Face，Fist，Open，Point）に対して，表 4.1 に示す 4 種類のパラメータで識別器を学習した．学習アルゴリズムは DAB，特徴セットは図 2.1 (a) に示した基本セットである．$F1$ は Viola と Jones の方法によって学習された識別器，すなわち各弱識別器が単一の特徴を使用する場合である．$F3$，$F5$，$F7$ は，それぞれ 3 つ，5 つ，7 つの特徴の共起性を用いた提案手法による識別器である．すべての弱識別器は同数の特徴を組み合わせる．なお，組み合わせる特徴の数を学習中に自動決定する場合の評価は，次節の実験 (2) において実施する．

図 4.10 に，$F1$ と $F3$ の学習で選択された特徴を示す．具体的には，学習開始の時点から最

表 4.1. 実験 (1) で比較する 4 つの識別器．F は各弱識別器で使用する特徴の数，T は弱識別器の数である．強識別器で使用した特徴の総数は $F \times T$ となる．弱識別器の数と特徴の総数が等しくなるのは $F1$，すなわち Viola と Jones の識別器のみである．特徴の総数が 1,000 に到達した段階で，Boosting による学習を停止し，これらの識別器を得た．

識別器	F	T	特徴の総数
F1	1	1,000	1,000
F3	3	333	999
F5	5	200	1,000
F7	7	142	994

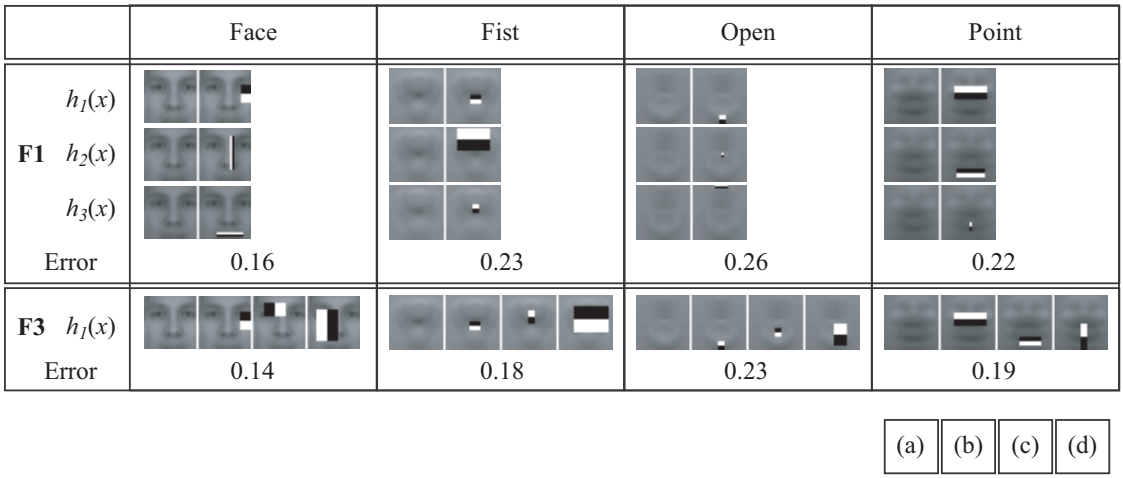

図 4.10. 学習結果と識別誤り率．Viola と Jones の方法による識別器 $F1$ と提案手法に基づく識別器 $F3$ の最初の 3 つの特徴を示す．1 つ目の特徴は同一であるが，2 つ目と 3 つ目の特徴は異なる．識別器 $F3$ の誤り率は常に $F1$ より小さい．

初に選択された 3 つの特徴を (b)，(c) および (d) に示す．(a) には，重み分布 D_t における検出対象カテゴリの重みつき平均画像を示す．平均画像 m_t の k 番目の画素値は次式によって計算される．

$$m_{t,k} = \sum_i D_t(i) x_k(i) \tag{4.13}$$

ここで，$x_k(i)$ は i 番目のサンプル画像の k 番目の画素値である．最初に選択された特徴は $F1$ と $F3$ で同一である．しかし，2 番目，3 番目に選択された特徴は異なっていることが分かる．この結果は上述のアルゴリズムの仕組みから予測されたとおりである．4 種類の検出対象に対して選択された特徴は互いに異なっている．例えば，顔の検出では目や鼻孔といった顔部品が重要であり，手形状 Open の検出ではシルエットの情報がより重要視されていることを読み取れる．図 4.10 には，これら 3 つの特徴を使用した識別器の識別誤り率を併記している．$F3$ の誤り率は一貫して $F1$ より小さい．これは，Boosting によって逐次的に選択された 3 つの特徴を使うよりも，3 つの特徴の共起性を利用した方が精度が高いということを示している．

表 4.2. 4 種類の識別器の処理時間．Intel® Xeon™ 3.2 GHz のプロセッサ 1 個を使って 1 つのテストサンプルを識別する平均処理時間を計測した．

識別器	特徴の総数	テストサンプル 1 個あたりの処理時間（秒）
F1	1,000	0.000259
F3	999	0.000226
F5	1,000	0.000223
F7	994	0.000220

図 4.11 に，4 種類の検出対象と対応する非検出対象を識別した際の誤り率を示す．4 本の曲線は $F1$，$F3$，$F5$ と $F7$ の誤り率を識別に使用した特徴の数に対してプロットしたものである．誤り率は，10-fold クロスバリデーションによって算出されている．基本特徴セットを用いた今回の実験では，特徴の数と識別処理の計算コストをほぼ同一と考えてよい．提案手法による識別器の誤り率は，$F1$ の誤り率に比べて小さい．$F1$ と $F3$ の最小誤り率を比較すると，$F1$ の誤り率の 30% から 50% が削減されている．$F5$ と $F7$ の誤り率は，$F3$ よりも高い．これは $F5$ や $F7$ は弱識別器で多くの特徴を組み合わせているため，学習サンプルに対して過剰適合（overfit）したことによる．組み合わせる特徴の数 F を適切に選択することは，識別精度を高めるにあたって重要であることが分かる．

表 4.2 に，$F1$ から $F7$ の 4 種類の識別器の計算時間を示す．Intel® Xeon™ 3.2GHz のプロセッサ 1 個を用いて，1 つの入力画像を識別する処理時間の平均値を求めた．処理時間はほぼ同等であるが，$F1$ に比べると各弱識別器において複数の特徴を組み合わせる方がやや処理時間が短い．これは，弱識別器の総数が少なくなり，その結果 α_t の加算回数が減少するためである．

図 4.1 に示した 3 つのフレームワークの違いを確認するため，さらに表 4.3 に示す実験を行った．15 個の特徴をそれぞれ異なるフレームワークで選択し，3 つの識別器を学習する．$F1$ は Viola と Jones のフレームワークによって学習した識別器で，各弱識別器が単一の特徴を使用する．$F15$ は，SFS によって組み合わせた 15 個の特徴を使用する弱識別器 1 つだけの識別器である．$F3$ は，5 つの弱識別器それぞれが 3 個ずつ特徴を組み合わせる提案手法による識別器である．$F15$ の識別誤り率は最も大きい．学習サンプルへの過剰適合と，多くの特徴を組み合わせたことによって共起特徴のとり得る場合の数が増えすぎ，統計的な信頼性が低下したことが原因である．$F3$ は最も誤り率が小さく，識別性能が優れている．

以降の実験では，$F1$ と $F3$ のみを比較する．これは $F3$ が $F5$ や $F7$ と比べても最も誤り率が小さく，識別精度が良いからである．また，提案するフレームワークにおいて，DAB の代わりに RAB を用いた場合や，基本特徴セットでなく拡張特徴セットを用いた場合に識別性能が改善されることを示す．最後に示す実験では，遮蔽に対する提案手法の頑健性を評価する．

表 4.3. 3 種類の識別器の誤り率比較．各識別器は 15 個の特徴を使用するが，それぞれ異なるフレームワークによって学習されている．$F1$ の弱識別器は単一の特徴を使用する．すなわち，15 個の弱識別器から構成される識別器である．$F15$ は，15 個の特徴を 1 つの弱識別器で組み合わせる．すなわち，弱識別器 1 個から構成される識別器である．$F3$ は提案するフレームワークによって学習された識別器で，それぞれ 3 つの特徴を組み合わせる 5 つの弱識別器から構成される．$F15$ は過剰適合により誤り率が最大となっている．一方，$F3$ の誤り率が最も小さく，優れた識別性能を示している．

識別器	F	T	Face	Fist	Open	Point
F1	1	15	0.10	0.18	0.22	0.17
F15	15	1	0.12	0.19	0.24	0.21
F3	3	5	0.07	0.12	0.18	0.12

4.3.4　実験 (2)：組み合わせる特徴数の決定

ここでは，3 つの識別器 $F1$，$F3$ および H の性能比較実験を行う．$F1$ と $F3$ は，各弱識別器で事前に設定された個数の特徴を使用する．一方，H は 4.2.4 節で説明した hold-out 法を用いて，特徴の数を自動選択する．特徴数の選択にあたっては，特徴数 F を 1 から順に，検証用サンプルにおける損失が減少から上昇に転じるか，$F^{max} = 7$ に到達するまで増やす．他の実験とは異なり，本実験では 10-fold クロスバリデーションは使用できない．これは 10 通りの識別器を学習すると，個々の弱識別器が異なる数の特徴を使用するため，識別誤り率の特徴数に対する平均値を算出できないからである．実際には，4 つのデータグループを学習に，2 つのグループを hold-out した検証用として用いた．学習された識別器は，残った 4 グループ上でテストした．顔検出器の場合は，4,000 枚の学習サンプルと 2,000 枚の検証用サンプルを使って学習し，残りの 4,000 枚のテストサンプルに対する識別性能を評価した．3 つの手検出器については，学習，検証，テストに使用したサンプル数はそれぞれ 2,000，1,000，2,000 である．すべて DAB に基づく識別器である．

図 4.12 に示す誤り率の曲線を見ると，H と $F3$ は同等の精度でいずれも $F1$ より良い．この結果から，hold-out 法による特徴数の決定方法は識別精度の改善において有効であるといえる．

4.3.5　実験 (3)：RAB による識別器の精度比較

図 4.13 は，DAB でなく RAB に基づく識別器の識別誤り率である．$F3$ は一貫して $F1$ より小さい誤り率を達成できている．この結果から，提案するフレームワークは，異なる Boosting アルゴリズムを用いても有効であることが分かる．図 4.11 と図 4.13 を比べると，DAB に基づく識別器より RAB に基づく識別器の精度が高い．RAB の弱識別器が対数尤度に基づく信頼度を出力することによる効果である．

4.3.6　実験 (4)：拡張特徴セットと DAB による識別器の精度比較

図 4.14 は，図 2.1 の (a) から (d) すべてからなる拡張特徴セットを用い，DAB により学習した識別器の精度を示す．前述の実験が (a) のみの基本特徴セットを用いていたのとは異なる．拡張特徴セットでは，方形の数が異なる特徴を含み特徴抽出処理の手順が異なるため，特徴数と識別処理に要する計算コストは常に等しいわけではないことに注意する．一方，基本特徴セットは，方形の数が 2 であるため，特徴数と計算コストは等価である．識別器 $F3$ は $F1$ より識別誤り率が小さい．このことから，人手で拡張した特徴セットに対しても提案手法は有効であるといえる．図 4.11 と図 4.14 を比べると，拡張特徴セットを用いた場合の方が精度が良い．人手で設計した特徴セットは，識別器の性能改善に効果的であることを示している．提案手法を用いれば，人手による改善を超える性能向上が可能である．

