Linear Distance Metric Learning for Large-scale Generic Image Recognition

（線形距離計量学習による大規模一般画像認識）

Hideki Nakayama

中山 英樹

Graduate School of Information Science and Technology

The University of Tokyo

A thesis submitted for the degree of

Doctor of Philosophy

本論文を愛する家族へ捧げます.
I would like to dedicate this thesis to my beloved family.

Acknowledgements

本論文の執筆，および私の研究室生活は多くの方々の多大な支えを頂きました．主査の國吉康夫教授には，学部の頃から 6 年の長きに渡り，研究者としての技術や考え方，哲学のご指導を賜りました．また，非常に恵まれた環境の中で自由に研究をさせて頂き，多くの経験をさせて頂きました．原田達也准教授には，研究内容から日常生活に至るまで，さまざまな面でのご指導・激励を頂きました．大村吉幸助教には，研究に関する本質的な指摘を度々頂きました．また，産業技術総合研究所の大津展之教授には，本論文の研究内容の核となる数理的技術について，東大にいらっしゃった頃からご指導を頂きました．先生方のご指導のもと，本論文を完成させたことを誇りに感じております．深く深く感謝をいたします．

また，学位の審査をしてくださった佐藤洋一教授，稲葉雅幸教授，森武俊特任准教授に心より感謝いたします．先生方に頂いたご指摘により，本論文をより良いものへ仕上げることができました．

研究室の森裕紀研究員には，研究者としてのあり方・考え方など多くの事を学ばせて頂きました．また，同じ年に学位審査を受けたこともあり，多くの場面で励ましを頂きました．同期の竹中一仁君，尾形邦裕君，Hassan Alirezaei 君には，さまざまな面で大変お世話になりました．皆様のおかげで，孤独を感じることもなく，大変楽しく博士課程を過ごすことができました．特に，竹中君とは研究に関する興味が近いこともあり，しばしばディスカッションをして頂きました．また，優秀な後輩にも恵まれ，刺激を受けながら快適な研究生活を送ることができました．特に，OG の松本理恵さん，博士一年の金崎朝子さん，修士二年の牛久祥孝君には，共著論文の執筆や学会出張などで大変お世話になりました．研究室スタッフの永井おりが技術専門職員，菊地万里研究補助員，都丸美緒子技術補佐員，細井久美子技術補佐員には，研究生活の上で必要なさまざまなご支援を頂きました．みなさまの支えがなければ本論文の完成はあり得ませんでした．深く感謝を致します．

日本学術振興会には，博士課程三年間に渡り十分な金銭的支援を頂き，恵まれた研究生活を送る事ができました．心より感謝致します．

最後に，私の研究生活を常に暖かく見守り応援してくれた両親・弟に最大の感謝をいたします．

Abstract

制約のない実世界の画像を計算機に認識させ，言語により記述させる技術を一般画像認識 (generic image recognition) と呼ぶ．一般画像認識は扱う画像や認識対象が多種多様であるため，極めて難しいタスクであると認知されている．汎用性の高い一般画像認識を実現するためには，大量の事例データからの学習が鍵となる．しかしながら，従来の手法は学習サンプル数に対するスケーラビリティを欠いていたため，大規模な画像コーパスを用いて学習・認識を行うことは著しく困難であった．

本研究では，学習サンプル数に対しスケーラブルかつ高精度な一般画像認識（画像アノテーション）アルゴリズムの開発に取り組む．高精度な画像アノテーションを行うためには，ボトムアップな画像特徴量と最終的に求める“意味”の間の隔たり (semantic gap) を緩和する必要がある．Semantic gap の問題に対処するためには，以下の 2 つのプロセスが重要となる．

1. 多様な画像特徴量の抽出
2. サンプル間距離計量の統計的学習

スケーラブルなシステムを実現するためには，両者の相性を考慮しそれぞれを設計することが極めて重要である．大規模コーパスを前提とする場合，学習サンプル数に対し線形オーダの計算コストで学習を実現することが望ましい．そこで本研究では，バイモーダルの線形次元圧縮手法である正準相関分析 (CCA) に着目しサンプル間距離計量の学習を行う．CCA の確率構造を利用することで，理論的に最適なサンプル間距離計量の導出を行い，これを Canonical Contextual Distance (CCD) と名付ける．CCD に基づく画像アノテーション手法は相対的に少ない計算コストで学習・認識が可能であり，かつ先行研究と遜色ない認識精度を達成できる．

一方，CCD を有効に用いるためには，画像特徴量が線形な性質を有していることが必要不可欠である．すなわち，画像特徴空間における内積が，特徴量が前提とする生成モデルの類似度を適切に近似している必要がある．そこで，本研究ではこの要件を満たす新しい画像特徴抽出の枠組みを開発する．まず，画像の局所特徴分布を単一のガウシ

アンによってモデル化する global Gaussian approach を提案する．さらに，ガウシアンを情報幾何の手法により近似的にコーディングした大域的特徴ベクトルである Generalized Local Correlation (GLC) を導出する．

CCD と GLC の組み合わせにより，目的とするスケーラブルな画像アノテーションシステムが完成する．最終的に，提案システムを 1,200 万枚の画像データセットへ適用し，その有効性を示す．

Contents

List of Figures

List of Tables

Chapter 1

序論

1.1 研究の背景

制約のない実世界の画像を計算機に認識させ，言語により記述させる技術を一般画像認識 (generic image recognition) [158; 235] と呼ぶ[1]（図 1.1）．我々人間も，外界の多くの情報を視覚から認識し行動決定を行っているように，実世界で行動する知能システムにおいて一般画像認識は必要不可欠な機能の一つであるといえる．

一般画像認識は，科学的側面・工学的側面（アプリケーション）の両面において価値の高い課題であるため，さまざまな分野の研究者の興味を引いてきた．科学的な価値は，人間の画像認識機能を機械的に実現することへの興味そのものにある．人間の外界認識能力は，認知心理学・脳科学など幅広い研究分野において議論の対象となっており，計算機科学の立場からのアプローチも重要な示唆を与えうると考えられる．

さらに，一般画像認識は人工知能分野における根本的な問題の一つである記号接地問題を担うものであるから，その解決がもたらす工学的価値は計り知れない．アプリケーションとしては，ロボット・自動車など実世界で行動するエージェントの実世界認識機能としての応用がまず考えられる．また，ライフログやサーベイランスシステムへの応用，インターネット画像の検索・フィルタリングなども挙げられる．

しかしながら，一般画像認識はその長い歴史にも関わらず未だ実現されておらず，コンピュータビジョンにおける究極の目標の一つと捉えられている．一般画像認識の難しさは，対象とする画像の多様性と物体の多さに起因する．まず，ある一つの特定物体（インスタンス）の画像であっても，視点による写り方や背景の変化，照明条件の変化，他物体との干渉などによりその見え方（アピアランス）は大きく変化する（図 1.2）．さらに，一般的な物体カテゴリは多様なインスタンスを包含するため，考慮すべきアピアランスの幅はさらに広がる．例えば，

[1] 一般物体認識と呼ばれることもある．なお，一般物体にはいわゆる物体のみならずシーンや形容詞などの抽象的な事物も含まれる．

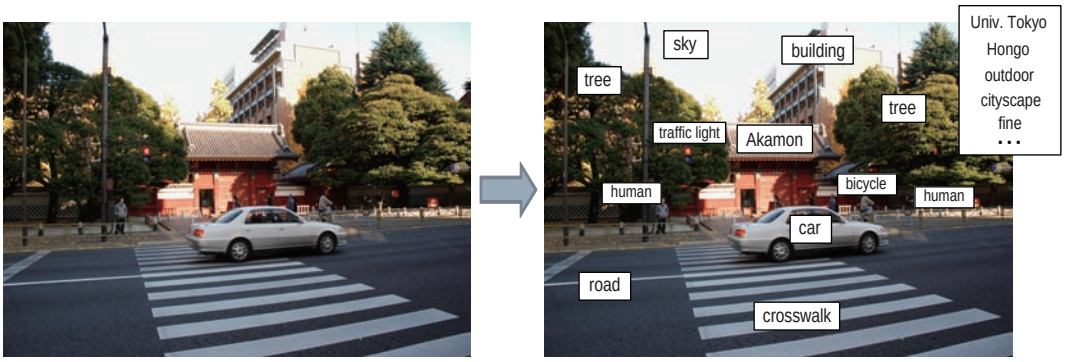


Figure 1.1: Illustration of generic image recognition. Several meanings (symbols) can be extracted from a single image.

我々人間は図 1.3の画像を見てこれらが全て“椅子”であると判断できるが，その色や形は変化に富んでいる．また，一般画像認識は特定のカテゴリだけでなく，人間と同様に広く世界に存在する事物を認識することを目標とするものである．一説によれば，人間は視覚のみから数万種類の一般物体を認識可能であると言われており [16]，同様の機能を計算機上で実現するためにはさらに多様な画像に対応する必要があることが分かる．

2 章で詳しく議論するが，一般画像認識は扱う問題の広範さと抽象性の高さから，明示的に認識のためのプロトタイプを設計することは困難であることが歴史的に示されている．このため，現在は統計的機械学習に基づくアプローチを中心に発達しており，大量の事例データからの学習が実用化への鍵と考えられている．しかしながら，従来開発されてきた一般画像認識の手法はスケーラビリティに乏しく，大量のデータからの学習は著しく困難であった．この問題が，一般画像認識の実用化を阻む最大の障壁であった．例えば，現在一般画像認識において標準的に用いられる識別器は非線形 SVM であるが，その計算コストは表 1.1のようになる．特に，学習にかかる計算コストが学習サンプル数に伴って著しく増大することが分かる．また，特に大規模な問題においては全ての学習サンプルをメモリ上に保持することは難しいため，評価関数の反復計算の度に全サンプルに対するストレージアクセスが発生する．この場合，学習の速度はさらに著しく低下する事が容易に予想される．このように，大規模一般画像認識は単にデータ数が増えるだけの問題ではなく，質的なブレークスルーを必要とする新しい研究領域であるといえる．

1.2 研究目的

本研究では，学習サンプル数に対しスケーラブルかつ高精度な一般画像認識アルゴリズムの開発を目的とする．より具体的には，まず画像全体への複数単語のラ

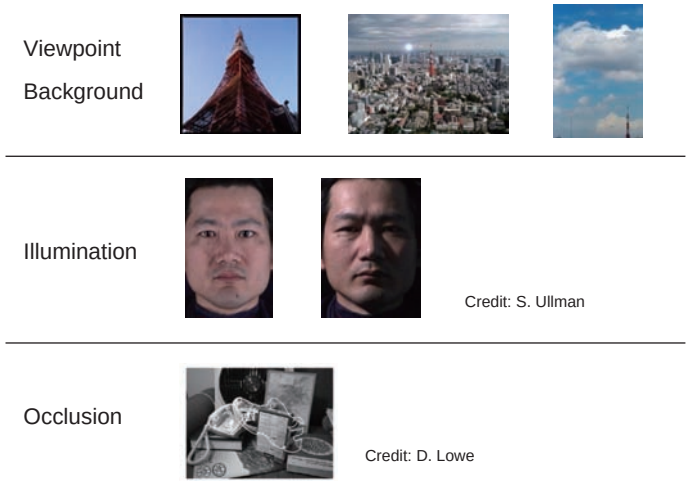


Figure 1.2: Appearance changes due to various real-world conditions.

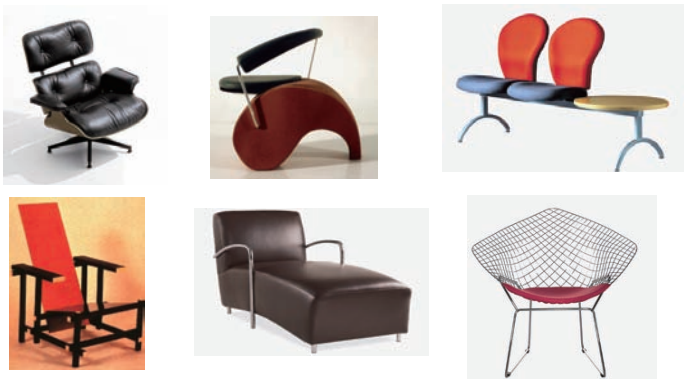


Figure 1.3: A variety of "chairs". Credit: Li Fei-Fei *et al.* CVPR'07 object recognition tutorial slides.

Table 1.1: Computational complexity of a non-linear SVM. N is the number of training samples.

	Complexity	Memory
Training	$O(N^2) \sim O(N^3)$	$O(N^2)$
Recognition	$O(N)$	$O(N)$

ベルづけ（画像アノテーション）を行う画像認識手法の開発を行う．さらに，前記認識手法との相性を考慮した，強力な画像特徴量を抽出する枠組みの開発を行う．

1.3 本論文の構成

本論文の構成について述べる（図 1.4）．1 章では，本論文の背景と目的を述べた．2 章では，一般画像認識研究の歴史と現状について述べ，本論文で開発するシステムの設計を行う．具体的には，画像アノテーションとよばれるタスクに取り組むことを述べる．3 章では，画像アノテーションに関する先行研究についてのサーベイを行う．4 章では，学習サンプル数に対してスケーラブルな画像アノテーション手法の開発を行う．5 章では，開発した画像アノテーション手法の評価実験を行う．6 章では，画像アノテーション手法と相性のよい画像特徴量の開発を行う．これにより，提案する一般画像認識システムが完成する．7 章では，実際に大規模なデータセットを用い，提案システムの評価実験を行う．8 章では，本論文の結論と展望について述べる．

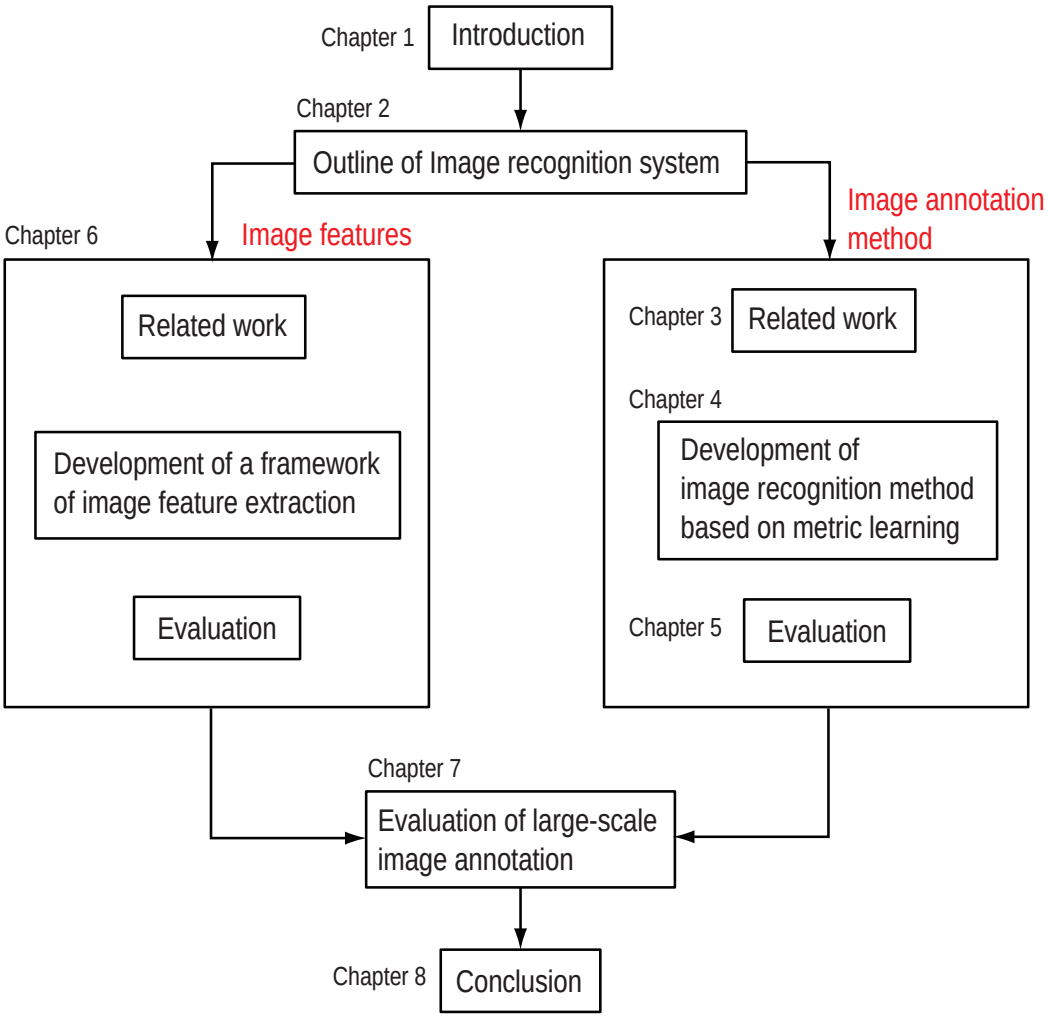


Figure 1.4: Structure of the thesis.

Chapter 2

画像認識アルゴリズムの構成

2.1 画像認識の歴史と現在の潮流

まず，画像認識研究の歴史を簡潔に述べる．人間は画像から数万種類の物体カテゴリを認識可能であると言われる [16] 一方，計算機にとってはたった一つの物体であっても正確に認識することは困難である．このため，計算機による物体認識（画像認識）の実現は古くから研究者の関心を引いてきたテーマであり，1950 年代から現在に至るまで半世紀以上の歴史がある．

1950 年代は，文字や指紋などの 2 次元パターンの認識から研究が始まった．この時代は，統計的パターン認識によるアプローチが主流であり，モーメント特徴など幾何的な不変性を持つ特徴量を利用するものが多く見られる [84]．このアプローチはその後，数十年にわたりモデルベーストなアプローチに取って代わられるが，1990 年代に再び注目を集めることになる．

1950 年代半ばには，Marvin Minsky と John McCarthy による人工知能のパラダイムが浸透し，統計に基づくアプローチは急速に廃れていった．この新しいパラダイムにおいては，人間の認知機能を数学的なツールによって十分に定式化するために，まず世界の記述を徹底的に単純化することから出発する．ビジョンにおいては，“積木の世界”(the blocks world) [162] から研究が始まった．この世界では，扱う物体は多面体に限定され，背景は一定である．このように単純化された世界において，任意の視点における 2 次元画像から元の 3 次元物体配置を推定する計算論を構築することが目的であった．その後，blocks world のような考え方は，より一般的な曲面形状を扱えるように，線画解釈 [74] の研究へ発展していく．しかしながら，一般的な実世界画像においてはそもそも線画をロバストに抽出すること自体が困難であった．このため，一般化円筒 [17] により対象を要素分解するアプローチが注目されるようになったが [26; 222]，ボトムアップなセグメンテーションにより要素抽出を行うことは依然として困難であった．これは，実画像認識における本質的な問題であり，その後も多くの試行錯誤がなされたが，いずれも解決には至らなかった．

一般化円筒による物体認識手法の多くは，モデルベースドな物体認識 [159] に

分類される．モデルベースド物体認識は，対象物体の 3 次元幾何形状モデルを知識として用意しておき，取得画像とマッチングを行うことにより行われる．しかしながら，形状を直接認識の手掛かりとするため，基本的にはある特定の物体の認識しか行えない．このため，一般的な物体カテゴリの認識を行うためには，無数に存在する形状モデルを全て与える必要があり事実上不可能である．また，“海”や“道路”などそもそも明確な形状モデルを定義できない一般物体については対処できない．知識ベースのアプローチによる物体認識の他の例としては画像エキスパートシステム [39] などが挙げられるが，同様の問題によりいずれも成功することはなかった．

このように，幾何形状に基づくモデルベースド認識が閉塞する中，1990 年代の初頭には再び統計的な手法が盛んに研究されるようになり，現在に至るまで中心的なアプローチとなっている．背景として，計算機の著しい発達により，1950 年代には計算困難であった統計解析手法が一般的に利用可能になったことや，SVM などの強力な識別手法が普及したことが挙げられる．鍵となったのが，3 次元復元は基本的に考慮せず，2 次元の外観のみで認識を行う appearance-based と呼ばれる考え方である．モデルベースド認識では人間がモデルを設計していたのに対し，統計的な認識手法では事例集合から自動的に重要な情報の抽出を行う．代表的なものとして，画素値のベクトルを固有空間法を用いて圧縮し，圧縮されたベクトルを特徴量とみなす固有顔法 [185] や，これを一般物体へ応用したパラメトリック固有空間法 [140] も提案された．また，画像のテクスチャや色などの統計的傾向をよく表現する画像特徴量の開発も行われた．代表的なものとして，カラーヒストグラム [165; 180] が挙げられる．カラーヒストグラムはシンプルな特徴であり非常に高速に抽出可能であるため，内容に基づく画像検索 (Content-Based Image Retrieval, CBIR) など大量の画像を扱うタスクにおいて広く用いられてきた [176]．

1990 年代の手法の多くは単純な大域的画像特徴を用いていたため，オクルージョンやスケール・向きなどの変化に弱いという欠点があった．2000 年代入ると，局所特徴に基づくアプローチが成功し，この問題はある程度解決されるようになった．これを可能としたのは，SIFT 特徴 [124; 125] に代表される局所特徴記述手法の進歩である．局所特徴はある注目点（特徴点）周りの小領域を記述するものであり，一般に回転，スケール，照明などの変化に不変性（またはロバスト性）を持つように設計される．特徴点検出の手法自体は，コーナー点検出手法 [11; 76] を中心に古くから研究されているが，SIFT 特徴は特徴点の検出から正規化，特徴記述，サブピクセル推定に至る一連の仮定を緻密に設計した初めての手法である．SIFT 特徴により，実画像において回転，スケール，照明などの変化に対してロバストに特徴点をマッチングさせることが可能となった．また，特徴点ベースのマッチングであれば，オクルージョンに対してもある程度のロバスト性がもたらされる．このような利点から，局所特徴による画像表現は現在に至るまでコンピュータビジョンの広範な領域において革新的な進歩をもたらしている．SIFT 特徴は，その後も多くの研究者によりさまざまな改良を加えられている [98; 179]．また，SIFT の他にも多数の局所特徴記述子が提案されている [10; 28; 190]．

もともとアピアランスベースドな画像認識は，インスタンスレベルでの物体認識である特定画像認識[1]を中心にスタートしており，前述の局所特徴も基本的には特定画像認識向けに設計された技術である．一方，一般画像認識においても局所特徴は有効であることが分かり，多くの研究で活用されるようになった．まず，物体の局所的なアピアランスとその位置関係で物体をモデル化する part-based アプローチが試みられた．代表的なものとして，数個の局所特徴の値と位置関係を用いる constellation model [55; 62] が挙げられる．このモデルでは，物体カテゴリごとに局所特徴の大まかな空間配置を学習することにより，個々の画像のプロトタイプからの変形（deformation）に対してロバストに認識が行える．しかしながら，一般画像認識においては数個の局所特徴のみでは各画像の情報を安定に表現することは難しい．これは，局所特徴は基本的に視覚的顕著性（saliency）のみに基づいて検出されるため，必ずしも認識において重要な点を検出するとは限らないためである．また，constellation model の学習では，局所特徴の空間配置を膨大な状態空間の中から brute-force に学習するため，学習にかかる計算コストが極めて高いという問題点もあった．

一方，位置情報を捨てた大量の局所特徴の統計量を用いた画像表現の有効性が示されるようになってきた．代表的なものは，量子化に基づくアプローチである bag-of-visual-words (BoVW) [40] である．これは，もともと自然言語処理の分野で用いられてきた bag-of-words (BoW) [130] の考え方をコンピュータビジョンへ応用したものである．まず，各学習画像から大量の局所特徴（数百個〜数千個）を抽出する．学習データセットの局所特徴をクラスタリングすることにより，いくつかのクラスタ中心 (visual words) が得られる．最終的に，一枚の画像は visual words のヒストグラムによって表現される．BoVW を用いた画像認識は汎用性が高く，さまざまなタスクで安定によい性能が得られることが分かり，その後精力的に研究がなされている．BoVW の成功の理由の一つは，個々の局所特徴の位置情報を無視するという発想の転換にあるが，一方で画像全体のおおまかな位置情報は認識に有効であると考えられる[2]．そこで，spatial pyramid matching (SPM) [110] では，画像を階層的にグリッド分割し，領域ごとの BoVW を比較することで位置情報を認識に活用した．シンプルな手法であるが，オリジナルの BoVW から大きく性能向上を行うことに成功している．現在，SIFT BoVW + SPM + SVM は一般画像認識におけるデファクトスタンダードのアルゴリズムとなっている．

これらは一般画像認識における必要不可欠なツールとして，現在も盛んに研究され着実に改良がなされているが，基本的には 2000 年代に一通りの完成を見ている．2010 年代に入り，これらのツールの上にどのように他の手法やリソースを統合していくか，というフェーズに関心が移りつつある．例えば，画像中の物体がなす文脈（context）の利用 [82; 184]，カテゴリ間階層構造の利用 [45; 71]，未知カ

[1]特定物体認識とも呼ばれる．一般画像認識，特定画像認識の差異については次節で詳しく議論する．

[2]例えば，空は画像上部に出現し，海は下部に出現する可能性が高い，など．

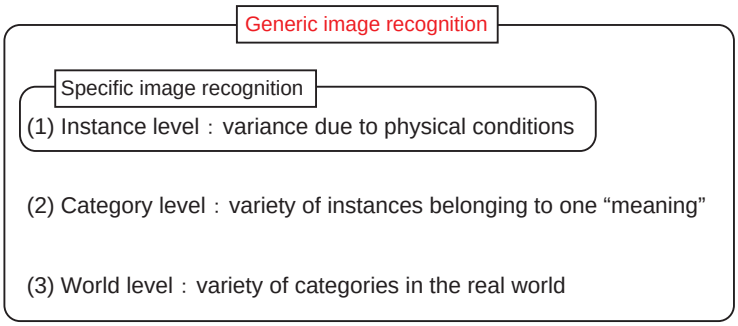


Figure 2.1: Three levels of variance in generic images.

テゴリの発見・適応 [112]，WordNet [59] などの外部リソースの利用 [46; 103; 182]，ユーザフィードバックの利用 [174] などの研究がなされている．2000 年代の初期の研究では，機械学習や自然言語処理などの他分野の学習手法を画像認識向けに改良するものが多かったのに対し，近年では画像ならではの問題を扱う研究が増えてきているといえる．

2.2 一般画像認識

2.2.1 一般画像認識 vs 特定画像認識

ここでは，一般画像認識と対を為す概念である特定画像認識について触れ，一般画像認識との差異を述べる．また，一般画像認識の持つ本質的な難しさについて説明する．特定画像認識とは，インスタンスレベルでの同一物体の認識を指す．例えば，一般画像認識における“車”の認識では，さまざまな乗用車，トラック，バスなどを対象として車か否かを判断するのに対し，特定画像認識では“この物体はトヨタカローラか否か”のみを判断する．

両者の本質的な違いは，アピアランスの変動の範囲により説明される．1 章で述べたように，これは大きく分けて 3 つのレベルへ分割することができる（図 2.1）．このうち，特定画像認識が対象とするのは (1) の問題である（図 1.2）．例えば，画像を取得する視点の変化や背景などの状況，他の物体とのオクルージョン，照明条件の変動など多くの不確定要素が考えられる．しかしながら，いずれも基本的に実世界における物理的な条件により説明できることが特徴である．このような物理的変動に対しては，例えば前述の SIFT 特徴などのようにある程度頑健な局所特徴量をトップダウンに設計することが可能であり，大きな成功を収めている．近年では，Google Goggles [1]などの商用に近いアプリケーションも登場している．

[1]http://www.google.com/mobile/goggles/

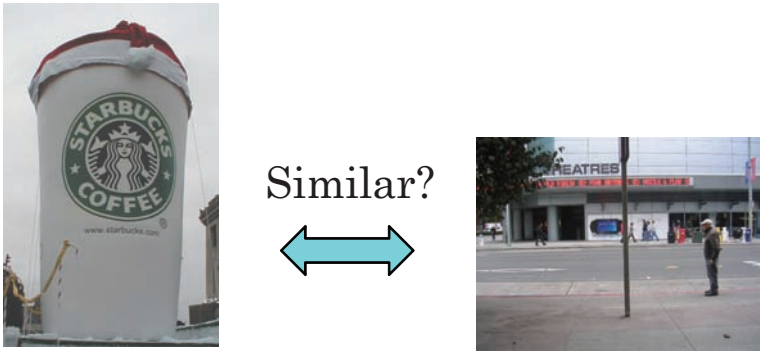


Figure 2.2: (a) A query image. (b) The closest image in terms of the color histogram. Credit: Jing et al. [94].

これに加えて一般画像認識では (2)，(3) の問題も考慮しなければならない．計算論的に特に難しいのは (2) の問題である．例えば，図 1.3が示すように，同じ“椅子”であっても世の中には多種多様な椅子が存在しており，それぞれアピアランスは大きく異なる．我々人間がこれらを同じ“椅子”として認識できるのは，経験に基づき“椅子”とはどのようなものであるかを学習しているためであり，単純に物理的な特性や拘束条件のみでは説明できない．このような，画像の物理的なアピアランスと人間が欲する意味との乖離は semantic gap [176] と呼ばれており，一般画像認識を特徴づける課題となっている．

2.2.2 Semantic Gap

図 2.2の二つの画像を例にとる．人間がこれらの画像から読み取る意味は大きく異なると考えられるが，カラーヒストグラムなどの単純な画像特徴量をとると両者はほぼ同じ値となる．これは，両者の持つ意味の弁別が著しく困難であることを意味する．これが semantic gap の問題に他ならない．Semantic gap を緩和するためには，以下の二つの要素が重要である．

表現能力の高い画像特徴の利用

まず，インスタンスレベルでの画像の弁別性能が十分でなければ，カテゴリの認識も困難である．例えば，図 2.2の二つの画像を弁別するためには，エッジヒストグラムなど形状を表す特徴量を同時に用いる必要があると考えられる．もちろん，実世界には無数の物体が存在するため，対象物体に依存しない汎用的な画像認識システムを実現するためには，異なる特性を持つできるだけ多くの画像特徴を用いる必要がある．最終的に，一枚の画像を表す画像特徴は非常に高次元とな

る．この時，視点・照明変化など実世界の物理的拘束条件に対する不変性・頑健性を有する画像特徴量を用いれば，特定画像認識は実現できる．

統計的学習手法による距離計量学習

一般画像認識はインスタンスではなくカテゴリとの近さを測るものであるため，カテゴリ内サンプルの分散の大きさが問題となる．カテゴリ内サンプルの分散は物理的条件とは必ずしも関係ないため，あらかじめ不変な特徴量を設計することは困難である．図 1.3の“椅子”カテゴリを例にとる．椅子の弁別には，座面の平面的な形状が本質的に重要な特徴であると予想できる．しかしながら，色など他の特徴はサンプルによって大きく異なっている．このため，単純に代表サンプルと画像特徴量によるマッチングを行うだけでは，色などのカテゴリの弁別には寄与しない成分が邪魔をするため，“椅子”カテゴリとの概念的な近さを測ることはできない．この問題が一般画像認識において解決すべき重要な課題である．

最も単純には，画像特徴空間を生成的に満たすように大規模な学習サンプルを用いるアプローチが考えられる．例えば，可能な限りあらゆる種類の椅子の画像を登録しておくことで，任意の椅子画像について画像特徴空間上で近接する学習サンプルが存在する確率が向上する．すなわち，データを十分に増やすことにより，画像特徴間の距離がそのまま意味的な距離に近づくと期待できる．ナイーブに知識データを増やすことで問題解決を測るアプローチは，前述の画像エキスパートシステムなどにおいては一度破綻しているものの，現在は Web の発達により当時とは比較にならないほど大規模かつ良質なデータを得られるようになり，再びその有効性が示されるようになっている [182; 203].

しかしながら，前述のように画像特徴空間は非常に高次元な空間であるため，これを生成的に満たすほどの教師付きサンプルを用意することは依然として現実的であるとは言いがたい．そこで，統計的機械学習の手法によりあらかじめ重要な特徴を選択しておくことが重要となる．すなわち，画像の意味的な距離が定義される新しい低次元の空間を生成するアプローチである．先の“椅子”の例を再び用いる．まず，椅子の正例と負例の画像をそれぞれ一定数用意する．これらを，判別的な識別器（例えば SVM など）に入力することで，正例と負例を弁別するために有効な特徴が自動的に選択され，判別のための超平面が張られる．識別器の内部は一般にブラックボックスであるが，図 1.3の“椅子”カテゴリの場合，椅子・非椅子を分離する超平面までの距離は形状特徴によって測られ，色特徴には依存しない構造が得られると予想できる．これは，形状特徴に基づく低次元の新しい空間が定義されたことを意味しており，この空間における距離は“椅子”カテゴリへの近さを反映していることが分かる．このように，機械学習による特徴選択を行うことで，画像の意味的な類似度を反映させた新しいサンプル間の距離計量を得ることができる．

2.2.3 一般画像認識の諸問題

一般画像認識は大きく分けると，(1) 画像全体へのラベル付与，(2) 画像中の特定領域へのラベル付与，の 2 つの問題に分割できる（図 2.3）．(2) では，領域とラベルの対応関係が明確な，比較的堅い物体認識を対象とする場合が多い．これに対し (1) では，物体認識のみならずシーン認識などのように，必ずしも領域とラベルの対応関係が明確でない対象を扱うことも多い．また，(1) における教師データは画像と対応するラベルのみ与えればよいが，(2) では対象の切り出しを行う必要があるため，学習データの作成が一般に高コストである．

画像全体へのラベル付与

この枠組みでは，システムは画像領域とラベルの直接的な関連性を推定する必要はなく，画像全体に対し付与するラベルを決定すればよい．このうち，画像に対し一つのカテゴリラベルのみを排他的に与えるタスクを**画像カテゴライゼーション**と呼ぶ．これは，一般画像認識において最初期から研究されてきたテーマである．最もシンプルな枠組みであり評価が明快であるため，画像特徴量や学習手法などの基本的な技術を発展させる土壌となった．

これに対し，一枚の画像に対し複数のラベルを与えるタスクを**画像アノテーション**と呼ぶ．アノテーションはカテゴライゼーションを包含する枠組みであり，より一般的な問題設定であるといえる．カテゴライゼーションでは，目的のラベル以外のサンプルを全て負例として扱えばよいのに対し，アノテーションではラベル間の関係性を考慮する必要がある．このため，アノテーションはカテゴライゼーションに比べやや複雑な問題であり，さまざまなアプローチからなる学習手法が提案されている．詳しくは，3 節にて述べる．

カテゴライゼーション・アノテーションはともに，閉じた小規模データセットからの学習については一定の完成を見せており，現在は，大規模なデータセットへの適用 [46; 182; 203] に向けた手法の研究が進んでいる．これらのタスクでは比較的グラウンドツルースが得やすく，近年では Web 上の大量の弱ラベル付きデータを用いることが出来るため，研究は飛躍的に発展しつつある．

画像中の特定領域へのラベル付与

物体検出（ディテクション）では，画像中の各物体のカテゴリとその存在する領域を判定する．物体の位置座標の取得が主な目的であり，領域は大まかに長方形や凸多角形などで近似される場合が多い．近年急速に実応用が進んでいる正面顔検出 [194] などもこの枠組みに含まれる．基本的には，位置・大きさを変えながら検出窓を走査し，各窓内で対象の有無を判断する sliding windows のアプローチにより実現可能である．しかしながら実際には，同一物体の 2 重検出を防ぐ non-maxima suppression や，物体の種類による領域干渉の可・不可の判断など，ディテクションならではの課題が多く存在する [47]．また，総当たりで全ての窓

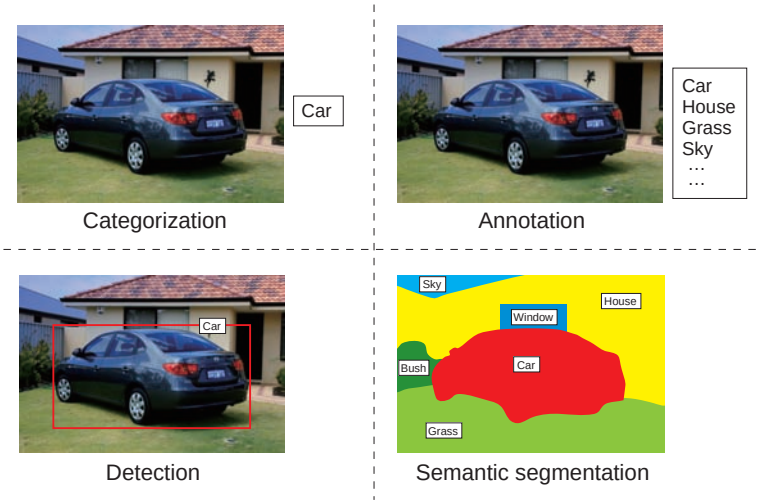


Figure 2.3: Various tasks of generic image recognition.