4.3.7　実験 (5)：拡張特徴セットと RAB による識別器の精度比較

図 4.15 は，拡張特徴セットと RAB による識別性能である．やはり $F3$ の方が $F1$ より優れた性能を示している．

4.3.8　実験 (6)：遮蔽に対する頑健性

ここでは，遮蔽に対する頑健性を確かめるための実験を行う．複数の特徴の共起性を利用する提案手法と Viola と Jones の方法の違いは，提案手法が特徴の空間的配置に関してより強い拘束を利用した弱識別器を使用していると考えることができる．このような強い拘束を用いると，遮蔽による影響を受けやすく頑健性が失われるのではないかという疑問が生じる．そこで，実験 (5) で使用したのと同じ識別器 $F1$ と $F3$ を用い，遮蔽を加えたテストサンプルを使って識別性能を比較する．図 4.16 に示すように，テストサンプルにはランダムパターンによって部分的な遮蔽を加えている．遮蔽領域の大きさは 8×25 画素であり，約 $1/3$ の領域が遮蔽されていることになる．ランダムパターンの各画素値は，0 と 255 の間で均等に分布する乱数によって決定した．図 4.17 に，$F1$ と $F3$ の識別誤り率を示す．Viola と Jones の方法に比べて，提案手法が遮蔽に弱いという結果ではなく，両者に顕著な差はない．なお，遮蔽が存在する場合の識別性能は，遮蔽パターン，位置や大きさ，対象物体のパターンによって変わるため，本実験だけであらゆる遮蔽の影響を予測することはできない．

4.4　まとめ

局所特徴量の 1 つである rectangle feature の共起性を利用した新しいオブジェクト検出のフレームワークを提案した．詳細な実験結果から，提案手法は Viola と Jones の方法に比べて高い識別性能を有することが分かった．提案手法の主な利点を以下にまとめる．

- 共起性を利用することで計算コストを増加させずに識別精度を向上させる．言い換えれ

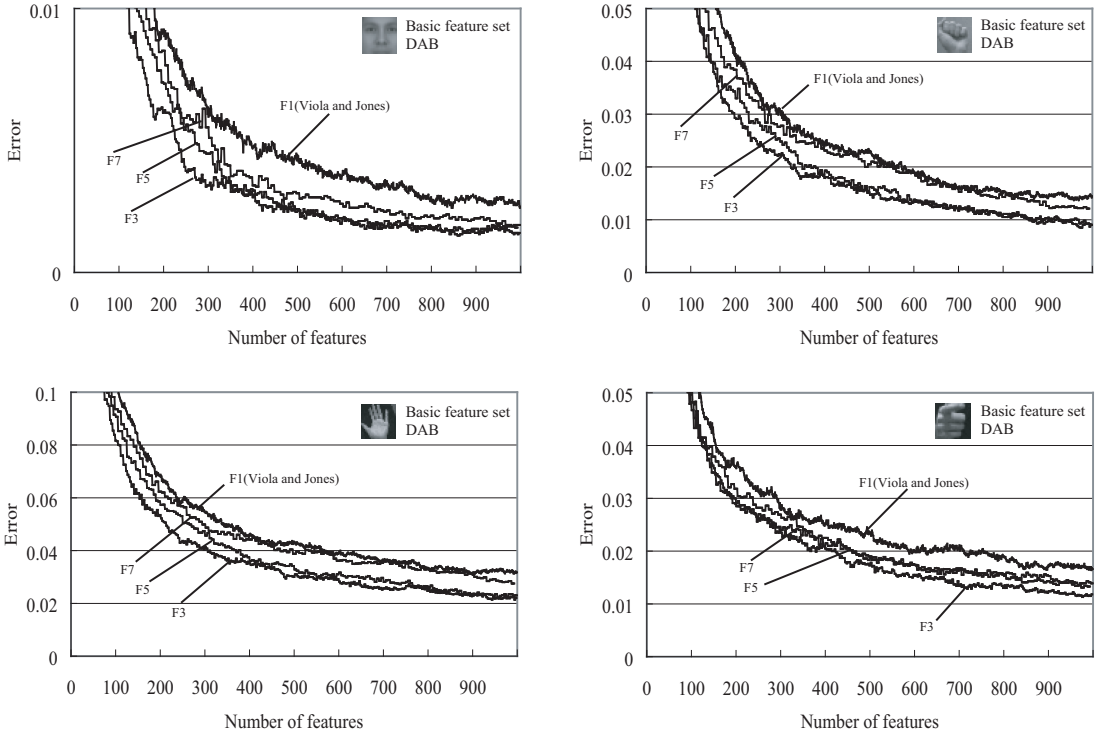

図 4.11. 実験 (1) : DAB による識別器の精度比較．左上 : Face 検出器，右上 : Fist 検出器，左下 : Open 検出器，右下 : Point 検出器

ば同等の識別精度をより小さい計算コストで達成できる．

- 複数の特徴の共起性は多様なオブジェクトの検出に役立つ可能性が高い．なぜならどんなオブジェクトも何らかの空間的な構造を有するため，共起性を見出すことができるからである．
- 提案手法には，Read AdaBoost のような異なる Boosting アルゴリズムを導入できる．また，特徴空間の拡張も可能であり，さらに識別性能を高める上で有効である．

ここでは，rectangle feature によって抽出される明度の勾配情報のみを使用した．Gabor 特徴や，色あるいは動きのような異なる特徴空間における rectangle feature の導入もオブジェクトによっては有効と考えられる．特徴空間の拡張については，今後の課題である．また，提案手法は多数カテゴリの識別にも応用できる可能性がある．Torralba ら [74] は，複数の識別器の間で特徴を共有 (share) するアイデアを提案している．Huang ら [33] は複数カテゴリの検出を効率よく行う Vector Boosting と呼ぶアルゴリズムを提案している．複数のカテゴリに対して共有可能な特徴の共起関係が存在すれば，提案手法とこれらの先行研究を組み合わせることによって，さらなる性能改善が可能と考えられる．

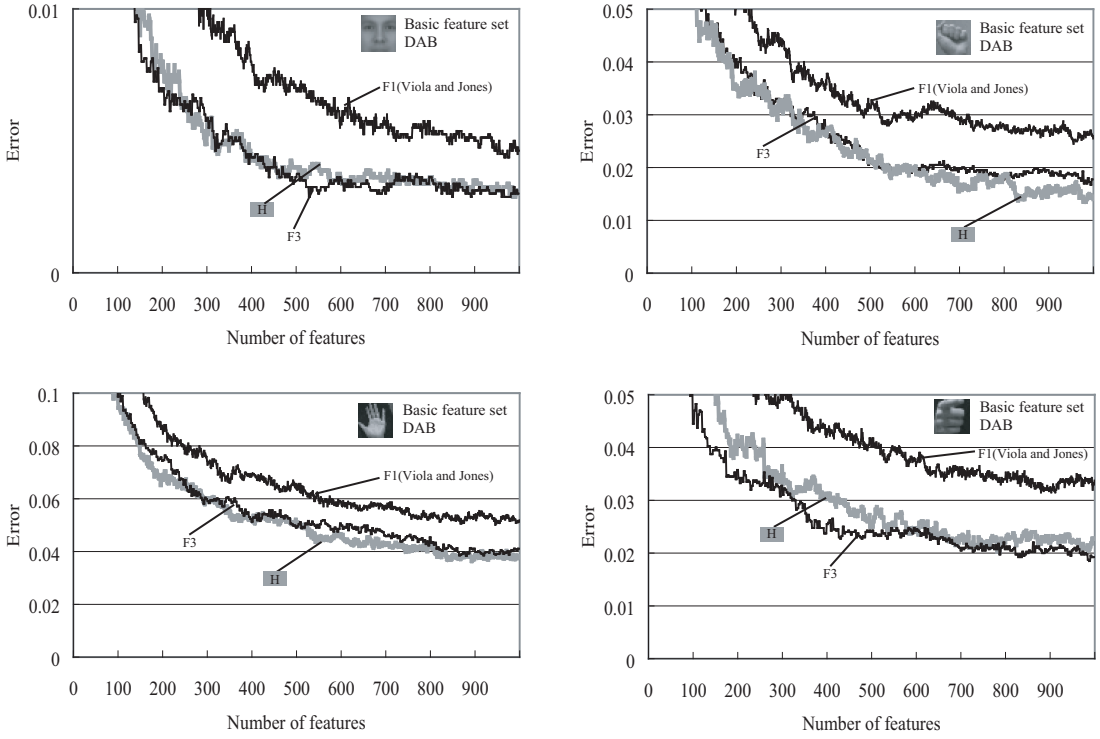

図 4.12. 実験 (2) : Hold-out 法によって組み合わせる特徴数を自動決定した識別器の精度比較.
それぞれの弱識別器で異なった数の特徴を組み合わせて識別器を学習.