を探索すると，組み合わせの数が膨大になるため計算が現実的に不可能となる．このため，枝がりを用いた効率のよい探索方法 [106] や，ハフ変換による投票を用いた高速な検出 [113; 128] など，さまざまなアプローチが提案されている．一方，**画像セグメンテーション**では，各物体のピクセルレベルの領域推定を行う．ここでのセグメンテーションとは，一般的なセグメンテーションが指すボトムアップな領域分割に加え，領域の意味の認識まで含むものであり，難易度は高い．近年では conditional random fields [105] を用いたアプローチが標準となっている [104; 172]．

2.3 学習用画像コーパス

2.3.1 小規模データセット

一般画像認識は，認識対象の選び方や学習画像，テスト画像のセットアップ等により難易度が大きく変化する．2000 年代初頭までは各研究者がそれぞれセットアップを行っていたが，分野の発展に伴い，いくつかの標準的なベンチマークデータセットが登場するようになった．2006 年以前のデータセットについては，[157] に詳しく述べられている．

画像アノテーションの分野において古くから用いられてきたのは，Corel5K [51] と呼ばれるデータセットである．これは，Corel 社が商用で販売していた Corel Stock Photo Library という画像ライブラリを利用したものである．ライブラリには合計で 8 万枚の画像が含まれ，各画像には検索のための複数のキーワードが付与されている．画像はおおまかなカテゴリごとに 100 枚ずつ用意されている．こ

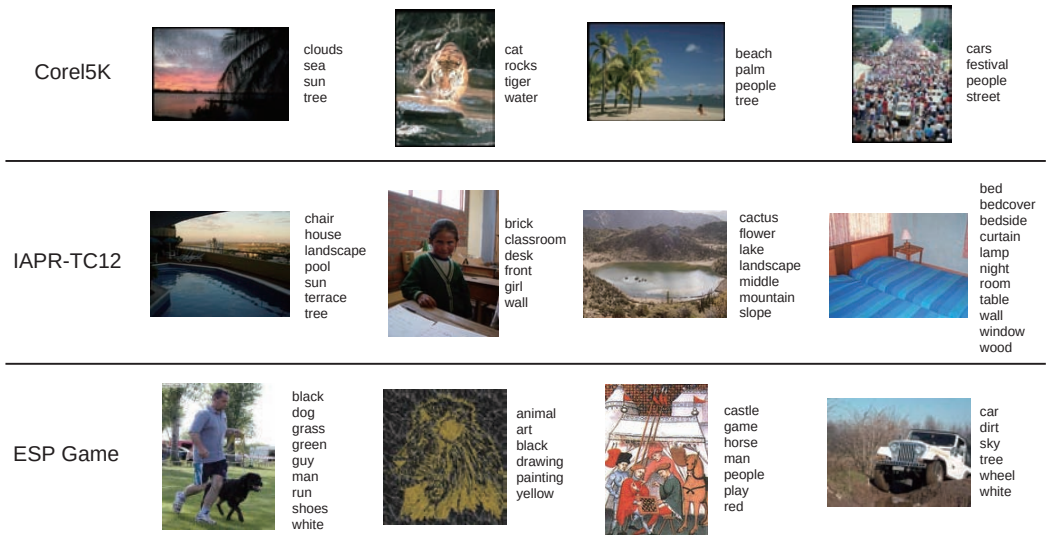


Figure 2.4: Three standard benchmarks for image auto-annotation. Top: Corel5K [51]. Middle: IAPR-TC12 [129]. Bottom: ESP Game [129].

のうち 50 カテゴリ，合計 5000 枚の画像を選んだものが Corel5K である．学習する単語は 371 種類であり，画像数に対し比較的単語数が多いセットアップとなっている．図 2.4上段に例を示す．Corel5K は標準的に用いられてきたベンチマークではあるが，比較的小規模なことや，画像が似通っていることなどから容易なタスクであると考えられており，現在では単独で評価に用いられることは少ない．2008 年の ECCV で発表された Makadia らの研究 [129] では，Corel5K に加えて新たに IAPR-TC12（図 2.4中段）, ESP game（図 2.4下段）と呼ばれるデータセットを評価に用いており，現在ではこの 3 つで性能を評価するのが一般的なプロトコルとなっている．IAPR-TC12 は，ImageCLEF [1] と呼ばれる，異種言語間での画像検索のワークショップで用いられたデータセットである．また，ESP game は，後述する ESP collaborative image labeling task [195] と呼ばれるオンラインゲームにより蓄積されたラベル付き画像の一部を用いたデータセットである．

一方，カテゴライゼーションにおいては，カリフォルニア工科大学の Caltech-101 [56; 57] というデータベースがデファクトスタンダードであった．これは，主に Google Image Search を用いて人手で集めた 9144 枚の画像から構成される．101 種類の物体クラスと背景クラスから構成され，それぞれについて 31 枚から 800 枚のサンプル画像が与えられている．Caltech-101 の例を図 2.5に示す．このように，各画像が一意に特定の単語クラスに属するという枠組みになっている．物体のスケールや向きなどはおおむね揃えられているが，色やテクスチャはバラエティに富んでいると言える．2007 年には，対象クラス数を 256 クラスへ増やした Caltech-256 が登場している．Caltech-101 と比べてクラス数・画像サンプ

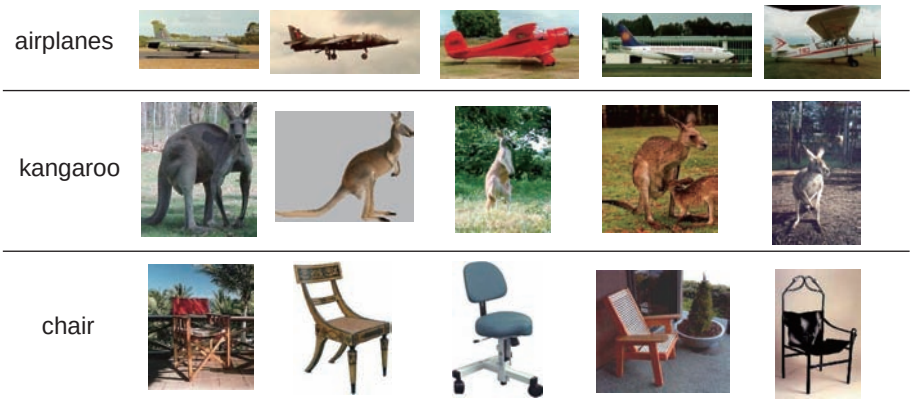


Figure 2.5: Example images from Caltech-101 dataset [56; 57].

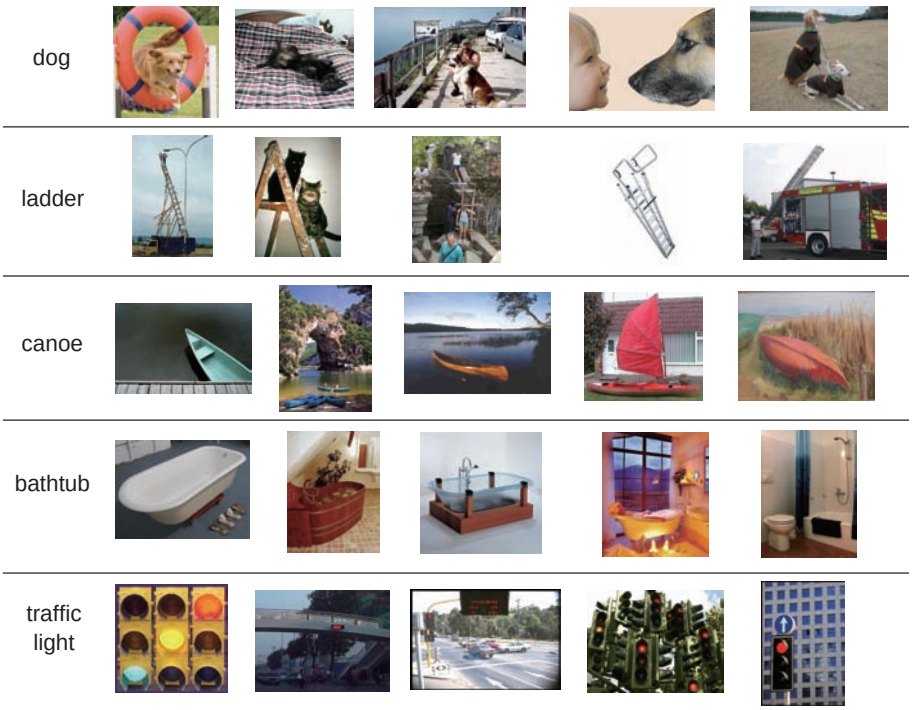


Figure 2.6: Example images from Caltech-256 dataset [70].

ル数が増えたのみならず，各クラス内の画像のばらつきも大きくなっており，難易度は上がっている（図 2.6）．現在の最新の手法では，Caltech-101 では 80%，Caltech-256 では 50%程度の認識率が達成されている [66; 217]．もうひとつの標準的なデータセットとして，PASCAL Visual Object Classes (VOC) [52; 53] が挙げられる．これは，VOC Challenge と呼ばれる画像認識のワークショップで用いられるベンチマークである．対象クラス数は 20 クラスと少ないが，カテゴライゼーションに加えディテクションのタスクも行われる．現在では，主にディテクションアルゴリズムのベンチマークとして用いられることが多い．

2.3.2 大規模データセット

Crowd sourcing

前述の小規模データセットは手法自体の評価を主な目的としており，学習後に得られる識別システムの有用性は必ずしも考慮していない．一般画像認識の実現には，実世界の広範なアピアランスをカバーするデータセットを構築する必要があり，膨大な人的コストを必要とする．このため，データセットの構築に多数の人間を取り込む必要があると考えられる．

このようなプロジェクトのさきがけとしては，LabelMe プロジェクト [163] が挙げられる．インターネットを介して提供されるアノテーションツールを用い，不特定多数のユーザが画像中の複数の物体の領域ラベリングを行う．また，画像自体もユーザが自由にアップロードし，データ数を増やしていくことが可能である．2009 年には，動画像を対象とした LabelMe Video [221] も登場している．しかしながら，画像ラベリングを行う動機はユーザの善意に依存している点が問題である．大量の画像にラベリングを行う作業は時間と忍耐を要するため，画像認識関連の研究者以外の参加を期待するのは難しいといえる．

これに対し，ラベリング自体をゲーム化することで，ユーザの動機付けを行うアプローチが提案されている．ESP game [195] では，以下のような画像アノテーションのゲームを行う．まず，ネットを介して匿名の二人のユーザが選ばれる．システムは，二人のユーザに同時に同じ画像を見せる．ユーザは，画像に適切な単語を自由にラベル付けするが，この時パートナーと同じ単語を選ぶと得点が得られるルールとなっている．このようにして二人のユーザが共通にアノテーションしたラベルは信頼性が高いラベルであると期待できる．このようなゲームをユーザを変えて繰り返し，高い頻度でラベルづけされた単語を最終的にグラウンドツルースとして採用する．同様に，Peekaboom [196] では画像の物体領域抽出のゲームを提案している．このように，ラベリング作業に娯楽性を持たせることで，人的コストを無償で確保することに成功しており，数百万枚規模のデータセットが構築されている．しかしながら，ユーザの目的はあくまでゲームであるため，精度の高いアノテーションを得ることは難しい．また，基本的にユーザはアノテーションを自由に行うため，ラベルづけが基本的な単語に偏り多様性が出にくい傾向がある点も問題となる．

Lotus Hill dataset [219] は，ImageParsing.com が提供する大規模データセットであり，有償で雇用された専門の技術職員により画像ラベリングがなされている．ラベリングに対する金銭的な対価を与えることで質の高い作業を保証しており，他のデータセットに比べて精度の高い，行き届いたアノテーションがなされている．しかし，当然ながらサービスを支える人的資源には限界があり，より大規模なデータを処理するのは困難である．また，提供するデータの大部分は有償であるため，学術研究を促進させる上では必ずしも有用であるとはいえない．

近年，大きな注目を浴びているのが，Amazon Mechanical Turk (AMT) の利用である [178]．AMT はインターネット上でのジョブポスティングサービスであり，不特定多数のユーザに画像ラベリングタスクを依頼することが可能である．タスク内容と支払い額などから興味を持ったユーザがラベリングに参加することになる．AMT を利用することにより，十分な賃金を用意すれば，多くの人間に高いモチベーションを持ってラベリングを行わせることが可能となる．また，タスクの内容は自由に設計できるため，研究の用途に応じて様々なデータセットを構築できる．AMT は既に多数の研究に利用されているが，一般画像認識において現在もっとも大規模なプロジェクトは ImageNet [46] である．ImageNet は WordNet [59] の概念構造に従って構築が進められているデータセットであり，2010 年 7 月現在，15,589 クラス 1120 万枚のサンプルが既に蓄積されている．2010 年の ECCV では，その一部を用い，10,000 クラスのカテゴリ識別を行った例が報告されている [45]．同時に，1,000 クラスのカテゴリ識別の認識性能を競うワークショップも開催されている [13]．このような大規模な一般画像認識が定量的に評価可能となったことは意義深い．AMT を利用した他のデータセットとしては，800 クラスの一般シーンカテゴリからなる SUN データセット [214] が挙げられる．

Web Image Mining

Crowd sourcing の発達により，比較的良質かつ大規模な一般画像コーパスを得ることが可能となってきている．しかしながら，データセットの構築は人手に依存している点には変わりなく，処理できるデータ量には限界がある．Web 上には，更に桁違いの数のデータが存在しており，その数は指数的に増加している．例えば，2010 年現在，Flickr という写真共有サイトには，既に 40 億枚以上の画像が蓄えられている [203]．また，Google は更に多くの画像をインデックス化している．このような Web 上の画像には，テキストなど何らかの意味的な情報が付与されている場合が多く，これらをマイニングすることで汎用性の高い一般画像認識を実現することが検討されている（Web 画像マイニング）[61; 213; 234]．Web 上には，様々な環境・条件下で取得された画像が無数に存在し，広範なアピアランスを学習することが期待できる．また，付加される意味情報も異なる多くの人間が用意したものであるため，データ全体としては個人の主観に依存しない，バイアスの少ないものになっていると期待できる．

Web 上の大量データを用いた統計的学習は，自然言語処理の分野では 2000 年代初頭に始まっており，用いるデータを増やすにつれて対数スケールで性能が向

上することが示されている [6]．同様のアプローチを大量の Web 画像に適用し，一般画像認識を行った初期の研究としては，AnnoSearch [202] が挙げられる．これは，未知画像と共にユーザが最低一つのキーワードを入力することで，半自動的にラベリングを行うシステムである．

Web からの画像知識獲得を目指す研究の多くは，既存のテキストベース画像検索エンジンに立脚する．すなわち，学習させたい単語を検索エンジンにかけることで関連する大量の画像を取得し，これをもとに識別器を構築する方法論である．この場合，システムに習得させる単語群をあらかじめ人間が設計する必要があり，従来は比較的小規模な単語数での基礎的な実験が試みられてきた [61; 234]．近年では，データセットの構築そのものに WordNet [59] のシソーラスを利用し，大規模に Web データを収集することで，広範な知識の獲得を目指すものが増えている [46; 182]．さらに，Normalized Google Distance [38] や Flickr Distance [213] のように，自律的に Web から構築・更新されるオントロジーを用いることで，自律的な画像知識獲得を可能とする研究もあらわれている [122]．

問題点は，テキストベース画像検索エンジンの性能がシステムのボトルネックとなり得る点である．現在の画像検索エンジンでは，オントロジーを有効に活用した検索は難しく，各単語を単純にキーワードとして検索する場合が多い．実際には，同綴異義語・ノイズなどの影響により望む対象と関わりのない画像も多くヒットするため，質の良いデータセットを得ることが難しい．このため従来研究では，検索上位画像の利用などの経験則の利用 [61; 122]，クラスタリングによるノイズ除去 [234]，画像類似度を用いた重みづけ [122] などの，検索結果のフィルタリングによる質の向上が図られているが，本質的な問題は検索段階の性能限界であるため，アドホックな後処理で解決を図るのは難しいといえる．このため，ユーザとのインタラクションの利用 [14]，半教師付き学習の枠組みによる逐次的な識別器の構築 [114]，画像・テキスト両方の類似度を用いたリランキング [167]，スパムタグの除去 [54] など現在に至るまでさまざまなアプローチが模索されている．

このように，Web 画像マイニングにおいて得られる画像コーパスはノイジーなものとならざるを得ないが，そのかわり極めて大規模である点が最大のメリットである．Torralba ら [182] は 8,000 万枚の画像を検索エンジンを用いてダウンロードし，単純な画像特徴量を用いた k 最近傍法による画像認識を行った．データセットは多くのノイズを含む雑多なものであるにも関わらず，学習に用いるデータ量に対して対数スケールで識別性能が向上することを実証した．彼らの報告によれば，データ量が増えるにつれて，クエリ入力に対してアピアランスが十分に類似した画像が存在する可能性がコンスタントに上昇する．さらに，類似度がある閾値を超えると，それらの画像が同じ概念（クラス）に属する可能性が飛躍的に高まるという知見が得られている．同様の結果は，20 億枚の Web 画像を用いてアノテーションを行った ARISTA プロジェクトにおいても報告されている [203]．このことは，画像特徴空間を十分に満たすサンプルデータを用意することで，画像のアピアランスの類似度が意味的な類似度に近づくことを示唆しているといえる．すなわち，Web スケールの大量データを用いることは semantic gap を解消

するための重要なアプローチの一つであると考えられる．

2.4 開発するアルゴリズムの要求機能・設計

本章では，一般画像認識を中心に画像認識研究の現状を論じた．一般画像認識においては semantic gap が本質的かつ未解決の問題であり，大規模な教師データを用いた統計的学習がその解決に必要不可欠である．従って，学習データと学習手法は車の両輪であり，システムの設計にあたり両者を考慮する必要がある．

一般画像認識にはさまざまなレベルの問題があるが，いずれも実用段階へは到達していないのが現状である．これは，汎用性の高い認識システムを構築するための大規模な学習データがこれまで存在していなかったことが最大の原因と考えられる．しかしながら，前述の crowd sourcing や Web 画像マイニングの発達により，画像全体に対し何らかの単語が紐づけられた弱ラベル付データセットは近年指数的に拡大している．このため，画像全体のラベリング手法は実応用への可能性が急速に開けつつある．

また，画像全体のラベリングは，ディテクションなどの領域ラベリング手法にとっても大きな助けとなる．一般にディテクションは非常に高コストであるため，全ての対象物体の検出器を走査することは現実的ではなく，画像全体のコンテクストから対象物体や走査範囲を限定する前処理が重要である．例えば，画像全体のラベリングからシーンを判断するとともに，可能性の高い物体のみをピックアップすることで効率よくディテクションが行えると期待できる．

本研究ではこのような背景を踏まえ，実用的な一般画像認識システム実現へ向けた第一段階として，画像全体へのラベリングへ取り組む．開発する手法の要求機能としては，以下が挙げられる．

1. 一枚の画像に対する，複数のラベル付けを許容する枠組みであること．その際，弱ラベリング[1]された学習データからこれを実現すること．

2. 複数ラベルが表す画像のコンテクストを用い，semantic gap を緩和すること．

3. 学習サンプル数に対してスケーラブルに学習・認識が行えること．

これらを実現するために，本研究では以下の 2 つを具体的な課題として取り組む．

2.4.1 画像アノテーション問題への取り組み

本研究で目標とするのは，制約のない一般的な実世界での画像シンボル化である．このような一般的な環境で得られる画像は多種多様な物体やシーンが映った雑然

[1] (1) ある単語ラベルが教師として与えられていなくとも，必ずしもそれが指し示す概念が画像中に存在しないとは限らない．(2) 単語ラベルと画像領域との対応関係は与えられない．

としたものになるため，画像は多義的な意味を持つことになる．例えば，図 2.4の左上隅の画像を例にとる．この画像に対し，人はどんな単語を付けるであろうか．大まかな傾向は合致するであろうが，実際に付ける単語は人によって“sunset, water”など様々な可能性がある．また，Flickr などの写真投稿サイトにおいては，“beautiful, impressive”などの形容詞，印象語などもしばしば付けられる．実際にグラウンドツルースとして与えられているのは“clouds, sea, sun, tree”の 4 つである．これらはいずれも画像内容を表しており，正解であるといえる．このように，一般的な画像認識においては多義性と冗長性の問題が本質であり，これを許容する枠組みが適切である．これは，画像アノテーションと呼ばれる研究分野において扱われてきた問題に他ならない．

なお，カテゴライゼーションはアノテーションの特殊なセットアップとなっていることに注意されたい．すなわち，アノテーションにおいて各サンプルに与えるラベルの数を一つに限定した場合がカテゴライゼーションの問題となる．従って，アノテーションの手法は一般性を損なわずカテゴライゼーションの問題へ適用可能である．逆に，カテゴライゼーションの手法は，“一つのサンプルが持つラベルは一つ”という制約条件へ特化しているため，アノテーションの問題へ適応することは難しい．この点については，3.1 節で詳しく述べる．

以上の理由から本研究では，画像アノテーションを本質的かつ最も重要な問題設定と考え，手法の開発を行う．

2.4.2 大規模学習データへの対応

Crowd sourcing や Web 画像マイニングなどの新しいアプローチにより得られる大量のデータを用いた学習が，汎用的な一般画像認識システム構築のための鍵であることを述べた．システムは Web 上の膨大な数のサンプルから学習を行わなければならない．さらに，学習サンプルの増加・変化に応じ再学習を行う必要も生じるであろう．しかしながら，これまでの一般的な画像認識手法は，小規模データセットにおいて性能向上を行うことを目的としてきたため，スケーラビリティに乏しい．このような手法では，Web スケールの大規模データに適用することは困難である．特に，学習を行うことは事実上不可能であるといえる．従って，スケーラビリティを向上させた手法を開発することが必須の要件となる．一般に計算コストと認識精度はトレードオフの関係にあるため，これらの背反する要件を高い水準で満たすことが重要となる．

Chapter 3

画像アノテーションの関連研究

3.1 先行研究

3.1.1 Region-based Generative Model

一般的に画像アノテーションは画像全体へのラベルづけを目的としており，領域と各ラベルの対応を推定する必要はない．しかしながら，画像アノテーションの歴史はまず領域に基づくアプローチから始まっており [51; 137]，画像中の各領域に適切な単語ラベルを付与することが興味の対象であった．

先駆的な研究として，word-image co-occurrence model [137] が挙げられる．まず，各画像をいくつかの粒度でグリッド分割をし，それぞれの領域からカラーヒストグラム，エッジ方向ヒストグラムなどの基本的な画像特徴をとる．以下，これを領域特徴と呼ぶ．次に，学習サンプル全体の領域特徴をクラスタリングし，いくつかのクラスタへ分ける．各クラスタにはアピアランスが似通った領域特徴が帰属すると期待できる．各クラスタに所属するサンプル（領域特徴）と教師ラベルの共起を見ることで，各単語の事後確率を推定する．この手法は非常にシンプルであるが，その後急速に発展する領域ベースの画像アノテーション手法の基本的な構造を備えている．

その後，一枚の画像をいくつかの領域特徴 (blob) で表現する”blobworld”[32] のアプローチが画像アノテーションにおいても応用されている．Co-occurrence モデルのとるアプローチと似ているが，グリッド分割ではなく画像セグメンテーション手法に基づく点が異なる．初期の研究として最も有名なものは，word-image translation model [51] である．これは，統計的機械翻訳の手法 [27] を画像アノテーションへ応用したモデルである．領域の分割は，Normalized-cut 法 [170] により行い，分割された領域から blob を抽出する．その後，co-occurrence model と同様に領域特徴のクラスタリングを行い，ベクトル量子化を行う．Translation model では，量子化された blob を画像側の “ 単語 ” とみなし，教師ラベル側の単語への “ 翻訳 ” を試みる．[7; 9] などでは，blob を連続量のまま扱っていたが，ここでの blob は量子化後のシンボルを示す点に注意されたい．Co-occurrence model

では，各 blob からの単語の事後確率の推定を行う際，単純に共起の頻度のみを用いていたが，Translation model では EM アルゴリズムによる推定を行う．同様に，最大エントロピーに基づく統計翻訳の手法 [15] を応用した研究例もあり [93]，translation model よりも高い画像アノテーション精度が報告されている．

Translation model が用いた機械翻訳のアプローチは，各 blob はなんらかの単語と一意に対応するという仮定の上に成り立っている．しかし，blob はあくまでアピアランスベースに生成されるものであり，必ずしも単語レベルにおいて明確な対応がとれるとは限らない．このため，一つの画像が持つ複数の blob と，ラベルづけされた複数単語の全体的な関連性をモデル化することが重要である．Cross-Media Relevance Model (CMRM) [92] では，学習サンプルを介した結合によりこれを行う．CMRM も，異種言語間の関連性をモデル化する手法 [108] を画像アノテーションへ応用したものである．各画像は，含まれる blob の出現頻度を示したヒストグラムとして表現される．新規画像は，類似した blob ヒストグラムを持つ学習サンプルが持つ教師ラベルの重みづけによりアノテーションされる．直感的には，blob ヒストグラムを大域的な画像特徴と解釈した場合の k 最近傍識別に近い．この点において，現在主流となっているノンパラメトリックな画像アノテーションのさきがけであるとも言える．なお，translation model などと異なり，各領域（blob）と単語の明示的な関係は得られないため，領域ラベリングは行えない．しかしながら，逆にそのようなアプローチをとることで画像の全体的な印象からサンプル間の類似度を測ることにつながり，アノテーション性能の向上に寄与していると考えられる．

Co-occurrence mode, translation model, CMRM はいずれも blob ベースの手法であり，各画像の領域特徴はあらかじめ blob としてベクトル量子化されていた．しかしながら，実際には量子化誤差に伴う性能低下が懸念される．大きな転換点となった手法として，Continuous Relevance Model (CRM) [109] が挙げられる．CRM も CMRM と基本的には同じアプローチをとっており，図 3.1のように画像と単語の関係性を学習サンプルのインスタンスを用いてモデル化する．大きな違いとして，CRM ではサンプル間類似度を測る際に，blob を介さずに直接領域特徴を用いる．具体的には，クエリ画像（新規入力画像）の各領域特徴ごとに学習サンプルの領域特徴との比較を行い，その積によりサンプルとの類似度を計算する．Blob 作成のためのクラスタリングやベクトル量子化が必要ないため，実装上はシンプルになっているにも関わらず，大幅な性能向上を果たしている点が興味深い．

CRM においても，最初の領域分割は従来研究と同様に Normalized-cut 法 [170] により行っていた．しかし，一般的な問題として，前処理である画像セグメンテーションの性能によってその後の認識性能が大きく影響されるという問題がある．このため，セグメンテーションをやめ，単純に画像をグリッド分割した CRM-Rectangles [60] が提案されており，CRM と比べてさらに性能向上を果たしている．その後，CRM にさらに改良が加えられた Multiple Bernoulli Relevance Model (MBRM) [60] も，CRM-Rectangles と同様に画像をタイル上に分割する方法を用いている．また，ユーザの検索クエリを構成するオペレータ (AND, OR など) を

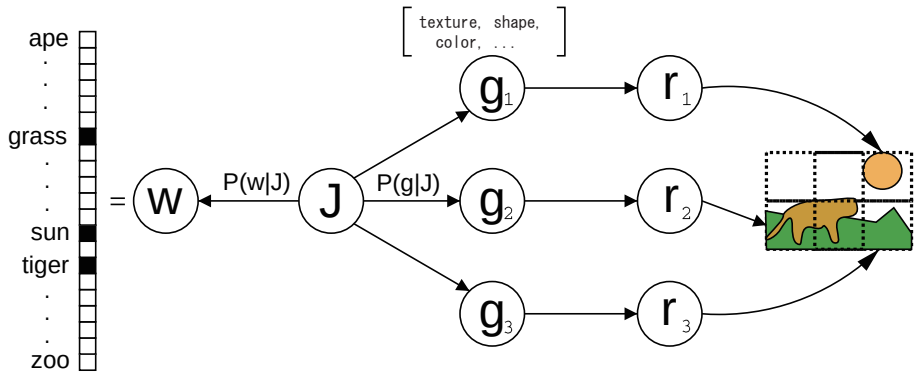


Figure 3.1: Graphical model of the CRM and MBRM. Credit: Feng *et al.* [60].

明示的に確率モデルとして表現する Inference Network (InfNet) [186] の枠組みを組み合わせた例も存在する [131].

なお，CRM と MBRM は，上述のように領域ベースのアプローチの流れを汲むものであるが，translation model の興味の対象であった領域ラベリングを行うことはできない．また，実装においても単純なサンプルベースのモデルをとっていることから，内容的には後述する non-parametric approach に近いとも言える．実際，最近の研究においては，CRM と MBRM は non-parametric なアプローチによる画像アノテーションの最初期の研究と解釈されることが多い．いずれにせよ，画像アノテーションの研究分野において，blob を用いた領域ベースのアプローチから，直接インスタンスの画像特徴を用いる non-parametric なアプローチへ向かう転換点となった研究であり，歴史的に意義深い．

以下，CRM と MBRM のアルゴリズムについて説明する．本手法では，画像を適当な数の領域に分割し，それぞれの領域特徴を用いる．この分割数を n とし，画像全体の特徴を $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2 ..., \boldsymbol{x}_n\}$ と書くことにする．実験では，画像を 5×5 のタイル状に分割し，$n = 25$ としている．また，単語群を $\boldsymbol{w} = \{w_1, ..., w_q\}$ と表す．w_i は各単語である．ある画像が画像特徴 X を持ち，同時に単語群 $\boldsymbol{w}$ がラベル付けされる同時確率 $P(X, \boldsymbol{w})$ は，学習サンプルを用いて周辺化することにより

$$P(X, \boldsymbol{w}) = \sum_{J \in T} P(J) P(X, \boldsymbol{w}|J) = \sum_{J \in T} P(J) P(X|J) P(\boldsymbol{w}|J), \tag{3.1}$$

と表すことにする．ここで，T は学習サンプルの集合，J はその中の各サンプルである．また，X，$\boldsymbol{w}$ は J について条件付独立であるという仮定を用いている．簡単のため，学習サンプルの事前確率は一定であるとする．N を学習サンプルの総数とすると，

$$P(J) = \frac{1}{N}, \tag{3.2}$$

となる．また，画像特徴 X の学習サンプル J に対する事後確率は

$$P(X|J) = \prod_{i=1}^{n} P(\boldsymbol{x}_i|J), \tag{3.3}$$

とする．つまり，各領域特徴を J について条件付独立とみなし，領域特徴の事後確率の積により画像特徴 X の事後確率を定義する．領域特徴の事後確率は

$$P(\boldsymbol{x}|J) = \frac{1}{n}\sum_{j=1}^{n} \frac{\exp\left\{-(\boldsymbol{x}-\boldsymbol{x}_j^J)^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{x}_j^J)\right\}}{\sqrt{2^k\pi^k|\Sigma|}}, \tag{3.4}$$

ただし，$\Sigma = \beta I$ とする．β はカーネルのバンド幅を決定するパラメータである．

以上までは CRM，MBRM で共通であるが，言語モデルである $P(\boldsymbol{w}|J)$ の実装が異なる．これを以下に示す．

$$P_{CRM}(\boldsymbol{w}|J) = \prod_{w\in\boldsymbol{w}} P(w|J), \tag{3.5}$$

$$P_{MBRM}(\boldsymbol{w}|J) = \prod_{w\in\boldsymbol{w}} P(w|J) \prod_{w\notin\boldsymbol{w}} (1-P(w|J)). \tag{3.6}$$

$P(w|J)$ は両手法で共通であり，

$$P(w|J) = \mu\frac{\delta_{w,J}}{N_J} + (1-\mu)\frac{N_w}{N_W}, \tag{3.7}$$

$\delta_{w,J}$ は学習サンプル J の正解ラベルに単語 w が含まれれば 1，そうでなければ 0 の値をとる変数，N_J は J の正解ラベルの単語数，N_w は全学習サンプルに含まれる w の数，N_W は全学習サンプルにラベル付けされた単語の総数である．また，μ は 0 から 1 までの値をとるパラメータであり，1 に近づけるほど各サンプルのラベルを重視し，0 に近づけるほどサンプル全体における出現頻度を重視することになる．

3.1.2 Local Patch Based Generative Model

領域ベースのアプローチにおいては，領域特徴と単語の生成モデルを構築したが，本アプローチでは局所特徴と単語の生成モデルを考える．局所特徴は抽出領域を小さくした領域特徴とも解釈できるため，基本的な枠組みは領域ベースの手法に比較的近い．しかしながら，領域ベースのアプローチでは一枚の画像が数個の領域特徴ベクトルで表されるのに対し，局所特徴ベースでは数百から数千個の局所特徴ベクトルで表されることになるため，実装には工夫が必要となる．

代表的な手法である Supervised Multiclass Labeling (SML) [29; 30; 31] の概要を図 3.2に示す．まず各画像サンプルごとに，局所特徴を混合正規分布により

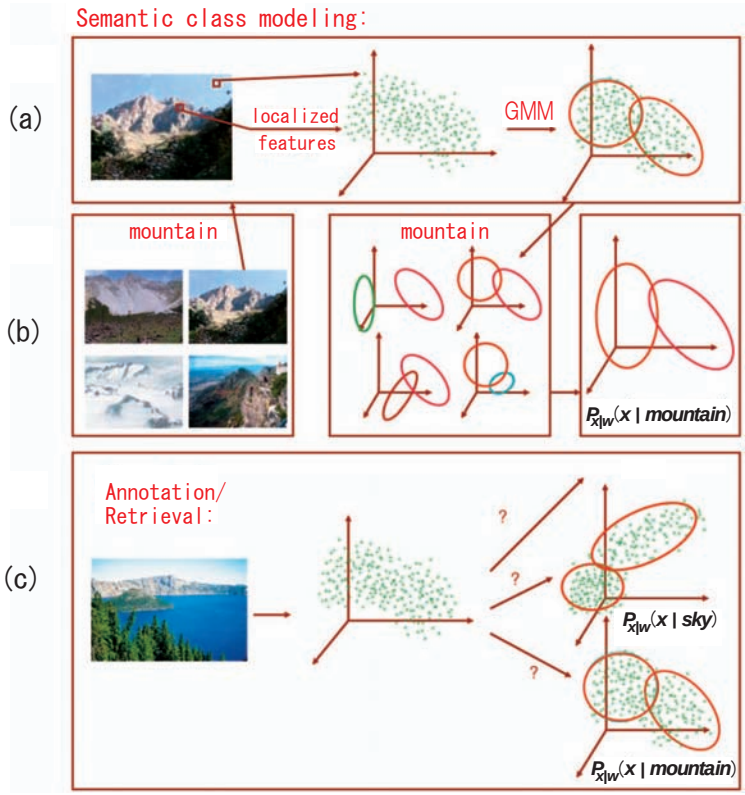


Figure 3.2: Illustration of SML. Credit: Carneiro *et al.* [29].

モデリングする．次にこの結果を用い，ある単語を教師として持つ画像すべての分布を平均化することにより，各単語からの領域特徴の事後確率をモデリングする．SML は大規模なモデルの推定を行うため，学習データ数が増えると実行困難となることが予想され，必ずしも実用的な手法であるとは言い難い．しかしながら，単語固有の局所特徴分布を推定するという，多くの研究者の興味の対象でありながら実現困難であった課題に挑戦した，科学的に意義深い研究と言える．

3.1.3 Binary Classification Approach

領域ベースのアプローチと並び，古くから研究されている方法として，2 値クラス識別に基づくアプローチが挙げられる．つまり，各単語についてそれぞれ独立に，画像がその単語クラスに属するか否かを判定する識別器を構築するものである．最終的なアノテーション結果は，最も高い反応を示した識別器の単語を順に出力する．このような他クラス識別の方法論は 1-vs-all と呼ばれ，排他的なカテゴライゼーションのタスクにおいてしばしば用いられる．画像アノテーションにおいても，Support Vector Machine (SVM) を用いたもの [41], ベイズポイントマ

シン [34] を用いたもの等がある．なお，各単語ごとに独立に識別器を構築する点では SML も同じであるが，SML は完全に生成的なモデルであるのに対し，これらは判別的なアプローチである点が異なる．

問題点として，これらの手法では単語クラス間の相関を考慮していない点が問題点として挙げられる．アノテーションの枠組みでは，一つの画像に複数のラベルが与えられており，これらは互いに関連している．例えば，“空”，“雲”，“太陽”といったラベルは同時に付与されやすいと考えられる．2 値クラス識別のアプローチはこの関係性を無視するものであり，一般的には適切とは言えない．近年，画像中の複数物体が為すコンテクストの利用が重要視されるようになってきたこともあり，現在ではアノテーションにおいて単純な 2 値クラス識別のアプローチがとられることは少ない．

Loeff らは，matrix factorization を利用し，複数ラベル情報を利用する手法を提案している [123]．この手法では，画像特徴空間において直接 2 値クラス識別器を生成するのではなく，一旦各クラスで共有される部分空間を生成し，その部分空間で最終的に用いる 2 値クラス識別器を構築することでアノテーション性能を大きく向上させている．

なお，アノテーションに関連する分野として，近年急速に注目を浴びている attribute base の物体認識では，2 値クラス識別において物体を表す複数のプロパティ（クラス名，色，テクスチャ）の共起を利用する例が存在する [199]．しかしながら，この手法は各プロパティが画像内の同じ一定領域と対応することを前提としている．アノテーションにおいて画像に与えられる単語は，必ずしも画像内の同一物体に関連するものでなく，また画像領域との対応が明確でない形容詞や印象語などのシンボルも含むため，このような手法を直接適用することは難しい．

3.1.4 Graph-based Approach

Graph-based image captioning (GCap)[150; 151] では，まず各インスタンスに所属する領域特徴と教師ラベルがインスタンス自身と結ばれる．更に，各領域特徴がそれぞれ学習データ中の近接する領域特徴と結ばれ，グラフが構築される．ラベルのついていない未知画像は，まずこのグラフへ接続された後，自身を出発点とするランダムウォークによってアノテーションが行われる．すなわち，定常確率が大きい単語ノードの順に単語が出力される．

Adaptive graph-based annotation method (AGAnn) [121] では，領域特徴ではなく，サンプル全体の画像類似度をもとにグラフを構築する．この際，データの局所的な分布に適応するため，nearest spanning chain (NSC) と呼ばれるグラフ構築の手法を提案している．これは，一般的な k-NN ベースのグラフ構築手法に比べ，パラメータの影響を受けにくことが特長である．各サンプルは，結合グラフ上のノードとして表される．このグラフ上で，教師サンプルのラベルを伝播させることによりアノテーション結果が推定される．Two-phrase Graph Learning Method (TGLM) [120] では，画像類似度に基づくグラフに加え，単語間類似度に基づくグラフを利用し，アノテーションの推定精度を向上させている．単語間類

似度は，学習データベースから推定するのみならず，Web からマイニングすることも考慮されている．

グラフベースの学習手法の本質は，クエリと類似した画像サンプルの教師ラベルを伝播させることにあり，この点では CRM [109] などのノンパラメトリックな方法に比較的近いと考えられる．

3.1.5 Regression Approach

画像特徴量を入力とする単語ラベルの回帰モデルを構築できれば，射影による画像アノテーションが実現できる．[230] では，正準相関分析を用いた線形射影によるアノテーションを行う．また，[231] では，画像中で物体が占める面積を線形回帰により近似し，カテゴリの共起を考慮した物体認識を行った．しかしながら，線形モデルは特徴分布に正規分布をあてはめることに等価であり，画像と単語の関連性を十分に表現することは難しい．また，シンボル情報である単語を線形近似することは厳密には適切ではない．

[75] では Kernel Canonical Correlation Analysis (KCCA) を用い，非線形の回帰モデルを構築している．さらに，[216] では multiple kernel learning [107] を KCCA，Kernel Multiple Linear Regression (KMLR) などに応用し，強力な回帰モデルの構築を行うとともに，回帰後の推定点を用い non-parametric にアノテーションを行う手法を提案している．しかしながら，一般にナイーブなカーネル化により解決が期待できるのは距離計量の問題のみであり，分布の対応関係を推定するのは依然として容易ではないといえる．

3.1.6 Topic Model Approach

トピックモデルは自然言語処理の分野を中心に発達しており，文書ドキュメントのクラスタリング・データマイニングにおいて大きな成果を収めている．代表的な手法として，Latent Semantic Analysis (LSA) [44], probabilistic Latent Semantic Analysis (pLSA) [80], Latent Dirichlet Allocation [20] が挙げられる．画像認識の分野においても，これらの手法は盛んに応用されている．自然言語処理の分野においては，テキストという単一モダリティの圧縮を行っていたのに対し，画像認識においては画像と単語（テキスト）という二つのモダリティを考慮する必要がある．

この枠組みでは，図 3.3のようなグラフィカルモデルを考える．このように，画像と単語の上位に，陽に観測されない (unsupervised な) 潜在変数 $\boldsymbol{l}$ (latent node) を仮定する．まず上位にある潜在変数が選択され，これに依存した形で画像特徴，単語特徴が生成される．この際，潜在変数 $\boldsymbol{l}$ に対し，画像特徴 $\boldsymbol{x}$ と $\boldsymbol{w}$ は条件付独立であるとする naive Bayes の仮定を置いている．画像特徴と単語特徴の同時確

率は，次のように潜在変数を用い平均化した形で表せる．

$$P(\boldsymbol{x}, \boldsymbol{w}) = \int P(\boldsymbol{x}, \boldsymbol{w}|\boldsymbol{l})\, P(\boldsymbol{l})\, d\boldsymbol{l} \tag{3.8}$$

$$= \int P(\boldsymbol{x}|\boldsymbol{l})\, P(\boldsymbol{w}|\boldsymbol{l})\, P(\boldsymbol{l})\, d\boldsymbol{l}, \tag{3.9}$$

ただし，モデルがもつ条件付独立の仮定から，$P(\boldsymbol{x}|\boldsymbol{w}, \boldsymbol{l}) = P(\boldsymbol{x}|\boldsymbol{l})$ の関係を式変形に用いている．

潜在変数は，画像と単語の本質的な関係を捉えた，「トピック」を表すものと考えればよい．画像と単語は，各トピックが生成する確率密度分布の混合によってモデル化される．このように，適切な潜在変数を用いた平均化計算を行うことで，画像と単語の複雑な関係性を比較的シンプルな密度関数を用いて表現できると期待できる．

例えば，“魚”という単語が付けられる画像が所属するトピックは，“海”トピック，“料理”トピックなど多くの可能性がある．同じ“魚”の画像であっても，トピックが異なればアピアランスは大きく異なるため，直接的に“魚”画像識別モデルを構築しようとすると，非常に複雑な分布を推定しなければならない．これに対し，各トピック内の“魚”画像のアピアランスの分散は小さいため，それぞれ比較的シンプルな確率モデル（例えばガウシアンなど）によって表現できる．

問題は，何を潜在変数として定義し，それをどのように推定するかであり，さまざまな実装法が提案されてきた．[7; 9] では，blob とラベルを EM アルゴリズムを用いて階層的にクラスタリングする．各クラスタが潜在変数として定義され，それぞれ正規分布により画像特徴，多項分布により単語群を生成する．各潜在変数は画像特徴と単語を直接的に関連付けている点で判別的であるといえる．しかしながら，サンプル内の全ての領域特徴と単語が一つのトピックから生成される構造であるため，モデル化のための柔軟性に欠ける．

この問題に対し，[8; 19] では LDA を応用したモデルを提案している．基本となる Gaussian-Multinomial LDA (GM-LDA) においては，サンプルの各領域特徴，各単語について潜在変数を多項分布を用いサンプリングする．多項分布のパラメータは，ハイパーパラメータによりチューニングされるディリクレ分布からのサンプリングによって得られる．このモデルにより，一つのサンプル内の複数の領域特徴・単語を異なるトピックの混合として表現することが可能になる．この結果，生成モデルとして表現能力を向上させている．また，GM-LDA では不可能であった領域ラベリングが可能である．しかしながら，画像特徴と単語の潜在変数は完全に独立に生成され，両者の依存関係は考慮されていない．従って，画像アノテーションを行う際に必要となる，画像が与えられた際の単語の事後確率の推定には必ずしも適していない．このため，クロスモーダルな潜在変数の依存関係を陽にモデル化した Multi-Modal Latent Dirichlet Allocation (MoM-LDA) を提案しており，GM-LDA よりもアノテーション性能を向上させている．

LDA はその後も画像認識において応用・改良が精力的になされている．例えば [58] では，13 種類のシーン画像のクラス識別に LDA を応用している．また，

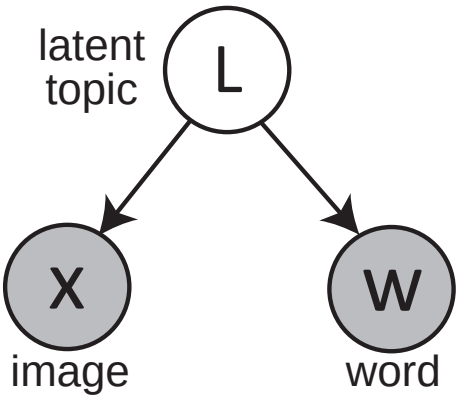


Figure 3.3: A topic model for image annotation.

[215] では MoM-LDA をさらに階層化したモデルを提案し，画像アノテーションへ適用している．

自然言語処理分野における，トピックモデルのもう一つの代表的な手法として pLSA が挙げられる．オリジナルの手法は pLSA の方が LDA よりも早く提案されているが，画像認識への応用はほぼ同時期である．[133] では，画像特徴とテキスト特徴を単純に結合し一つの特徴ベクトルとし，オリジナルの LSA，pLSA をそのまま適用している．しかしながら，当然ながらアノテーションすべき未知画像にはテキストが存在しないため何らかの工夫が必要である．この手法では，単純にテキスト特徴部に 0 を当てはめているが，その根拠は必ずしも明らかではない．また，一般に画像特徴よりもテキスト特徴の方がはるかに意味的な弁別性が強いため，両者を単純に結合し pLSA の学習を行うと，結局テキスト特徴のみを扱う場合とほとんど等価な結果となることが指摘されている．このため [134] では，各ドキュメントのトピックへの帰属を求める際にテキスト情報のみを用いる非対称なモデルを提案し，性能が大きく向上することを示した．これらの改良については，[135] に詳しくまとめられている．

pLSA はその後も多くの研究で用いられており，特に bag-of-visual-words [40] によるアプローチの確立に伴い注目を浴びるようになった．これは，画像を visual word の集合と解釈できるようになったことで，pLSA を適用する意義がより明確になったためであると考えられる．例えば，[23; 25] では，pLSA による次元圧縮をシーン認識へ応用し，良好な性能を得ている．[118] では，大規模なデータベースに pLSA を適用する方法について検討し，画像検索へ応用した．[117] では，pLSA の確率モデルを多層化し，各モーダルで構築された pLSA モデルを結合させるマルチモーダル pLSA を提案している．

3.1.7 Non-parametric Approach

古典的なk最近傍識別に代表されるアプローチであり，クエリ画像の近傍学習サンプルの教師ラベルを直接利用しアノテーションを行う．3.1.1 節で述べたように，ノンパラメトリックな画像アノテーション手法のさきがけは，CRM と MBRM である．直感的には最もシンプルな方法論であるが，これらの手法により画像アノテーションに有効であることが示され，現在に至るまで中心的なアプローチとなっている．その後も，いくつかのノンパラメトリックな手法が提案されている．Non-parametric Density Estimation (NPDE) 法 [220] では大域的画像特徴量を用いカーネル密度推定を行っている．Correlated Label Propagation (CLP) [96], Context-Based Keyword Propagation (CBKP) [126] では，共起を考慮した教師ラベルの伝播方法を提案している．また，Dual Cross-Media Relevance Model (DCMRM) [122] では，単純に学習サンプルの教師ラベルを用いるだけでなく，外部から与えられるオントロジーを有効に活用するための工夫がなされている．Multi-label Sparse Coding (MSC) [197] では，学習サンプルとの類似度の算出に sparse coding を用いている．

近年では，複数特徴を組み合わせることで非常に精度の高いアノテーションが行えることが示されている．そのさきがけとなった手法として，Makadia らの提案した Joint Equal Contribution (JEC) [129] が挙げられる．この手法では，カラーヒストグラム，Haar ウェーブレットなどの複数の基本的な画像特徴量を用いている．それぞれの画像特徴量について，まず適切な距離計量を用いてサンプル間の base distance を求める（例えば，カラーヒストグラムについてはカイ 2 乗距離，Haar ウェーブレットについては L1 距離など）．その後，これらの base distance を正規化し，等価な重みにより足し合わせることで最終的な距離計量とし，近傍サンプルの検索を行う．このように，JEC はシンプルな手法であるが，2008 年時点におけるベストスコアを記録している．現在のベストスコアを保持している手法は，Guillaumin らの提案した TagProp [72] である．この手法では，bag-of-visual-words (BoVW) [40], GIST 特徴 [148] など，15 種類のさまざまな大域特徴・局所特徴を用いている．さらに，JEC と異なり，leave-one-out におけるアノテーションスコアを最大化する基準に従い，それぞれの base distance の適切な重みを学習する．また，group sparsity を利用し，効率よく複数特徴の選択・重みづけを行う手法も提案されている [226]．これは，複数の画像特徴量間の冗長性と，同一画像特徴量内の特徴要素間の冗長性の両方を考慮し学習を行うものである．複数特徴量を用いる画像アノテーション手法の成功は興味深く，カテゴライゼーションにおける multiple kernel learning [107] の成功と相似をなしているといえる．

3.1.8 まとめ

まず，先行研究のアノテーション精度について述べる．画像アノテーションにおいては，Corel5K [51] と呼ばれる画像データセットが標準的なベンチマークとし

Table 3.1: Performance of previous works using Corel5K.

	Year	MR	MP	F-m	N+	MAP	MAP (R+)
Co-occurrence [137]	1999	0.02	0.03	0.02	19	-	-
Translation [51]	2002	0.04	0.06	0.05	49	-	-
CMRM [92]	2003	0.09	0.10	0.09	66	0.17	-
Maximum Entropy [93]	2004	0.12	0.09	0.11	-	-	-
CRM [109]	2003	0.19	0.16	0.17	107	0.24	-
NPDE [220]	2005	0.18	0.21	0.19	114	-	-
InfNet [131]	2004	0.24	0.17	0.20	112	0.26	-
CRM-Rectangles [60]	2004	0.23	0.22	0.23	119	0.26	0.30
Independent SVMs [123]	2008	0.22	0.25	0.23	-	-	-
MBRM [60]	2004	0.25	0.24	0.25	122	0.30	0.35
AGAnn [121]	2006	0.27	0.24	0.25	126	-	-
SML [29]	2007	0.29	0.23	0.26	137	0.31	0.49
DCMRM [122]	2007	0.28	0.23	0.26	135	-	-
TGLM [120]	2009	0.29	0.25	0.27	131	-	-
MSC [197]	2009	0.32	0.25	0.28	136	0.42	**0.79**
Matrix Factorization [123]	2008	0.29	0.29	0.29	-	-	-
JEC [129]	2008	0.32	0.27	0.29	139	0.33	0.52
CBKP [126]	2009	0.33	0.29	0.31	142	-	-
Group Sparsity [226]	2010	0.33	0.30	0.31	146	-	-
TagProp [72]	2009	**0.42**	**0.33**	**0.37**	**160**	**0.42**	-

て用いられおり，性能競争が続いている．Corel5K の詳細については，5 章を参照されたい．性能評価には，Mean Recall，Mean Precision，F-measure と呼ばれる指標を用いる．これらはいずれも値が大きいほど性能が良いことを示す．詳細については Appendix A を参照されたい．

表 3.1に先行研究の性能をまとめる．ここでは，F 値の昇順に先行研究を並べてあり，ノンパラメトリックなアプローチに基づく手法を太字で表記してある．このように，ノンパラメトリックな手法は画像アノテーションにおいて中心的なアプローチとなっており，歴史的に見てもよい成績を挙げていることが分かる．もちろん，各手法において用いる画像特徴量が異なるため，一概に手法自体の性能差を論じることはできないが，特によいスコアを記録している JEC 以降の手法がいずれもノンパラメトリックな手法である点は興味深い．

画像アノテーションにおいてノンパラメトリックな手法が効果的であるのは，複数ラベリングというタスクの性質に起因する．排他的なカテゴリ分けとは異なり，アノテーション問題においてはラベルは互いに関連しており，ラベルの共起が表すコンテクストを有効に用いる必要がある．ノンパラメトリックなアプロー

チでは，サンプルベースに共起情報を利用しており，コンテクストを陰にモデル化していると解釈できる．また，アノテーション問題は一般性が高く，複雑な分布を扱うため，パラメトリックなモデルでは多くのパラメータの推定を行う必要が生じ，安定に学習を行うことが難しくなる傾向にある．これに対し，ノンパラメトリックな手法は基本的にサンプルベースに分布の推定が行われるため，比較的チューニングが容易であり実用性が高いといえる．以上の考察により，本研究においてもノンパラメトリックなアプローチをベースに手法の開発を行うこととする．

3.2 ノンパラメトリック画像アノテーションのための semantic gap の緩和方法

前節のサーベイにより，画像アノテーションにおいては，事例ベースによるノンパラメトリックなアプローチが有効であることを示した．しかしながら，解決すべき重要な課題が 2 点存在する．

第一に，2.2 節で議論した semantic gap への対処である．一般に，サンプルの類似度評価は何らかの画像特徴間の距離により行われる場合が多いが，low-level な画像特徴と画像の持つ意味は必ずしも直接関連せず，大きな隔たりがある．この問題に対処するためには，できるだけ多くの学習サンプルを用いると同時に，人間が与える教師情報に対する判別性の高い特徴を選択し（すなわち，大きい重みを与え），新たな距離計量を得る必要がある．

第二に，学習データセットの規模に伴い，認識にかかる計算コストが計算量・メモリ使用量の両面において大きく増大するという問題である．一般に，汎用性の高いシステムを構築するためには高次元な画像特徴を用いる必要がある[1]．しかしながらノンパラメトリックな手法では，全ての学習サンプルをメモリ上に展開し，クエリとの距離計算を行う必要がある．したがって，用いる特徴の次元数が大きい場合，学習サンプル数の多い大規模なシステムにおいては実現困難となる．

以上から，特徴次元数を削減しつつ，意味的な判別性を高めたサンプル間距離計量を学習する手法が必須であるといえる．本節では，この話題に関する既存研究をまとめる．

3.2.1 Distance Metric Learning

元の特徴空間（画像特徴空間）における入力ベクトルを $\boldsymbol{x} \in R^p$ とする．簡単のため，ここでは特徴空間がユークリッド空間であることを仮定する[2]．タスクに関する事前知識がない場合，2 つのサンプル点 i, j 間の距離は特徴ベクトル間の

[1] 例えば，TagProp では 15 種類，合計 37,000 次元の画像特徴を用いる．
[2] 一般性を損なわず，カーネル法により非線形化を行うことが可能である．

ユークリッド距離によって測られる．

$$dist_E(i,j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)}. \tag{3.10}$$

Mahalanobis distance metric learning (MDML) は，以下のように，半正定値実対称行列 M により定められるマハラノビス距離を学習するものである．

$$dist_M(i,j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T M(\boldsymbol{x}_i - \boldsymbol{x}_j)}. \tag{3.11}$$

M はコレスキー分解により $M = WW^T$ と書きかえられる．つまり式 3.11は，元の特徴 $\boldsymbol{x}$ を W により射影した新しい空間（部分空間）上におけるユークリッド距離に等しい．これはもともと PCA などの古典的な次元圧縮手法が行っている操作に他ならず，MDML は一般化された枠組みであると解釈できる．

実際にどのように M を決定するかはタスクによって異なり，目的に応じて設計される評価関数に従い最適化することで学習される．例えば，Locality Preserving Projections (LPP) [77] などの多様体構造学習の手法は，もとの特徴空間における局所的な近接構造の保存を行う．我々の目的は，$\boldsymbol{x}$ と対で与えられるラベル情報を用い，semantic gap を緩和した新しい距離計量を学習することである．ここでは，この話題に関する先行研究を中心に挙げる．

基本的に，MDML の手法は k 最近傍法への適用を念頭においている．最初期の研究である Neighborhood Components Analysis (NCA) [68] では，学習データ中の leave-one-out による k 最近傍識別の精度を評価関数として定式化し，gradient descent により解を推定する．Maximally Collapsing Metric Learning (MCML) [67] では，同じクラスに属する全ての学習サンプルを同じ点にマッピングし，それ以外のサンプルを無限遠へ飛ばすように凸な評価関数を設計している．同様に，Large Margin Nearest Neighbor (LMNN) 法 [206] では，各学習サンプルの k 最近傍点が同じクラスに属し，かつ違うクラスのサンプルができるだけ遠くへ配置されるように最適化を行う．また，LMNN を高速化した fast-LMNN [207] も提案されている．Information-Theoretic Metric Learning (ITML) [43] では，M に対応するガウス分布を考え，線形の制約条件のもので LogDet divergence を最小化することで学習を行い，情報論的に自然な形で事前知識を利用している．

その他にも，タスクに応じたさまざまな評価関数による MDML が行われている．例えば ranking-based distance metric learning [198] では，類似画像検索への適用を念頭に置き，検索におけるランキングの精度自体を評価関数として学習を行う．また，[36; 173] などでは，検索におけるユーザのログデータを用い，オンラインにメトリックの学習を行う．その後も現在に至るまで研究が進んでおり，画像アノテーション [198; 212]，類似画像検索 [36; 81; 88; 198]，顔画像認識 [73] などへの応用が広く行われている．なお，本節での焦点からはやや外れるが，特徴空間上の局所領域ごとに異なるマハラノビス距離を学習することにより，さらに判別性を高める local distance metric learning のアプローチも盛んに研究されている [64; 65; 161; 198].

上述の MDML の手法の利点は，実際のデータ分布の局所構造に対してアダプティブに学習できる点であるが，問題点も多い．第一に，スケーラビリティに乏しい点である．多くの手法は，学習サンプルのペアワイズ・トリプレットワイズの距離計算を必要としており，学習手法は必然的に $O(N^2) \sim O(N^3)$ となる（N は学習サンプル数）．第二に，次元圧縮が陽に考慮されていない点である．W のランクに制約を加えることで次元削減を行うこと自体は可能であるが，次元数を変化させるたびに再学習が必要となる．第三に，多くの手法は最適化の反復計算の度に全ての学習サンプルを用いる必要がある点である．大規模な問題ではサンプルをメモリ上に保持することが困難であるため，ストレージアクセスを行わざるを得ず，学習の速度は著しく低下する．以上の問題を考慮し，本研究では基本的な線形次元圧縮手法による MDML に着目し，次節で詳しく述べる．

3.2.2 バイモーダル次元圧縮手法

ここで注目する次元圧縮手法は，グローバルな評価関数に基づく固有値問題により定式化されるシンプルな MDML と解釈できるが，以下のような特長から本研究の目的に適している．

- 学習サンプル数 N に対し，$O(N)$ の計算オーダで学習が可能である
- 最大固有値に対応する固有ベクトルを順に選択することで，学習後に圧縮次元数を任意に設定可能である
- 大域的最適解を，比較的安定かつ解析的に求めることが可能であり，反復的なデータアクセスが生じない．

画像特徴量を $\boldsymbol{x} \in \mathcal{R}^p$, ラベル特徴量を $\boldsymbol{y} \in \mathcal{R}^q$ と表記する．N 枚のラベル付き画像データセットから，N 個の画像特徴とラベル特徴のペア $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$ を抽出する．これらの共分散行列を，$C = \begin{pmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{pmatrix}$ と表記する．ここで，

$$C_{xx} = \frac{1}{N}\sum_i^N (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T, \tag{3.12}$$

$$C_{yy} = \frac{1}{N}\sum_i^N (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T, \tag{3.13}$$

$$C_{xy} = \frac{1}{N}\sum_i^N (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T, \tag{3.14}$$

$$C_{yx} = C_{xy}^T, \tag{3.15}$$

である．ただし，$\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ はそれぞれ画像特徴，ラベル特徴の学習サンプルにおける平均である．目的は，新しい d 次元 $(d \ll p)$ の部分空間を学習することで

ある．この部分空間上の点を $\boldsymbol{r} \in R^d$ と表記し，圧縮変数と呼ぶことにする．2 つのサンプル i, j 間の距離は，対応する部分空間上の 2 点間のユークリッド距離 $dist(i,j) = ||\boldsymbol{r}_i - \boldsymbol{r}_j||$ として求められる．

Partial Least Squares (PLS)

Partial least squares (PLS) [210] は，マルチモーダルな次元圧縮手法として最も基本的なものである．PLS は，画像特徴，ラベル特徴についてそれぞれ線形変換 $\boldsymbol{s}_{PLS} = V_x^T(\boldsymbol{x} - \bar{\boldsymbol{x}}), \boldsymbol{t}_{PLS} = V_y^T(\boldsymbol{y} - \bar{\boldsymbol{y}})$ を考える．この時，変換後の新変量 $\boldsymbol{s}_{PLS}$ と $\boldsymbol{t}_{PLS}$ の間の共分散を最大化する規準により V_x と V_y を決定する．このような変換 V_x と V_y は，次の固有値問題の解のうち最大固有値に対応する上位 d 本の固有ベクトルを選択することで得られる．

$$C_{xy}C_{yx}V_x = V_x\Theta \quad (V_x^TV_x = I_d), \tag{3.16}$$
$$C_{yx}C_{xy}V_y = V_y\Theta \quad (V_y^TV_y = I_d). \tag{3.17}$$

ここで，Θ は固有値を要素として持つ対角行列である．圧縮変数は $\boldsymbol{r}_{PLS} = \boldsymbol{s}_{PLS} = V_x^T(\boldsymbol{x} - \bar{\boldsymbol{x}})$ により得られる．したがって，PLS により得られるマハラノビス距離は次のようになる．

$$dist_{PLS}(i,j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T V_x V_x^T (\boldsymbol{x}_i - \boldsymbol{x}_j)}. \tag{3.18}$$

PLS の結果は，元の特徴量の分散の大きさに強くされる．例えば，一方の変量のスケールが他方の変量に比べ極端に大きい場合には，前者の主成分分析に近い結果となる．このため本研究では，通常の PLS に加えあらかじめ各変量の特徴要素の分散を正規化した後に PLS を行う場合も考慮する．例えば，画像特徴量については，

$$\boldsymbol{x}' = \Sigma_X^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}), \tag{3.19}$$

により正規化を行う．ここで，Σ_X は画像特徴量の各特徴要素の標準偏差を要素に持つ対角行列である．ラベル特徴に関する分散正規化についても同様である．分散正規化後に行う PLS を，本研究では normalized PLS (nPLS) と呼ぶことにする．

PLS は古典的な手法であるが，最新の人検出の手法においても活用されており [168]，その有効性が見直されている．この手法では PLS を用い，17 万次元に及ぶ画像特徴量を 20 次元にまで，精度をほとんど損なうことなく圧縮している．その結果，従来は現実的に不可能であった大規模な学習問題への適用を容易なものとしている．なお，[168] においては，ラベル側の情報 (y-view) は人か/人でないかの 2 値であるが，アノテーションの問題では一つのサンプルに複数の単語がついているため，より豊富な情報を PLS において活用できると期待される．

Canonical Correlation Analysis (CCA)

CCA は Hotelling [83] により 1936 年に考案されて以来，多変量解析の代表的な手法の一つとして用いられてきた．PLS が新変量間の共分散を最大化するように射影を求めるのに対し，CCA では相関を最大化する規準でこれを求める．すなわち，線形変換 $\boldsymbol{s}_{CCA} = U_x^T(\boldsymbol{x} - \bar{\boldsymbol{x}})$, $\boldsymbol{t}_{CCA} = U_y^T(\boldsymbol{y} - \bar{\boldsymbol{y}})$ によって得られる新変量 $\boldsymbol{s}_{CCA}$ と $\boldsymbol{t}_{CCA}$ の相関が最大となるように射影行列 U_x および U_y を求める．射影行列 U_x，U_y は次の一般化固有値問題の解として得られる．

$$C_{xy}C_{yy}^{-1}C_{yx}U_x = C_{xx}U_x\Lambda^2 \quad (U_x^TC_{xx}U_x = I_d), \tag{3.20}$$
$$C_{yx}C_{xx}^{-1}C_{xy}U_y = C_{yy}U_y\Lambda^2 \quad (U_y^TC_{yy}U_y = I_d). \tag{3.21}$$

ただし，Λ は大きい順に d 個 $(\min\{p, q\} \geq d \geq 1)$ の正準相関係数を並べた対角行列である．圧縮変数は，$\boldsymbol{r}_{CCA} = \boldsymbol{s}_{CCA} = U_x^T(\boldsymbol{x} - \bar{\boldsymbol{x}})$ として得られる．したがって，CCA により得られるマハラノビス距離は次のようになる．

$$dist_{CCA}(i, j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T U_x U_x^T (\boldsymbol{x}_i - \boldsymbol{x}_j)}. \tag{3.22}$$

CCA は，いくつかの画像アノテーションの先行研究においても用いられている [75; 216; 230]．しかしながら，これらの手法では直接的な回帰モデルの構築を目的としており，我々が目的とする次元圧縮の観点からの研究はなされていない．CCA による次元圧縮に関連した研究としては correlational spectral clustering [18] が挙げられる．この研究では，CCA・カーネル CCA を，画像とテキストからなるドキュメントの教師なしクラスタリングへ応用している．圧縮された部分空間上でクラスタリングを行うことで，元の特徴空間に比べ潜在的なトピックがよりよく分離されることが示されている．

Multiple Linear Regression (MLR)

MLR は PLS と CCA の中間的な手法であり，2 変量のうち片方のみを正規化（白色化）した非対称な構造となる．名前が示す通り，通常は回帰に用いられる手法であり，説明変数側に正規化が行われる．画像アノテーションにおいては，ラベルを目的変数，画像特徴を説明変数にとることが自然である．この場合，以下の固有値問題として定式化される．

$$C_{xy}C_{yx}W_x = C_{xx}W_x\Omega \quad (W_x^TC_{xx}W_x = I_d), \tag{3.23}$$
$$C_{yx}C_{xx}^{-1}C_{xy}W_y = W_y\Omega^2 \quad (W_y^TW_y = I_d). \tag{3.24}$$

ここで，Ω は大きい順に d 個の固有値を要素に持つ対角行列である．圧縮変数は，$\boldsymbol{r}_{MLR} = W_x^T(\boldsymbol{x} - \bar{\boldsymbol{x}})$ として得られる．したがって，MLR により得られるマハラノビス距離は次のようになる．

$$dist_{MLR}(i, j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T W_x W_x^T (\boldsymbol{x}_i - \boldsymbol{x}_j)}. \tag{3.25}$$

Table 3.2: Relationship between dimensionality reduction methods. All methods can be interpreted as special cases of PLS.

Image features	Label features	Method
-	-	PLS
	$\Downarrow$	
variance normalization	variance normalization	nPLS
whitening	-	MLR
whitening	variance normalization	nMLR
whitening	whitening	CCA

なお，PLS と同様に，MLR の結果は説明変数の分散の大きさに強く影響される．そこで，あらかじめラベル特徴の分散を正規化した後に MLR を行う場合も考慮する．これを，normalized MLR (nMLR) と呼ぶことにする．

PLS, CCA, MLR の関係性

PLS, CCA, MLR は互いに密接に関連した手法である [22]．CCA や MLR は，変量に正規化を行った後に PLS をかけた場合と等価な形になっている．表 3.2に，この関係性をまとめる．また，表 3.3に，各手法の計算コストをまとめる．特徴次元数を一定にとる場合，これらの線形手法の計算コストは学習サンプル数に対し線形オーダとなる．この性質は，特に大規模な問題において有益であると考えられる．

大規模学習データにおける効率的な実装

基本的に，前述の手法の固有値問題を解くために必要となるのは共分散行列のみである．従って，共分散行列のみを逐次的に計算すれば，学習サンプルをメモリ上に保存しておく必要はない．例えば，C_{xx} は以下のように変形できる．

$$C_{xx} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T, \tag{3.26}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^T - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T. \tag{3.27}$$

これは，$\boldsymbol{x}_i$ を逐次的に足していくことで容易に計算できることが分かる．アルゴリズム全体としては，(1) 共分散行列の計算，(2) 学習サンプルの射影，の合計二回のみデータアクセスすればよい．

Table 3.3: Computational complexity of PCA, PLS, and CCA based methods: (1) calculating covariances, (2) solving eigenvalue problems, and (3) projecting training samples using the learned metric.

	(1)	(2)	(3)
PCA	$O(Np^2)$	$O(p^3)$	$O(Npd)$
PLS	$O(Npq)$	$O(min\{p^2(p+q),(p+q)q^2\})$	$O(Npd)$
MLR	$O(N(p^2+pq))$	$O(p^3+p^2q)$	$O(Npd)$
CCA	$O(N(p^2+pq+q^2))$	$O(p^3+q^3+p^2q+pq^2)$	$O(Npd)$

Chapter 4

サンプル数にスケーラブルな画像アノテーション手法の開発

本章では，3.2.2 節で考察した次元削減手法に着目し，学習時・認識時の両方において計算効率の高い，ノンパラメトリック画像アノテーション手法の開発を行う．提案手法は，以下の特長を有する．

- サンプル数に対し線形オーダの計算コストで学習が可能．
- 解析解が求まるため，学習時に反復的にデータアクセスする必要が生じない．
- 認識において，サンプル間距離計算のコストが相対的に小さい．

提案手法において核となるのは，正準相関分析によるセマンティックな次元圧縮とサンプル間距離計量の学習である．3.2.2 節で示したように，意味的な次元削減手法として PLS，MLR の利用も考えられるが，この点に関する比較実験は 5 章を参照されたい．

4.1 ノンパラメトリック画像アノテーション

学習用データセットとして，N 個の画像とラベルのペア $T_i = \{I_i, L_i\}$ $(1 \leq i \leq N)$ が与えられるものとする．ここで，I は画像，L は対となるラベルを示す．新規画像（クエリ）I_Q が入力された際，適切な複数ラベルを推定する識別器をサンプルベースに構築する．本研究では，k 最近傍識別，MAP 推定の 2 つの手法を検討する．

4.1.1 k 最近傍識別

k 最近傍法はノンパラメトリックな識別則として最も基本的なものである．今，クエリ I_Q と学習サンプル T_i の間の距離 $DIST(I_Q, T_i)$ が定義されているとする（具

体的な実装は 4.2 節以降で述べる）．この距離を用い，学習サンプル中で最もクエリに近い上位 k 個のサンプルを検索する．これらのサンプルにおける出現頻度の高い順に単語を出力する．

4.1.2 MAP 識別

より一般的な定式化として，学習サンプルの一つ一つを弱識別器と考え，サンプルベースに MAP 識別器を構築するアプローチが考えられる．単語 w についての MAP 識別器による事後確率は，N を学習サンプル数として

$$P(w|I_Q) = \sum_{i=1}^{N} P(w|T_i)P(T_i|I_Q), \tag{4.1}$$

と表せる．先行研究の多くがこのモデルによって説明できる．

また，上述の k 最近傍識別も，式 4.1の特殊なケースと解釈できる．すなわち，

$$P(T_i|I_Q) = \begin{cases} 1/k & \text{If } T_i \text{ is in the top } k \text{ nearest neighbors of } I_Q, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

$$P(w|T_i) = \begin{cases} 1 & \text{If } w \text{ is given to sample } T_i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

とおいた場合がこれに相当する．

以下，提案手法における実装について述べる．各学習サンプル（弱識別器）への重みを与える $P(T_i|I_Q)$ の項は，次節以降で詳しく述べる．$P(w|T_i)$ は，先行研究である CRM [109] のように言語モデルをトップダウンに設計する．本研究では，次のように各サンプルの持つラベルと単語の逆頻度 (IDF) の重みづけ和を用いる．

$$P(w|T_i) = \mu\delta_{w,T_i} + (1-\mu)\frac{\log\left(N/N_w\right)}{\log N}, \tag{4.4}$$

ただし，N_w は単語 w をラベルに持つ学習画像の数, δ_{w,T_i} は，学習サンプル $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ に単語 w がラベル付けされていれば 1，そうでなければ 0 をとる．μ は 0 から 1 までの値をとるパラメータであり，実験的に決定する．また，複数単語の事後確率 $\boldsymbol{w}$ は次のように定義する．

$$P(\boldsymbol{w}|T_i) = \prod_{w\in\boldsymbol{w}} P(w|T_i). \tag{4.5}$$

式 4.5が示すように，各サンプルが成す弱識別器は単語クラスを独立に扱い共起を考慮していない．しかしながら，式 4.4のモデル設定から，サンプル内に同時に与えられている単語について大きな事後確率を示す．従って，式 4.1のように多数のサンプルの成す弱識別器を足し合わせることで，学習データ全体におけるラベルの共起を陰に表現していると期待できる．

4.2 確率的正準相関分析による距離計量学習

4.2.1 正準相関分析

学習用データセット $\{T_i\}_{i=1}^N$ から p 次元の画像特徴 $\boldsymbol{x} = (x_1, \cdots, x_p)^T$, q 次元のラベル特徴 $\boldsymbol{y} = (y_1, \cdots, y_q)^T$ を抽出し，学習サンプル $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$ を構築する．CCA の詳細については，3.2.2 節を参照されたい．ここでは，本章で必要となるいくつかの言葉の定義を行う．

CCA は，線形変換 $\boldsymbol{s} = U_x^T(\boldsymbol{x} - \bar{\boldsymbol{x}})$, $\boldsymbol{t} = U_y^T(\boldsymbol{y} - \bar{\boldsymbol{y}})$ によって得られる新変量 $\boldsymbol{s}$ と $\boldsymbol{t}$ の相関が最大となるように射影行列 U_x および U_y を求める．以下，$\boldsymbol{s}, \boldsymbol{t}$ をそれぞれ画像側正準変量，ラベル側正準変量と呼び，これらが射影される空間をそれぞれ画像側正準空間，ラベル側正準空間と呼ぶことにする．また，Λ を，大きい順に d 個 $(\min\{p, q\} \geq d \geq 1)$ の正準相関係数を並べた対角行列とする．正準空間の学習過程では，画像特徴・ラベル特徴が相補的に教師として作用する．この結果，アピアランス・セマンティクスの両方において本質的な特徴をとらえた部分空間が得られる．従って，正準空間の構造を利用しサンプル間の距離を測ることで，意味的に類似した近傍サンプルの検索が可能になると期待できる．

CCA により得られる構造を利用する方法最も簡単な方法として，単純に画像側正準空間においてサンプル間距離を測る方法が考えられる．この枠組みを CCAsim [145; 233] と呼ぶことにする．

k 最近傍法で用いる距離尺度として，画像側正準空間におけるユークリッド距離を用いる．

$$DIST_{CCA}(I_Q, T_i) = ||U_x^T \boldsymbol{x}_Q - U_x^T \boldsymbol{x}_i||. \tag{4.6}$$

これは，式 3.22と同一である．

MAP 識別で用いる，各学習サンプルの事後確率は，クエリを中心とするガウシアンにより定義する．

$$P_{CCA}(T_i|I_Q) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{s}_i - \boldsymbol{s}_Q)^T\Sigma^{-1}(\boldsymbol{s}_i - \boldsymbol{s}_Q)\right)}{\sum_{j=1}^N \exp\left(-\frac{1}{2}(\boldsymbol{s}_j - \boldsymbol{s}_Q)^T\Sigma^{-1}(\boldsymbol{s}_j - \boldsymbol{s}_Q)\right)}. \tag{4.7}$$

ただし，$\Sigma = \alpha I$ とする (I は単位行列)．分母は $\sum_{i=1}^N P_{CCA}(T_i|I_Q) = 1$ を満たすための規格化定数である．α はカーネルのバンド幅を決定するパラメータであり，設計者によるチューニングが必要である．

4.2.2 確率的正準相関分析

通常の CCA で得られるのはあくまで線形変換 U_x, U_y とこれに対応して得られる 2 つの正準空間のみである．画像特徴とラベル特徴の潜在的な表現として 2 つの正準空間をどう用いるべきか，またサンプル間の距離計量がどうあるべきかという知見は得られない．このため，$DIST_{CCA}$，P_{CCA} などでは，アドホックに画像側正準空間におけるユークリッド距離を距離計量として用いていた．

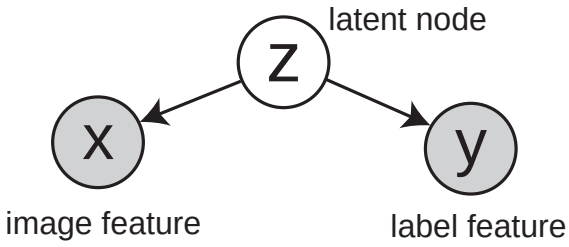


Figure 4.1: Graphical model of PCCA.