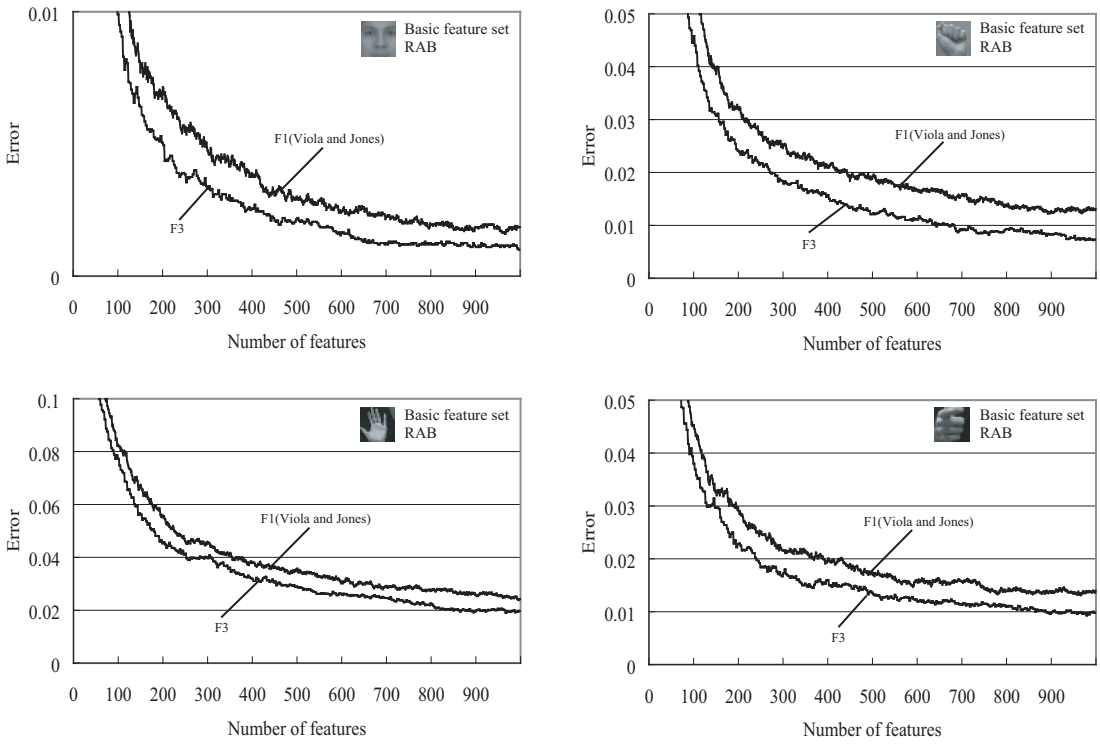

図 4.13. 実験 (3)：RAB による識別器の精度比較

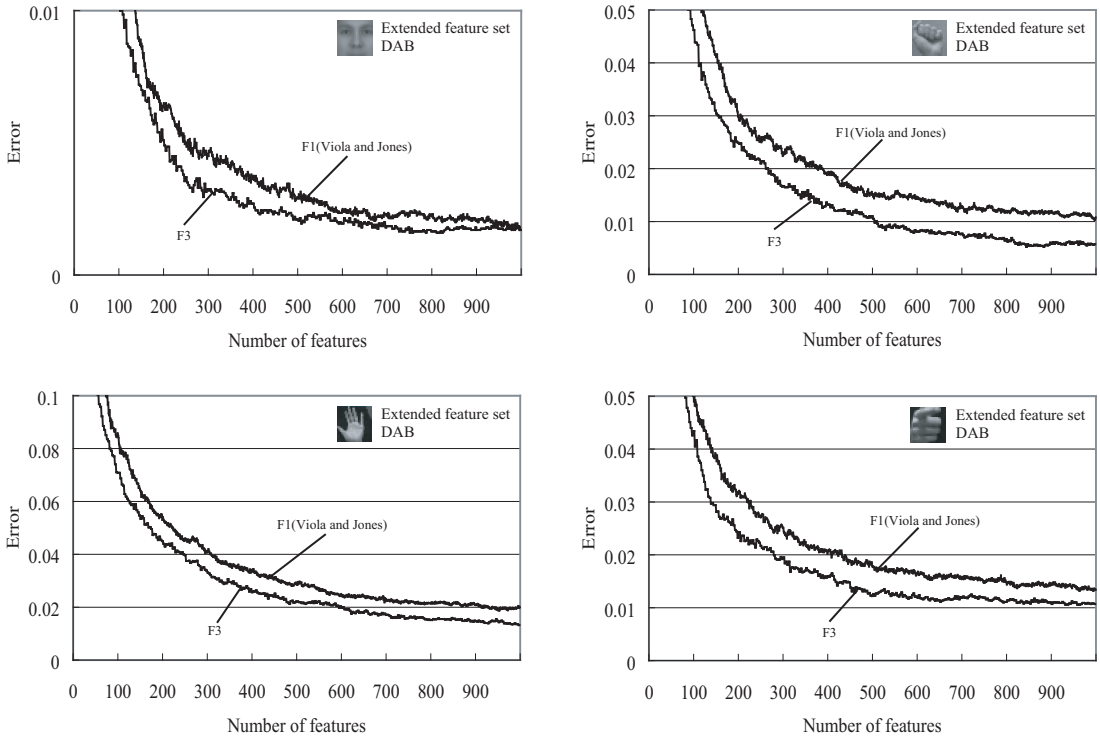

図 4.14. 実験 (4)：拡張特徴セットと DAB による識別器の精度比較

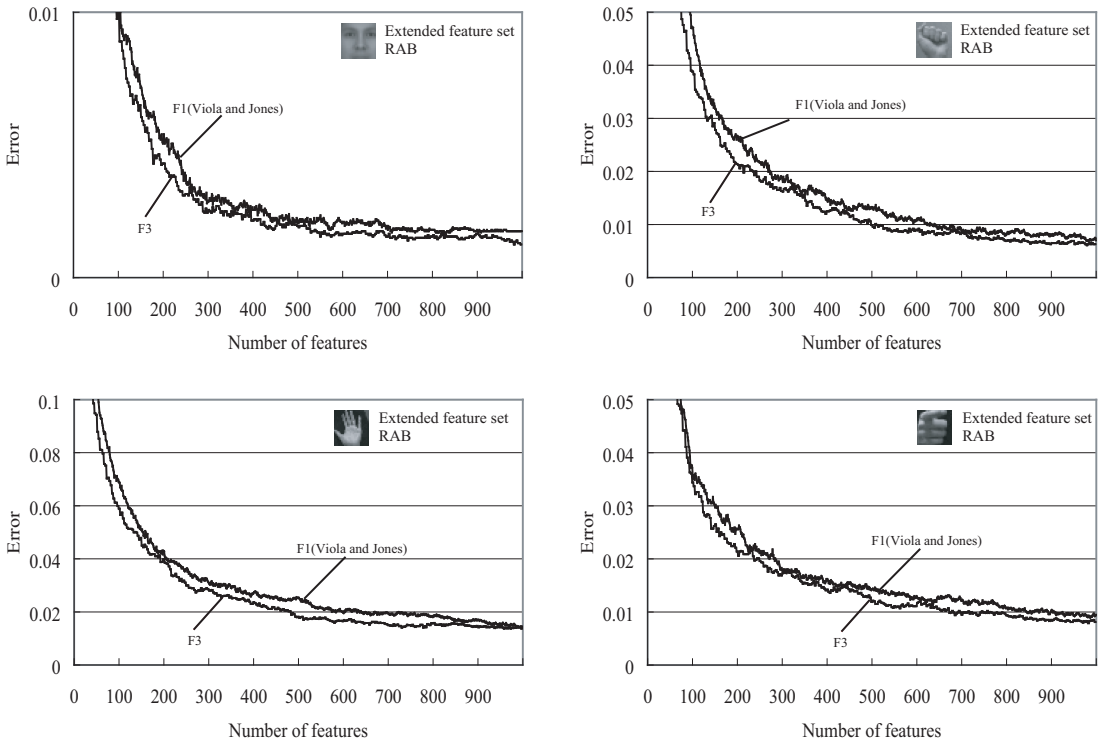

図 4.15. 実験 (5)：拡張特徴セットと RAB による識別器の精度比較

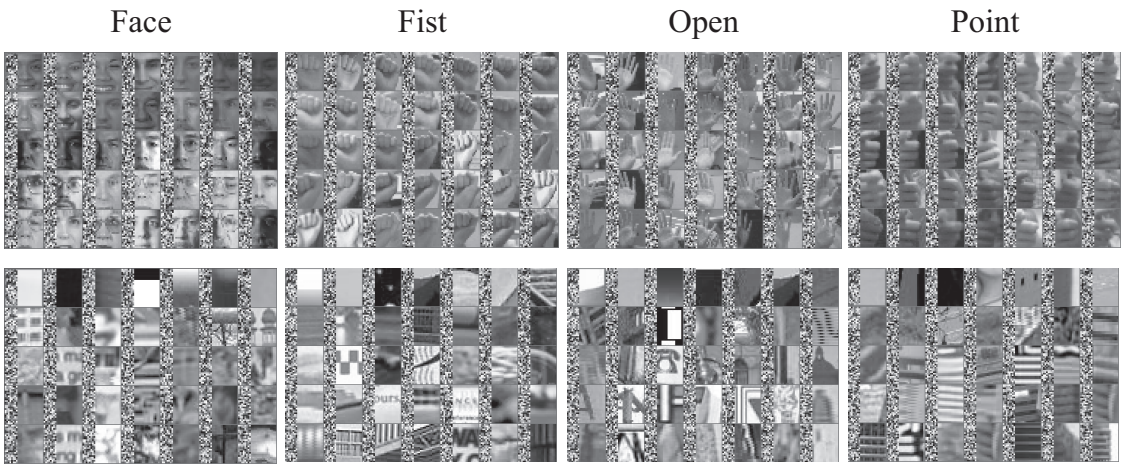

図 4.16. ランダムパターンによって部分的に遮蔽されたテストサンプル．上段：正事例，下段：負事例.

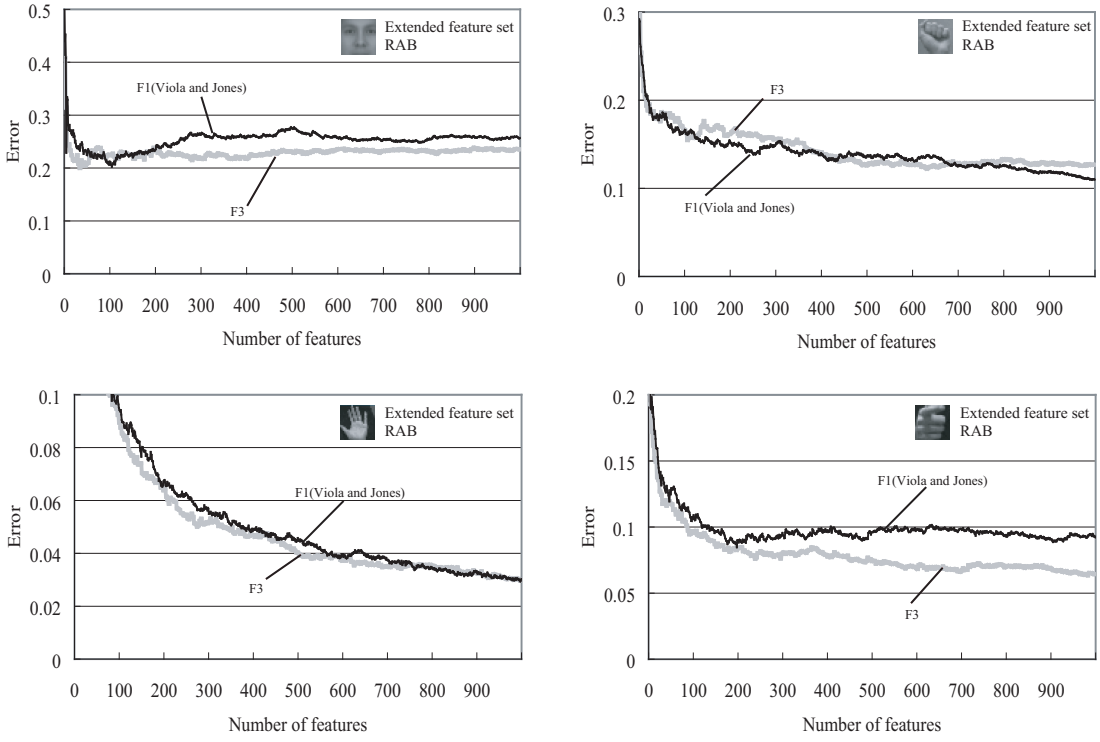

図 4.17. 実験 (6)：ランダムパターンで部分的に遮蔽された状況での精度比較．拡張特徴セットと RAB による識別器を使用．$F1$ と $F3$ で顕著な差はない．

第 5 章

局所特徴量の共起性を利用したサブカテゴリ識別手法

前章では，空間的に異なる位置で観測された rectangle feature の共起関係を利用したオブジェクト検出方法を提案した．ここでは，同一位置から抽出した複数の異なる局所特徴量の共起性を利用したサブカテゴリ識別手法を提案する．前章のオブジェクト検出手法は，対象カテゴリと非対象カテゴリの 2 カテゴリを識別するアルゴリズムであるが，本章では多数のカテゴリを識別するフレームワークを扱う．