この点について，PCCA のアプローチにより，CCA の背後には以下のモデルで示される確率構造が備わっていることが示されている [5]（図 4.1）.

$$\begin{aligned} \boldsymbol{z} &\sim \mathcal{N}(0, I_d),\ \min\{p,q\} \geq d \geq 1, \\ \boldsymbol{x}\,|\,\boldsymbol{z} &\sim \mathcal{N}(W_x\boldsymbol{z} + \boldsymbol{\mu}_x, \Psi_x), W_x \in \mathcal{R}^{p\times d}, \Psi_x \succeq 0, \\ \boldsymbol{y}\,|\,\boldsymbol{z} &\sim \mathcal{N}(W_y\boldsymbol{z} + \boldsymbol{\mu}_y, \Psi_y), W_y \in \mathcal{R}^{q\times d}, \Psi_y \succeq 0. \end{aligned} \tag{4.8}$$

ここで，$\mathcal{N}$ は正規分布を示す．また，$\Psi_x \succeq 0, \Psi_y \succeq 0$ は，Ψ_x, Ψ_y がそれぞれ半正定値行列であることを示す．$\boldsymbol{z}$ は $\boldsymbol{x}, \boldsymbol{y}$ を条件付き独立の仮定の下で生成する潜在変数であり，d は $\boldsymbol{z}$ の次元数を示す (式 3.21, 式 3.21の d と同じものである)．このモデルの最尤推定解は解析的に得ることが可能であり，基本的な構造は通常の CCA の解と一致する [5]．すなわち，$\boldsymbol{x}, \boldsymbol{y}$ から $\boldsymbol{z}$ へのマッピングは，まず正準変量 $\boldsymbol{s}, \boldsymbol{t}$ への射影によって行われる．ここまでは通常の CCA と同様であるが，PCCA では相関係数の重みを用いて 2 つの正準変量を混合し，最終的な潜在変数 $\boldsymbol{z}$ への射影を得る．この射影は確率的な形で行われ，事後確率密度分布を正規分布の形で与える．

以下，詳細を述べる．$M_x, M_y \in \mathcal{R}^{d\times d}$ を，$M_x M_y^T = \Lambda$ かつ spectral norm がそれぞれ 1 未満という条件を満たす任意の行列とする．あるインスタンスの画像特徴 $\boldsymbol{x}$ のみが与えられた場合，その潜在変数 $\boldsymbol{z}$ の条件付き事後確率 $p(\boldsymbol{z}|\boldsymbol{x})$ は正規分布をなし，その中心 $\dot{\boldsymbol{z}}$ と分散 Φ_x はそれぞれ

$$\dot{\boldsymbol{z}} = E(\boldsymbol{z}\mid\boldsymbol{x}) = M_x^T U_x^T(\boldsymbol{x} - \bar{\boldsymbol{x}}), \tag{4.9}$$

$$\Phi_x = var(\boldsymbol{z}\mid\boldsymbol{x}) = I - M_x M_x^T, \tag{4.10}$$

と表される．同様に，画像特徴 $\boldsymbol{x}$ とラベル特徴 $\boldsymbol{y}$ の両方が与えられた場合はそれぞれ次のようになる[1]．

$$\ddot{\boldsymbol{z}} = E(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}) = \begin{pmatrix} M_x \\ M_y \end{pmatrix}^T \begin{pmatrix} (I-\Lambda^2)^{-1} & -(I-\Lambda^2)^{-1}\Lambda \\ -(I-\Lambda^2)^{-1}\Lambda & (I-\Lambda^2)^{-1} \end{pmatrix} \begin{pmatrix} U_x^T(\boldsymbol{x} - \bar{\boldsymbol{x}}) \\ U_y^T(\boldsymbol{y} - \bar{\boldsymbol{y}}) \end{pmatrix}, \tag{4.11}$$

[1] [5] と一部結果が異なるが，本論文が正しい．

$$\Phi_{xy} \;=\; var(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}) = I - \begin{pmatrix} M_x \\ M_y \end{pmatrix}^T \begin{pmatrix} (I-\Lambda^2)^{-1} & -(I-\Lambda^2)^{-1}\Lambda \\ -(I-\Lambda^2)^{-1}\Lambda & (I-\Lambda^2)^{-1} \end{pmatrix} \begin{pmatrix} M_x \\ M_y \end{pmatrix}. \quad (4.12)$$

M_x と M_y の設定には，回転とスケールの自由度が存在するが，本研究では最も単純に次の対角行列でそれぞれ与えることにする[1]．

$$M_x = \Lambda^{\beta},\; M_y = \Lambda^{1-\beta} \quad (0 < \beta < 1). \quad (4.13)$$

この定義の結果，Φ_x と Φ_{xy} は共に対角行列となる．β は，潜在空間の学習において画像特徴とラベル特徴の寄与を調整するパラメータとであり，β を 0 に近づけるほど画像特徴を重視し，1 に近づけるほどラベル特徴を重視することを意味する．

4.2.3 提案手法: Canonical Contextual Distance

前述の通り，各サンプルの潜在変数は潜在空間上においてガウシアンによる事後確率分布を為す．これらの確率分布を考慮することで，確率的 CCA が仮定するモデルにおいて適切な類似度評価指標を導出できると考えられる．このようにして導出される類似度評価指標，ならびにその枠組みを Canonical Contextual Distance (CCD) [142; 143; 232] と呼ぶことにする．

学習サンプルは画像と対になる教師ラベルから構成されるが，事後確率分布の推定において画像側のみを考慮するアプローチ (1-view CCD) と画像側・ラベル側双方を考慮するアプローチ (2-view CCD) の 2 つが考えられる．以下，順に説明する．

1-view CCD (CCD1)

k 最近傍識別で用いる距離尺度として，クエリ $\boldsymbol{x}_Q$ と学習サンプル $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ が潜在空間上でなす分布の Kullback-Leibler (KL) ダイバージェンスを用いる．学習サンプルの画像側 (x-view) のみを考慮する場合 (図 4.2(a)), 以下のようになる．

$$DIST_{CCD1}(I_Q, T_i) \;=\; KL\left(p(\boldsymbol{z}|\boldsymbol{x}_Q), p(\boldsymbol{z}|\boldsymbol{x}_i)\right) \quad (4.14)$$

$$\;=\; (\dot{\boldsymbol{z}}_Q - \dot{\boldsymbol{z}}_i)^T \Phi_x^{-1} (\dot{\boldsymbol{z}}_Q - \dot{\boldsymbol{z}}_i). \quad (4.15)$$

一般に，KL ダイバージェンスは非対称な計量であり距離とは異なるが，ここでは式 4.15から明らかなように単純なユークリッド距離と等価になっていることに注意されたい．すなわち，

$$\boldsymbol{r}_{CCD1} = \Phi_x^{-1/2} \dot{\boldsymbol{z}}, \quad (4.16)$$

とすれば，$\boldsymbol{r}$ のユークリッド距離として求められる．

[1] それぞれ，対角行列である Λ の各要素を β 乗，$1-\beta$ 乗し，並べたものである．

MAP 識別に用いる，クエリに対する各学習サンプルの事後確率は，$p(\boldsymbol{z}|\boldsymbol{x}_Q)$ と $p(\boldsymbol{z}|\boldsymbol{x}_i)$ の Bhattacharyya 距離により定義する．

$$
\begin{aligned}
P_{CCD1}(T_i|I_Q) &= \frac{\int \sqrt{p(\boldsymbol{z}|\boldsymbol{x}_i)p(\boldsymbol{z}|\boldsymbol{x}_Q)}d\boldsymbol{z}}{\sum_{j=1}^{N}\int \sqrt{p(\boldsymbol{z}|\boldsymbol{x}_j)p(\boldsymbol{z}|\boldsymbol{x}_Q)}d\boldsymbol{z}} \\
&= \frac{\exp\left(-\frac{1}{8}(\dot{\boldsymbol{z}}_Q-\dot{\boldsymbol{z}}_i)^T\Phi_x^{-1}(\dot{\boldsymbol{z}}_Q-\dot{\boldsymbol{z}}_i)\right)}{\sum_{j=1}^{N}\exp\left(-\frac{1}{8}(\dot{\boldsymbol{z}}_Q-\dot{\boldsymbol{z}}_j)^T\Phi_x^{-1}(\dot{\boldsymbol{z}}_Q-\dot{\boldsymbol{z}}_j)\right)},
\end{aligned}
\tag{4.17}
$$

ただし，分母は $\sum_{i=1}^{N} P_{CCD1}(T_i|I_Q) = 1$ を満たすための規格化定数である．

2-view CCD (CCD2)

基本的には CCD1 と同様であるが，学習サンプルのラベル側の寄与を陽に考慮する点が異なる (図 4.2(b))．

k 最近傍法で用いる距離尺度は以下のようになる．

$$
\begin{aligned}
DIST_{CCD2}(I_Q,T_i) &= KL\left(p(\boldsymbol{z}|\boldsymbol{x}_Q), p(\boldsymbol{z}|\boldsymbol{x}_i,\boldsymbol{y}_i)\right) \\
&= \frac{1}{2}\log\frac{|\Phi_{xy}|}{|\Phi_x|} - \frac{d}{2} + \frac{1}{2}\mathrm{Tr}(\Phi_{xy}^{-1}\Phi_x) + \\
&\qquad (\dot{\boldsymbol{z}}_Q-\ddot{\boldsymbol{z}}_i)^T\Phi_{xy}^{-1}(\dot{\boldsymbol{z}}_Q-\ddot{\boldsymbol{z}}_i).
\end{aligned}
\tag{4.18}
$$

最初の 3 項は定数であるため，$DIST_{CCD2}$ も単純なユークリッド距離計算により評価することができる．すなわち，

$$
\boldsymbol{r}_{CCD2}^{Q} = \Phi_x^{-1/2}\dot{\boldsymbol{z}}_Q, \tag{4.19}
$$

$$
\boldsymbol{r}_{CCD2}^{i} = \Phi_{xy}^{-1/2}\ddot{\boldsymbol{z}}_i, \tag{4.20}
$$

のように，クエリと学習サンプルをそれぞれ変換すればよい．

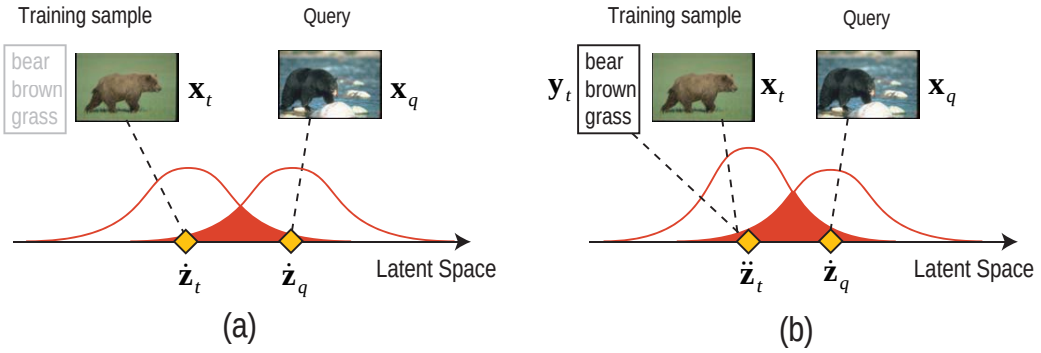


Figure 4.2: Illustration of canonical contextual distances. Estimation of distance between a query and training sample: (a) from the x-view only (CCD1); and (b) considering both the x- and y-views (CCD2).

同様に，MAP 識別に用いる事後確率は以下のようになる．

$$P_{CCD2}(T_i|I_Q) = \frac{\int \sqrt{p(\boldsymbol{z}|\boldsymbol{x}_i,\boldsymbol{y}_i)p(\boldsymbol{z}|\boldsymbol{x}_Q)}d\boldsymbol{z}}{\sum_{j=1}^{N}\int \sqrt{p(\boldsymbol{z}|\boldsymbol{x}_j,\boldsymbol{y}_j)p(\boldsymbol{z}|\boldsymbol{x}_Q)}d\boldsymbol{z}}$$
$$= \frac{\exp\left(-\frac{1}{8}(\dot{\boldsymbol{z}}_Q-\ddot{\boldsymbol{z}}_i)^T\left(\frac{\Phi_x+\Phi_{xy}}{2}\right)^{-1}(\dot{\boldsymbol{z}}_Q-\ddot{\boldsymbol{z}}_i)\right)}{\sum_{j=1}^{N}\exp\left(-\frac{1}{8}(\dot{\boldsymbol{z}}_Q-\ddot{\boldsymbol{z}}_j)^T\left(\frac{\Phi_x+\Phi_{xy}}{2}\right)^{-1}(\dot{\boldsymbol{z}}_Q-\ddot{\boldsymbol{z}}_j)\right)}, \tag{4.21}$$

4.3 画像特徴の非線形距離計量の埋め込み

提案手法の核である CCA や，3.2.2 節で述べた PLS や MLR などは，教師ラベルの表す意味情報を考慮した次元圧縮を効率的に行うことができる．画像特徴には，基本的には任意の特徴ベクトルを用いることが可能である．しかしながら，これらの線形手法では，元となる特徴量が非線形な距離計量[1](base distance) を持つ場合に対応することが困難であり，著しい性能低下につながる場合がある．実際に，現存する多くの画像特徴量は，カイ 2 乗距離や L1 距離など非線形な距離計

[1] ここでの距離計量とは，画像特徴が背後に持つ生成モデルの観点における数理的に適切な距離の測り方を意味しており，本章全体で議論しているサンプル間の類似度評価指標とは区別していることに注意されたい．

量を用いる必要があることが知られている．適切な画像特徴量の選択は本研究全体の重要なテーマの一つであり，5 章で詳しく検証を行う．ここでは，画像特徴量が非線形な性質を持つ場合の一般的な対処法について述べる．

このような場合，カーネル PCA (KPCA) [166] を用い，あらかじめ元の特徴空間における非線形な距離計量を新たなユークリッド空間に埋め込むことを考える．あるカーネル関数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$ が与えられるとする．ここで，$\phi : \boldsymbol{x} \rightarrow \phi(\boldsymbol{x})$ は入力特徴ベクトル $\boldsymbol{x}$ を明示されない高次元の空間へマッピングする射影である．N 個の学習サンプルのうち，ランダムに n_K 個 $(n_K \leq N)$ のサンプルを選び，以下のカーネルベースベクトルを定義する．

$$\boldsymbol{k}_x = \big(K(\boldsymbol{x}, \boldsymbol{x}_1), ..., K(\boldsymbol{x}, \boldsymbol{x}_{n_K})\big)^T. \tag{4.22}$$

カーネルトリックにより，KPCA は $\boldsymbol{k}_x$ 座標系における線形問題として解くことが可能である．KPCA による埋め込み後の新変量は，KPCA の射影行列 B を用い，$\tilde{\boldsymbol{x}} = B^T \boldsymbol{k}_x$ によって得られる．詳しくは，Appendix B を参照されたい．このようにして線形に埋め込まれた $\tilde{\boldsymbol{x}}$ を新たな特徴ベクトルとして，PLS, CCA などの線形手法に適用する．

カーネル関数には，適切な距離計量を指数関数に入れ定義する (GRBF カーネル)．この方法は，多くの距離計量において良好な性能を得られることが報告されている [193; 225]．

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{1}{2P}\mathrm{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right). \tag{4.23}$$

$\mathrm{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ は χ^2 distance や L1 distance などの base distance である．バンド幅である P の設定が問題となるが，本論文では n_K の学習サンプル中の全ての 2 点間距離の平均値を用いる．この設定方法は，クロスバリデーションなどで P を最適化した場合に匹敵する性能が得られることが報告されている [225]．

理論的には，多くの学習サンプルをカーネル化に用いるほど（すなわち，n_K が大になるほど）埋め込みの精度は向上する．このため，標準的な KPCA では与えられる全ての学習サンプルをカーネル化に用いる $(n_K = N)$．しかしながら，新規に入力されるサンプル（クエリ）の認識を行うためには，n_K 個の学習サンプルを用いたカーネルベースを求める必要がある．これを行うためには，n_K 個のサンプルの生の特徴ベクトルをメモリ上に常時保持しておき，クエリとの距離計算を行わなければならない．n_K が大きい時，この処理の計算コストは生の画像特徴量を用いた全探索に近づくため，ノンパラメトリックな画像アノテーションを効率化するという当初の目的から逸脱する．また，学習の際には n_K 次元の固有値問題を解く必要が生じるため[1]，現実的に実行不可能となる．

そこで，本論文では，学習サンプルからランダムに抽出した少数のサンプル $(n_K = 300)$ をカーネル化に用いる近似的な実装を行う．詳しくは，Appendix B

[1] 一般に，計算コストは $O(n_K^3)$ となる．

を参照されたい．これは，大規模半正定値行列の近似的スペクトル分解の手法である Nyström 法 [209] や column sampling [48] と基本的に同じ発想に基づくものである．このアプローチにより，現実的な計算コストで効率よく非線形距離計量の埋め込みを行うことが可能である．

なお，本論文の焦点からは外れるが，KPCA は多様体構造学習の最も一般的な枠組みであるため，大規模な問題で KPCA を解くことそのものが大きな研究領域となっている [181]．上述の Nyström 法を更に大規模な問題へ適用するための改良 [102; 116]，additive kernel を対象とした効率のよい解法 [155] などさまざまな研究がなされており，これらの応用も重要な将来課題である．

4.4 ラベル特徴の設計

ラベル特徴としては，各単語の存在を示すバイナリベクトルを用いる．特徴ベクトルの各要素が各候補単語に対応し，あるサンプルにその単語がついていれば 1，そうでなければ 0 をとる．この定義の結果，ラベル特徴ベクトルの内積は，2 つのサンプルの教師ラベルが共通に持つ単語の数となる．これは，ラベルの意味的な近さを表す指標の一つとして妥当性があると考えられる．従って，ラベル特徴の座標系は内積が自然な形で定義されており，線形性を仮定するために適した空間になっていると解釈できる．しかしながら，一般にラベル特徴はスパースなベクトルとなるため，共分散行列 C_{yy} が特異になりやすい．これは，CCA など共分散行列に正定性を必要とする手法を用いる場合問題となる．このような場合には，正則化項を共分散行列に加えることにより安定化が行える．例えば，C_{yy} の代わりとして $C_{yy} + \gamma I$ を用いることが考えられる．γ は正の実数であり，汎化性能を決定するパラメータとなる．

4.5 キーワードベース画像検索への応用

画像アノテーションの応用アプリケーションとして，キーワードに基づく画像検索（リトリーバル）が考えられる．メタデータが付与されていない画像に対し，アノテーションにより内容を示すキーワードを付与しておくことにより，既存のテキストベース画像検索と同じ枠組みでの検索が可能となる．しかしながら，ユーザにとって実用性の高い画像検索を行うためには，キーワードに適合する確実性の高い画像をより上位にランキングする必要がある．その実現のためには単純にアノテーションの結果を用いるだけでは不十分であり，確率的な枠組みを考える必要がある．ここでは，最尤推定，MAP 推定の 2 つのアプローチを考える．

$\boldsymbol{w}_Q$ を検索単語（群）とする．まず，最尤推定によるリトリーバルについて解

説する．$\boldsymbol{w}_Q$ に対する，候補画像 I_c の尤度を g_l とすると，

$$g_l = P(\boldsymbol{w}_Q|I_c) \tag{4.24}$$
$$= \sum_{i=1}^{N} P(\boldsymbol{w}_Q|T_i)P(T_i|I_c), \tag{4.25}$$

と計算できる．これは，画像 I_C を入力した際の単語群 $\boldsymbol{w}_Q$ に関するアノテーションのスコアと解釈できる．g_l の値の大きい順に候補画像をランキングすることによりリトリーバルを行う．

同様に，MAP 推定では事後確率を用いてランキングを行う．$\boldsymbol{w}_Q$ に対する，候補画像 I_c の事後確率を g_{pp} とすると，

$$g_{pp} = p(I_c|\boldsymbol{w}_Q) \tag{4.26}$$
$$= \frac{P(\boldsymbol{w}_Q|I_c)p(I_c)}{P(\boldsymbol{w}_Q)} \tag{4.27}$$
$$= \frac{\left(\sum_{i=1}^{N} P(\boldsymbol{w}_Q|T_i)P(T_i|I_c)\right)p(I_c)}{P(\boldsymbol{w}_Q)} \tag{4.28}$$
$$\propto g_l p(I_c), \tag{4.29}$$

となる．ここで，入力である $\boldsymbol{w}_Q$ の事前確率 $P(\boldsymbol{w}_Q)$ は定数扱いできる点に注意されたい．

このように，MAP 推定におけるスコア g_{pp} は，最尤推定におけるスコア g_l に各候補画像の事前確率 $p(I_c)$ を掛け合わせたものになっている．$p(I_c)$ の設計にはなんらかの事前知識が必要であり重要な課題であるが，本研究ではスコープからはずし一定確率とする．この場合，g_{pp} によるランキングは，g_l によるそれと一致する．以下，本論文では特に断りのない限り，最尤推定によるリトリーバルを行う．

4.6 考察

4.6.1 提案手法のまとめ

提案手法 (CCAsim, CCD1, CCD2) はいずれも CCA により得られる部分空間をノンパラメトリックに利用するものであるが，CCAsim と比べ CCD (CCD1, CCD2) ではより厳密に PCCA の枠組みを適用し，理論的に最適な潜在空間と距離計量を得ている．CCAsim では，入力特徴空間から潜在空間への射影は基本的に線形射影による点推定で行われ，その上にユークリッド距離を距離計量としてヒューリスティックに用いていた．これに対し，CCD1 と CCD2 では射影そのものが確率的に与えられ，最適な距離計量が自動的に決定される．この結果，潜在空間の

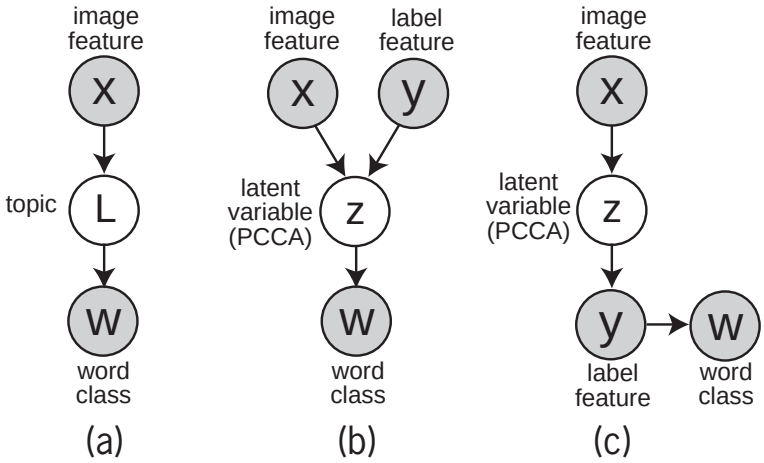


Figure 4.3: (a): Typical topic model approach. (b), (c): Approaches to the annotation problem using PCCA.

各次元が正準相関係数の大きさによって重みづけられることになり，より意味的な判別性の高い構造が得られている．

さらに，CCAsim と CCD1 では，各学習サンプルの潜在変数の導出においてラベル特徴の寄与を考慮していないが，CCD2 では画像特徴とラベル特徴の両方を陽に利用したより表現能力の高い潜在変数を得ていると期待できる．

4.6.2 Topic model に基づく他手法との関連性

図 4.1から明らかなように，(確率的）正準相関分析は topic model と同様の確率構造を備えている．実際，PCCA は確率モデルとして正規分布を用いた場合の pLSA であると解釈できる．一般的に pLSA や LDA などでは，シンボル情報である単語群を扱えるように多項分布などを用い定式化されているため，構築された topic model がそのまま識別器となる（図 4.3(a)）．しかしながら，一般に学習の際は EM アルゴリズムや変分ベイズ法などの逐次計算による推定が必要であり，局所的な最適解しか保証されず，性能は初期値に強く依存する．また，学習サンプル数や単語数の増加に伴い，計算コストも大きく増大する．一方，PCCA は正規分布に基づくシンプルなモデルであるため，解析的に大域的最適解を求めることができる点が大きなメリットである．しかしながら，単語情報は本来シンボルであり正規分布には適用できないため，topic model の構築は直接識別器の構築に結びつかない．この点が，一般的な topic model に基づく画像アノテーション手法との大きな相違点である．

そこで提案手法では，図 4.3 (b) のように，あらかじめ複数単語情報をラベル特徴として数量化し，まず画像特徴とラベル特徴から topic model の構築を行う．ここでは，画像特徴とラベル特徴を用いたおおまかな次元圧縮（特徴選択）を行

うことが目的である．その後，学習サンプルの一つ一つを“topic”と解釈し，事例ベースに各サンプルの単語情報を再利用することで識別器を構築する．各サンプルは弱識別器として解釈でき，これらを入力クエリとの距離に応じて重みづけることでベイズの最適識別器を構築する．このように，提案手法では次元圧縮が中間的な表現であるラベル特徴 $\boldsymbol{y}$ を用いて行われ，その後 $\boldsymbol{w}$ を用いてノンパラメトリックに識別器を構築する 2 段階のステップを踏むことに注意されたい．

CCA によりアノテーションを行うアプローチとして，topic model から出力されるラベル特徴の推定点を直接利用することも考えられる（図 4.3(c)）．すなわち，$\hat{\boldsymbol{y}} = argmax \int p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z}$ からアノテーション結果を出力する識別則を設計する方法論である．しかしながら，CCA 自体は線形の誤差最小化手法にすぎないため，上記の方法論は単純な線形回帰とほぼ等価な結論となる．CCA による線形回帰を画像アノテーションに利用した先行研究としては，[75; 230] が挙げられる．しかし，画像アノテーションが扱う一般的な問題では，画像と単語の関連性を線形モデルのみで十分に表現することは難しく，情報の損失が大きいと言える．

これに対し提案手法は，次元圧縮自体は線形の枠組みで近似し大まかな特徴選択を行ったのち，潜在空間上 ($\boldsymbol{z}$) に残存する非線形構造をノンパラメトリックに活用する，より実用性の高いアプローチをとっているといえる．

Chapter 5

画像アノテーション手法の性能評価実験

5.1 データセット

本章では，4つのデータセットを用いて提案する画像アノテーション手法の性能を比較評価する．画像アノテーションの研究分野においては，Corel5K [51] と呼ばれるデータセットがベンチマークとして用いられてきた．近年ではこれに加え，IAPR-TC12，ESP Game と呼ばれるデータセットも用いられており，計3つのデータセットで評価を行うことが標準的なプロトコルとなっている [72; 129]．本研究でも，この3つのデータセットを用い，先行研究との比較を行う．更に，より規模の大きいデータセットである NUS-WIDE [37] も評価の一部に用いる．以下にそれぞれ説明する．また，各データセットの画像数，単語数等を表 5.1にまとめる．

Corel5K

Corel5K [51] は画像アノテーションの研究分野におけるデファクトスタンダードのベンチマークである．これは，Corel stock photo library の一部を利用して作られたデータセットであり，5000 枚のラベル付き画像からなる．このうち，4500 枚が学習用のデータ，残りの 500 枚がテストデータとして指定されている．データセットには 260 種類の単語が含まれる．

IAPR-TC12

IAPR-TC12 は，もともと異種言語間の画像検索タスクにおいて用いられたデータセットである．Makadia ら [129] は，この中から一般的な名詞のみを抽出し，画像アノテーションのベンチマークとしてセットアップを行った．17,665 枚の学習サンプルと 1,962 枚のテストサンプルからなり，268 種類の単語を含む．

Table 5.1: Statistics of the training sets of the benchmarks.

	Corel5K	IAPR-TC12	ESP Game	NUS-WIDE
dictionary size	260	291	268	81
# of images	4,500	17,665	18,689	161,789
# of words per image (avg/max)	3.4/5	5.7/23	4.7/15	1.9/12
# of images per word (avg/max)	58.6/1004	347.7/4999	362.7/4553	3721.7/44255

ESP Game

ESP Game は，オンラインの画像ラベリングゲーム [195] により収集されたラベル付き画像を用いたデータセットである．Makadia ら [129] により，一般に公開されている 60,000 枚のサンプルのうちの一部を用いて構築された．18,689 枚の学習サンプルと 2,081 枚のテストサンプルからなり，291 種類の単語を含む． 実画像のみならず，ロゴや絵などの多様な画像から構成される．

NUS-WIDE

NUS-WIDE [37] は比較的大きなデータセットであり，161,789 枚の学習サンプル，107,859 枚のテスト画像によって構成される．これらは，大規模な写真投稿サイトである Flickr からダウンロードされた画像である．扱う単語数は 81 種類であるが，全てのサンプルが人手によりチェックされ，ラベルづけされている．

なお，データセット中には，81 種類の単語ラベルがいずれも存在しない画像が多く存在する．本論文では，学習サンプルについては 161,789 枚全てを用いるが，テストサンプルについては最低一つのラベルがついている画像をランダムに 2,000 枚選択し，評価に用いる．

5.2 基礎評価

本節では，Corel5K, IAPR-TC12, NUS-WIDE の 3 つのデータセットを用い，提案手法の有効性を議論する．まず，提案手法を他の次元削減手法と比較し，ノンパラメトリック画像アノテーションにおける性能を比較調査する．同時に，さまざまな画像特徴量を用い，基本となる線形次元圧縮と，KPCA による非線形埋め込み後の次元圧縮の両方を検証し，それぞれどのような場合に有効であるかを考察する．

5.2.1 画像特徴量

Corel5K と IAPR-TC12 については，以下の 5 つの画像特徴量を用いる．

1) Densely-sampled SIFT [124] bag-of-visual-words (BoVW) (1000 dim)

2) Densely-sampled Hue [190] BoVW (100 dim)

3) GIST [148] (512 dim)

4) HSV color histogram (4096 dim)

5) HLAC (2956 dim)

HLAC 特徴を除き，これらの特徴は TagProp [72] において利用されたものであり，著者らのホームページで公開されている[1]．HLAC 特徴の詳細については Appendix C を参照されたい．

NUS-WIDE については，以下の 4 つの画像特徴量を用いる．

1) Edge histogram (73 dim)

2) Color correlogram (144 dim)

3) Grid color moment (225 dim)

4) SIFT BoVW (500 dim)

これらの特徴量は，データセットの提供者によりあらかじめ抽出されたものであり [37]，ホームページ上で公開されている[2]．ベースラインとして，いくつかの代表的な base distance (χ^2 距離, L1 距離など) をそれぞれの画像特徴量に直接適用した場合の評価を行う．また，これらの base distance をカーネル PCA によりあらかじめ埋め込んだ場合についても評価し，単純に線形手法を適用した場合との比較を行う．

5.2.2 セットアップ

提案手法を含むいくつかの次元圧縮手法を比較し，ノンパラメトリックな画像アノテーションにおける性能（F 値）を調査する．ここでは，最も単純な識別手法である k 最近傍法によりアノテーションを行う．アノテーションの精度は，それぞれの次元圧縮手法により得られる距離計量の優劣を端的に示すと考えられる．

まず，提案手法については CCD1(式 4.15), CCD2(式 4.18) の両方を評価する．CCD2 は CCD1 と異なり，サンプル間類似度の算出において学習サンプルのラベルを陽に考慮する．詳しくは，4 節を参照されたい．

比較する次元圧縮手法として，次のものを評価する．いずれも，圧縮後の部分空間におけるユークリッド距離をサンプル間類似度として用いる．

[1] http://lear.inrialpes.fr/data
[2] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

- PCA
- PCAW
- PLS
- nPLS
- MLR
- nMLR
- CCA

ここで，PCAW は PCA with whitening を示し，PCA により得られる主成分の分散を正規化した場合を示す．PLS(nPLS), MLR(nMLR) の詳細については，3.2.2 節を参照されたい．

k 最近傍法において，関連する近傍サンプルが同数の場合は，データセット全体での頻度が少ない単語を優先して出力する．Corel5K と IAPR-TC12 については，パラメータ k を $k = 1, 2, 4, 8, 16, 32$ にとり，最もよいスコアを採用する．同様に，NUS-WIDE については，$k = 50, 100, 150, 200$ をとる．

評価は，基本的に先行研究のプロトコル [51] に従う．各テスト画像に 5 単語ずつアノテーションを行い，単語ごとに定義される recall と precision の平均と，その F 値を計算する．ここでは，F 値を性能指標として用いる．これは，値が大きいほどアノテーションが正確であることを示すものである．詳しくは，Appendix A を参照されたい．

5.2.3 実験結果

まず，Corel5K と IAPR-TC12 の結果について述べる．これらのデータセットについては，画像特徴次元数に比べ学習サンプルの数が相対的に少ないため，MLR や CCA など逆行列計算を必要とするアルゴリズムを解くことが困難である．そこで，本実験では MLR (nMLR)，CCA (CCD1, CCD2) については，あらかじめ画像特徴を PCA により 200 次元に圧縮して用いる[1]．カーネル PCA を用いた非線形距離計量の埋め込みを行う際は，最初の 200 次元のカーネル主成分をとり，これを新しい画像特徴として用いる (4.3 節を参照のこと)．BoVW やカラーヒストグラムなどのヒストグラム特徴については，χ^2 距離の埋め込みを行う．GIST 特徴については，最適な距離計量は明らかでなく，経験的に L2 距離が用いられることが多い．しかしながら，本実験では L1 距離の方がよい精度を示しているため，L1 距離の埋め込みを行う．HLAC 特徴については，χ^2 距離・L2 距離・L1 距離とも著しく低い精度となったため，カーネル法は用いず線形手法の適用のみ行う．

[1] ただし，Hue BoVW についてはもともとの特徴が 100 次元であるため，そのまま用いる．

図 5.1から図 5.10に比較結果を示す．PLS, MLR, CCA などのラベル情報を用いる次元圧縮手法 (linear) は，unsupervised な手法である PCA に比べアノテーション精度を大きく向上させている．特に，nPLS や CCD が安定に高い性能を示しており，$d = 10 \sim 20$ 程度で元の画像特徴量の L2 距離と同等以上の F 値を得ている．元の画像特徴空間にユークリッド性を仮定する条件のもとでは，これらの線形手法により次元削減と精度向上の両方が期待できる．しかしながら，多くの画像特徴では，ユークリッド性を仮定することは適切でない（Hue BoVW, HSV color histogram など）．このような場合，単純な線形手法による次元圧縮は，元の画像特徴の L2 距離を上回るものの，カイ 2 乗距離などの非線形の距離計量には遠く及ばない結果となる（例えば図 5.4など）．これは，元の画像特徴空間にユークリッド性を仮定することの損失が大きすぎるためであると考えられる．このような場合，カーネル PCA による距離計量の埋め込みが有効に働くことが分かる．ここでは，学習サンプルの一部 ($n_K = 300$) のみを用いてカーネル化を行っているにも関わらず，大きく精度を向上させていることは興味深い．一方，GIST 特徴は，多くの先行研究で L2 距離を用いていることからも示されるように，特徴空間にユークリッド性を仮定することは経験的にある程度妥当性がある．このため，図 5.7, 図 5.8が示すように，L1 距離の埋め込みは線形の場合に比べてそれほど大きな性能向上につながっておらず，場合によってはむしろ性能が低下することが分かる．

以下，その他に得られた知見と考察をまとめる．

- 線形 (linear) の場合，CCD > nPLS >（その他）となる傾向にあり，場合によっては nPLS が CCD を上回る（例えば図 5.6）．これは，もともとの画像特徴量の多様体構造（非線形距離計量）を無視する上で，PLS の持つ安定性が有利に働いているためであると考えられる．反面，KPCA により埋め込みが行われた場合は，CCD が常に優位となっている．このことは，空間が適切にユークリッド化された前提のもとでは，CCD がより適切な距離計量を与えることを示している．

- CCA 関連の手法では，CCD2 > CCD1 > CCA となる場合が多い．CCA は全ての次元を等価に扱うため，d が大きくなると性能が著しく低下する場合がある（例えば，図 5.7）．これに対し，CCD は各次元を寄与に応じて重みづけるため，圧縮次元数 d の設定に対して比較的安定に性能を保っている．また，CCD2 は CCD1 よりも総じて性能がよく，サンプル間距離計算において陽にラベル特徴の寄与を考慮することが奏功しているといえる．

- HLAC 特徴は，単体の特徴として非常によい性能を持つことが分かる．特筆すべきは線形手法と相性がよい点であり，他の特徴で KPCA を用いた場合と同等以上の性能を示している．総じて，PCAW, MLR, CCA, CCD など，元特徴に白色化を行う手法を適用した場合に高いスコアを示している．しかしながら，元特徴での L2 距離・L1 距離を用いた場合や，PCA, PLS など元特徴の分散を保存する手法を用いた場合は著しく低いスコアとなって

いる．これは，HLAC 特徴の持つ，特徴要素ごとの分散の偏りが極めて大きく，かつ特徴要素間の相関が大きいという性質に起因する．このことが，一般的な画像特徴と異なり，特徴自体の距離計量の設計を難しいものにしている．HLAC 特徴は強力な画像特徴量であるが，このような特殊な性質に留意して用いる必要がある．

次に，NUS-WIDE における結果を図 5.11から図 5.14にまとめる．提供されている特徴はいずれも正規化されているため，ここでは L1 距離と L2 距離のみベースラインとして比較に用いる（BoVW を除く）．全体として，Corel5K，IAPR-TC12 における線形 (linear) の場合と同様の傾向を示しているといえる．CCD はどの特徴においても安定によい精度を示しており，$d = 10$ 程度で元の画像特徴を用いた距離計算と同等の性能を得ている．しかしながら，Corel5K, IAPR-TC12 における実験と異なり，CCD1 と CCD2 の性能に有意な差は認められない．NUS-WIDE は比較的大規模なデータセットであるが，ラベルとして与えられるのは 81 種類の基本的な概念のみである．これらは，次元圧縮の過程では有効であるものの，実際のサンプル間距離計算には寄与していない可能性がある．

最後に，NUS-WIDE における実際の計算時間を報告する．認識にかかるコストは，ある定められた d についてどの手法も同じであるため，ここでは学習に要する時間を表 5.2にまとめる．ここでは，圧縮次元数を $d = 20$ とした場合の計算時間を示す．計算機は，8 コアのデスクトップ計算機 (Xeon 3.20 GHz) を用いた．PLS, MLR や CCD は，PCA よりも大きな計算コストを必要とするものの，計算時間の差は大きくても 2 倍程度であることが分かる．特に，画像特徴の次元数が，単語数に対し相対的に大きい場合 ($p \gg q$) にその差は小さくなる．例えば，500 次元の BoVW においては，PLS の方が PCA よりも早く解けている．この結果は，表 3.3における分析から説明される．なお，nPLS，nMLR などは特徴の分散正規化にかかる計算コストが大きく，計算時間が長くなっていることが分かる．

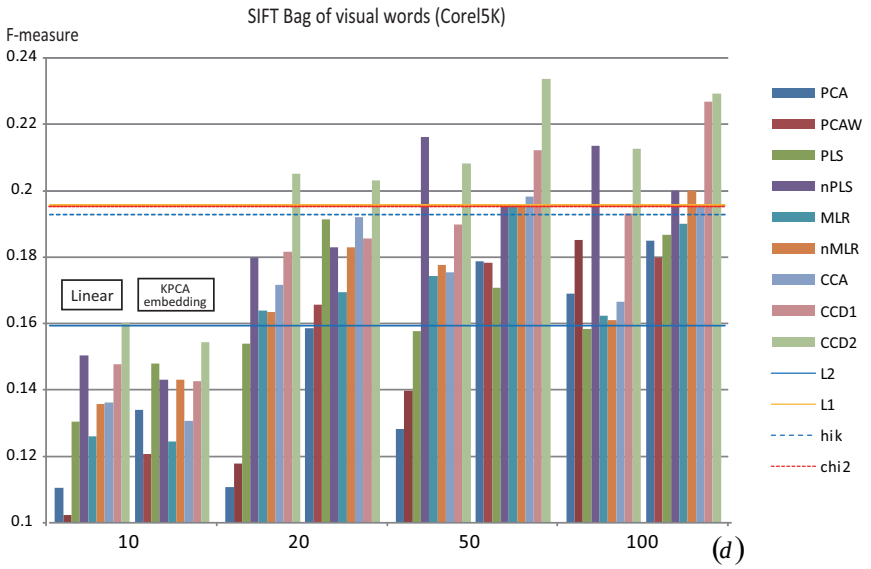


Figure 5.1: Results for the Corel5K dataset (1000-dimensional SIFT BoVW). Methods are compared using different features with designated dimensionality (d). For each entry, the left set of bars corresponds to normal linear methods, while the right set corresponds to those with KPCA embedding.

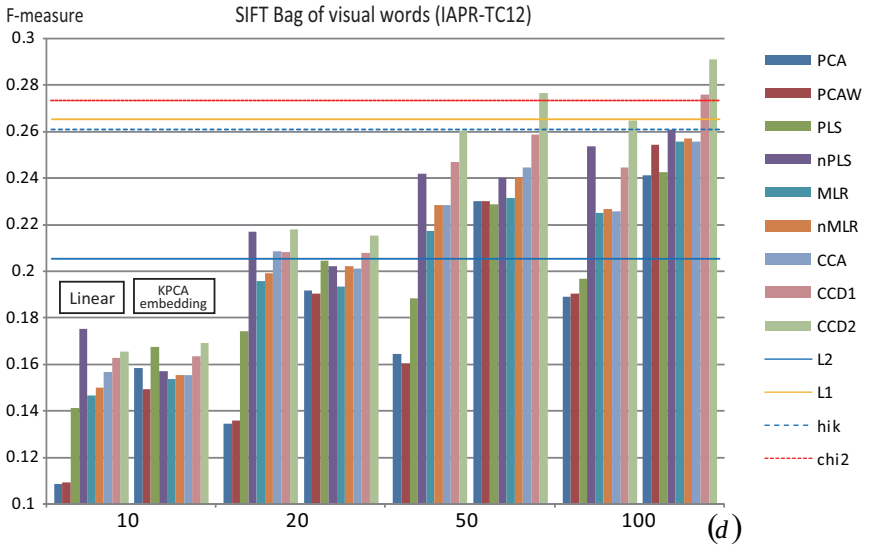


Figure 5.2: Results for the IAPR-TC12 dataset (1000-dimensional SIFT BoVW).

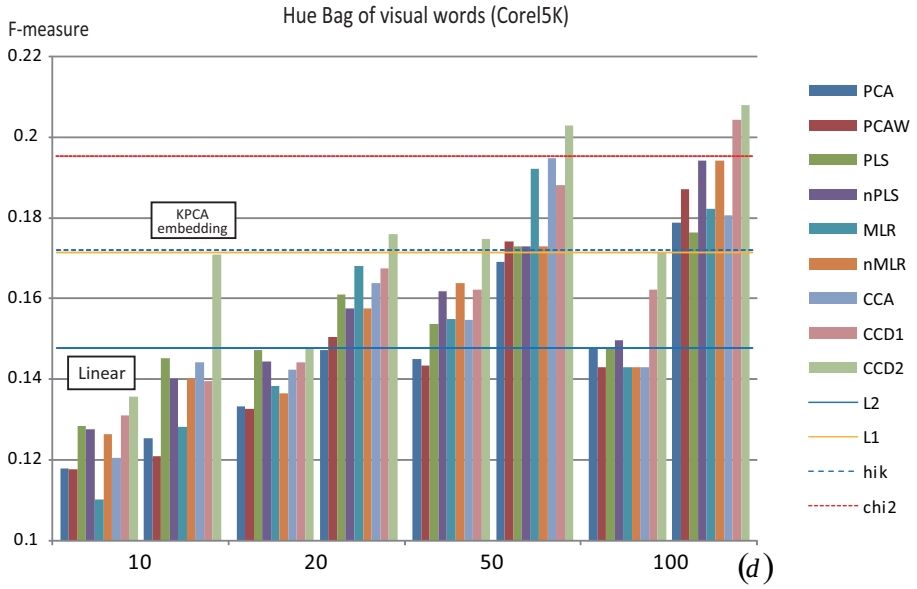


Figure 5.3: Results for the Corel5K dataset (100-dimensional hue BoVW).

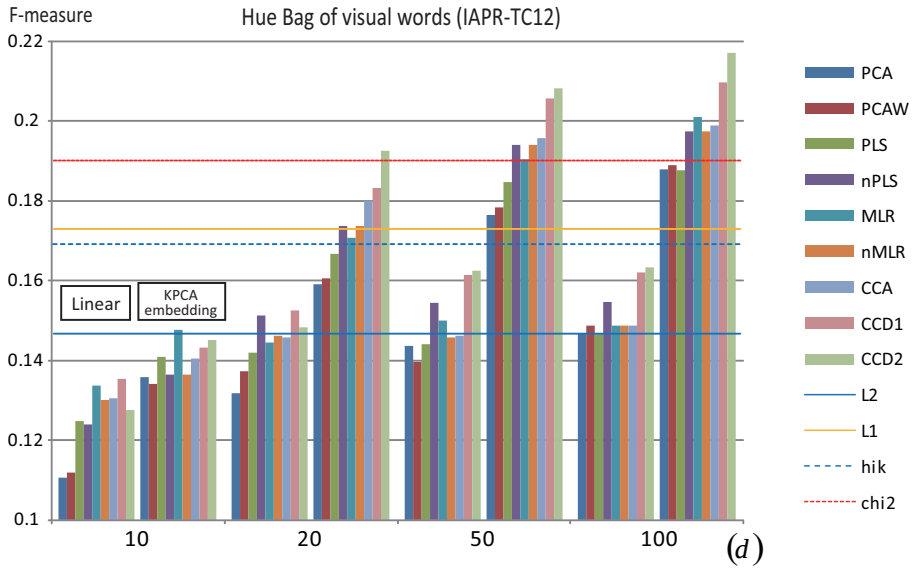


Figure 5.4: Results for the IAPR-TC12 dataset (100-dimensional hue BoVW).

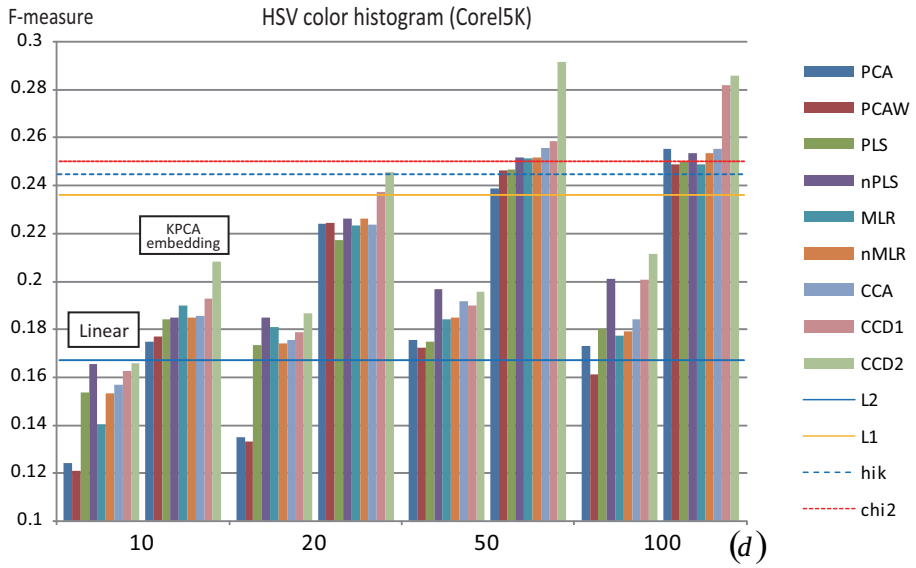


Figure 5.5: Results for the Corel5K dataset (4096-dimensional HSV color histogram).

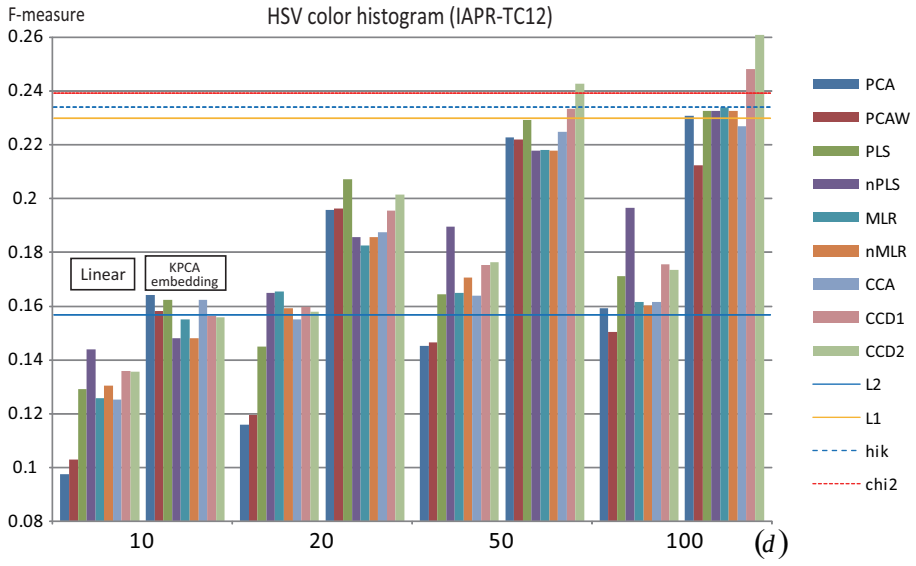


Figure 5.6: Results for the IAPR-TC12 dataset (4096-dimensional HSV color histogram).

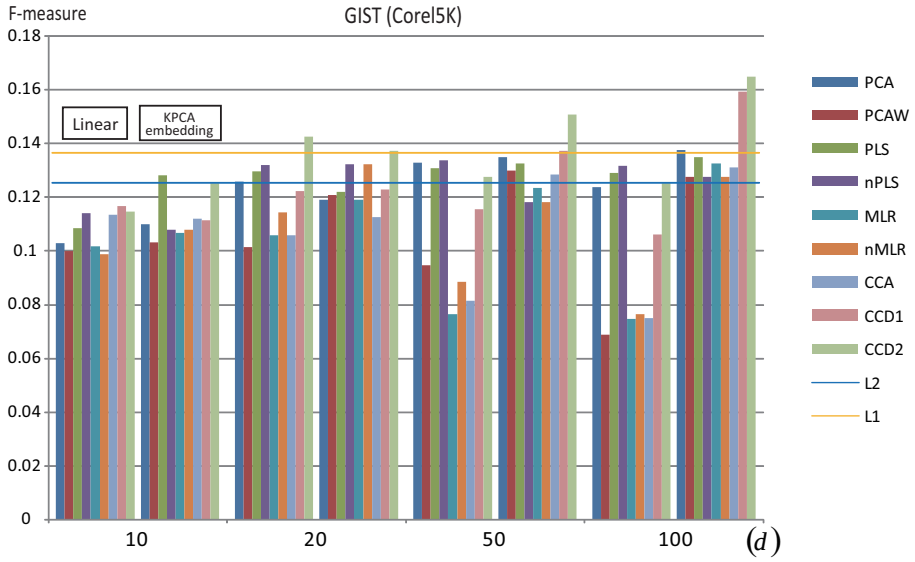


Figure 5.7: Results for the Corel5K dataset (512-dimensional GIST).

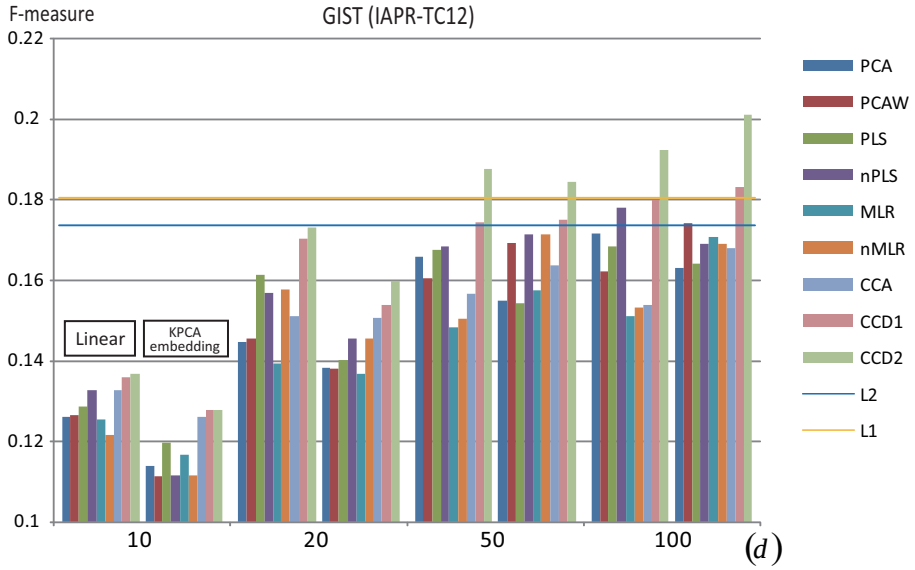


Figure 5.8: Results for the IAPR-TC12 dataset (512-dimensinal GIST).

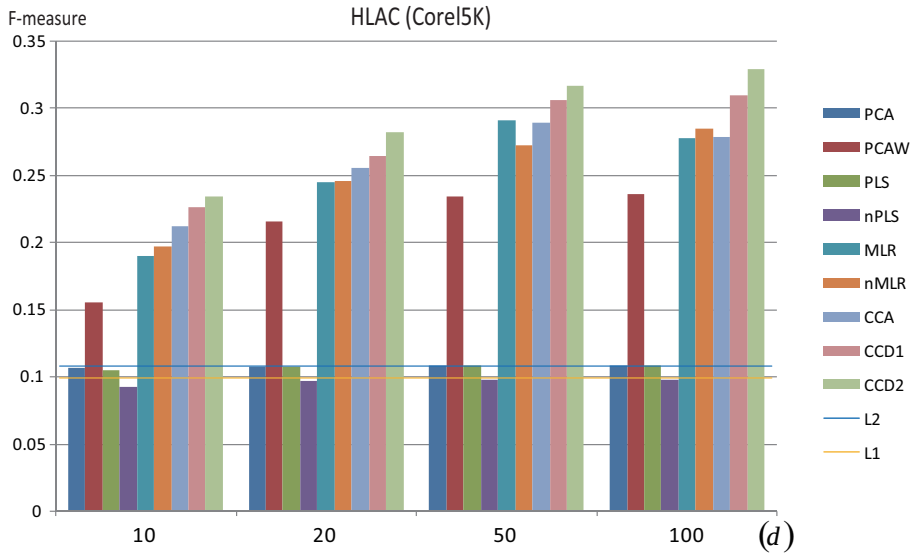


Figure 5.9: Results for the Corel5K dataset (2956-dimensional HLAC). Only linear methods are compared.

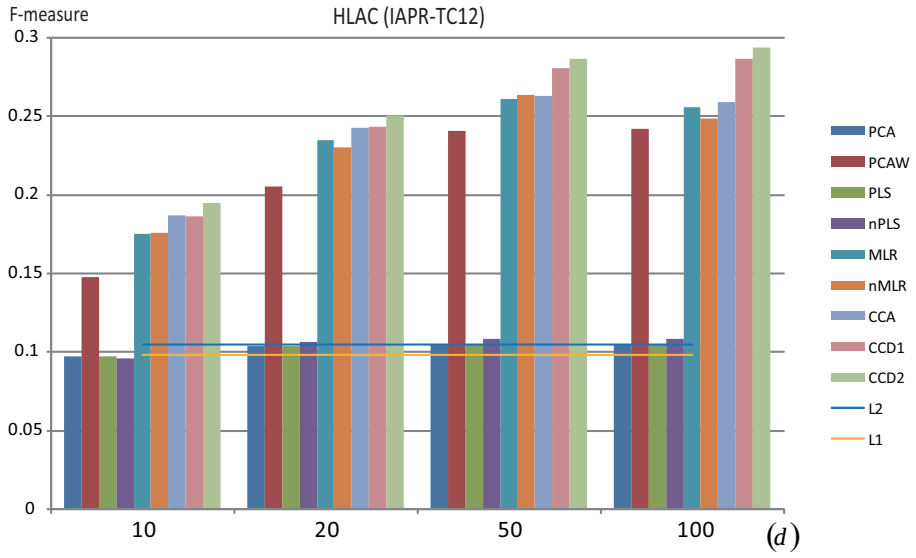


Figure 5.10: Results for the IAPR-TC12 dataset (2956-dimensional HLAC).

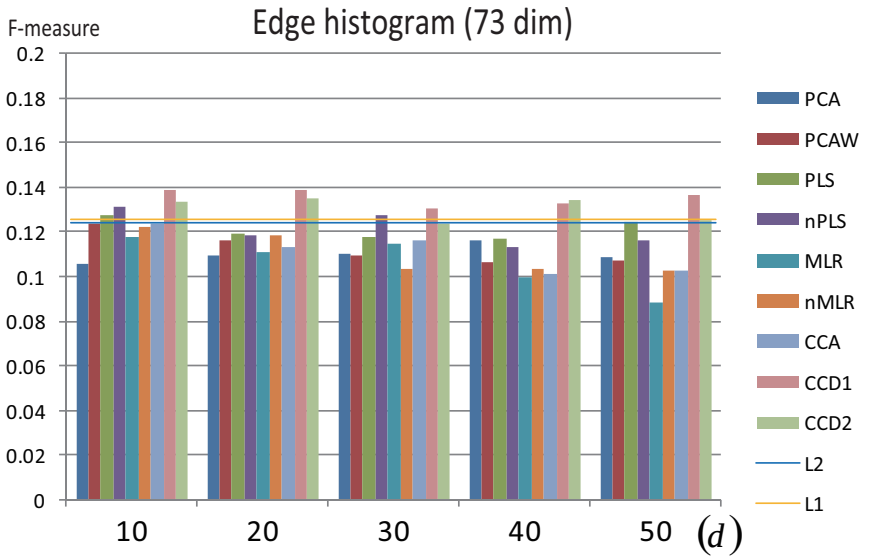


Figure 5.11: Results for the NUS-WIDE dataset (edge histogram). Methods are compared using different features with designated dimensionality (d).

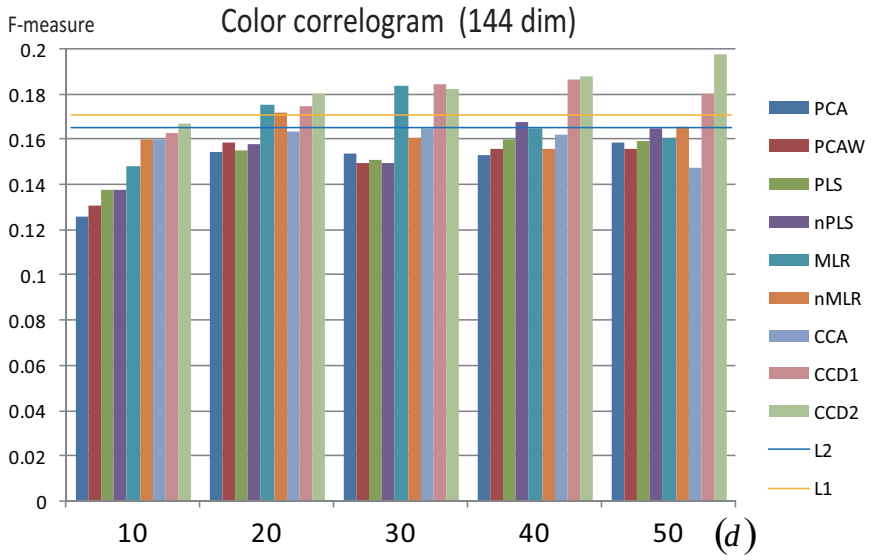


Figure 5.12: Results for the NUS-WIDE dataset (color correlogram).

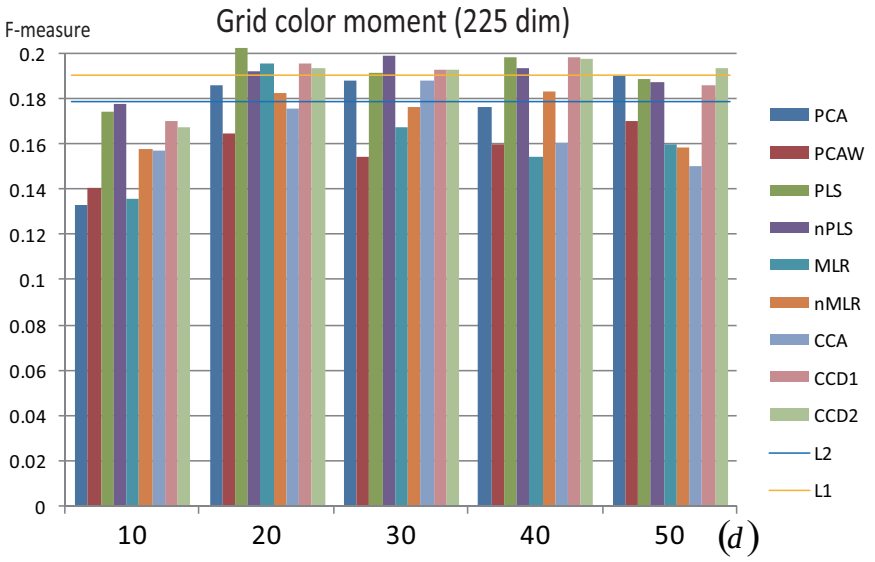


Figure 5.13: Results for the NUS-WIDE dataset (grid color moment).

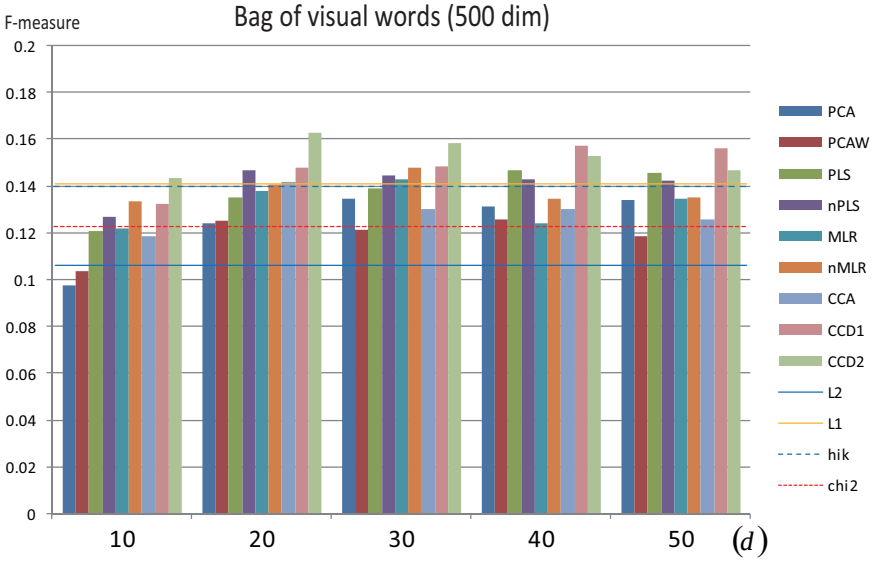


Figure 5.14: Results for the NUS-WIDE dataset (SIFT BoVW).

Table 5.2: Computation times for training the system on the NUS-WIDE dataset using each method[s]. We found that the differences in running times between PCA and PCAW, and between CCA and CCD are negligible for a small d.

	NUS-WIDE (161,789 samples, 81 words)			
	EDH (73 dim)	Cor. (144 dim)	C. mom. (225 dim)	BoVW (500 dim)
PCA (PCAW)	1.2	2.0	3.4	8.0
PLS	1.9	2.6	3.6	6.7
nPLS	3.5	5.2	7.4	14.6
MLR	2.0	2.7	4.0	8.3
nMLR	2.7	3.4	4.8	9.0
CCA (CCD)	2.1	3.0	4.5	10.1

5.3 先行研究との比較

次に，Corel5K, IAPR-TC12, ESP Game の 3 つのデータセットを用い，他の画像アノテーション手法との性能比較を行う．ここでは，キーワードベース画像検索（リトリーバル）の性能についても比較するため，リトリーバルの確率的な定式化が行いやすい MAP 推定を実装に用いる．

5.3.1 画像特徴量

以下の 3 つの場合について調べる．

(a) HLAC 特徴のみを用い，直接提案手法を適用する場合．

(b) TagProp が用いる 15 種類の画像特徴を等価な重みでカーネルベースに結合し KPCA で埋め込みを行った後，提案手法を適用する場合．

(c) TagProp が用いる 15 種類の画像特徴を異なる重みでカーネルベースに結合し KPCA で埋め込みを行った後，提案手法を適用する場合．

(a) で用いる HLAC 特徴は前節の実験と同じものである．詳細については Appendix C を参照されたい．ここでは，(b)，(c) の場合について説明を行う．TagProp では，以下の画像特徴を用いる．

1) SIFT bag-of-visual-words (Dense sampling)

2) SIFT bag-of-visual-words (Harris detector)

3) Hue bag-of-visual-words (Dense sampling)

4) Hue bag-of-visual-words (Harris detector)

5) RGB color histogram

6) HSV color histogram

7) LAB color histogram

8) GIST feature

1), 3), 6), 8) は前節の実験で用いたものと同一である．さらに，TagProp では 8) を除く全ての特徴について，画像の y 軸を 3 分割し，それぞれの領域から個別に特徴記述を行い結合したものも用いる．従って，最終的に用いる特徴量は 15 種類となる．提案手法では，GIST については L1 距離，それ以外の特徴についてはカイ 2 乗距離による GRBF カーネルを作成する．次に，それぞれのカーネルを線形に結合して最終的なカーネルとし，KPCA に用いる．

$$K_{all} = \sum_{i=1}^{N_F} \alpha_i K_i. \tag{5.1}$$

ここで，N_F は特徴の数（ここでは $N_F = 15$），K_i は i 番目の特徴由来のカーネルを示し，α_i はその相対的な重みである．各カーネルにかける重みの学習は，multiple kernel learning [107] と呼ばれる機械学習の分野において盛んに議論されている話題である．(b) では，最も単純に全てのカーネルの重みを等価に設定する．

$$K_{all}^{average} = \frac{1}{N_F} \sum_{i=1}^{N_F} K_i. \tag{5.2}$$

(c) では，よりタスクに適したカーネルの重みを学習することを考える．KCCA における MKL も提案されている [216] が，最適化の反復計算のステップごとに固有値問題を解く必要があるため，計算量が多い．ここでは，support vector regression (SVR) [50] によるラベル特徴の回帰をタスクとして MKL を行い，最適化された重みを式 5.1に用いる．SVR による回帰は，最終的な目的であるアノテーションとは必ずしも直結しないが，大まかに重要なカーネルを選択できると期待できる．MKL SVR の実装は Shogun Library [177] を用いる．

以後，(a) の場合を CCD (HLAC)，(b) の場合を CCD (15F: average+KPCA)，(c) の場合を CCD (15F: SVRMKL+KPCA) と表記する．

5.3.2 実験結果

表 5.3, 表 5.4, 表 5.5 に，それぞれ Corel5K, IAPR-TC12, ESP Game における結果をまとめる．各スコアの意味については，Appendix A を参照されたい．ここでは，カーネル化に用いる基底サンプル数を $n_K = 300$ とした．まず，提案手法のうち単純に HLAC 特徴を用いる CCD (HLAC) は，TagProp を除く最新の手法

Table 5.3: Performance comparison using Corel5K.

	MR	MP	F-m	N+	MAP	MAP (R+)
Co-occurrence [137]	0.02	0.03	0.02	19	-	-
Translation [51]	0.04	0.06	0.05	49	-	-
CMRM [92]	0.09	0.10	0.09	66	0.17	-
Maximum Entropy [93]	0.12	0.09	0.11	-	-	-
CRM [109]	0.19	0.16	0.17	107	0.24	-
NPDE [220]	0.18	0.21	0.19	114	-	-
InfNet [131]	0.24	0.17	0.20	112	0.26	-
CRM-Rectangles [60]	0.23	0.22	0.23	119	0.26	0.30
Independent SVMs [123]	0.22	0.25	0.23	-	-	-
MBRM [60]	0.25	0.24	0.25	122	0.30	0.35
AGAnn [121]	0.27	0.24	0.25	126	-	-
SML [29]	0.29	0.23	0.26	137	0.31	0.49
DCMRM [122]	0.28	0.23	0.26	135	-	-
TGLM [120]	0.29	0.25	0.27	131	-	-
MSC [197]	0.32	0.25	0.28	136	0.42	0.79
Matrix Factorization [123]	0.29	0.29	0.29	-	-	-
JEC [129]	0.32	0.27	0.29	139	0.33	0.52
JEC (15F) [72]	0.33	0.29	0.30	140	-	-
CBKP [126]	0.33	0.29	0.31	142	-	-
GS [226]	0.33	0.30	0.31	146	-	-
TagProp [72]	0.42	0.33	0.37	160	0.42	-
CCD (HLAC)	0.36	0.32	0.34	149	0.42	0.63
CCD (15F: average+KPCA, $n_K = 300$)	0.38	0.34	0.36	151	0.42	0.64
CCD (15F: SVRMKL+KPCA, $n_K = 300$)	**0.41**	**0.36**	**0.38**	**159**	**0.43**	**0.65**

(JEC, GS など) と同等の認識精度を有することが示された．これらの先行研究では，複数の画像特徴を用い性能を向上させているのに対し，提案手法は 1 つの特徴量のみで比肩する精度を達成している点に注意されたい．さらに，KPCA により複数特徴を埋め込んだ場合はより高いアノテーション・リトリーバル精度を得ている．

次に，カーネル化に用いる基底サンプル数 n_K と認識精度の関係を図 5.15に示す．このように，より多くのサンプルをカーネル化に用いるほど認識精度は向上する．また，複数のカーネルを等価な重みで結合させた場合 (15F: average) よりも，MKL により重みを学習した場合 (15F: SVRMKL) の方が精度が向上するケースが多い．この傾向は，特に n_K が小さい場合に顕著であるが，n_K が大きい場合は必ずしも優位ではなく，15F: average の方が安定によい認識精度を示す場

Table 5.4: Performance comparison using IAPR-TC12.