図 5.1 に提案手法の概要を示す．まず，注目画素およびその画素を中心とする周辺領域から抽出された複数の特徴量の共起関係を Random Forest [12] によって学習し，それを用いてそれぞれの特徴の組み合わせを量子化する．これにより，例えば色と形状を表現する 2 つの特徴から，特定の色がその周辺領域でどのような形状で分布しているかという情報を評価することが可能となる．例えば，図 5.1 の注目画素 A 周辺からは目の輪郭と同色の領域の空間的分布，B 周辺からは羽の一部の色の広がりを表現できる．次に，量子化された特徴のヒストグラムから算出した新たなカーネルと，個々の特徴から独立に算出された Spatial Pyramid Kernel（以下，SPK）とを Multiple Kernel Learning（以下，MKL）により結合する．これにより，例えば，図 1.6(a) と (b) のように類似したカテゴリや (a) と (c) のように大きく異なるカテゴリの識別にそれぞれ適するようにカーネルの重み調整が行われ，識別精度が向上する．

以下では，共起特徴を導入した識別器の学習方法について述べた後，計算コストの増加に関して検討する．それにより，共起特徴の導入によって，識別精度向上と計算コスト低減の両立につながる可能性を示す．最後に，200 種類の鳥類を含むデータベース “CUB-200” を用いた大規模なサブカテゴリ識別実験を行い，提案手法と従来手法を詳細に比較し，識別精度や計算コストで提案手法が優れていることを示す．

5.1　共起性を利用した識別器の学習

ここでは，共起性を利用した識別器の学習手順について述べる．まず，鳥類の識別に用いる特徴について述べる．次に，Random Forest によりそれらの特徴間の共起性を抽出する方法

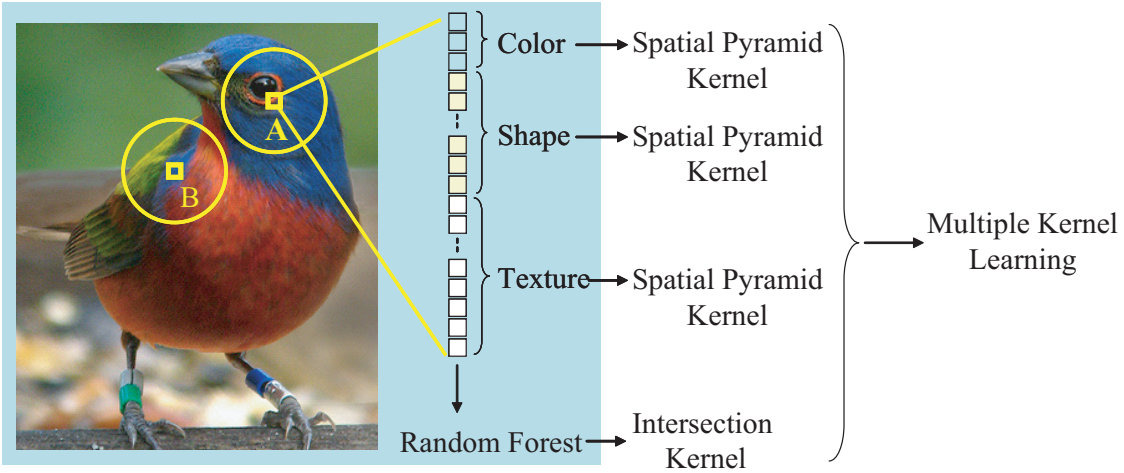

図 5.1. 提案手法の概要.

を説明する．最後に，MKL により，個別の特徴から得られたカーネルと Random Forest から得られたカーネルを結合し，SVM 識別器を学習する．

5.1.1　特徴

鳥類を識別するにあたって重要な画像特徴は，色，形，テクスチャ[*1]という知見がある [52]．そこで，それらに対応する特徴として，以下の 3 種を用いる．

- 色 : YCbCr．RGB 値をそのまま使用することも考えられるが，明るさ成分と色成分を分離した YCbCr を選択した．
- 形 : Self-similarity [70]．まず，注目画素を囲む小領域とその周辺の同じ大きさの小領域とで SSD により定義される類似度を求め，類似度マップを作成する．次に，図 5.2 に示すような log-polar ビンごとに類似度の最大値を求め，特徴記述子とする．局所領域における物体形状あるいは幾何学的なレイアウトを抽出できるため，形に対応する特徴として位置づける．
- テクスチャ : Opponent SIFT [78]．照明光の明るさ変化に対して不変となるような変換 (2.2) を RGB 各成分に適用した後，各成分 (O_1, O_2, O_3) について SIFT 特徴量 [43] を算出する．Sande ら [78] により，様々な色特徴の中で最も良い性能を示したと報告されているため選択した．なお，SIFT は各画素における勾配をヒストグラム化したものであるので，本稿ではテクスチャに対応する特徴として位置づけているが，色の情報も反映されている．

[*1] [52] では color, shape, pattern と表現されている．

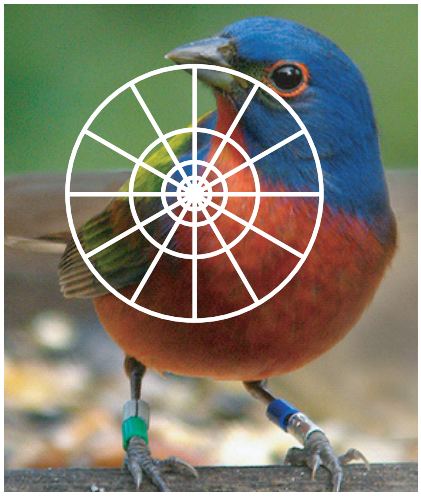

図 5.2. Self-similarity 特徴量の計算に用いられる log-polar ビン．注目画素から離れるに従ってビンが大きくなる．これにより，形状変化を許容できる．

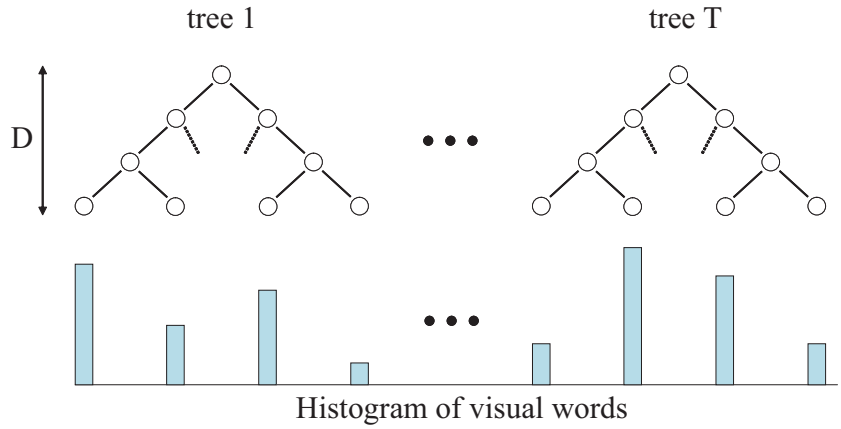

図 5.3. Random Forest による visual words ヒストグラムの生成．深さ D の二分木を T 個使い，ヒストグラムを求める．

5.1.2　Random Forest による複数の特徴の共起関係の抽出

Breiman [12] により提案された Random Forest は，アンサンブル学習の一種であり，複数のランダム木を組み合わせた識別器である．その後，Moosmann ら [49] により，識別ではなく visual words を得るためのクラスタリングの方法として用いられた．図 5.3 に示すように，それぞれのランダム木に入力された特徴ベクトルは，どれか一つのリーフノードに到達するが，リーフノードのインデックス番号を visual word とすることで visual words のヒストグラムが得られる．

Moosmann らは単一の特徴のみを扱っていたが，以下の手順で Random Forest を学習することで複数の特徴の共起関係を表現するヒストグラムを計算できる．

準備： 注目画素およびその周辺から得られた色，形およびテクスチャに対応する特徴ベクトルをそれぞれ x^C，x^S，x^T とし，特徴ベクトルの次元数をそれぞれ D^C，D^S，D^T とする．さらに，それらを 1 次元に並べた特徴ベクトルを $x = (x^C, x^S, x^T)$ とする．x が抽出された学習用画像のクラスラベル c との組 $S = \{(x_1, c_1), \cdots, (x_N, c_N)\}$ を初期サンプル集合として準備する．ここで，N は学習画像から抽出された特徴ベクトルの総数である．

ランダム木の学習： 次に，個々のランダム木を学習する．初期サンプル集合 S から重複を許してランダムに N 個のサンプルを取り出し（ブートストラップ抽出），学習サンプル S_0 を作成する．このサンプルを起点として，図 5.4 に示す手順に従って，ノードを追加しながらランダム木を成長させる．ランダム木は二分木とし，各ノード n では boolean テスト：

$$T_n = \{x(d_n) < \theta_n\} \tag{5.1}$$

により，S_n を S_{n+1} と S_{n+2} という 2 つのサンプル集合に分割する．ここで，$S_n = S_{n+1} \cup S_{n+2}$ かつ $S_{n+1} \cap S_{n+2} = \emptyset$ である．d_n は，複数の特徴ベクトルを並べて作ったベクトル x において，最もクラス間の分離がよくなる次元のインデックス番号である．具体的には，$D^C + D^S + D^T$ 次元の特徴ベクトル x の一部，$\sqrt{D^C + D^S + D^T}$ 次元をランダムに選択し，そのうちで最も分離が良くなる次元を 1 つ選ぶ．なお，D^C，D^S，D^T の値が大きく異なると，候補として選択される次元が次元数の多い特徴に偏る恐れがあるため，各次元数で正規化した確率にしたがって候補次元を決定した．θ_n は，選択された次元の特徴量の大小関係を評価するしきい値であり，最もクラス間の分離が良くなる値を選択する．ここで，分離の良さの規準は，

$$H(T_n) = \frac{|S_{n+1}(T_n)| H_{n+1}(T_n) + |S_{n+2}(T_n)| H_{n+2}(T_n)}{|S_{n+1}(T_n)| + |S_{n+2}(T_n)|} \tag{5.2}$$

$$H_{n+1}(T_n) = -\sum_c P_{n+1}(c|T_n) \log_2 P_{n+1}(c|T_n) \tag{5.3}$$

$$H_{n+2}(T_n) = -\sum_c P_{n+2}(c|T_n) \log_2 P_{n+2}(c|T_n) \tag{5.4}$$

で定義されるエントロピーとする．ここで，$|S_{n+1}(T_n)|$ と $|S_{n+2}(T_n)|$ はテスト T_n によって分割されたサンプル集合に含まれるサンプルの数である．また，$P_{n+1}(c|T_n)$ と $P_{n+2}(c|T_n)$ は，各サンプル集合に含まれるクラス c の生起確率である．このエントロピーが最小となるような d_n と θ_n を選択し，テスト T_n を決定する．つまり，クラス間の分離が最も良くなるように，各ノードにおける識別パラメータ d_n と θ_n を決定しながら，学習サンプル S_n を分割するという手続きを繰り返す．以下の条件を満たした場合は，リーフノードとし，ルート方向に遡って木の成長がとまるまで同様の手続きを行う．