	MR	MP	F-m	N+	MAP	MAP (R+)
MBRM [129]	0.23	0.24	0.23	223	0.24	0.30
JEC [129]	0.29	0.28	0.30	250	0.27	0.31
JEC (15F) [72]	0.19	0.29	0.23	211	-	-
TagProp [72]	0.35	0.46	0.40	266	0.40	-
GS [226]	0.29	0.32	0.30	252	-	-
CCD (HLAC)	0.26	0.35	0.30	249	0.32	0.38
CCD (15F: average+KPCA, $n_K = 300$)	0.28	0.43	0.34	**251**	0.37	0.43
CCD (15F: SVRMKL+KPCA, $n_K = 300$)	**0.29**	**0.44**	**0.35**	**251**	**0.39**	**0.44**

Table 5.5: Performance comparison using ESP game dataset.

	MR	MP	F-m	N	MAP	MAP (R+)
MBRM [129]	0.19	0.18	0.18	209	0.18	0.24
JEC [129]	0.25	0.22	0.23	224	0.21	0.25
JEC (15F) [72]	0.19	0.24	0.21	222	-	-
TagProp [72]	0.27	0.39	0.32	239	0.28	-
CCD (HLAC)	0.18	0.27	0.22	221	0.19	0.22
CCD (15F: average+KPCA, $n_K = 300$)	**0.24**	0.33	0.28	**236**	0.26	0.30
CCD (15F: SVRMKL+KPCA, $n_K = 300$)	**0.24**	**0.36**	**0.29**	232	**0.27**	**0.31**

合もある．このことは，認識精度を重視する場合にはカーネル化に用いるサンプル数を増やすことが特に重要であることを示唆している．逆に，認識に必要な計算コストを重視し少数サンプルによりカーネル化を行う場合は，MKL による学習は効果的であるといえる．

最後に，提案手法と TagProp [72] の認識性能の詳しい比較を行う．TagProp の高い認識性能は，複数特徴量を用いたサンプル間距離計量の学習のみならず，ロジスティック判別モデルの学習により単語ごとの学習サンプル数のバイアスを緩和している点によるところが大きい．[72] に倣い，距離計量学習のみの場合を TagProp ML，ロジスティック判別モデルを加える場合を TagProp σML と表記する．比較結果（F 値）を表 5.6にまとめる．提案手法は，少数のサンプル ($n_K = 300$) を用いたカーネル化においても，Corel5K については TagProp σML，他については TagProp ML を上回る性能を得ている．TagProp σML と同様のロジスティック判別モデルの導入により，提案手法においても更に性能向上が行えることが期待されるが，これは今後の課題とする．

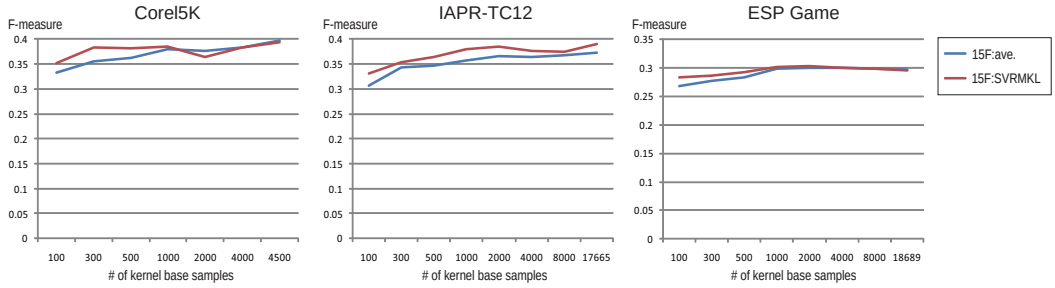


Figure 5.15: Annotation performance (F-measure) with a varying number of base samples for kernel PCA embedding.

5.3.3 計算コスト

表 5.7に，提案手法と先行研究の計算コストを，学習時と認識時に分けて示す．大規模なデータへの適用を念頭においた場合，データ数に対するスケーラビリティが特に重要であるため，学習サンプル数 N に対するコストについて考察する．また，n_K をカーネル化を行う場合に用いる基底サンプル数とする．ここでは，先行研究の中で特に優れた認識精度を示している JEC [129], GS [226], TagProp [72] との比較をまとめた．

これらの先行研究はすべて，提案手法と同じくサンプルベースの認識アルゴリズムをとっているため，アノテーションの計算コストはいずれも学習サンプル数に対し線形である．しかしながら，各サンプルとの類似度評価を行うコストが異なるため，実際の計算時間には差が表れる．JEC, TagProp は画像特徴空間においてサンプルとの類似度評価を行うためアノテーションのコストは $O(pN)$ となる．提案手法では PCCA の潜在空間においてこれを行うため，計算コストは $O(dN)$ となる．一般に，CCA による圧縮で $d < p$ となるため，計算コストは元の画像特徴空間で類似度評価を行う場合よりも軽減される．実際の計算量は画像特徴の選択によるため一概には議論できないが，例えば TagProp では 15 種類の特徴量を用いており，その次元数は合計で数万に及ぶ．これに対し提案手法で用いる潜在空間の次元数は 50 から 100 程度であり，高速に類似度評価を行うことが可能である．ただし，カーネル PCA による埋め込みを行う場合は，カーネルベースの計算を行う際に元の画像特徴空間における距離計算が必要となる．このため，カーネル化に用いる基底サンプル数が増えると，提案手法による高速化の効果は著しく低下する．同様に，GS では group sparsity を用いた学習により選択される d_{GS} 次元の特徴を用いて類似度評価が行われる．

次に，学習に必要なコストについて述べる．JEC は複数画像特徴を用い単純な最近傍識別を行う手法であるため，基本的に学習フェーズは存在しない．GS, TagProp は複数の特徴量の最適な重みを学習するために，学習データにおける全ての 2 点間距離を求める必要があるため，厳密には計算コストは $O(N^2)$ となる

Table 5.6: Comparison of annotation performance (F-measure) using TagProp.

	Corel5K	IAPR-TC12	ESP Game
TagProp ML	0.337	0.329	0.284
TagProp σML	0.369	**0.399**	**0.323**
CCD (HLAC)	0.341	0.297	0.217
CCD (15F: average+KPCA, $n_K = 300$)	0.355	0.342	0.277
CCD (15F: SVRMKL+KPCA, $n_K = 300$)	0.383	0.353	0.286
CCD (15F: SVRMKL+KPCA)	**0.394**	**0.391**	0.296

Table 5.7: Comparison of computational costs against the number of samples. N is the number of whole training samples, while n_K is the number of those used for kernelization.

	Training	Annotation
JEC [129]	-	$O(pN)$
GS [226]	$O(N^2)$	$O(d_{GS}N)$
TagProp [72]	$O(N) \sim O(N^2)$	$O(pN)$
Proposed (linear)	$O(N)$	$O(dN)$
Proposed (KPCA embedding)	$O(N + n_K^3)$	$O(dN + pn_K)$

が，[72] では近似的な実装を行い準線形オーダで学習を行っている．提案手法において中心となる CCA の学習は，共分散行列と一般化固有値問題の計算からなり，このうち学習サンプル数に依存するのは共分散行列の計算である．従って，提案手法 (linear) は学習サンプル数に対しては線形オーダで学習が実行できる．一方，KPCA の計算はカーネル化に用いるサンプル数 (n_K) の 3 乗オーダのコストがかかるため，KPCA による埋め込みは n_K が大きくなると非現実的になる．

5.4 考察

提案手法を標準的なベンチマークに適用した結果，先行研究と比べ遜色のないアノテーション精度が得られた．また，学習・認識の両面において計算コストを大きく削減できることが分かった．これは，提案手法が基本的に線形の学習手法を用いているためである．

しかしながら，画像特徴が非線形な距離計量を持つ場合，単純に提案手法を適用すると著しくアノテーション精度が低下することも確認された．このような場合は，カーネル法を用いてあらかじめサンプルをユークリッド空間に埋め込むこ

とで対処できるものの，十分な性能を得るためには多くのサンプルを用いてカーネル化を行う必要がある．結果的に，計算コストは線形の場合に比べて急激に増え，提案手法のメリットが失われる．現在，一般に普及している画像特徴量の多くは非線形な距離計量を持つため，このジレンマは多くの場面で問題となる．

一方，HLAC 特徴は線形手法と非常に相性がよく，単純に線形の枠組みで提案手法を適用するだけで，他の特徴をカーネル化した場合と同等以上の性能が得られた．これは，HLAC 特徴自体がある程度ユークリッド的な性質を有しているためであると考えられる．

このように，画像認識システムの開発にあたっては，認識手法と画像特徴量の相性を考えた設計を行うことが重要である．本章の検証により，提案手法に対しては HLAC 特徴のように，線形手法を直接適用可能な画像特徴量を用いることで，スケーラブルかつ高精度のアノテーションアルゴリズムが実現できることが示唆された．次章では，そのような条件を満たす汎用的な画像特徴抽出の枠組みを開発する．

Chapter 6

画像特徴記述手法の開発

1 本章では，CCD のような線形手法においても性能を発揮できる強力な画像特徴量を抽出する枠組みの開発を行う．

6.1 局所特徴分布からの大域特徴コーディング

大きく分けて，画像からの特徴ベクトル抽出は次の 2 つの工程に分けられる．

1. 多数の局所特徴の抽出

2. 抽出された局所特徴を一つの大域特徴ベクトルへコーディング

1 と 2 はそれぞれ最終的な特徴ベクトルの性能に関わる重要な要素であるが，最新の一般画像認識の研究では主に 2 のコーディングの工夫により大幅に認識性能が向上することが示されている [156; 201; 218; 228]．同時に，コーディング方法は識別器の設計と不可分な関係にある．そこで，本章ではコーディングの問題に取り組むことにする．コーディングにおいて鍵となるのは，画像から抽出される局所特徴分布の統計的傾向をいかにして効率よく利用するか，という問題である．

次節以降で詳しく述べるが，一般的な bag-of-visual-words (BoVW) [40] は，分布の高次の統計的特徴を疎に利用していると解釈できる．これに対し，提案手法では逆に，分布の低次の特徴を密に用いるアプローチをとる．具体的には，画像固有のガウス分布により，各画像の局所特徴の分布をモデル化し，適切なコーディング方法と距離計量を設計する．この際，情報幾何 [4] の枠組みを用いることで，スケーラブルかつ強力な線形近似を行うことができる．

提案手法はシンプルなアプローチであるが，様々なデータにおいて良好な性能を得ている．また，提案手法は BoVW と相補的な関係にあるため，両者を同時に用いることでさらに認識精度が向上する．

6.2 先行研究

ここでは，局所特徴分布の利用という観点から先行研究について論じる．一般的に，画像認識で用いる局所特徴は高次元である場合が多い．例えば，最も広く用いられている SIFT [124] は 128 次元である．しかしながら，画像一枚あたりから抽出される局所特徴数は通常高々数千個ほどである．このため，安定に分布情報を利用することは容易ではない．表 6.1に，分布モデルの複雑さ，コーディングの疎密を基準に先行研究を分類する．以下，それぞれ説明を行う．

6.2.1 Non-parametric method

分布を活用する上でもっとも直接的なアプローチは，コーディングのプロセスを陽に行わずに，生の局所特徴をノンパラメトリックに利用することである．Boiman ら [21] は，Naive-Bayes nearest neighbor (NBNN) 識別則を提案した．NBNN では，クエリ画像中のすべての局所特徴について，全学習画像の局所特徴の中から最近傍サンプルを選び，その投票により識別を行う．NBNN はシンプルな方法ながら，2008 年当時に他手法を大きく上回る認識精度を記録し，大きな注目を浴びた．これは，サンプルベースのアプローチをとることで，分布情報を陰な形で比較的安定に利用できているためであると考えられる．しかしながら，認識を行うためには全学習画像のすべての局所特徴を保持しておき，クエリ画像の局所特徴との距離計算を行う必要があるため，計算コストは極めて重い．このため NBNN は実用性には欠ける手法であると言わざるを得ないが，局所特徴の利用法に一石を投じた点で意義深い研究である．

6.2.2 Gaussian Mixtures

Gaussian mixture model (GMM) は代表的なパラメトリック確率モデルの一つであり，局所特徴分布のモデル化にも応用されている．最初の試みとしては，Vasconcelos らの研究 [136; 192] が挙げられる．彼らは，GMM により各画像の局所特徴分布の生成モデルを推定するとともに，分布間類似度を畳み込んだカーネル関数を SVM に適用することで識別に利用するアプローチを提案した．GMM を用いた方法は，分布の高次な統計的特徴を密に利用していると解釈でき，理想的には識別に有効な情報を最も多く与えうる．しかしながら，前述のように一枚の画像から抽出できる局所特徴数は限りがあるため，画像固有の GMM を安定に推定することは実際にはほぼ不可能である．Vasconcelos らの研究においても，実際の実験において GMM によるモデル化は実装しておらず，単一のガウシアンの利用に留まっている．

このため実際は，GMM は学習データセット全体から構築し，個々の画像の分布をそこからの差分として表現するアプローチが主流である [153; 227; 229]．Zhou ら [227] は，各画像の GMM をこのアプローチにより推定し，そのパラメータを画像の特徴ベクトルとして用いた．また，Perronnin ら [153] は，データセッ

トの GMM からの差分を Fisher スコアベクトルの形で表し，Fisher カーネルを用い画像識別に利用した．これらの手法は高い認識精度を示す一方，特徴ベクトルはデータセット全体の生成モデル (GMM) に強く依存した形になる．他のタスクへ適用するためには GMM の再構築が必要であるが，GMM の計算コストは非常に大きい点が問題となる．

6.2.3 Bag-of-Visual-Words

Bag-of-visual-words (bag-of-keypoints) [40] は，文書特徴の一つである bag-of-words [130] を画像認識へ応用したものであり，現在一般画像認識におけるデファクトスタンダードのアプローチとなっている．まず，GMM の場合と同様に，学習データセット全体の局所特徴をクラスタリングすることにより，いくつかのクラスタ中心 (visual words) を得る．一般的には，クラスタリングは k-means 法によって行われる．各画像の局所特徴は最も近い visual word へ割り当てられ，最終的に visual word のヒストグラムがその画像の特徴ベクトルとなる．この手法は，GMM において各ガウシアンの混合比のみを特徴として用いていると解釈できる．この点で，bag-of-visual-words は高次の統計的特徴を疎に利用しているといえる．

Bag-of-visual-words は，比較的低い計算コストで良好な認識精度を得られることから注目を浴びてきた．しかしながら解決すべき重要な課題がいくつかあり，精力的に改良がなされている．まず，量子化方法の問題が挙げられる．標準的な k-means 法では，データセット全体においてサンプルが密に分布している場所にセントロイドがおかれるため，必ずしも個々の画像を表現するために適した visual words が生成できない．この問題に対し，Jurie ら [95] は radius-based mean-shift clustering を利用することで，より適切なコードブック (visual words) が生成できることを示した．Wu ら [211] はクラスタリング時の距離計量として histogram intersection が一般的にすぐれた性能を発揮することを示した．また，局所特徴のコーディングの際に，最近傍の visual word に排他的に割り当てるのではなく，いくつかの visual word に重複を許して割り当てる soft assignment strategy [154; 191] のアプローチも研究されている．特に近年，少数の visual words に割り当てを行うスパースコーディングにより強力な特徴ベクトルが生成できることが示されており，大きな注目を浴びている [201; 218]．

もちろん，クラスタリング以外の量子化方法も検討されている．例えば，Tuytelaars ら [187] は局所特徴空間のグリッド分割とハッシュに基づく量子化方法を提案している．また，Shotton ら [171] は，random decision forest を利用した高速な量子化を実現した．

6.2.4 Covariance Descriptor

低次の統計量を利用する手法としては，Tuzel らの提案した covariance descriptor [188; 189] が挙げられる．これは，局所特徴の共分散行列を画像特徴ベクトルと

Table 6.1: Summary of previous work and our work from the viewpoint of local feature statistics.

	High-level	Low-level
Dense	Non-parametric [21] Gaussian mixture model [136; 227]	Covariance [188; 189] This work (single Gaussian)
Sparse	Bag-of-visual-words [40; 218]	

してコーディングするものであり，ガウス分布に基づく提案手法と密接に関係する．共分散行列は，リーマン多様体上の点として表される．さらに，微分幾何の手法により多様体の構造を活用し，LogitBoost による学習を行う。この手法は，人検出タスクにおいて高い認識精度を示している．共分散は代表的な低次統計量であり，各画像固有の限られた数の局所特徴からでも比較的安定に抽出可能であると期待できる．

6.3 提案手法： Global Gaussian Approach

各画像を，画像固有の局所特徴が為すガウス分布として表現する．これを global Gaussian approach [144] と呼ぶことにする．ここでの global とは，局所特徴空間全体において一つのガウス分布を張ることを意識したものであり，空間中の局所的な構造を推定する GMM や bag-of-visual-words と対比を為している．

6.3.1 情報幾何に基づくガウシアンのコーディング

ある画像 I_j から，D 次元局所特徴 $\{\boldsymbol{v}_k\}$ を抽出するものとする．I_j は，パラメータ $\boldsymbol{\theta}(j)$ を持つガウス分布 $p_j(\boldsymbol{v};\boldsymbol{\theta}(j))$ によって説明される．各サンプルの為すガウス分布は，リーマン多様体上の一点として表される．さらに，情報幾何の手法を用い，理論的に保証された計量をカーネル関数として用い，画像識別へ応用する．

確率分布から出発する自然な帰結として，最適なカーネルは Kullback-Leibler (KL) ダイバージェンスに基づくものであることが示される．これは，[136] において用いられたものと基本的に同一であり，global Gaussian approach の精度面における上限を与える．しかしながら，KL ダイバージェンスは高コストな非線形の計算を必要とするため，スケーラビリティに乏しい．そこで，情報幾何の手法を用いることで，KL ダイバージェンスを近似するコーディング方法を導出する．これにより，本章の最終的な目的である線形評価可能な特徴ベクトルが得られる．

6.3.2 情報幾何の紹介

情報幾何 [4] は微分幾何学に基づく体系であり，統計的学習手法の幾何的な理解を目的として始まった [4]. n 個の実数パラメータ $\boldsymbol{\theta} = (\theta^1, ..., \theta^n)$ を持つ確率変数 $\boldsymbol{v}$ の確率分布 $p(\boldsymbol{v}; \boldsymbol{\theta})$ を考える．$\boldsymbol{\theta} = (\theta^1, ..., \theta^n)$ を座標系と考えると，確率分布モデル全体はリーマン多様体とみなすことができる．個々の確率分布は，多様体上の一点に対応する．情報幾何は，この多様体に統計的に自然な構造を入れる．まず，フィッシャー情報行列をリーマン計量として利用する．

$$G^{\theta}_{lm}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\frac{\partial \log p(\boldsymbol{v}; \boldsymbol{\theta})}{\partial \theta^l} \frac{\partial \log p(\boldsymbol{v}; \boldsymbol{\theta})}{\partial \theta^m}\right]. \tag{6.1}$$

多様体上の各点のごく近傍はそれぞれユークリッド空間であるとみなせる．これはその点における接空間と呼ばれ，内積がリーマン計量によって定義される．次に，これらの接空間の接続を行う．情報幾何では，α-接続と呼ばれる接続を考える．α は多様体の構造を決定するパラメータである[1]．いくつかの特別な確率モデルに関しては，適切なパラメータ座標系 ξ をとることにより，接空間の接続係数が全て 0 となる「平らな」接続を行うことができる．そのような座標系 ξ が存在するとき，モデル空間 (多様体) は α-平坦であると定義され，ξ は α-アフィン座標系と呼ばれる．α-平坦な空間では，測地線は α-座標系上の線となる（α-測地線）．また，α-平坦な空間は常に $-\alpha$-平坦であることが知られており，ξ-座標系と双対な別のアフィン座標系（$-\alpha$-アフィン座標系）が存在する．以下でさらに詳しく述べるが，情報幾何では $\alpha = \pm 1$ の場合が特に重要となる[2]．実際，統計学習において広く用いられている確率モデルの多くに ± 1-座標系が存在することが知られている．このため，情報幾何はさまざまな学習手法の幾何的な分析と解釈に利用されてきた．例えば，EM アルゴリズム [3]，ブースティング [141]，変分ベイズ法 [85] に適用した例が挙げられる．詳しくは，[4] を参照されたい．

指数分布族は実用上最も重要な確率モデルであり，情報幾何においても中心的な役割を果たす．指数分布族は以下の確率関数により定義される．

$$p(\boldsymbol{v}; \boldsymbol{\theta}) = \exp\left(\sum_{i=1}^{n} \theta^i F_i(\boldsymbol{v}) - \psi(\boldsymbol{\theta}) + C(\boldsymbol{v})\right). \tag{6.2}$$

ここで，$\boldsymbol{\theta}$ はモデルのパラメータであり，F は観測される確率変数 $\boldsymbol{v}$ に依存する関数（測定関数）である．$\psi(\boldsymbol{\theta})$ はポテンシャル関数であり，$C(\boldsymbol{v})$ は $\boldsymbol{\theta}$ と独立な関数である．指数分布族は 1-平坦であり，$\boldsymbol{\theta}$ が対応する 1-アフィン座標系である．上述のように，$\boldsymbol{\theta}$ と双対な別のアフィン座標系 $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)$ をとることができ，$\eta_i = E_{\boldsymbol{\theta}}[F_i(\boldsymbol{x})]$ によって定義される．$\boldsymbol{\eta}$-座標系は十分統計量の空間であると解釈

[1] $\alpha = 0$ の場合が，物理学等でしばしば登場する Levi-Civita 接続である．

[2] 情報幾何では，1-接続や 1-平坦などの言葉は特別に e-接続，e-平坦 (e:exponential), -1-接続や-1-平坦などの言葉は m-接続，m-平坦 (m:mixture) とそれぞれ定義されているが，ここでは簡単のため省略する．

できる．$\boldsymbol{\eta}$-座標系のリーマン計量は $\boldsymbol{\theta}$-座標系の計量 (G^{θ}, 式 6.1) の逆行列となる．これは，以下の変換により明示的に導出することができる．

$$G^{\eta}_{lm} = \frac{\partial \theta^l}{\partial \eta_m}. \tag{6.3}$$

6.3.3 Generalized Local Correlation (GLC)

ガウシアンも指数分布族に属しており，$n = D + D(D+1)/2$ のパラメータを持つ．μ と Σ をそれぞれ分布の平均と共分散とする．式 6.2について，

$$\begin{aligned} C(\boldsymbol{v}) = 0, \quad F_i(\boldsymbol{v}) = v_i, \quad F_{ij}(\boldsymbol{v}) = v_i v_j \ (i \leq j), \\ \theta^i = \sum_{j=1}^{D} (\Sigma^{-1})_{ij}\mu_j, \quad \theta^{ii} = -\frac{1}{2}(\Sigma^{-1})_{ii}, \quad \theta^{ij} = -(\Sigma^{-1})_{ij} \ (i < j), \end{aligned} \tag{6.4}$$

のようにとれば，ガウス分布も同様の形で表される．

$$p(\boldsymbol{v};\boldsymbol{\theta}) = \exp\left[\sum_{1\leq i\leq D} \theta^i F_i(v) + \sum_{1\leq i\leq j\leq D} \theta^{ij} F_{ij}(v) - \psi(\boldsymbol{\theta})\right]. \tag{6.5}$$

ただし，

$$\psi(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} + \frac{1}{2}\log(2\pi)^D|\Sigma|. \tag{6.6}$$

また，$\boldsymbol{\theta}$-座標系と双対な $\boldsymbol{\eta}$-座標系は次のようになる．

$$\eta_i = \mu_i, \quad \eta_{ij} = \Sigma_{ij} + \mu_i\mu_j \ (i \leq j). \tag{6.7}$$

$\boldsymbol{\theta}$-座標系はモデルパラメータの空間であり，$\boldsymbol{\eta}$-座標系は十分統計量の空間である．分布から生成されるサンプル（局所特徴）が十分に観測できる理想的な条件においては，どちらの座標系を用いてもよい．しかしながら，現実的には各サンプル（画像）について限られた数の局所特徴しか観測することができない．このため，観測から推定される統計量を用い，各サンプルを $\boldsymbol{\eta}$-座標系へプロットする．$\boldsymbol{e}_i, \boldsymbol{e}_{ij}$ をそれぞれ η_i 軸，η_{ij} 軸に対応する基底ベクトルとする．$\boldsymbol{\eta}$-座標系は，以下のように記述される．

$$\begin{aligned} \boldsymbol{\eta} &= \sum_{1\leq i\leq D} \eta_i \boldsymbol{e}_i + \sum_{1\leq i\leq j\leq D} \eta_{ij}\boldsymbol{e}_{ij} \\ &= (\eta_1, ..., \eta_D, \eta_{11}, ..., \eta_{1D}, \eta_{22}, ...\eta_{2D}, ..., \eta_{DD})^T \\ &= \big(\hat{\mu}_1, ..., \hat{\mu}_D, \hat{\Sigma}_{11} + \hat{\mu}_1^2, ..., \hat{\Sigma}_{1D} + \hat{\mu}_1\hat{\mu}_D, \\ &\qquad \hat{\Sigma}_{22} + \hat{\mu}_2^2, ..., \hat{\Sigma}_{DD} + \hat{\mu}_D^2\big)^T. \end{aligned} \tag{6.8}$$

式 6.8が示すように，$\boldsymbol{\eta}$-座標は観測される局所特徴の各要素の平均と相関を列挙したものになっている．このコーディングを，generalized local correlation (GLC) と呼ぶことにする．これは，古典的な HLAC 特徴 [149] の自然な一般化になっており，任意の局所特徴記述子に対して汎用的に適用可能な表現形式であるといえる．$\boldsymbol{\eta}$-座標系のリーマン計量は以下のようになる．

$$
\begin{aligned}
G^{\eta}_{ij} &= (\Sigma^{-1})_{ij}(1+\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu})+ \\
&\qquad \sum_{k=1}^{D}\mu_k(\Sigma^{-1})_{ki}\sum_{k=1}^{D}\mu_k(\Sigma^{-1})_{kj}, \\
G^{\eta}_{i(pq)} &= -(\Sigma^{-1})_{pi}\sum_{k=1}^{D}\mu_k(\Sigma^{-1})_{kq}- \\
&\qquad (\Sigma^{-1})_{qi}\sum_{k=1}^{D}\mu_k(\Sigma^{-1})_{kp}\ \ (p<q), \\
G^{\eta}_{i(pp)} &= -(\Sigma^{-1})_{pi}\sum_{k=1}^{D}\mu_k(\Sigma^{-1})_{kp} \\
G^{\eta}_{(pq)(rs)} &= (\Sigma^{-1})_{ps}(\Sigma^{-1})_{qr}+(\Sigma^{-1})_{qs}(\Sigma^{-1})_{pr} \\
&\qquad\qquad (p<q, r<s), \\
G^{\eta}_{(pq)(rr)} &= (\Sigma^{-1})_{pr}(\Sigma^{-1})_{rq}\ \ (p<q), \\
G^{\eta}_{(pp)(rr)} &= \frac{1}{2}(\Sigma^{-1})^2_{pr}.
\end{aligned}
\tag{6.9}
$$

添え字は式 6.8と対応する．例えば，$G^{\eta}_{i(pq)} = \langle \boldsymbol{e}_i, \boldsymbol{e}_{pq}\rangle$，$G^{\eta}_{(pq)(rr)} = \langle \boldsymbol{e}_{pq}, \boldsymbol{e}_{rr}\rangle$ である．

6.3.4 カーネル関数

KL divergence based kernel

情報幾何では，2 つの確率分布 $f(\boldsymbol{v})$, $g(\boldsymbol{v})$ に対応する多様体の 2 点 $P : f(\boldsymbol{v})$, $Q : g(\boldsymbol{v})$ 間の α-ダイバージェンスが以下のように定義される．

$$
D^{(\alpha)}(P||Q) = \psi\big(\boldsymbol{\theta}(P)\big) + \varphi\big(\boldsymbol{\eta}(Q)\big) - \sum_{i=1}^{n}\theta^i(P)\eta_i(Q). \tag{6.10}
$$

ここで，$\varphi(\boldsymbol{\eta})$ は $\boldsymbol{\eta}$-座標系におけるポテンシャル関数である．α-ダイバージェンスは，情報幾何において重要な役割を果たす．直感的には，2 点 PQ 間の非類似度を示すものであるが，対象律が成り立たないため厳密には距離計量ではない．なお，双対な $-\alpha$-ダイバージェンスは $D^{(-\alpha)}(P||Q) = D^{(\alpha)}(Q||P)$ となる．指数分布族の場合，1-ダイバージェンス ($\alpha = 1$) は $f(\boldsymbol{x})$ と $g(\boldsymbol{x})$ の KL ダイバージェンスに一致する．

$$
k(f||g) = \int f(\boldsymbol{x})\left[\log f(\boldsymbol{x}) - \log g(\boldsymbol{x})\right] d\boldsymbol{x}. \tag{6.11}
$$

また，双対な -1-ダイバージェンス $(\alpha = -1)$ は $k(g||f)$ となる．我々は -1-平坦な $\boldsymbol{\eta}$-座標系を利用するため，-1-ダイバージェンスをカーネルに利用することを考える．[136] のアプローチに従い，逆向きのダイバージェンスと和をとることで対称化を行い，2 点間の距離を定義する．

$$\begin{aligned} & dist\big(\boldsymbol{\eta}(P), \boldsymbol{\eta}(Q)\big) \\ = \; & D^{(-1)}(P||Q) + D^{(-1)}(Q||P) \\ = \; & k(g||f) + k(f||g) \\ = \; & tr(\Sigma_P \Sigma_Q^{-1}) + tr(\Sigma_Q \Sigma_P^{-1}) - 2d + \\ & tr\left((\Sigma_P^{-1} + \Sigma_Q^{-1})(\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q)(\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q)^T\right). \end{aligned} \tag{6.12}$$

これを指数化することで，Mercer の条件を満たすカーネル関数が得られる．

$$K_{kl}(P, Q) = \exp\big(-a \; dist\big(\boldsymbol{\eta}(P), \boldsymbol{\eta}(Q)\big)\big). \tag{6.13}$$

ここで，a は平滑化パラメータである．KL ダイバージェンスを利用するためには，共分散行列の逆行列の計算が必要となり，画像から抽出される局所特徴数が少ない場合に不安定となる．このため，共分散行列に正則化項を加えることで安定性の向上を行う．すなわち，$\Sigma \to \Sigma + bI$ と置き換える．b は正の実数パラメータである．これは，観測される局所特徴にホワイトノイズを加える処理と等価である．

Ad-hoc linear kernel

まず，近似線形化の最もシンプルなベースラインとして，$\boldsymbol{\eta}$-座標系に直接線形カーネルを適用することを考える．これは多様体の計量を完全に無視する強い近似であり，局所特徴記述子の性質やスケーリングの影響を大きく受けると予想される．このカーネルを，ad-hoc linear kernel (ad-linear) と呼ぶことにする．

$$K_{ad}(P, Q) = \boldsymbol{\eta}(P)^T \boldsymbol{\eta}(Q). \tag{6.14}$$

Center tangent linear kernel

より適切な近似を行う為には，式 6.9のリーマン計量を利用する必要がある．この計量は，$\boldsymbol{\eta}$-座標系の各点で異なる値をとるため，効率のよい利用法を考える必要がある．ここでは，$\boldsymbol{\eta}$-座標系における学習サンプルの中心 $\boldsymbol{\eta}_c = \frac{1}{N}\sum_i^N \boldsymbol{\eta}(i)$ における計量のみで代表させる．これは，先行研究である e(m)-PCA [2] が初期化に用いる近似方法を参考にしたものである．

$$K_{ct}(P, Q) = \boldsymbol{\eta}(P)^T G^{\eta}(\boldsymbol{\eta}_c) \boldsymbol{\eta}(Q). \tag{6.15}$$

ここで，$G^{\eta}(\boldsymbol{\eta}_c)$ は点 $\boldsymbol{\eta}_c$ における計量である．これは，$\boldsymbol{\eta}_c$ の接空間によりモデル空間を近似していると解釈できる．このカーネルを center tangent linear kernel

(ct-linear) と呼ぶことにする．Ct-linear kernel の実装は，以下の線形変換を加えた $\boldsymbol{\zeta}$-座標系に通常の線形カーネルを適用することで実現できる．

$$\boldsymbol{\zeta} = \left(G^{\eta}(\boldsymbol{\eta}_c)\right)^{1/2}\boldsymbol{\eta}. \tag{6.16}$$

従って，学習のスケーラビリティを損なうことなく，ad-hoc linear kernel から近似の精度を向上させることができる．

6.4 カーネル学習器による厳密な評価

6.4.1 データセット

まず，情報幾何によって導出される GLC の有効性を確認するとともに，理論的上限である KL ダイバージェンスと比較するために，前節で開発したカーネル関数を用いカーネル学習器を構築し識別性能を評価する．なお，ここでは統一的な比較のため全てのカーネルについてカーネル学習器を用いるが，線形カーネルによる学習は元の座標系に直接線形学習器を適用することと基本的に等価である点に注意されたい．

本節では，シーン認識（カテゴライゼーション）に関連する 3 つのデータセットを用いる．一つ目は，Lazebnik ら [110] の 15 クラスのシーン画像データセットである (LSP15)．LSP15 は，シーン認識において多くの研究で標準的に用いられてきたデータセットであり，10 クラスの屋外シーン，5 クラスの室内シーン画像から構成される．二つ目は，Li ら [115] の 8 クラスのスポーツ画像データセットである (8-sports)．8-sports の画像は，背景の競技場所と前景のアスリートによって特徴づけられ，シーン認識と物体認識の両方の要素を備えている．三つ目は，Quattoni ら [160] の 67 クラスの室内シーン画像データセットである (Indoor67)．多くのクラスを有し，かつ画像のクラス内分散が大きいため従来よりも挑戦的なデータセットであるといえる．図 6.1に，それぞれのデータセットの画像例を示す．

実験は，先行研究と同じプロトコルに従って行う．LSP15 では，各クラスについてランダムに 100 個ずつ学習サンプルを選び，残りをテストサンプルとする．8-sports では，各クラスについてランダムに 80 個の学習サンプルと 70 個のテストサンプルを重ならないように選ぶ．同様に，Indoor67 では各クラスについて 80 個の学習サンプルと 20 個のテストサンプルを選ぶ．認識性能は，各クラスの認識率の全クラス平均によって測る[1]．このスコアを，学習サンプルとテストサンプルをランダムに入れ替えながら何度も測定し，その平均値をもって最終的な性能指標とする．本実験では，10 回の試行の平均値を用いた．

[1] すなわち，Confusion matrix の対角要素の平均値である．

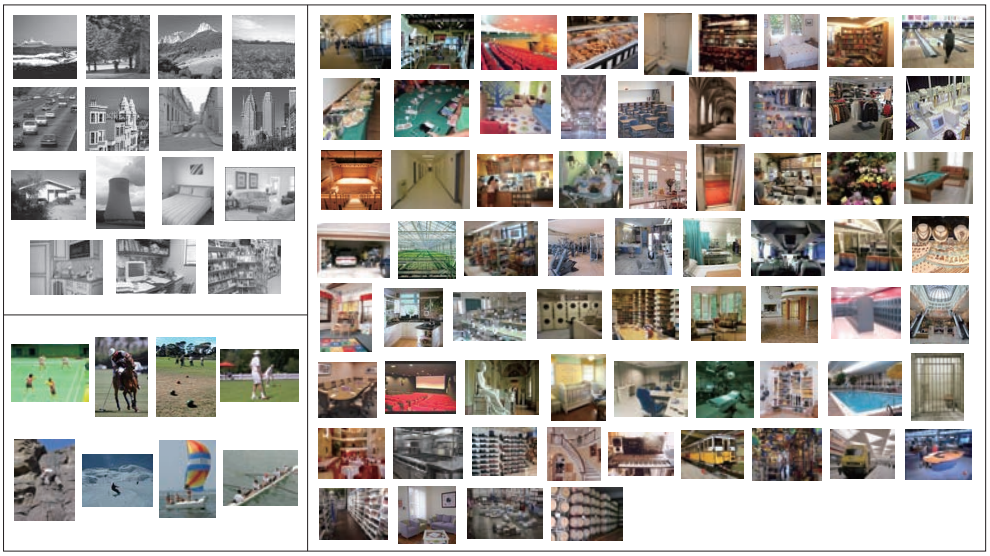


Figure 6.1: Images from benchmark datasets. Top left: LSP15 [110]. Bottom left: 8-sports [115]. Right: Indoor67 [160].

6.4.2 識別手法

本節の実験では，2つの識別器を用いる．一つ目は，support vector machine (SVM) である．SVM は，画像カテゴライゼーションタスクにおいて標準的に用いられる識別器である．二つ目は，probabilistic discriminant analysis (PDA) [87] による識別器である．PDA は，古典的な線形判別分析 (LDA) の確率的な解釈を行ったものであり，LDA の構造を利用した最適なベイズ識別器を構築することができる．他クラス識別器を構築するにあたり，SVM では複数の 2 クラス識別器の構築と組み合わせが必要となるが，PDA による識別器は一般化固有値問題を一度解くだけで構築できる．SVM と PDA のそれぞれについて，前節で開発したカーネル関数を適用しカーネル化を行う．SVM の実装には，LIBSVM [33] を用いる．以下，PDA を用いた識別器の構築について説明を行う．

Probabilistic discriminant analysis (PDA)

まず，基本となる線形の場合について述べる．Σ_w をクラス内共分散行列，Σ_b をクラス外共分散行列とする．LDA の解は次の一般化固有値問題を解くことによって得られる．

$$\Sigma_b W = \acute{\Sigma}_w W \Lambda \quad (W^T \acute{\Sigma}_w W = I). \tag{6.17}$$

ここで，$\acute{\Sigma}_w = \Sigma_w + \gamma I$ である．γ は，過学習を防ぐ正則化項の大きさを決定する正の実数パラメータであり，実験的に決定される．W は固有ベクトルを並べた

行列であり，Λ は，これに対応する固有値（判別規準）を大きい順に並べた対角行列である．

K をクラス数，$t = N/K$ を各クラスあたりの学習サンプル数，$\boldsymbol{\mu}_\eta$ を全学習サンプルの画像特徴の平均ベクトルとする．画像特徴 $\boldsymbol{\eta}$ は以下の射影により潜在空間（判別空間）へマッピングされる．

$$\boldsymbol{u} = \left(\frac{t-1}{t}\right)^{1/2} W^T(\boldsymbol{\eta} - \boldsymbol{\mu}_\eta). \tag{6.18}$$

また，潜在変数の分散は以下の式のようになる．

$$\Psi = \max\left(0, \frac{t-1}{t}\Lambda - \frac{1}{t}\right). \tag{6.19}$$

この確率構造を用い，新規入力画像 $\boldsymbol{\eta}_s$ を最尤推定によって識別する．$\boldsymbol{\eta}_s$ の推定点である $\boldsymbol{u}_s$ が，以下のようにクラス C から生成される確率を考える．

$$p(\boldsymbol{u}_s|\boldsymbol{u}^C_{1\ldots t}) = \mathcal{N}\left(\boldsymbol{u}_s|\frac{t\Psi}{t\Psi + I}\bar{\boldsymbol{u}}^C, I + \frac{\Psi}{t\Psi + I}\right). \tag{6.20}$$

ここで，$\boldsymbol{u}^C_{1\ldots t}$ はクラス C に属する t 個の独立な学習サンプルの潜在変数を示す．また，$\bar{\boldsymbol{u}}^C$ はその平均である．$\boldsymbol{\eta}_s$ は，式 6.20を最大とするクラスへ識別される．これは，潜在空間においてユークリッド距離を用いた nearest-centroid 識別とほぼ等価な識別則になっている[1]．

Kernelized PDA

カーネル判別分析 (KDA) は，カーネルトリックにより陰に生成される高次元特徴空間において LDA を行うものである．従って，上述の PDA による確率構造を同様に利用することが可能である．以下，これを KPDA を表記する．ある高次元空間への射影 $\phi : \boldsymbol{\eta} \to \phi(\boldsymbol{\eta})$ が，カーネル関数 $K(\boldsymbol{\eta}(i), \boldsymbol{\eta}(j)) = \langle\phi(\boldsymbol{\eta}(i)), \phi(\boldsymbol{\eta}(j))\rangle$ によって陰に与えられるとする．

N を全学習サンプル数，$\boldsymbol{\eta}^K = \big(K(\boldsymbol{\eta}, \boldsymbol{\eta}(1)), ..., K(\boldsymbol{\eta}, \boldsymbol{\eta}(N))\big)^T$ をそれらにより与えられるカーネルベースベクトルとする．KDA の定式化は結果的に，$\boldsymbol{\eta}^K$ をあらたな入力ベクトルと見なし，この上に LDA をかけた形になる．すなわち，Σ^K_w, Σ^K_b をそれぞれカーネルベースベクトルのクラス内共分散行列，クラス外共分散行列とすると，KDA は以下の一般化固有値問題として定式化される．

$$\Sigma^K_b V = \acute{\Sigma}^K_w V \Lambda^K \qquad (V^T \acute{\Sigma}^K_w V = I). \tag{6.21}$$

$\acute{\Sigma}^K_w$, V, Λ^K などの定義は，式 6.17と同様である．同様に，潜在空間へのマッピングは以下のように得られる．

$$\boldsymbol{u}^K = \left(\frac{t-1}{t}\right)^{1/2} V^T(\boldsymbol{\eta}^K - \boldsymbol{\mu}^K_\eta). \tag{6.22}$$

[1] $t \to \infty$ の時，両者は完全に一致する．

ただし，$\boldsymbol{\mu}_{\eta}^{K}$ は学習サンプルのカーネルベースベクトルの平均である．最終的に，式 6.20と同じ識別則を利用し，新規サンプルの識別を行う．

6.4.3 セットアップ

局所特徴抽出

一般的に，局所特徴抽出は 2 つの工程からなる．すなわち，特徴点の検出と，検出された特徴点における局所特徴抽出である．

特徴点検出の方法として，コーナー点検出 [76] や Difference of Gaussian フィルタ [124] を用いた，視覚的顕著性により特徴点を決定するアプローチが古くから用いられてきた．しかしながら，ボトムアップな視覚的顕著性は必ずしも画像の意味的な差異とは関連しない．Nowak ら [147] は，さまざまな特徴点検出手法に基づく bag-of-visual-words による画像認識性能を比較した結果，ランダムに特徴点をサンプリングするアプローチが最もよい性能を得ることを報告している．同時に，認識性能を決定する上で最も重要なのは抽出する局所特徴の数であることを指摘している．また，Fei-Fei ら [58] は 13 クラスのシーン画像データセットを用いた実験において，画像を均等にグリッド分割し各局所領域で特徴記述を行う方法が最もよい性能を得ることを報告している．これは，dense sampling と呼ばれる方法であり，一般画像認識における特徴記述方法として現在標準的に用いられているアプローチである [25; 58; 147; 211; 227]．以上を踏まえ，本実験においても dense sampling による特徴記述を行う．具体的には，16×16 ピクセルのセルを 5 ピクセルずつスライドさせながら，各セルから局所特徴を抽出する．

局所特徴記述子としては，SIFT [124] (128-dim) と SURF [10] (64-dim) の 2 つを用いる．Mikolajczyk ら [132] は，さまざまな画像認識のタスクにおいて SIFT 特徴が平均的に最も良い性能を得ることを報告している．また，SURF 特徴は，計算コストが大きく低減しているにも関わらず SIFT に匹敵する性能を持つ特徴記述子として広く用いられている．

空間情報の利用

一般画像認識において標準的に用いられる spatial pyramid kernel [110] にならい，画像のおおまかな位置情報を識別に利用することを考える．まず，画像を第 0 層から第 L 層まで階層的にグリッド分割する．第 l $(0 \leq l \leq L)$ 層の画像はそれぞれ $2^l \times 2^l$ の小領域へ分割される．それぞれの小領域において独立に $\boldsymbol{\eta}$ 座標系を生成し，K_{kl} や K_{ct} 等のカーネル関数を生成する．最終的に，これらを以下のように一つのカーネル関数へ統合する．

$$K^{GG}(P,Q) = \frac{1}{\sum_{i=0}^{L}\beta^i} \sum_{l=0}^{L} \frac{\beta^l}{2^{2l}} \sum_{k=1}^{2^{2l}} K^{(l,k)}(P,Q). \tag{6.23}$$

Table 6.2: Basic results of the global Gaussian approach with the LSP15 and 8-sports datasets using different kernels (%). No spatial information is used here.

		LSP15		8-sports	
		SIFT	SURF	SIFT	SURF
KPDA	ad-linear	77.3	75.9	77.9	72.4
	ct-linear	78.8	78.5	79.7	78.1
	KL div.	80.4	**81.5**	**81.7**	79.6
SVM	ad-linear	69.9	72.1	70.6	70.2
	ct-linear	75.7	77.7	75.5	73.3
	KL div.	76.3	**78.3**	**78.3**	74.9

β は階層の相対的な重みを決定する正の実数パラメータである．また，添え字 (l, k) はその要素が第 l 層の k 番目の小領域に属するものであることを示す．

なお，K_{ct} カーネルの実装に関して，各小領域ごとのリーマン計量を導出することは計算コストが大きく困難であるため，第 0 層 $(L = 0)$ の計量を全ての小領域で代用する．

Bag-of-Visual-Words の実装

ベースラインとして，Global Gaussian に用いるものと同じ局所特徴を用い，bag-of-visual-words (BoVW) の実装を行う．ここでは，最も基本的な k-means クラスタリングによりコードブックを生成する．Visual words の数は，200 と 1000 の 2 通りを用いる．識別器に用いるカーネル関数は，BoVW による画像認識のデファクトスタンダードである spatial pyramid matching [110] を適用した histogram intersection kernel を用いる．以下，このカーネルを K^{BoVW} と記述する．

さらに，いくつかの実験においては，提案手法によるカーネル関数 (式 6.23) と BoVW のカーネル関数を両方識別に用い，性能向上を図る．ここでは，以下のシンプルな線形結合により 2 つのカーネルを統合して用いる．

$$K^{GG+BoVW} = \frac{1}{1+\kappa} K^{GG} + \frac{\kappa}{1+\kappa} K^{BoVW}. \tag{6.24}$$

κ は結合の重みを決定するパラメータである．なお，K^{BoVW} の理論的上限値は 1 であるが，K^{GG} は正規化されていないため，重み κ の値は必ずしも直感的に両者の重要度を表さないことに注意されたい．

Table 6.3: Performance comparison with spatial information for LSP15 (%). The SURF descriptor is used.

		L=0	L=1	L=2
GG	KPDA (ad-linear)	75.9	78.8	79.8
	KPDA (ct-linear)	78.5	81.6	82.3
	KPDA (KL div.)	**81.5**	**84.8**	**86.1**
	SVM (ad-linear)	72.1	73.2	74.3
	SVM (ct-linear)	77.7	80.1	80.7
	SVM (KL div.)	78.3	82.2	83.1
BoVW200	KPDA	71.9	78.5	81.1
	SVM	70.6	76.3	78.6
BoVW1000	KPDA	77.1	80.7	82.5
	SVM	74.9	78.0	79.4

Table 6.4: Performance comparison with spatial information for the 8-sports dataset (%). The SIFT descriptor is used.

		L=0	L=1	L=2
GG	KPDA (ad-linear)	77.9	79.3	80.2
	KPDA (ct-linear)	79.7	81.5	82.9
	KPDA (KL div.)	**81.7**	**83.2**	**84.4**
	SVM (ad-linear)	70.6	71.6	71.7
	SVM (ct-linear)	75.5	77.2	78.8
	SVM (KL div.)	78.3	80.2	81.4
BoVW200	KPDA	72.0	76.9	79.6
	SVM	71.7	76.3	77.7
BoVW1000	KPDA	77.8	80.6	81.5
	SVM	76.2	78.1	79.1

Table 6.5: Performance of the global Gaussian, BoVW, and combined approach (%). An $L = 2$ spatial pyramid is implemented. Kernel PDA is used for classification. The SURF descriptor is used for LSP15, while the SIFT descriptor is used for the 8-sports dataset.

	LSP15	8-sports
GG (KL)	**86.1**±0.5	**84.4**±1.4
GG (ct-linear)	82.3±0.4	82.9±1.0
BoVW200	81.1±0.7	79.6±1.1
BoVW1000	82.5±0.7	81.5±1.7
GG (ct-linear) + BoVW200	85.0±0.5	83.2±0.9
GG (ct-linear) + BoVW1000	85.3±0.5	83.4±0.7

6.4.4 実験結果

LSP15, 8-sport における詳細な検証

まず，LSP15 と 8-sports データセットを用い，提案する Global Gaussian アプローチの有効性を詳しく検証する．表 6.2に，位置情報 (SPM) を用いない最も基本的な場合の性能を示す．“Ad-linear” は ad-hoc linear kernel, “ct-linear” は center tangent linear kernel, “KL div.” は KL divergence based kernel をそれぞれ示す．. ここでは，SIFT，SURF の両方をそれぞれ試し比較した．認識性能は，KL div., ct-linear, ad-linear の順となり，理論的な考察と合致する結果となった．Ct-linear は，同じ線形の枠組みでありながら ad-linear から大きく性能を向上させており，適切な計量を用いることの有効性が示された．また，LSP15 においては SURF，8-sports では SIFT による局所特徴記述がそれぞれ優位な結果となった．

次に，SPM を用い位置情報を加え，ベースラインである BoVW と比較を行う．提案手法と BoVW の両方において，第 2 層 ($L = 2$) までの spatial pyramid を用いる．表 6.2の結果を踏まえ，LSP15 では SURF，8-sports では SIFT 記述子をそれぞれ用いる．表 6.3に LSP15, 表 6.4に 8-sports における実験結果をそれぞれ示す．提案手法は，1000 次元の BoVW と同等以上の認識精度が得られることが示された．また，BoVW と同様に，SPM の利用により提案手法においても認識精度が大きく向上することが確認された．

最後に，提案する Global Gaussian と BoVW の統合を行う．KL divergence based kernel は高い性能を得ることができるが，計算コストが大きく実用性に乏しい．また，非線形カーネルに依存した実装となるため，大規模なデータにおいて学習を行うことが著しく困難である．ここでは，線形カーネルである ct-linear kernel と BoVW の histogram intersection kernel を式 6.24に従い結合する．これは，将来的に histogram intersection kernel の近似線形化手法 [127; 193] などを統合することで完全に線形の枠組みとなり，線形識別器による学習が可能となるこ

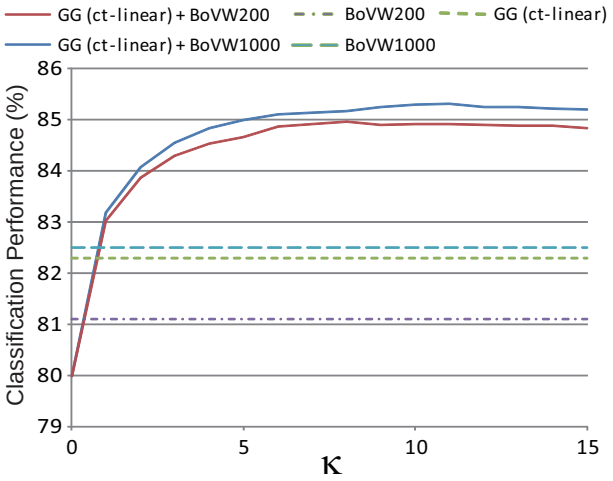


Figure 6.2: Merging the global Gaussian and BoVW approaches for use with the LSP15 dataset. κ is the parameter for weighting the kernels (Eq. 6.24).

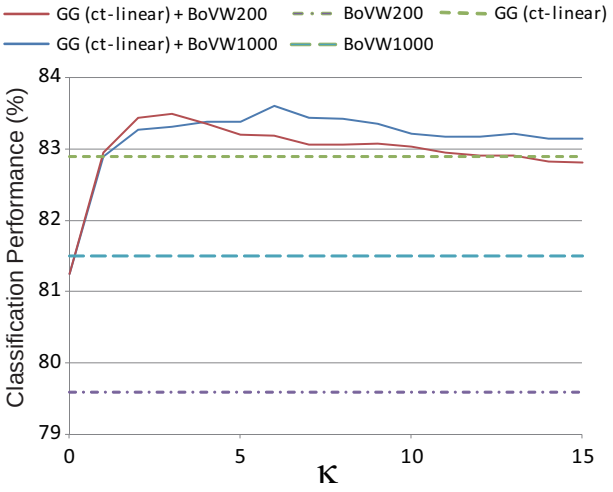


Figure 6.3: Merging the global Gaussian and BoVW approaches for use with the 8-sports dataset. κ is the parameter for weighting the kernels (Eq. 6.24).

Table 6.6: Performance comparison with previous work (%). For our method, an $L = 2$ spatial pyramid is implemented, and kernel PDA is used for classification. We used the SURF descriptor for LSP15 and Indoor67, and the SIFT descriptor for the 8-sports dataset.

Method	LSP15	8-sports	Indoor67
GG (KL-div.)	**86.1**±0.5	**84.4**±1.4	**45.5**±1.1
GG (ct-linear) + BoVW1000	85.3±0.5	83.4±0.7	44.9±1.3
Previous	85.2 [227]	84.2 [211]	39.6 [138]
	85.2 [139]	73.4 [115]	25.0 [160]
	84.1 [211]		
	83.7 [25]		
	83.4 [138]		

とを見越した実装である（7 章で実際にこれを行う）．

表 6.5が示すように，Global Gaussian と BoVW がカバーする異なる統計的特徴を用いることで，認識性能がさらに向上できることが示された．また図 6.2，図 6.3に，重みパラメータ κ の性能に対する影響を示す．κ の変化に対し，認識精度は比較的安定に推移しているといえ，multiple kernel learning による最適化も実行可能であると期待できる．

先行研究との比較

表 6.6に LSP15, 8-sports, Indoor67 における先行研究との性能比較結果をまとめる．近年の一般画像認識の研究では，複数の特徴記述子を用いることで識別性能が大きく向上することが知られている [24; 214; 217] が，これは本章のスコープ外であるため，ここでは，単一の特徴記述子を用いる研究のみ示す．LSP15 では，GMM ベースの手法である hierarchical Gaussianization [227] と，sparse coding をペアワイズなコードブック生成に利用した directional local pairwise bases (DLPB) [139] が現在の最高スコアである 85.2%を記録している．提案手法は，KL divergence based kernel を用いる場合に 86.1%，よりスケーラブルな ct-linear + BoVW のアプローチにおいても 85.3%の識別率を得ている．8-sports では HIK-codebook [211] が 84.2%を達成しており，提案手法では KL div., ct-liner + BoVW がそれぞれ 84.4%，83.4%となった．なお，[211] では局所特徴を元画像と Sobel フィルタによるエッジ画像の両方から，異なる 5 つのスケールで特徴抽出を行っているのに対し，提案手法では元画像から単一スケールで特徴抽出を行っている点に注意されたい[1]．Indoor67 においては，local pairwise codebook (LPC) [138] が 39.6%を達成している．これは，前述の DLPB の前身となった手法であり，ペアワイズに

[1]Sobel フィルタによるエッジ画像を利用しない場合，[211] の識別率は 81.9%である．

近接する局所特徴を結合し BoVW ヒストグラムを生成するものである．提案手法の実装には，LSP15 における良好な性能を鑑み，SURF 特徴を用いた．スコアは，KL div., ct-liner + BoVW がそれぞれ 45.5%，44.9%となり，共に LPC を上回る結果となった．

以上のように，提案手法はシーン認識における 3 種のベンチマークにおいて，いずれも良好な結果を得ている．特に，シンプルかつスケーラブルな ct-liner + BoVW によるアプローチにおいても，既存研究以上の識別正解率を得た点が興味深い．

6.4.5 考察

本章の目的は，線形手法に直接適用可能な特徴のコーディング方法の開発である．ここまでの実験により，GLC と ct-linear カーネルが良好な識別性能を示すことが確認された．これは，式 6.16の $\boldsymbol{\zeta}$-座標系に線形識別器を適用可能であることを示唆している．ここで，$\boldsymbol{\zeta}$-座標系は GLC（$\boldsymbol{\eta}$-座標系）にある正則なアフィン変換を加えたものであることに注意されたい．これは，LDA や CCA など，特徴空間のアフィン変換に対する不変性を有する手法へ適用することを念頭に置いた場合には，式 6.16の変換は必要なく，GLC を直接特徴ベクトルとして利用可能であることを意味している．前章までで開発した CCD は CCA に基づく手法であるため，CCD にとって GLC は理想的な特徴表現であるといえる．さらに，この考え方は GLC がプロットされる $\boldsymbol{\eta}$-座標系をユークリッド空間とみなすことであるから，部分空間の利用などにより効率よくコーディングの計算コストが削減できる可能性がある．次節では，この点について検証を行う．

6.5 GLC と判別的線形学習器によるスケーラブル化

6.5.1 GLC の圧縮

ここでは，式 6.8で定義される GLC のコーディング方法について改めて詳しく述べるとともに，いくつかの応用形を考える．N 枚の学習画像があるとする．各画像 $I^{(j)}(j \leq N)$ からそれぞれ $p^{(j)}$ 個の D 次元局所特徴 $\boldsymbol{v}_k^{(j)}(k \leq p^{(j)})$ を抽出する．これらの平均を $\boldsymbol{\mu}^{(j)} = \frac{1}{p^{(j)}} \sum_k^{p^{(j)}} \boldsymbol{v}_k^{(j)}$ とする．これは，局所特徴の 0 次の自己相関と解釈することもできる．また，$R^{(j)} = \frac{1}{p^{(j)}} \sum_k^{p^{(j)}} \boldsymbol{v}_k^{(j)} \boldsymbol{v}_k^{(j)T}$ を $I^{(j)}$ の自己相関行列とする．式 6.8の GLC は，0 次の自己相関と 1 次の自己相関を列挙したものである．すなわち，

$$\boldsymbol{\eta}_{0th+1st}^{(j)} = \begin{pmatrix} \boldsymbol{\mu}^{(j)} \\ upper(R^{(j)}) \end{pmatrix}, \tag{6.25}$$

が最も基本的な GLC のコーディングである．ここで，$upper()$ は対称行列の上三角部分の要素を列挙したベクトルである．例えば，$upper(R^{(j)})$ は $D(D+1)/2$ 次元のベクトルとなる．

次に，部分空間を利用した簡便な GLC のコーディング方法を考える．最もシンプルなものは，特徴要素の一部のみを用いるものである．例えば，0 次の自己相関のみを用いる場合は

$$\boldsymbol{\eta}_{0th}^{(j)} = \boldsymbol{\mu}^{(j)}. \tag{6.26}$$

となり，これは平均ベクトルに他ならない．エッジヒストグラム，カラーヒストグラムなど多くの基本的な画像特徴はこれに含まれる．また，1 次の自己相関のみを用いる場合は，

$$\boldsymbol{\eta}_{1st}^{(j)} = upper(R^{(j)}). \tag{6.27}$$

GLC の問題は，局所特徴の次元数 D が大きい時，1 次相関の数が大きくなりその後の学習にかかる計算コストが増大することである．このため，あらかじめ局所特徴の次元数を PCA により圧縮することを考える．R を全ての学習画像の局所特徴の自己相関行列とする．すなわち，

$$R = \frac{1}{\sum_j^N p^{(j)}} \sum_j^N p^{(j)} R^{(j)}. \tag{6.28}$$

以下の固有値問題により，局所特徴を圧縮する射影行列 U が得られる．

$$RU = U\Omega \quad (U^T U = I). \tag{6.29}$$

ここで，Ω は固有値を要素に持つ対角行列である．主成分を適切な次元 m で打ち切り，対応する上位 m 本の固有ベクトルからなる射影行列を m とする．この射影行列により局所特徴は m 次元へ圧縮されるため，1 次相関の数は $m(m+1)/2$ へおさえられる．最終的に，主成分の 1 次相関に基づく GLC は以下のようになる．

$$\tilde{\boldsymbol{\eta}}_{1st}^{(j)} = upper({U_m}^T R^{(j)} U_m). \tag{6.30}$$

また，平均（0 次相関）ベクトルを加える場合は，

$$\tilde{\boldsymbol{\eta}}_{0th+1st}^{(j)} = \begin{pmatrix} \boldsymbol{\mu}^{(j)} \\ upper({U_m}^T R^{(j)} U_m) \end{pmatrix}, \tag{6.31}$$

となる．これらは，式 6.25にある線形変換を加え，部分空間へ射影を行っているに他ならない．

6.5.2 データセット

本節の実験では，前節で用いた LSP15 に加え，OT8 [148] と呼ばれる 8 クラスのシーン画像データセット，Caltech-101 [56] と呼ばれる 101(+1) クラスの物体画像データセットを用いる．OT8 は全部で 2,688 枚のカラー画像からなるデータセットであり，図 6.4に示す 8 つのカテゴリにより構成される[1]．

[1] なお，LSP15 は文献 [58; 110] の著者らが OT8 に 7 つのクラスを追加したものである．

Figure 6.4: Sample images from the OT8 dataset.

Caltech-101 は，一般物体認識におけるデファクトスタンダードのベンチマークとして広く用いられているデータセットである．101 種類の様々な物体と背景画像の計 102 クラスの識別がタスクである．詳しくは，2.3.1 節，図 2.5を参照されたい．

前節の実験同様，評価は平均クラス識別率によって行われる．OT8，LSP15 については各クラス 100 画像，Caltech-101 については各クラス 30 画像ずつランダムに選び学習サンプルとし，残りをテストサンプルとする．学習サンプルとテストサンプルをランダムに入れ替えながら識別率を 10 回測定し，その平均値をスコアとする．

6.5.3 セットアップ

局所特徴抽出

本節の実験では，以下の 4 種類の局所特徴記述子を用いる．

1. SIFT [124]
2. RGB-SIFT [25]
3. Local edge histogram
4. Local HSV color histogram

RGB-SIFT はカラー画像で用いられる記述子であり，R, G, B の各チャンネルの白黒画像からそれぞれ抽出される SIFT 特徴を結合するものである．したがって，RGB-SIFT 特徴は $128 \times 3 = 384$ 次元となる．Local edge histogram は，局所特徴抽出窓内で記述される 72 次元の単純な勾配方向ヒストグラムである．同

様に，Local HSV color histogram は 84 次元[1]の色ヒストグラムである．これらの基本的な局所ヒストグラムにおいても, GLC は汎用的に利用可能である．

各局所特徴は，前節と同様に dense sampling により抽出する．ここでは，$P \times P$ ピクセルのセルを M ピクセルずつスライドさせながら，各セルから局所特徴を抽出する．パラメータ P と M の認識精度に関する影響は実験で調査する．

識別手法

本節では，基本的に線形の PDA 識別器（6.4.2 節）を用いる．PDA の核である LDA はアフィン不変性を有するため，GLC を直接特徴ベクトルとして利用可能であり，本節の検証に適した識別器である．ただし，位置情報の利用は spatial pyramid matching ではなく，より簡便な方法を用いる．また，いくつかの実験においては SVM (LIBSVM) を比較に用いる．

PDA 識別器を用いる場合，式 6.23の spatial pyramid matching も原理的にはシンプルな形で実現できる．すなわち，各小領域の GLC を直列に結合し一つの特徴ベクトルとして入力に用いれば，領域の重みも含め学習が行われる．しかしながらこの方法では，特徴ベクトルの次元が大きくなる点が問題である LDA の固有値問題の計算コストは特徴次元数の 3 乗に比例するため，この方法では学習のコストが非常に大きくなる．

そこで，ここでは領域ごとに独立に識別器を構築し，重み付き対数尤度により識別を行う近似的なアプローチをとる．これを，SP-PDA と呼ぶことにする．まず，6.4.3 節と同様に，画像を階層的にグリッド分割する[2]．画像の第 l 層 $(0 \leq l \leq L)$ を $(l+1) \times (l+1)$ の小領域に分割し，それぞれの小領域で独立に特徴抽出と PDA 識別器の構築を行う．新規サンプルの重み付き対数尤度は次のようになる．

$$\mathcal{L} = \sum_{l=0}^{L} \alpha^l \sum_{i=1}^{(l+1)^2} \log p(\boldsymbol{u}_s^{(l,i)} | \boldsymbol{u}_{1\ldots t}^{(l,i)C}). \tag{6.32}$$

ここで，添え字 (l,i) はその要素が第 l 層の i 番目の領域のものであることを示す．$p(\boldsymbol{u}_s^{(l,i)} | \boldsymbol{u}_{1\ldots t}^{(l,i)C})$ は，領域 (l,i) の PLDA 識別器による出力（尤度）である．α^l は第 l 層の重みを決定するパラメータであり，実験的に決定される．新規サンプルは，$\mathcal{L}$ を最大とするクラスへ識別される．これは結局，以下の重み付き距離を最小とするクラス $\hat{C}$ へ識別することに等価である．

$$\hat{C} = \underset{C}{\operatorname{argmin}} \sum_{l=0}^{L} \alpha^l \sum_{i=1}^{(l+1)^2} (\tilde{\boldsymbol{u}}_s^{(l,i)C})^T (\Theta^{(l,i)})^{-1} (\tilde{\boldsymbol{u}}_s^{(l,i)C}), \tag{6.33}$$

[1] H:36 次元, S:32 次元, V:16 次元.
[2] 分割方法が前節と異なることに注意されたい．

Table 6.7: Baseline performance for OT8 (%) using GLC in different types. Classification is conducted via PDA and SVM. Regarding the results for the SVM, the plain number indicates the classification score using a linear kernel, while the italic number in parenthesis indicates that using the RBF kernel. The best score for each descriptor is shown in bold.

	0th (Mean)		1st (Cor.)		0th+1st	
	PDA	SVM	PDA	SVM	PDA	SVM
Edge Hist	66.5	70.3 (*71.0*)	**74.5**	73.6 (*72.7*)	**74.5**	73.6 (*72.8*)
Color Hist	45.2	47.4 (*50.8*)	54.1	55.3 (*55.9*)	54.2	55.3 (***56.3***)
Gray-SIFT	73.1	72.5 (*73.5*)	84.8	80.9 (*81.1*)	**85.0**	80.9 (*81.0*)
RGB-SIFT	77.7	75.2 (*76.2*)	86.4	81.4 (*81.6*)	**86.8**	81.7 (*81.9*)

ただし，

$$\tilde{\boldsymbol{u}}_s^{(l,i)C} = \boldsymbol{u}_s^{(l,i)} - \frac{t\Psi^{(l,i)}}{t\Psi^{(l,i)} + I}\bar{\boldsymbol{u}}^{(l,i)C}, \tag{6.34}$$

$$\Theta^{(l,i)} = I + \frac{\Psi^{(l,i)}}{t\Psi^{(l,i)} + I}. \tag{6.35}$$

この方法は各小領域の独立性を仮定しており，領域特徴の共起性を考慮していない．したがって，厳密な spatial pyramid matching に比べると近似的なアプローチであるといえる．しかしながら，識別器の学習は各小領域において独立に行われるため，重み α^l がオンラインにチューニング可能である点が大きなメリットである．

6.5.4 実験結果

まず，さまざまな局所特徴記述子において GLC を抽出し，その効果を検証する．また，いくつかのパラメータと認識精度の関係を精査する．その後，先行研究と定量的な性能比較を行う．全ての実験は，8 コアのワークステーション (dual Xeon 3.20GHz) 上で行う．

ベースライン

ここでは，4 種類の局所特徴からそれぞれ GLC を抽出し，効果を調べる．Edge/color histogram については，dense sampling のパラメータを $P = 10$, $M = 5$ とした (6.5.3 節)．式 6.25に従い，基本的な GLC の抽出を行う．SIFT, RGB-SIFT については，パラメータを $P = 16$, $M = 5$ とした．これらの記述子の次元は比較的

大きいため，式 6.31のように，PCA による次元圧縮を予め行い 1 次の GLC を抽出する．主成分の次元数は $m = 30$ とする．また，PDA と SVM の識別器の性能比較を行う．識別器のパラメータは，実験的に最良のパラメータを与えることとする．

表 6.7に，各局所特徴記述子を用いた際の認識精度をまとめる．まず，異なる統計的量に基づく GLC を比較し，それぞれの有効性を調べる．“0th” は局所特徴の平均（0 次相関）のみを用いる場合（式 6.25）であり，一般的な大域的画像特徴量とほぼ同じ形となる．“1st” は局所特徴の 1 次相関のみを用いる場合であり，edge/color histogram については式 6.27，SIFT, RGB-SIFT については式 6.30により抽出する．“0th+1st” は両方を用いる場合であり，edge/color histogram については式 6.25，SIFT, RGB-SIFT については式 6.31により抽出する．SVM のスコアは，線形カーネルを用いた場合を通常表記し，RBF カーネルを用いた場合を斜体で括弧中に表記した．

これらの結果が示すように，どの局所特徴記述子においても，1 次の GLC を利用することで大きく認識精度が向上する．さらに，edge histogram を用いた場合を除き，0 次と 1 次の GLC を併用することでわずかに認識精度が向上する．しかしながら全体としては，1 次と 0 次+1 次の場合に大きな性能差は見られなかった．理論的には，0 次（平均）と 1 次の相関は異なる統計量であるため，両方用いることで識別のために有効な情報が増えると考えられる．しかしながら，自己相関（$R^{(j)}$ の対角要素）と平均は意味的に冗長性があるため，実際に平均を加えることが有効であるかは，タスクや記述子の性質に依存すると考えられる．

また，PDA は color histogram の場合を除いて SVM を上回る性能を示した．特に，SIFT/RGB-SIFT の 1 次相関を用いる場合に良好な認識精度を示しており，SVM との性能差も顕著である．これは，PDA の持つアフィン不変性が，PCA に起因するスケール変化を吸収しているためであると考えられる．

Bag-of-Visual-Words との比較

ここでは，SIFT 記述子を用い，同じ局所特徴量を利用した際の GLC と BoVW の性能を比較する．Dense sampling のパラメータを $P = 16$, $M = 5$ とする．GLC は式 6.31を用い，用いる局所特徴の主成分数は $m = 30$ とする．BoVW の実装は，標準的な k-means 法により行う．Visual words の数は，200, 500, 1000, 1500 の 4 通りについて調べる．識別器は，PDA と SVM の両方を試し，特徴との相性を調べる．SVM の実装には，線形カーネル，RBF カーネルに加え，histogram intersection カーネル (HIK) と χ^2 カーネル [224] を用いる．後者の 2 つは，ヒストグラム特徴について設計されたものであり，GLC には直接適用できない点に注意されたい．

表 6.8に結果をまとめる．GLC と PDA の組み合わせが他を大きく上回る認識精度を示しており，その相性の良さが再び示される結果となった．また，BoVW においては，線形手法である PDA や，線形カーネルによる SVM を適用した場合には十分な性能が出ないが，HIK や χ^2 カーネルなどの非線形カーネルを用いる

Table 6.8: Classification performance of GLC and bag-of-visual-words (BoVW) for OT8 (%). We implement BoVW with 200, 500, 1000, and 1500 visual words.