- あらかじめ設定した木の深さ D になる．
- $|S_n|$ が所定の個数以下になる．
- S_n に含まれるサンプルのクラスラベルがすべて同じになる．

以上の手順により，複数の特徴から識別に有効な次元としきい値の組み合わせが得られ，共

Extract S_0 from S and set $n = 0$.
Repeat until all nodes are processed.
  If the following conditions are met:
    Depth of the node reaches D.
    $|S_n|$ is smaller than the predefined value.
    All sample labels in S_n become the same.
  Then create a leaf node.
  Else
    (A) Determine T_n by choosing d_n and θ_n
      that minimize the entropy $H(T_n)$.
    (B) Split S_n into S_{n+1} and S_{n+2}
      according to T_n.
    (C) $n = n + 1$.
  End

図 5.4. ランダム木の成長手順

起関係を取り出せる．たとえば，例えば特定の色が注目画素の周辺でどのように分布しているか，あるいは周辺にどのようなテクスチャがあるかという情報を抽出できる．

Random Forest の学習： ランダム木を所定の個数 (T) 学習し，Random Forest が得られる．ブートストラップ抽出されたサンプルから学習した複数のランダム木はそれぞれ異なるため，汎化性が保たれる．

識別時には，各特徴ベクトルをランダム木に入力し，どのリーフノードに到達したかを調べる．リーフノードのインデックス番号を visual word と見なしてヒストグラムを作成する．

なお，従来の k-means クラスタリングによって得られる visual words を組み合わせることによっても複数の特徴の共起関係を表現できるが，次に述べる問題があるため Random Forest によるクラスタリングの方が実用的である．例えば，色，形，テクスチャの 3 種類の特徴がそれぞれ M 個の visual words で表現されている場合を考える．visual words の組み合わせに応じて M^3 のビンを持つヒストグラムに投票すれば共起関係を表現できる．しかし，通常 M の値は数百 [82] [38] となるため，ヒストグラムのビン数は数百万と大きくなり，カーネルの計算コストが増大するという問題がある．また，1 枚の画像から得られる特徴の数に対してヒストグラムのビン数が圧倒的に大きくなるため，ヒストグラムはスパース（疎）となり，統計的な信頼性が低下する恐れがある．Random Forest によるクラスタリングでは，木の深さ D と木の数 T を適切に選択してヒストグラムの長さを調節することにより，このような問題を回避することができる．

5.1.3　MKL による結合

MKL によって学習される SVM 識別器の識別関数は,

$$f(h) = \sum_{i \in SV} y_i \alpha_i K(h, h_i) \tag{5.5}$$

で定義される．ここで，h は visual words のヒストグラム，h_i と $y_i \in \{+1, -1\}$ は，それぞれ学習時に選択されたサポートベクターとそのクラスラベル，K は正定値カーネルである．K は,

$$K = \sum_f w_f K_f \tag{5.6}$$

のように複数のカーネルの線形結合である．MKL は，係数 α_i および各カーネルの結合重み w_f を同時に学習する．w_f の値はクラスによって異なり，対象クラスを識別するのに適したカーネルには大きな重み，そうでないカーネルには小さな重みが与えられる．

提案手法では,

$$K = \sum_j w_j K_j^{SP} + \sum_k w_k K_k^{RF} \tag{5.7}$$

のように，Spatial Pyramid Kernel (SPK) : K^{SP} と Random Forest Kernel (RFK) : K^{RF} の線形結合として識別器を学習する．これにより，visual words の空間的な配置を考慮に入れた SPK と，複数の特徴の共起性を反映した RFK とを併用する．従来の SPK だけでは取り出すことができなかった識別に有効な共起関係を RFK によって抽出できていれば，MKL によって重み w_k に大きな値が与えられる．逆に，RFK より SPK が有効に働くようなクラスに対しては，SPK の重み w_j が相対的に大きくなるように重みが調節される．RFK が全く有効に作用しないような極端な場合には，SPK にのみ重みが与えられ，従来手法と同等の識別能力となる．

SPK は,

$$K_L^{SP} = \frac{1}{2^L} K_l + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} K_l \tag{5.8}$$

により，対象物体を囲む外接四角形を順次 4 分割，16 分割というように細分化しながら，分割領域ごとにカーネルを計算する [38]．ここで，L は分割する階層の数である．各階層では,

$$K_l(h_1, h_2) = \sum_d \min(h_1(d), h_2(d)) \tag{5.9}$$

によりヒストグラムインタセクションとしてカーネルを計算する．

RFK は，Random Forest のリーフノードに到達した特徴の数を計数したヒストグラム間で，上述のインタセクションを計算する．

5.1.4　Schroff ら [69] の方法との違い

Random Forest を用いて共起特徴を抽出する方法は，一般物体のセグメンテーションで用いられた例 [69] があり，RGB ヒストグラムや HOG の共起関係を求めることで，輪郭部分のセグメンテーション精度が高まったことが報告されている．しかしながら，共起特徴のみによる画素単位での識別を行っている点や，MKL による共起特徴の重み最適化はしていない点が提案手法と異なる．CUB-200 で例示したように互いに類似したクラスや大きく異なるクラスが混在するような大規模な識別課題では，共起特徴を単独で用いるより，クラスごとに重みを制御する方がよいと考えられる．実験で提案手法の有効性を示す．

5.1.5　識別処理における計算コスト

テストサンプルが入力されたときの識別処理に要する計算コストについて，SPK のみを MKL で結合する従来手法と，SPK と RFK を結合する提案手法とを比較する．

Random Forest によって visual words ヒストグラムを生成する計算コストは，ランダム木の深さを D，木の数を T とすると，1 個の特徴について $O(DT)$ である．これは，d 次元の特徴ベクトルを M 個の k-means クラスタに割り当てる計算コスト $O(dM)$ に対して，通常かなり小さくできる．例えば，先行研究 [49, 82] ではそれぞれ $D = 10$，$T = 5$，$d = 60$，$M = 300$ などが使用されており，計算回数が大幅に少なくなることが分かる．また，Random Forest によるクラスタリングは，特徴ベクトルの次元数 d に依存しない．したがって，visual words ヒストグラムの計算コストは従来の k-means クラスタリングに比べ一般に小さいと言える．

次に，ヒストグラム間の類似度すなわちカーネルの計算コストについて議論する．ヒストグラムの次元数は最大 $2^D \times T$ となるので，カーネルの計算コストは $O(2^D T)$ となる．$O(2^D T)$ を提案手法による計算コストの増加とみなすかどうかは，従来手法をどう位置づけるかによって変わってくる．例えば，従来手法で 3 種類の特徴を組み合わせた識別器があり，提案手法によって識別精度をさらに高めるようとする場合を考える，提案手法では RFK を追加し，新たなカーネル計算が必要となるため，計算コストは増加する．しかし，4 種類目の特徴を従来手法の枠組みのまま追加することによって精度を高めようとする場合と比較すると，むしろ計算コストを削減できる．もし，提案手法によって 3 種類の特徴から識別に有効な共起関係を抽出でき，4 種類の特徴を使う従来手法と同等以上の識別精度が得られたとする．従来手法では 4 種類目の特徴抽出処理が新たに必要となるが，提案手法ではすでに抽出済みの特徴の共起関係を用いるため，特徴抽出処理の計算コスト増加はない．また，上述の議論から，ヒストグラム計算のコストも大幅に小さい．従来手法と提案手法でヒストグラムの長さ（次元数）が同等であれば，カーネル計算のコスト増加もないため，提案手法が計算コスト面で有利である．次の実験において，この議論を裏付けるデータを示す．

(a) Original image　(b) Manual segmentation　(c) Final segmentation

図 5.5. 前処理によるセグメンテーション結果．(a) 原画，(b)CUB-200 データセットに付帯している人手によるおおまかなセグメンテーション，(c)GrabCut によるセグメンテーション．

5.2　実験

以下では，複数の特徴から独立に算出した SPK を MKL により結合する従来手法と，共起関係を導入した提案手法の識別精度を比較し，提案手法によって識別精度が向上することを示す．

5.2.1　鳥類データセット

実験に使用する鳥類データセット “CUB-200” [87] は，主として北米に生息する 200 種類の鳥類の画像を合計 6,033 枚集めたものである．図 1.6 のような互いによく似たクラスも含む．各画像には，人手でおおまかな輪郭を入力したセグメンテーション画像が付帯されている．実験では，このセグメンテーション画像を用いて識別対象以外の画像領域を取り除いた．

5.2.2　前処理

より正確なセグメンテーション画像を得るために GrabCut [62] による前処理を施す．図 5.5 に一例を示す．前節で述べたように，原画 (a) に対して，(b) のセグメンテーションが付与されている．(b) を用いて前景と背景の色モデルを学習し，GrabCut によって色モデルを更新しながらセグメンテーションを行う．セグメンテーション結果を (c) に示す．さらに，(c) の前景領域を囲む外接四角形を切り出して，長辺が 500 ピクセルとなるようスケールを調整し，すべての画像で鳥の大きさが同程度となるようにした．

5.2.3　特徴抽出

特徴抽出は，いわゆる dense sampling により，1 画素おきに配置した格子上で，なおかつ前節の前処理で前景と判定された画素において行った．各画像からは，平均数千個の特徴ベクトルが得られた．特徴ベクトルの次元は，YCbCr，Self-similarity，OpponentSIFT がそれぞれ 3 次元，60 次元，384 次元である．Self-similarity は，log-polar ビンの角度方向を 20 分割，

表 5.1. Random Forest によって組み合わせた YCbCr と Self-similarity 特徴の識別精度.

T	D	ヒストグラムの次元数	1 位正解率 (%)
3	11	6,144	32.1
5	10	5,120	32.6
10	9	5,120	32.5
20	8	5,120	32.2

半径方向を 3 分割して計算した．Self-similarity および OpponentSIFT の算出は，それぞれ Gulshan [2] と Sande [3] のコードを使用した．

5.2.4　識別器の学習

まず，k-means 法を用いて各特徴ベクトルをクラスタリングした．学習用の画像は $N = 3,000$ すなわち，各クラス 15 枚ずつ選び，それらから抽出された特徴ベクトルから visual words を得た．visual words の数 M は，先行研究 [82] で用いられた値 (300) に近い $M = 250$ とした．この場合，階層数 $L = 3$ とした Spatial Pyramid によるヒストグラムの次元数は $5,250 (= M \times (1 + 4 + 16))$ となる．

また，学習画像の同じ画素位置から抽出された複数の特徴ベクトルから Random Forest を学習した．まず，Random Forest の木の数 T と木の深さ D を変えて，識別精度の変化を調べた．このとき，Random Forest によって得られるヒストグラムの次元数が，上述の Spatial Pyramid ヒストグラムの次元数 $5,250$ とほぼ同じになる範囲で T と D を選ぶようにした．これは，SPK と RFK の計算コストを揃えて識別精度を比較するためである．表 5.1 に，YCbCr と Self-similarity を Random Forest によって組み合わせた場合の識別精度（1 位正解率）を示す．たとえば，$T = 5$，$D = 10$ のときヒストグラムの次元数は $5,120 (= 2^D \times T)$ となり，最も $5,250$ に近い値となる．ヒストグラムの次元数がほぼ一定に保たれるように，D と T を変えたところ，識別精度はほとんど変化しなかった．そのため，以降の実験では最も識別精度の高かった $T = 5$，$D = 10$ に固定した．

以上の処理によって，すべての学習画像が visual words のヒストグラムによって表現される．個別の特徴から得た 5,250 次元のヒストグラムが 3 種類と，Random Forest により共起関係が表現された 5,120 次元のヒストグラムを 4 種類（3 種類の特徴すべてを組み合わせた場合と，任意の 2 つを組み合わせた場合）得た．これらから算出したカーネルの重みを MKL によってクラスごとに調整し，SVM 識別器を学習する．one-against-all で合計 200 個の識別器を学習し，最大の評価値が得られたクラスに分類する．なお，MKL は Vedaldi と Varma のコード [82] を利用した．

表 5.2. 各特徴単体の識別精度

Feature	1 位正解率 (%)	3 位正解率 (%)	10 位正解率 (%)
Feature 1 (F1): YCbCr	26.6	40.2	57.8
Feature 2 (F2): Self-similarity	19.4	31.7	46.6
Feature 3 (F3): Opponent SIFT	25.8	40.1	56.4

表 5.3. MKL により結合された SPK の識別精度（従来手法）

Feature	1 位正解率 (%)	3 位正解率 (%)	10 位正解率 (%)
F1+F2	33.7	50.2	67.1
F1+F3	35.0	50.7	68.9
F2+F3	27.8	42.9	59.2
F1+F2+F3	35.8	52.0	69.5

5.2.5　識別精度

各クラスから学習画像 15 枚をランダムに選定して識別器を学習し，残りの画像に対してテストするという試行を 5 回繰り返す交差検証により識別精度を比較する．識別精度は，クラスごとの識別率を平均し，さらにそれを 5 回の試行で平均した値とする．対象カテゴリに対する類似度が最高となる 1 位正解率，対象カテゴリが 3 位，10 位までに入る 3 位正解率と 10 位正解率もあわせて示す．

表 5.2 は，各特徴単体の識別精度である．YCbCr が 26.6%（1 位正解率）と最も良く，ついで OpponentSIFT，最後に Self-similarity の順となっている．以降では，この 3 種類の特徴を組み合わせた場合の識別精度を示すが，表記を簡略化するために，各特徴を F1，F2，F3 で表す．また，MKL によるカーネルの結合を $F1 + F2$ のように + で記述し，Random Forest による共起特徴の抽出を $F1 \times F2$ のように × で記述する．後者がより積極的に特徴間の共起関係を評価していることを反映させ，このような記述方法を便宜的に用いる．

表 5.3 に，MKL によって SPK の結合重みを調整する従来手法の識別精度を示す．特徴を単独で用いる場合に比べて大幅に精度が向上している．特に，色と形 $(F1 + F2)$ あるいは色とテクスチャ $(F1 + F3)$ の併用の効果が高いことが分かる．

表 5.4 は，RFK（Random Forest による共起特徴）のみを用いた場合の識別精度を示す．特徴を単独で用いる場合に比べて識別精度を改善しているが，表 5.3 に示す従来手法に比べてやや識別精度が低い．これは，従来手法ではクラスごとに個々のカーネルの重みを最適化できるが，共起特徴に基づくカーネル単独ではそのような自由度がないことや，SPK と異なり，特徴の空間的な配置を考慮していないことが要因である．また，3 種類の特徴すべてを用いた場合よりも，2 種類の特徴を用いた場合の方が識別精度が高い．これは，特徴を多く組み合わせ

表 5.4. RFK のみの識別精度

Feature	1 位正解率 (%)	3 位正解率 (%)	10 位正解率 (%)
F1×F2	32.6	48.0	65.5
F1×F3	30.4	44.9	63.0
F2×F3	24.1	39.4	56.1
F1×F2×F3	29.8	47.2	65.6

ると，特徴ベクトルの次元数が大きくなり，同じ木の深さでは十分な共起関係が抽出できなくなるためである．Moosmann ら [49] も高次元特徴を用いた場合に同様の現象が起きたことを報告している．したがって，木の深さを増加させれば識別精度が高まる可能性があるが，本稿では SPK と RFK で同等の計算コストをかけた場合の識別精度を比較するために，木の深さを一定に保って実験を行う．

表 5.5 に RFK と SPK を結合する提案手法による識別精度を示し，表 5.3 と比較する．まず，すべての特徴の組み合わせにおいて，Random Forest によって求められたカーネルを追加することによって識別精度が向上している．例えば，$F1 + F2$ が 33.7% であるのに対し，$(F1 + F2) + (F1 \times F2)$ は 36.8% に向上している．このとき，図 1.6 に示した (a)Blue Grossbeak を (b)Indigo Bunting と混同した誤りは 12.5% 減少し，(c)Artic Tern を (d)Caspian Tern と混同した誤りは 62.5% 減少していた．また，$F1+F2+F3$ が 35.8% であるのに対し，それに任意の Random Forest による組み合わせを追加した $(F1+F2+F3)+(*)$ は，すべてで識別精度が向上している．すなわち，提案手法は従来手法からさらに識別精度を向上できているといえる．

また，$F1 + F2 + F3$ が 35.8% であるのに対し，$(F1 + F2) + (F1 \times F2)$ が 36.8%，$(F1 + F3) + (F1 \times F3)$ が 36.6% である．すなわち，従来手法の枠組みのままで 3 種類目の特徴 $F3$ あるいは $F2$ を追加するよりも，提案手法によって特徴を 2 種類だけ組み合わせた場合の方が高精度である．このとき，5.1.5 節での議論から，計算コストの面でも提案手法の方が優れている．表 5.6 に，$F1+F2+F3$ と $(F1+F2)+(F1\times F2)$ の計算コストが $F1+F2$ に比べてどれだけ増加しているかを示す．従来手法では，$F3$ を新たに追加することにより，特徴抽出処理が増加，その特徴のヒストグラムを k-means クラスタリングで計算するコストが $O(dM)$ 増加，Spatial Pyramid ヒストグラムの次元数（5,250）に基づくカーネル計算コストが増加する．一方，提案手法では，すでに抽出済みの特徴 $F1$ と $F2$ の共起性を評価するため，特徴抽出処理の計算コスト増加はゼロ，Random Forest によるヒストグラム計算は $O(DT)$ と小さく，カーネル計算の次元数は 5,120 と従来手法と同等である．つまり，提案手法では計算コストの削減と高精度化を両立しているとみなせ，共起関係を利用することの有効性を示している．なお，SVM による識別ではサポートベクターの数に比例して識別処理のコストは増加するが，今回の実験では従来手法，提案手法ともにサポートベクターの数は N すなわち全学習サンプルがサポートベクターとなっており，差はなかった．

表 5.5. RFK と SPK を結合した場合の識別精度（提案手法）

Feature	1 位正解率 (%)	3 位正解率 (%)	10 位正解率 (%)
(F1+F2) + (F1×F2)	36.8	53.5	70.1
(F1+F3) + (F1×F3)	36.6	53.1	69.9
(F2+F3) + (F2×F3)	29.3	45.5	62.0
(F1+F2+F3) + (F1×F2)	37.9	53.9	71.5
(F1+F2+F3) + (F1×F3)	37.3	54.1	71.3
(F1+F2+F3) + (F2×F3)	36.3	52.9	69.9
(F1+F2+F3) + (F1×F2×F3)	37.4	54.3	71.8

表 5.6. 識別器 $F1 + F2$ からの計算コストの増加量比較

	従来手法 $F1 + F2 + F3$	提案手法 $F1 + F2 + (F1 \times F2)$
特徴抽出	F3	0
ヒストグラム計算	$O(dM)$	$O(DT)$
カーネル計算	5,250 次元	5,120 次元

表 5.7. MKL によって求められた各カーネルの結合重みの例

Feature	Weight
F1	0.328
F2	0.134
F3	0.306
F1×F2	0.231

図 5.6 に表 5.2〜5.5 をまとめた棒グラフを示す．最も高い識別精度は，$(F1 + F2 + F3) + (F1 \times F2)$ の 37.9% であった．CUB-200 データセットでの識別精度は，Branson ら [11] による 19% という報告しかない．彼らは，CUB-200 に付帯されているセグメンテーション領域の外接四角形内で特徴抽出を行っており，本稿とは条件が異なるため単純な比較はできないが，参考として示しておく．

また，表 5.7 に識別器 $(F1 + F2 + F3) + (F1 \times F2)$ の各カーネルの結合重みを示す．これはクラスごとに算出された式 (5.7) の重み w_j および w_k を全クラスで平均し，さらに 5 回の試行で平均した値である．共起性を反映した特徴 $F1 \times F2$ にも一定の重みが付与され，識別に寄与していることが分かる．

なお，$(F1 + F2 + F3) + (F1 \times F2) + (F1 \times F3)$ のように複数の RFK を結合する実験も行ったが，$(F1 + F2 + F3) + (F1 \times F2)$ と同等の識別精度であったため割愛する．

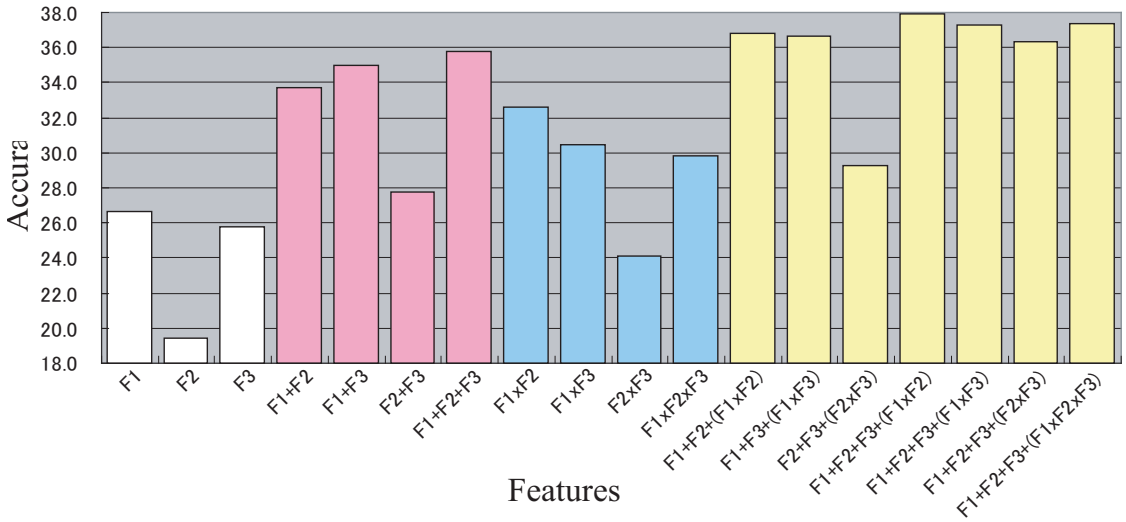

図 5.6. 1 位正解率の比較．提案手法による $F1 + F2 + F3 + (F1 \times F2)$ の識別率が最も高い．

5.3　まとめ

本章では，複数の局所特徴の共起性を抽出することによって，サブカテゴリを精度良く識別する手法を提案した．様々な特徴の組み合わせを評価し，提案手法によれば，従来よりも小さい計算コストで同等以上の識別精度が得られることを示した．実験の結果から，注目画素の色とその空間的配置の組み合わせが高い識別精度を示していることから，鳥類の識別には色とそのレイアウトを表現するような特徴記述子を設計すればよいという指針が得られた．

第 6 章

結論

6.1 本論文のまとめ

本論文では,

- 高精度かつ高速なオブジェクト検出とオブジェクト認識技術の開発
- サブカテゴリ識別による専門家の知識への容易なアクセスの実現

の 2 つの目的を掲げ，その達成に向けて解決すべき技術課題，および課題解決へのアプローチを示した．具体的には,

1. 照明条件の変動や撮影時に付加されるノイズなどの外乱に対する頑健性確保
2. 個体差や姿勢変化などのカテゴリ内の変動に対する許容性確保
3. 互いに類似したカテゴリを見分けるカテゴリ間の識別能力確保
4. 短時間に識別処理を実行できる高速性確保

の技術課題において特に 2 と 3 を重要視し，特徴量の確率分布をテンプレートとする新たな照合法と，局所特徴量の共起性を利用したオブジェクト検出手法およびサブカテゴリ識別手法を提案した．2 と 3 の解決に注力しながら，1 と 4 も満たす方法を模索し，通常では二律背反となる高精度化と高速化の両立を図った．使用している特徴量は増分符号や rectangle feature などの既存手法であり，また学習アルゴリズムも AdaBoost や SVM のように先行研究で実績のある手法である．しかし，提案手法は，特徴量の確率分布を利用することや共起性を導入することにおいて新規性と有効性を備えている．

数多くの実験を通じて，提案手法は計算コストを増やすことなく，識別精度を高められることを示した．具体的には，確率的増分符号相関と呼ぶ新しいテンプレート照合法は，既存の照合法の中で照合精度と処理速度のバランスが良い増分符号相関に対し，顔の検出や顔向きの識別において優れた性能を示した．また，識別に有効な rectangle feature の共起関係を自動選択する学習アルゴリズムは，オブジェクト検出手法で主流となっている Viola と Jones の方法を上回る識別精度を達成した．いずれも比較手法に対して計算コストを保ったまま高精度化したことがポイントである．最後に，局所特徴量の共起性を利用したサブカテゴリ識別手法は,

200 カテゴリの鳥類の識別において，一般物体認識で最も良いとされている既存手法に比べて高精度であった．鳥類の写真をアップロードするとカテゴリ名を検索できるシステムに利用すれば，専門家しか知らない知識に容易にアクセスする可能性が拓ける．

検出対象に合わせて人手で特徴量を設計することは有効である反面，他の対象への適用可能性を狭めてしまう恐れがある．提案手法は，学習サンプルを変更することによって，異なったオブジェクトの検出や認識が可能となる汎用的な手法である．

6.2　今後の課題

オブジェクト認識において残された課題のうち最大のものは，扱えるカテゴリ数の増加である．人間が識別可能なカテゴリ数は 3,000 から 30,000 と言われている [8]．本論文では 200 カテゴリの鳥類という現時点では最大規模のカテゴリ数を扱ったが，まだ人間の識別能力には及ばない．人間の知的作業の支援や代行を目的とすれば，数多くのカテゴリを効率よく，また確実に識別する手法が必要である．Griffin と Perona [29] による Taxonomy は，効率化に有効となるアイデアである．また，カテゴリ数を増やすには，追加学習が可能な識別方式も必要である．AdaBoost や SVM のように discriminative な学習アルゴリズムは，あらかじめ識別したいカテゴリすべてが与えられていることが前提であり，カテゴリを追加すると識別器の再学習が必要となる．これらの方法は，追加学習に要する計算時間の観点では generative なアルゴリズムに比べて不利である．Boiman ら [9] の最近傍識別ベースの手法もアプローチの一つとして有望である．

オブジェクト検出は，正面顔の検出では精度や速度において実用的なレベルに達したと言える．しかし，顔向きの変化に対応しようとすると，45 度回転，90 度回転といった回転量に応じたカテゴリを新たに設定し，カテゴリごとに識別器を学習するという方法がとられる．回転量の刻み幅をどのように設定すれば，効率と精度のバランスが良くなるかについては，実験的に決定する方法が主流で解決方法の決定打はないのが現状である．アピアランス（見え）ベースの方法を使用する限り，直面する問題である．多様な学習サンプルを識別精度と処理コストのバランスが最も良くなるように，自動的にクラスタリングするアルゴリズムが必要である．

発表文献リスト

論文

1. 三田雄志，佐藤誠，“特徴の共起性利用による鳥類の識別性能向上”，電子情報通信学会論文誌，Vol.J95-D, No.1, pp.67-75, 2012
2. T. Mita, T. Kaneko, B. Stenger and O. Hori, “Discriminative feature co-occurrence selection for object detection,” IEEE Trans. on PAMI, Vol.30, No.7, 1257-1269, 2008
3. 三田雄志，金子敏充，堀修，“顔検出に適した共起に基づく Joint Haar-like 特徴”，電子情報通信学会論文誌，Vol.J89-D-II, No.8, pp.1791-1801, 2006
4. 三田雄志，金子敏充，堀修，“個体差のある対象の画像照合に適した確率的増分符号相関”，電子情報通信学会論文誌，Vol.J88-D-II, No.8, pp.1614-1623, 2005

国際会議

1. T. Mita, T. Kaneko, and O. Hori, “A probabilistic approach to fast and robust template matching and its application to object categorization,” International Conference on Pattern Recognition (ICPR), pp.597-601, 2006.
2. T. Mita, T. Kaneko, and O. Hori, “Joint Haar-like features for face detection,” International Conference on Computer Vision (ICCV), pp.1619-1626, 2005.

Technical Report

1. P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, “Caltech-UCSD Birds 200,” California Institute of Technology, CNS-TR-2010-001, 2010.

国内学会・研究会

1. 三田雄志，金子敏充，堀修，“顔検出に適した Joint Haar-like 特徴の提案”，画像の認識・理解シンポジウム (MIRU2005), pp104-111, 2005（優秀論文賞受賞）
2. 三田雄志，金子敏充，堀修，“クラス内変動を有する対象の照合に適した確率的増分符号相関”，画像の認識・理解シンポジウム (MIRU2004), Proc.I, pp.571-576, 2004
3. 三田雄志，金子敏充，堀修，“微少な差異を含む画像の照合に適した空間差分確率テンプレートの提案”，第 9 回画像センシングシンポジウム, J-2, pp.561-566, 2003

書籍・解説

1. 倉爪 亮 (著), 石川 博 (著), 加藤 丈和 (著), 佐藤 淳 (著), 三田 雄志 (著)，八木康史 (編)，斎藤英雄 (編)，“コンピュータビジョン　最先端ガイド　1”，(株) 新技術コミュニケーションズ，第 5 章 AdaBoost
2. 三田雄志，“画像特徴量 [III] -輝度に着目した画像特徴量と顔検出-”，電子情報通信学会誌, pp.58-62, Jan. 2011
3. 三田雄志，“AdaBoost の基本原理と顔検出への応用”，情処研報，2007-CVIM-159, pp.265-272, 2007
4. 三田雄志，“Boosting 学習を利用した画像認識”，画像センシングシンポジウムチュートリアル講演, 2007

謝辞

私を社会人ドクターとして受け入れて下さり，学会への論文投稿や本論文の執筆にあたってあたたかいご指導をいただいた佐藤誠教授に心より感謝いたします．

本論文をまとめるにあたり有益なご助言をいただいた長橋宏教授，熊澤逸夫教授，山口雅浩教授，張暁林准教授に深く感謝いたします．

本研究の共同研究者である株式会社東芝研究開発センターの堀修氏と金子敏充氏には，多くの有益な助言をいただきました．深く感謝いたします．同じく共同研究者である Toshiba Research Europe の Björn Stenger 氏には，データの提供や英文論文の校正などで数多く手助けしていただきました．心より感謝いたします．

本論文で使用した鳥類データベースの構築や鳥類の識別という研究課題の着想には，California Institute of Technology の Pietro Perona 教授や Peter Welinder 氏をはじめ，Vision Lab のメンバーとのディスカッションが役立ちました．Caltech で貴重な時間を過ごさせていただいたことも含めて感謝いたします．

2 年間に渡る博士課程の学生生活では，研究や論文執筆のために週末も家にこもる機会が多くありました．暖かくサポートしてくれた妻かおりと，本論文の執筆を頻繁にほほえましく邪魔してくれた息子謙信に心をこめて「ありがとう」と申し上げます．また，十分な教育環境や進学についての助言など学位取得までの道筋を整えてくれた父達雄と母美智子に感謝します．

三田雄志

参考文献

[1] *http://en.wikipedia.org/wiki/Indigo_Bunting*.

[2] *http://www.robots.ox.ac.uk/~vgg/software/SelfSimilarity/*.

[3] *http://koen.me/research/colordescriptors/*.

[4] D. I. Barnea and H. F. Silverman. A class of algorithms for fast digital image registration. *IEEE Trans. Comput.*, Vol. C–21, No. 2, pp. 179–186, 1972.

[5] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. on PAMI*, Vol. 24, No. 4, pp. 509–522, 2002.

[6] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. *Proc. of CVPR*, 2005.

[7] A. Berg and J. Malik. Geometric blur for template matching. *Proc. of CVPR*, 2001.

[8] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychological Review*, Vol. 44, No. 2, pp. 115–147, 1987.

[9] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. *Proc. of CVPR*, 2008.

[10] A. Bosch, A. Zisserman, and X. Muoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Trans. on PAMI*, Vol. 30, No. 4, pp. 712–727, 2008.

[11] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. *Proc. of ECCV*, 2010.

[12] L. Breiman. Random forests. *Machine Learning*, Vol. 1, pp. 5–32, 2001.

[13] L. G. Brown. A survey of image registration techniques. *ACM Computing Survey*, Vol. 24, pp. 325–376, 1992.

[14] M. Burl and P. Perona. Recognition of planar object classes. *Proc. of CVPR*, pp. 223–230, 1996.

[15] M. Burl and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. *Proc. of ECCV*, pp. 628–641, 1998.

[16] M. B. Clowes. On seeing things. *Artificial Inteligence*, Vol. 2, No. 1, pp. 79–116, 1971.

[17] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of

keypoints. *In Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 1–22, 2004.

[18] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *Proc. of CVPR*, Vol. 1, pp. 886–893, 2005.

[19] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. *Proc. of ECCV*, 2006.

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. *http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html*.

[21] L. Fei-Fei, R. Fergus, and P. Perona. A Bayesian approach to unsupervised one-shot learning of object categories. *Proc. of CVPR*, pp. 1134–1141, 2003.

[22] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *IEEE. CVPR 2004, Workshop on Generative-Model Based Vision*, 2004.

[23] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. *Proc. of CVPR*, pp. 380–387, 2004.

[24] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Computational Learning Theory: Eurocolt*, pp. 23–37, 1995.

[25] J. Friedman, T. Hastie, and R. J. Tibshirani. Additive logistic regression: a statistical view of boosting. *Technical report, Department of Statistics, Sequoia Hall, Stanford Univerity*, July 1998.

[26] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: generative models for recognition under variable pose and illumination. *Proc. of IEEE Conf. on Automatic Face and Gesture Recognition*, pp. 277–284, 2000.

[27] K. Grauman and T. Darrell. Pyramid match kernels: discriminative classification with sets of image features. *Proc. of ICCV*, pp. 1458–1465, 2005.

[28] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. *Technical Report 7694*, 2007.

[29] G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. 2008.

[30] A. Hadid, M. Pietikainen, and T. Ahonen. A discriminative feature space for detecting and recognizing faces. *Proc. of CVPR*, Vol. 2, pp. 797–804, 2004.

[31] C. Harris and M. Stephens. A combined corner and edge detector. *In Alvey Vision Conference*, pp. 147–151, 1998.

[32] B. Heisele, T. Poggio, and M. Pontil. Face detection in still gray images. *A.I. Memo*, No. 1687, 2000.

[33] C. Huang, H. Ai, Y. Li, and S. Lao. High-performance rotation invariant multiview

face detection. *IEEE Trans. on PAMI*, Vol. 29, No. 4, pp. 671–686, 2007.

[34] A. Jain and D. Zongker. Feature selection: evaluation, application, and small sample performance. *IEEE Trans. on PAMI*, Vol. 19, No. 2, pp. 153–158, 1997.

[35] T. Joutou and K. Yanai. A food image recognition system with multiple kernel learning. *Proc. of ICIP*, 2009.

[36] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. *Proc. of CVPR*, 2004.

[37] M. Kolsch and M. Turk. Robust hand detection. *Proc. of IEEE Conf. on Automatic Face and Gesture Recognition*, pp. 614–619, 2004.

[38] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. *Proc. of CVPR*, 2006.

[39] S. Z. Li and Z. Q. Zhang. FloatBoost learning and statistical face detection. *IEEE Trans. on PAMI*, Vol. 26, No. 9, pp. 1112–1123, 2004.

[40] R. Lienhart and J. Maydt. An extended set of Haar-like features for rapid object detection. *Proc. of ICIP*, Vol. 1, pp. 900–903, 2002.

[41] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, Vol. 30, No. 2, pp. 79–116, 1998.

[42] C. Liu and H. Y. Shum. Kullback-Leibler boosting. *Proc. of CVPR*, pp. 587–594, 2003.

[43] D. G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, Vol. 60, No. 2, pp. 91–110, 2004.

[44] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: the extended M2VTS database. *Proc. of 2nd Conf. on Audio and Video-based Biometric Personal Verification (AVBPA99),* `http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb`, 1999.

[45] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, Vol. 60, No. 1, pp. 63–86, 2004.

[46] T. Mita, T. Kaneko, and O. Hori. Joint Haar-like features for face detection. *Proc. of ICCV*, pp. 1619–1626, 2005.

[47] T. Mita, T. Kaneko, and O. Hori. A probabilistic approach to fast and robust template matching and its application to object categorization. *Proc. of ICPR*, pp. 597–601, 2006.

[48] T. Mita, T. Kaneko, B. Stenger, and O. Hori. Discriminative feature co-occurrence selection for object detection. *IEEE Trans. on PAMI*, Vol. 30, No. 7, pp. 1257–1269, 2008.

[49] F. Moosmann, B. Triggs, and F. Jurie. Randomized clustering forests for image classification. *IEEE Trans. on PAMI*, Vol. 30, No. 9, pp. 1632–1646, 2008.

[50] P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset

selection. *IEEE Tras. Computers*, Vol. 26, No. 9, pp. 917–922, 1977.

[51] M. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. *Proc. of Indian Conf. on Comp. Vision, Graphics & Image*, pp. 722–729, 2008.

[52] The Cornell Lab of Ornithology. All About Birds. *http://www.allaboutbirds.org/page.aspx?pid=1053*.

[53] Y. Ohta. Knowledge-based interpretation of outdoor natural color scenes. *Pitman Advanced Publishing Program*, 1985.

[54] E. J. Ong and R. Bowden. A boosted classifier tree for hand shape detection. *Proc. of IEEE Conf. on Automatic Face and Gesture Recognition*, pp. 889–894, 2004.

[55] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. *Proc. of CVPR*, pp. 130–136, 1997.

[56] C. P. Papageorgiou, M. Oren, and T. Poggio. A trainable system for object detection. *IJCV*, Vol. 38, pp. 15–33, 2000.

[57] P. Perona. Vision of a Visipedia. *Proc. of the IEEE*, Vol. 98, No. 8, pp. 1526–1534, 2010.

[58] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing J*, Vol. 16, No. 5, pp. 295–306, 1998.

[59] A. R. Pope. Model-based object recognition: A survey of recent research. *Technical Report TR-94-04, University of British Columbia, Computer Science Department*, 1994.

[60] P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, Vol. 15, No. 11, pp. 1119–1125, 1994.

[61] X. Ren and J. Malik. Learning a classification model for segmentation. *Proc. of ICCV*, 2003.

[62] C. Rother, V. Kolmogorov, and A. Blake. GrabCut: interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, Vol. 23, pp. 309–314, 2004.

[63] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. on PAMI*, Vol. 20, No. 1, pp. 23–38, 1998.

[64] F. Samaria and A. Harter. Parameterisation of a stochastic model for human face identification. *2nd IEEE Workshop on Applications of Computer Vision*, 1994.

[65] S. Satoh, Y. Nakamura, and T. Kanade. Name-It: naming and detecting faces in video by the integration of image and natural language processing. *Proc. of IJCAI-97*, pp. 1488–1493, 1997.

[66] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, Vol. 37, pp. 297–336, 1999.

[67] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans.*

on PAMI, Vol. 19, No. 5, pp. 530–535, 1997.

[68] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *IJCV*, Vol. 56, pp. 151–177, 2004.

[69] F. Schroff, A. Criminisi, and A. Zisserman. Object class segmentation using random forests. *Proc. of BMVC*, 2008.

[70] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. *Proc. of CVPR*, 2007.

[71] T. Sim, S. Baker, and M. Bsat. The CMU Pose, Illumination, and Expression (PIE) database. *Proc. of the 5th International Conference on Automatic Face and Gesture Recognition*, 2002.

[72] S. D. Strearns. On selecting features for pattern classifiers. *Proc. of ICPR*, pp. 71–75, 1976.

[73] K. K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Trans. on PAMI*, Vol. 20, No. 1, pp. 39–51, 1998.

[74] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. *Proc. of CVPR*, 2004.

[75] M. A. Turk and A. P. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, Vol. 3, No. 8, pp. 71–86, 1991.

[76] M. Turk and A. Pentland. 'face recognition using eigenfaces. *Proc. of CVPR*, pp. 586–591, 1991.

[77] F. Ullah, S. Kaneko, and S. Igarashi. Orientation code matching for robust object search. *Trans. IEICE on Inf. & Syst.*, Vol. E84–D, No. 8, pp. 999–1006, 2001.

[78] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. on PAMI*, Vol. 32, No. 9, pp. 1582–1596, 2010.

[79] J. van de Weijer, T. Gevers, and A. Bagdanov. Boosting color saliency in image feature detection. *IEEE Trans. on PAMI*, Vol. 28, No. 1, pp. 150–156, 2006.

[80] V. N. Vapnik. *The nature of statistical learning theory.* Springer, 1995.

[81] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. *Proc. of ICCV*, 2007.

[82] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. *Proc. of ICCV*, 2009.

[83] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, Vol. 57, pp. 137–154, 2004.

[84] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *IJCV*, Vol. 63, pp. 153–161, 2005.

[85] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. *Proc. of CVPR*, pp. 101–108, 2000.

[86] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. *Proc. of CVPR*, pp. 18–32, 2000.

[87] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[88] B. Wu, H. Ai, C. Huang, and S. Lao. Fast rotation invariant multi-view face detection based on Real AdaBoost. *Proc. of IEEE Conf. on Automatic Face and Gesture Recognition*, pp. 79–84, 2004.

[89] M. H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *IEEE Trans. on PAMI*, Vol. 24, No. 1, pp. 34–58, 2002.

[90] D. Zhang, S. Z. Li, and D. G. Perez. Real-time face detection using boosting in hierarchical feature spaces. *Proc. of ICPR*, pp. 411–414, 2004.

[91] エルッキオヤ著, 小川英光, 佐藤誠訳. パターン認識と部分空間法. 産業図書, 1986.

[92] 金子俊一. 実世界マシンビジョンのためのロバスト画像照合技術. 電気学会論文誌, Vol. 121-C, No. 5, pp. 830–834, 2001.

[93] 佐藤雄隆, 金子俊一, 五十嵐悟. 選択的正規化相関によるロバスト画像照合. 電気学会論文誌, Vol. 121-C, No. 4, pp. 800–807, 2001.

[94] 三田雄志, 金子敏充, 堀修. 微少な差異を含む画像の照合に適した空間差分確率テンプレートの提案. 第 9 回画像センシングシンポジウム, Vol. J-2, pp. 561–566, 2003.

[95] 三田雄志, 金子敏充, 堀修. クラス内変動を有する対象の照合に適した確率的増分符号相関. 画像の認識・理解シンポジウム (MIRU2004), Vol. 1, pp. 571–576, 2004.

[96] 三田雄志, 金子敏充, 堀修. 顔検出に適した Joint Haar-like 特徴の提案. 画像の認識・理解シンポジウム (MIRU2005), pp. 104–111, 2005.

[97] 三田雄志, 金子敏充, 堀修. 個体差のある対象の画像照合に適した確率的増分符号相関. 電子情報学会論文誌, Vol. J88-D-II, No. 8, pp. 1614–1623, 2005.

[98] 三田雄志, 金子敏充, 堀修. 顔検出に適した共起に基づく Joint Haar-like 特徴. 電子情報通信学会論文誌, Vol. J89-D-II, No. 8, pp. 1791–1801, 2006.

[99] 三田雄志, 佐藤誠. 特徴の共起性利用による鳥類の識別性能向上. 電子情報学会論文誌, Vol. J95-D, No. 1, pp. 67–75, 2012.

[100] 山口修, 福井和広. 定性的 3 値表現に基づく画像マッチング. 信学技報, PRMU2002-34, 2002.

[101] 斉藤文彦. ブロック照合投票処理を用いた遮蔽に強い画像マッチング. 電子情報学会論文誌, Vol. J84-D-II, No. 10, pp. 2270–2279, 2001.

[102] 村瀬一朗, 金子俊一, 五十嵐悟. 増分符号相関によるロバスト画像照合. 電子情報学会論文誌, Vol. J83-D-II, No. 5, pp. 1323–1331, 2000.

[103] 堀修. 統計的手法を用いた様々な背景からの顔領域の抽出. 情報学研報, CVIM-106-19, pp. 139–145, 1997.

[104] 堀修. 統計的手法による画像からの顔領域の抽出. 情報処理学会論文誌, Vol. 40, No. 8,

pp. 3281–3288, 1999.

[105] 柳井啓司. 一般物体認識の現状と今後. 情報処理学会論文誌 コンピュータビジョンとイメージメディア, Vol. 48, No. SIG_16(CVIM_19), pp. 1–24, 2007.

[106] 蔡篤儀, 李鎔範. ファジィ推論を用いた心臓超音波画像における心筋症のコンピュータ支援診断. 電子情報学会論文誌, Vol. J84-A, No. 12, pp. 1431–1438, 2001.