	GLC (0th+1st)	BoVW 200	BoVW 500	BoVW 1000	BoVW 1500
PDA	**85.0**	78.9	79.9	80.7	80.8
SVM (linear)	80.9	77.2	78.1	78.6	78.6
SVM (RBF)	81.0	77.5	78.3	78.8	78.7
SVM (HIK)	N/A	80.0	82.0	82.7	83.0
SVM (χ^2)	N/A	80.8	82.5	83.2	83.7

と大きく性能が向上することが分かる．これは，前章で得られた知見と合致する結果である．BoVW においても，多数の visual word と非線形カーネルを用いることにより良好な性能を得ることができるが，visual words の数が増えるにつれて特徴抽出の計算コストは増大する（6.5.4 節）．また，これまで述べてきたように，カーネル化を行うと，学習の計算コストは $O(N^2)$ から $O(N^3)$ となり，スケーラビリティが著しく損なわれる．本章での目的は線形学習器と相性のよい特徴の考察であり，GLC+PDA の組み合わせは理想的に条件を満たしているといえる．

パラメータに関する検証

次に，GLC のさまざまなパラメータについて精査する．特徴記述子は SIFT を用いる．識別は全て PDA によって行う．

まず，dense sampling のサンプリングステップ M と認識精度の関係を図 6.5に示す．ここでは，局所特徴記述子の窓幅を $P = 16$, 局所特徴の圧縮次元数を $m = 30$ に固定する．グラフの横軸は，PDA 識別器の正則化項の大きさ γ を示す（対数スケール）．図 6.5から，より密に局所特徴を抽出するほど認識精度が向上することが分かる．究極的には，全てのピクセルから局所特徴を抽出する場合 $(M = 1)$ に最もよい精度が得られることが分かる．この結果は，Nowak ら [147] の示した知見と合致する．しかしながら，サンプリングを密にするほど画像あたりの特徴抽出のコストは増大するため，認識精度と計算コストのトレードオフとなる．

次に，次元圧縮のパラメータ m の影響について調べる．ここでは，$P = 16$, $M = 5$ に固定する．PCA による次元圧縮はグラウンドツルースのカテゴリとは無関係に行われるため，m が小さすぎる場合は認識精度が低下することが容易に予想される．図 6.6に結果を示す．m が大きくなるほど認識精度が向上しており，直感に合致する結果となった．しかしながらここにもトレードオフが存在し，m が大きくなるほど特徴ベクトルの次元数が増え，PDA の計算コストも増大する．特徴抽出の時間を除き，GLC+PDA の学習は，$m = 10$ の場合 0.1 秒，$m = 30$ の

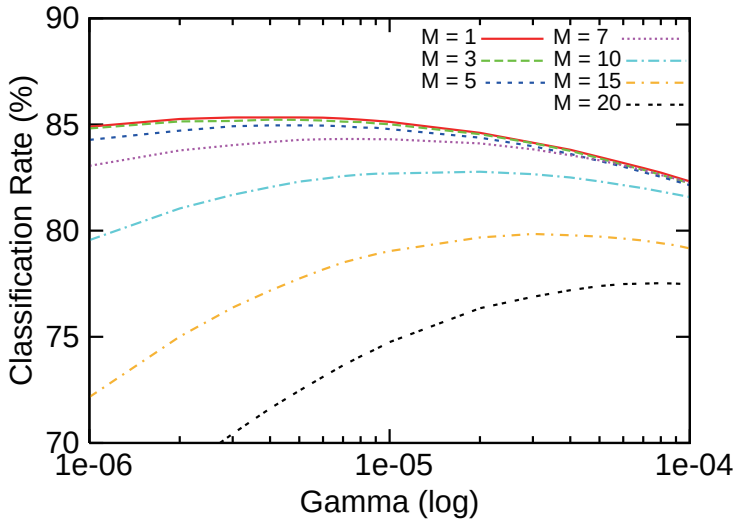


Figure 6.5: Effect of sampling density on performance ($P = 16$, $m = 30$).

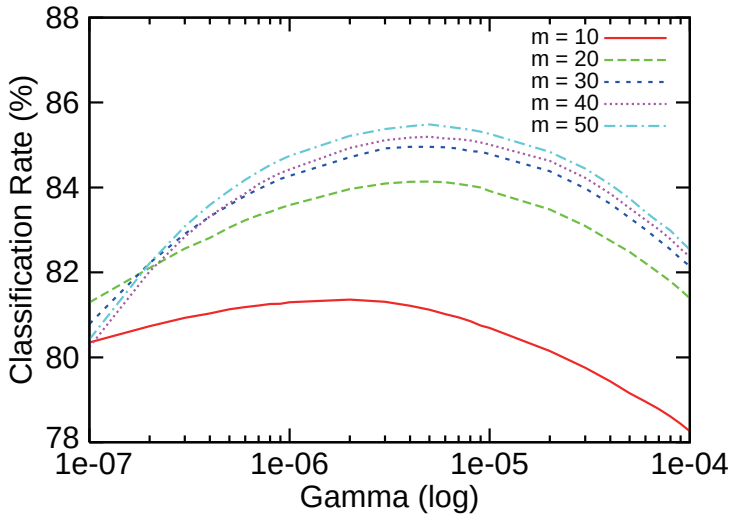


Figure 6.6: Effect of the dimensionality of PCA compression ($P = 16$, $M = 5$).

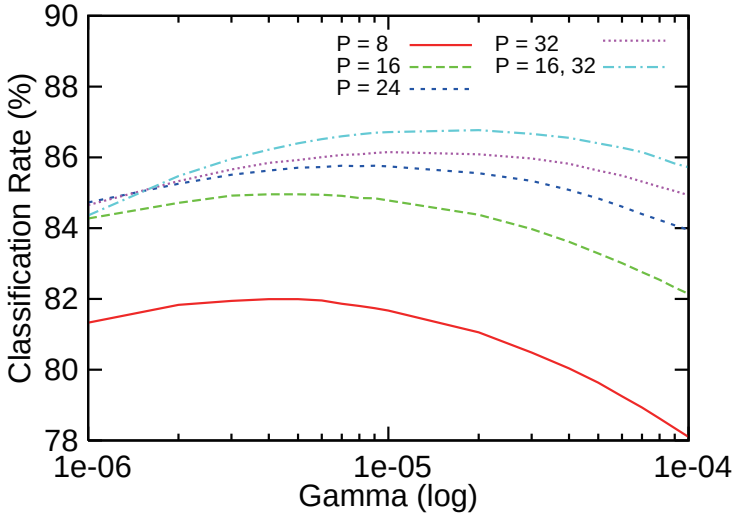


Figure 6.7: Effect of the scale parameter of the SIFT-descriptor ($m = 30$, $M = 5$).

場合 1 秒，$m = 50$ の場合に 10 秒程度を要する[1]．

さらに，局所特徴抽出窓のスケール P（SIFT 記述子の直径）の影響を調べる．図 6.7に，4 つの異なるスケールを用いた場合の認識精度を示す．ここでは，$m = 30$, $M = 5$ に固定した．なお，Bosch ら [25] は，複数のスケールで SIFT 特徴を抽出することで認識精度が向上することを示している．これを参考に，マルチスケールの GLC($P = 16, 32$) についても調べる．まず，それぞれのスケールで抽出される SIFT 特徴から GLC を通常どおりそれぞれ生成し，直列に結合し最終的な画像特徴ベクトルとして用いる．実験結果より，特徴記述子のスケールは認識精度に大きく影響する重要なパラメータであることが分かる．また，マルチスケールの GLC を用いることでさらに認識精度が向上することも示された．

空間情報の寄与

画像の位置情報を加える SP-PDA の効果について調べる．図 6.8に，第 1 層までの spatial pyramid を考慮する場合 (L1)，図 6.9に第 2 層までの spatial pyramid を考慮する場合 (L2) を示す．それぞれ，第 1 層に対する各層の相対的な重みに対する認識精度の変化をプロットした．どちらの場合も，オリジナルの PDA に比べ認識精度は向上している．認識精度の最高値は，L1 の場合 87.2%, L2 の場合 88.0% となった．また，グラフの挙動より，認識精度は各層の重みの変化に対して比較的安定に高い値を保っており，単純に等しい重みを与えた場合[2]($\alpha^1/\alpha^0 = \alpha^2/\alpha^0 = 1$)

[1] テストサンプルの識別に関しては，全てのサンプル (OT8 では 1,888 画像) の識別を 0.05 秒程度で行うことができる．

[2] これは，各領域で生成される PDA 識別器を naive Bayes により結合した場合と等価である．

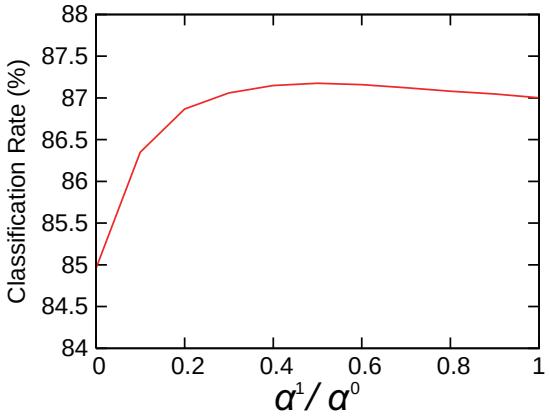


Figure 6.8: Effect of the weight parameter using at most the 2nd layer ($P = 16$, $m = 30$, $M = 5$, $\gamma = 5.0e - 06$).

でも良好な性能を示していることが分かる．この場合の認識精度は，L1 の場合 87.0%, L2 の場合 87.8%となり，最適な重みを与えた場合と比べても 0.2%程度の違いである．この安定性は，実用上好ましい特長であるといえる．

次元圧縮に関する議論

局所特徴の次元圧縮を行い GLC を抽出する場合には，学習サンプルから PCA の射影行列を求める必要がある．このコストは BoVW で必要なクラスタリングと比較して軽微ではあるが，さらに大規模なデータにおいては困難となる可能性がある．また，PCA の結果はデータセット依存となるため，他のデータへの汎化性は明らかではない．そこで，ここではより簡便な GLC の次元圧縮方法を 2 種類検証する．一つ目は，単純に 1 次相関の要素 (式 6.25の $upper(R^{(j)})$) をランダムにサンプリングするものである．これは，オリジナルの GLC の random subspace を利用しているに他ならない．これを R-GLC と呼ぶことにする．二つ目は，タスクとは無関係の大規模な画像レポジトリ[1]を用いて PCA を行うことにより，予め汎用的な射影行列を求めておく方法である．これは，PCA-SIFT [98] が局所特徴の圧縮に用いている方法と同じ考え方に基づく．

実験では，以下の 3 通りを比較する．

1. OT8 の学習データセットから PCA を解く場合（標準的な GLC の実装）．
2. Caltech-101 からランダムに 3,000 枚の画像を選択し PCA を解く場合．
3. R-GLC (random subspace を用いた次元圧縮)．

[1]例えば，web 上の画像など．

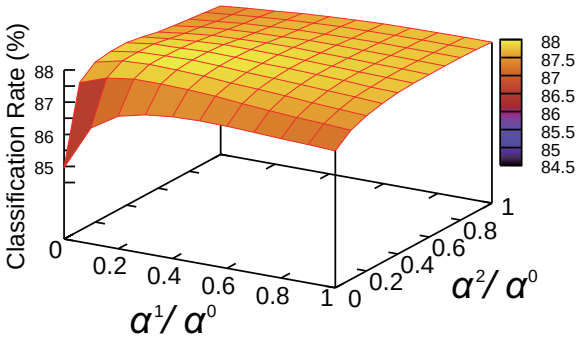


Figure 6.9: Effect of the weight parameter using at most the 3rd layer ($P = 16$, $m = 30$, $M = 5$, $\gamma = 5.0e-06$).

2 では，異なるデータセットである Caltech-101 で計算された PCA の射影が OT8 において有効であるか調べることで，pre-computed な射影の利用可能性を検証する．また，3 の R-GLC の実装では，特徴ベクトルの次元数を他の場合と揃えるため，$m(m+1)/2$ 個の 1 次相関をランダムにサンプリングするものとする．認識精度は 100 試行の平均によって評価する．

図 6.10に $m = 30$, $M = 5$ の場合の実験結果を示す．まず，R-GLC の認識精度は PCA を用いる他の 2 つよりも約 0.7%低い結果となった．R-GLC は特徴空間の分散を一切考慮しないため，当然の結果であるといえる．しかしながら，R-GLC が完全にランダムなアプローチであり全く事前の処理を必要としないことを踏まえると，0.7%程度の損失は軽微であるとも考えられる．この結果は，SIFT 記述子の性質に起因している可能性がある．SIFT 記述子は局所的なエッジヒストグラムによって構成されているため，特徴要素の 1 次相関は明示的にある形状パターンに対応しており，冗長性の少ない表現になっていると考えられる．また，Caltech-101 から得られる射影による GLC と，通常の GLC の性能差は約 0.05%とほぼ変わらない結果となった．これは，タスクに非依存な汎用的な射影を共有できる可能性を示唆するものであり，今後さらに検証すべきである．

以上まとめると，タスクに非依存な次元圧縮方法としては，あらかじめ外部の大規模画像データを用いて汎用的な射影を学習しておくことが有効であると考えられる．そのようなデータがない場合は，多少認識精度は劣るが R-GLC を次善の策として用いることができる．

先行研究との比較

OT8, LSP15, Caltech-101 の 3 つのデータセットを用い，提案する GLC+PDA (SP-PDA) の性能を先行研究と定量的に比較する．LSP15 では SIFT, OT8 と Caltech-101 では RGB-SIFT を局所特徴記述子として用いる．GLC は $P = 16, 32$ の 2 スケールから抽出し結合する．PCA の圧縮次元数は $m = 50$ とする．実験結果は，

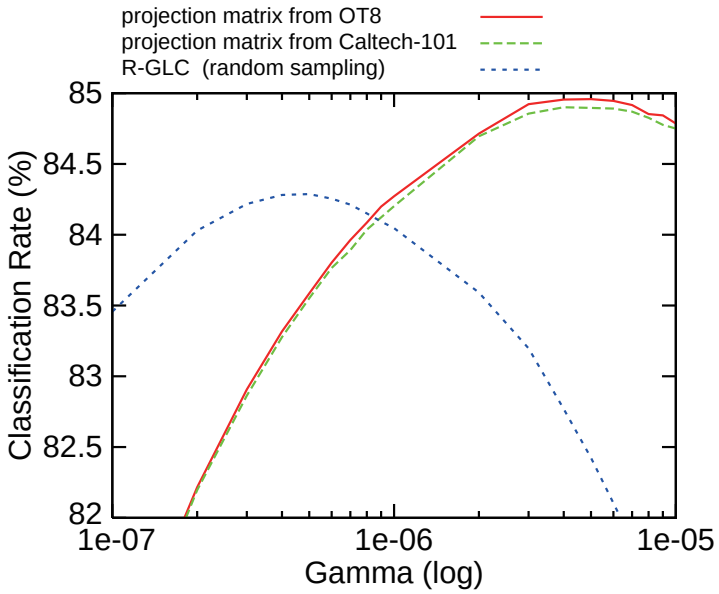


Figure 6.10: Results using different dimensionality compression methods ($P = 16$, $m = 30$, $M = 5$). We used two different projection matrices (one from OT8 and the other from Caltech-101), and random sampling.

画像の位置情報 (spatial information) を用いる場合 (with SI) と用いない場合 (no SI) に分けてまとめる．

表 6.9に認識精度の比較結果を示す．まず，OT8 と LSP15 の 2 つのシーン認識データセットにおける結果について考察する．[25; 110] では，SIFT 記述子を用いて BoVW ヒストグラムを抽出し，SVM による識別を行う．[204] は conditional random field (CRF) [105] による part-based な生成モデルを構築し，識別とセグメンテーションを同時に行う．しかしながら，その計算コストは BoVW+SVM と比較しても非常に高いものである．

提案手法は，位置情報を用いない場合 (L0) に，相対的に高い認識精度を示している．これは，画像の全体的な特徴記述として GLC が優れた性能を有することを端的に表しているといえる．位置情報を加える場合，OT8 では Perina ら [152] の手法が 92.8%, LSP15 では前節で開発した KL divergence に基づくカーネルによる Global Gaussian が 86.1%とそれぞれ現在の最高スコアを記録している．GLC+PDA は単純な線形モデルによる識別器であり，学習・認識とも極めて高速に行うことが可能であるが，これらとの認識精度差は 2%程度に留まっている．

次に，Caltech-101 における結果について考察する．[110] の “no SI” の場合が，BoVW+SVM による現在最もスタンダードなベースラインであり，41.2%となっている．また，[78] ではいくつかの基本的な大域的画像特徴量を結合し SVM により識別を行い，39.6%を記録している．提案手法 (L0) は，これらをいずれも大

きく上回る認識精度を示している．

位置情報を加える場合，最新の手法の多くは BoVW や GMM など高次統計量を利用するアプローチの改良に基づいており，非常に高い認識精度を示している [139; 201; 218; 227]．例えば [139] との比較から分かるように，LSP15 と比べて Caltech-101 における提案手法との性能差は顕著である．これは，主に二つの理由に起因すると考えられる．第一に，画像の位置情報の利用の精度である．Caltech-101 では，対称物体は概ね同じ向き・大きさに揃えられ，画像の中心に配置されている．このため，他のデータセットに比べ位置情報は特に有効な手掛かりであるといえる．提案手法の SP-PDA は位置情報の利用に関して近似的なアプローチであるため，他手法と差が開いていると考えられる．第二に，物体認識というタスク自体の性質である．抽象的なシーンと異なり，リジッドな物体の画像には物体固有の局所パターンが表れやすく，これを捉えることが認識精度を向上させる鍵となる．これを捉えるためには，局所特徴空間のローカルな構造を用いる BoVW や GMM の方が，Global Gaussian に基づく GLC よりも有効である可能性が高い．しかしながら，前節で述べたとおり，両者は対立する概念ではなく，同時に用いることでさらに性能向上が行える可能性があることに注意されたい．この点については，より実践的な大規模データを用いさらに検証を行うべきである．

計算コスト

最後に，GLC の計算コストを，最終的な特徴抽出（相関演算）と前処理 (PCA) のそれぞれについて順に述べる．p を画像一枚あたりの平均的な局所特徴の数，D を局所特徴の次元数，m を PCA の打ち切り次元数とする．また，V を BoVW における visual words の数とする．

まず，特徴抽出にかかるコストを述べる．基本的な GLC (式 6.25) の抽出コストは $O(pD^2)$ であり，PCA による圧縮を行った場合 (式 6.31) は $O(pm(D+m))$ となる．一般に $m < D$ であるため，計算コストは削減される．一方，BoVW によるヒストグラム抽出にかかるコストは $O(pVD)$ となる．多くの場合，$V \gg D$ である．例えば，標準的な SIFT 特徴の場合 $D = 128$ であるが，visual words の数 V は数千程度にとられる．BoVW による特等抽出コストは，kd-tree [12] や locality sensitive hashing [42] などの近似最近傍法の利用により削減可能であるが，認識精度とのトレードオフとなる．また，追加の学習コストが必要となる．

次に，前処理について必要なコストを考察する．基本的な GLC (式 6.25) では，前処理は必要ない．また，random subspace による特徴選択を行うか，pre-computed な射影を利用すれば，前処理のコストを一切かけずに次元圧縮を行うことができる．同様に，BoVW においても，ランダムに抽出した visual words でも比較的良好な性能が得られることが報告されている [147]．しかしながら，より表現能力の高い特徴を得るためには，それぞれ学習フェーズが必要となる．BoVW では，k-means アルゴリズムにより全ての学習サンプルの局所特徴のクラスタ

Table 6.9: Comparison of the performance using two scene datasets and Caltech-101 (%).

Dataset	GLC + PDA			Previous	
	L0	L1	L2	no SI	with SI
OT8	88.8	90.5	91.1		**92.8** [152]
				82.3 [204]	90.2 [204]
				82.5 [25]	87.8 [25]
LSP15	80.0	83.2	84.1	81.5 [144]	**86.1** [144]
					85.2 [227]
					85.2 [139]
					84.1 [211]
				72.7 [25]	83.7 [25]
				74.8 [110]	81.4 [110]
Caltech-101	55.0	63.3	64.8		**77.3** [139]
					73.4 [201]
					73.2 [218]
					73.1 [227]
					67.7 [25]
					66.2 [223]
				41.2 [110]	64.6 [110]
				58.2 [69]	
				39.6 [78]	

リングを行う[1]．この計算コストは，k-means の繰り返しステップ数を I として，$O(pNVDI)$ である．一般に，タスクの規模が大きくなるにつれて，N, V はともに増大する．また，収束に必要な繰り返しステップ数 I も増える．さらに，繰り返し演算を必要とするアルゴリズムであるから，高速に演算を行うためにはメモリ上に全ての局所特徴を保持しておく必要があり，メモリ使用量は $O(pND)$ となる．一方，GLC では PCA を一度解くだけであり，計算コストは $O(D^3 + pND^2)$ となる．また，メモリ上には共分散行列のみ保持しておけばよいため，メモリ使用量は $O(D^2)$ である．

最後に，OT8 における実際の計算コスト (1CPU) について報告する[2]．ここでは，SIFT 記述子を用い，局所特徴のサンプリングレートを $M = 10$ とした．また，BoVW の実装では，visual words の数を $V = 1500$ とした[3]．以上パラメータをまとめると，$N = 800, D = 128, p = 600, V = 1500$ である．PCA の計算は 90 秒ほどで終了したが，k-means による visual words の生成は 18 時間を要した．また，最終的な特徴ベクトルのコーディングでは，GLC は画像 1 枚あたり 60 ミリ秒程度で抽出可能であったが，BoVW は 3.2 秒ほど要した．

[1] 実際には，現実的に計算可能な数の局所特徴をランダムに選択しクラスタリングする場合が多い．

[2] 局所特徴抽出の時間は含めない．

[3] これは，[25] において最も優れた認識精度を示したパラメータである．

Chapter 7

大規模画像認識の定量的評価実験

本章では，4 章で開発した画像アノテーション手法，6 章で開発した画像特徴量を組み合わせ，スケーラブルかつ高精度な一般画像認識システムを実装する．1,200 万枚の大規模な画像データセットを用いた実験により，提案するシステムの有効性を示す．

7.1 データセット構築 (Flickr12M)

本章では，Flickr と呼ばれる写真共有サイト[1]にアップロードされている画像を用い，大規模なデータセットを構築する．

Flickr は最大の写真共有サイトであり，世界中のユーザがデジタルカメラなどで撮影した画像をアップロードしている．各画像はネットユーザに公開されており，自由にコメント・タグ付けなどを行うことができる．毎分数千枚の画像が世界中からアップロードされており，2010 年には既に 40 億枚以上の画像が蓄積されている [203].

Flickr 上の画像とタグの例を図 7.1に示す．本研究では，各画像に与えられたタグをラベル情報のグラウンドツルースとして用いる．基本的に，タグは何らかの形で画像内容に関連するものであると期待できる．しかしながら，実際には意味を為さない単語や，画像情報からはほぼ推定不可能なメタ情報も含まれる．このため，Corel5K などの統制されたデータセットと比較して非常にノイズの多い雑多なデータであるといえる．

7.1.1 画像サンプルの収集

Flickr では画像の検索にキーワードが必要であるため，“ All time most popular tags ”[2]にリストされている単語（表 7.1）をキーワードとして検索を行い，得ら

[1] http://www.flickr.com/
[2] http://www.flickr.com/photos/tags/

Rooster Days
Parade
Broken Arrow
Oklahoma
Old Car
Police
Police Car
Highway Patrol
Vintage
Retro

south africa
cape town
camps bay
cameraphone
mobile
k750i
beach
tidal pool
mountain
reflection
sky
landscape
panorama
autostitch
stitched
25
twenty five
mycapetown
BestofTableMountain

California
San Diego County
El Cajon
tree
kapok
Ceiba pentandra
leaf
leaves
flower
flowers
pink
bud
buds
flowering trees
Malvaceae
Ceiba speciosa
Ceiba
rosa
flor
bello
el arbol
arbol
Silk floss tree

Figure 7.1: Examples of Flickr data: images and corresponding social tags.

Table 7.1: The most popular 145 tags on Flickr. These tags were used for the initial download.

animals architecture art august australia autumn baby band barcelona beach
berlin bird birthday black blackandwhite blue boston bw california cameraphone
camping canada canon car cat chicago china christmas church city
clouds color concert cute dance day de dog england europe
fall family festival film florida flower flowers food football france
friends fun garden geotagged germany girl girls graffiti green halloween
hawaii hiking holiday home house india ireland island italia italy
japan july june kids la lake landscape light live london
macro may me mexico mountain mountains museum music nature new
newyork newyorkcity night nikon nyc ocean paris park party people
photo photography photos portrait red river rock rome san sanfrancisco
scotland sea seattle show sky snow spain spring street summer
sun sunset taiwan texas thailand tokyo toronto tour travel tree
trees trip uk urban usa vacation vancouver washington water wedding
white winter yellow york zoo

れた画像を全て用いる[1]．これらの検索単語そのものは，直接教師として用いないことに注意されたい．

本研究では合計 18,176,861 枚の画像をダウンロードした．この画像セットには，1,486,869 種類のタグが含まれている．このうち，出現回数が 2,000 回未満のタグを削除した後，1 つもタグを持たない画像サンプルを除外する．最終的に，12,283,296 画像，4,130 単語からなるデータセットを得た (以下，Flickr12M と記述する)．画像サイズは，おおよそ 512×384 の大きさにそろえてある．

テストデータも同様の方法により，Flickr からダウンロードした画像を用い構築する．ここで，Flickr 上には，同一のユーザが提供するほぼ同じ見た目とタグを持つ画像が多数存在する点に注意する必要がある（図 7.2）．例えば，記念撮影を行う場合，同じ地点において同じ構図で数枚の写真を撮影することは自然である．これらの画像が，学習データとテストデータに同時に存在することは避けるべきである．ここでは，学習データとテストデータに用いる画像のタイムスタンプをずらすことにより対処する．Flickr12M の持つ 4,130 単語のうち少なくとも一つをグラウンドツルースとして持つ画像をランダムに 10,000 枚選びテストデータとする．

[1] この方法はデータにバイアスを与える要因となるため，将来的にはランダムなクローリングに置き換えられるべきである．

Figure 7.2: Examples of near-duplicate images in the Flickr dataset. Each row corresponds to a duplicate set. These images are annotated with the same social tags.

Table 7.2: Statistics of the Flickr12M dataset.

dictionary size	4130
# of images	12,283,296
# of words per image (avg/max)	3.47/75
# of images per word (avg/max)	10325/491595

Table 7.3: Word frequencies in Flickr12M.

Frequency			# of words
200,001	-		16
100,001	-	200,000	53
50,001	-	100,000	75
30,001	-	50,000	80
20,001	-	30,000	134
10,001	-	20,000	414
5,001	-	10,000	844
2,000	-	5,000	2514

Table 7.4: Most frequently used words in Flickr12M.

	Frequency
wedding	491595
vacation	355111
travel	350101
party	274706
japan	273445
family	263835
beach	260641
summer	251521
italy	243073
trip	239890

7.1.2 Flickr12M データセットの構成

Flickr12M の詳細データを表 7.2にまとめる．一枚の画像が持つラベルの数は平均 3.47 と Corel5K などと同程度であるが，画像ごとのばらつきが大きい．また，単語の出現頻度を表 7.3，図 7.3に示す．単語の出現頻度は非常に偏っており，少数の単語が支配的になっているといえる．このようなバイアスは Web 画像マイニングにおける一般的な問題であり，いかにして多様なアノテーションを行うかが重要となる．表 7.4に，最も出現頻度の高い 10 単語を示す．

7.2 基礎評価実験

まず，160 万枚までのサブセットを用い，5.2 節と同様の評価実験を行い，大規模な問題における提案手法の有効性を確認する．また，いくつかの画像特徴量を比較し，GLC に基づく特徴量が特に有効であることを示す．

7.2.1 画像特徴量

ここでは，以下の画像特徴量を比較する．

1) Tiny image [182] (3072dim)

2) RGB color histogram (4096dim)

3) GIST [148] (960dim)

4) HLAC (2956dim)

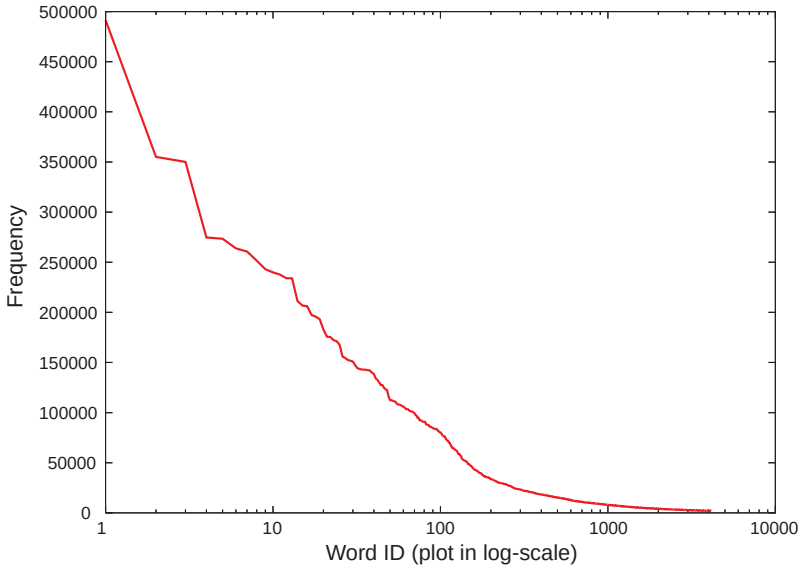


Figure 7.3: Word frequencies in the Flickr12M dataset.

5) SURF GLC (2144dim)

6) SURF BoVW (1000dim)

7) SURF BoVW-sqrt (1000dim)

8) RGB-SURF GLC (3432dim)

Tiny image は，32×32 ピクセルに縮小した画像のピクセル値をそのまま画像特徴量として用いるものである．従って，3 チャンネルの画像の場合，特徴次元数は $32 \times 32 \times 3 = 3072$ となる．RGB color histogram は，RGB の色次元をそれぞれ 16 分割したヒストグラムであり，TagProp で用いられた特徴の一つである．GIST は，基本的に 5.2 節で用いたものと同一であるが，タスクの規模を考慮し，特徴次元数を 960 次元へ増やしている．HLAC は，前章までで用いたものと同一である．GLC, BoVW の抽出に用いる局所特徴は SURF 特徴を利用し，Dense sampling による抽出を行う[1]．k-means クラスタリングにより 1000 個の visual words を生成し，これを用いて 1000 次元のヒストグラムを生成する (SURF BoVW)．なお，文献 [155] より，BoVW の Bhattacharyya カーネルは，BoVW の各要素の平方根をとったベクトル (BoVW-sqrt) の線形カーネルと等価であることが指摘されている．これは，BoVW-sqrt は BoVW よりも線形手法に適した表現であることを意味するものであるため，本実験で比較を行う．また, RGB-SURF は R, G, B の各

[1] ここでは，16×16 ピクセルのパッチを 8 ピクセルずつずらしながら抽出する．一枚の画像当たり，$1,200 \sim 1,300$ 点の局所特徴が得られる．

グレースケール画像から抽出した SURF 特徴を結合したものであり，その次元数は $64 \times 3 = 192$ となる．式 6.31に従い，RGB-SURF GLC を抽出する．用いる主成分の次元数は 80 とする．

7.2.2 評価方法

アノテーションは，5 章と同様に，各画像について上位 5 単語を出力するものとする．評価は，2 種類の F 値を用いて行う．一つ目は，単語ごとの Recall・Precision の全単語平均の F 値であり，5 章で用いたものと同一である．詳しくは，Appendix A を参照されたい．本章では，これを F_W と表記する．

F_W は，より多くの単語についてアノテーションに成功するほど値が向上するため，アノテーションの多様性を測ることに適した指標である．これに加え，二つ目の指標として画像ごとの Recall・Precision の全単語平均の F 値 (F_I) を用いる．あるテスト画像 I_j について，システムが正しくラベルづけした単語数を x, グラウンドツルース（タグ）の単語数を y, 正解・不正解に関わらずシステムが I_j にラベル付けした単語数を z とする（本実験では常に $z = 5$ である）．テスト画像 I_j についての Recall, Precision は以下のように定義される．

$$\text{Recall}(I_j) = x/y \tag{7.1}$$
$$\text{Precision}(I_j) = x/z. \tag{7.2}$$

これらの全テスト画像における平均を Image-centric Mean Recall (MRi)，Image-centric Mean Precision (MPi) とする．最終的に用いる F_I は次のように定義される．

$$F_I = \frac{2 \times \text{MRi} \times \text{MPi}}{\text{MRi} + \text{MPi}}. \tag{7.3}$$

F_I は，個々のテスト画像へのアノテーションの正確さを直接反映する指標である．従って，F_W とは逆に，比較的少数の基本的な単語の認識精度を強く反映すると考えられる．

7.2.3 実験結果

潜在空間の次元数 d は，20, 50, 100, 200, 300 次元の 5 通りにとる．それぞれについて，k 最近傍法による画像アノテーションを $k = 50, 100, 200, 400, 800, 1600$ について行い，ベストスコアを採用する．なお，5.2 節の実験結果では，MLR は PLS や CCA と有意な性能差は認められなかったため，本実験では割愛する．

それぞれの画像特徴量について，学習サンプル数を倍々に増やしながら手法の比較を行う．利便性を考慮し，実験結果を Appendix D へまとめた．基本的に，どの手法・圧縮次元数においても，学習サンプル数を増やすほど認識精度が向上していることが分かる．学習サンプル数を 10 万サンプル (100K) から 160 万サンプル (1.6M) に増加させた時，F_I は 20%程度，F_W は 100〜200%程度上昇してお

り，特に F_W のスコア向上が顕著である．このことは，アノテーションの多様性を向上させる（すなわち，認識可能な語彙数を増やす）ためには学習データセットの大きさが重要であることを示唆している．

HLAC 特徴においては，CCD2 が安定に最もよいスコアを出しており，5.2 節で得られた知見と一致する．また，SURF GLC や RGB-SURF GLC についても同様の傾向を示しており，HLAC 特徴と同様に提案手法との相性がよいことが確認された．他の特徴においては，CCD は必ずしも優位ではなく，nPLS などが CCD を上回る場合もある．RGB カラーヒストグラムの例が示すように，CCA・CCD のスコアはサンプル数が少ない場合他手法に劣るが，サンプル数が増えるにつれ次第に差を縮め，追い越している．これは，学習サンプル数が増えるほど CCA の安定性が向上するためであると考えられ，学習データが大規模になるほど提案手法が優位になることを示唆している．

図 7.4に，各特徴に CCD2 を適用した際のスコアを示す．各特徴について，最もよいスコアを示した圧縮次元数を選択している．HLAC と GLC は他を大きく上回る性能を示している．HLAC は局所特徴記述子として画素値そのものを用いた場合の GLC であると解釈できることから，GLC と同様の傾向を示すことは理にかなっており，GLC ベースの画像特徴の有効性を裏付けるものである．

また，BoVW-sqrt は BoVW よりも常によいスコアを示している．このことから，BoVW-sqrt は線形手法にとってより適した表現となっていることが裏付けられた．

これらの結果を踏まえ，次節では HLAC, SURF GLC, SURF BoVW-sqrt の 3 つの特徴量を用い，Flickr12M の全データを用いた学習を行う[1]．

7.3 本実験

7.3.1 定量的評価

まず，個々の画像特徴量の性能を評価する．図 7.5に，SURF BoVW-sqrt, SURF GLC, HLAC のそれぞれのスコアを示す．ここでは，PCAW (BoVW-sqrt については PCA)，CCA, CCD2 の比較を行っている．さらに，図 7.6に複数の特徴量を結合して用いた場合の結果を示す．各特徴量のスケールは異なるため，ここでは CCA と CCD2 のみ適用している．前節の予備実験と同様に，データ数に対しアノテーション精度が対数スケールで向上していることが確認できる．また，CCD2 は CCA よりも常に高いスコアを示している．

さらに，複数の特徴量を用いることにより，アノテーション精度が大きく向上することが示された．例えば，6 章での狙い通り，GLC と BoVW を併用することでそれぞれを独立に用いる場合に比べ有意にスコアが向上している．また，全ての特徴量を利用した場合に最も高いスコアを得ている．これは，2.2.2 節で議論したように，semantic gap を緩和するためには出来るだけ多様な特徴を用い

[1] RGB-SURF GLC は計算コストが大きいため，次節の実験では利用しない．

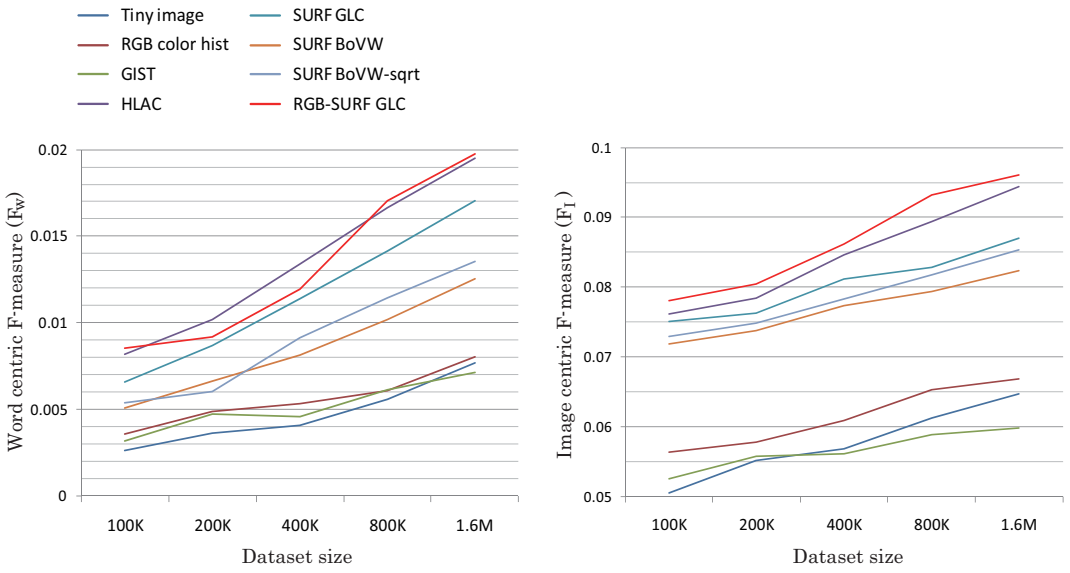


Figure 7.4: Annotation performance of each feature with CCD2 (<1.6M samples).

ることが重要であることを裏付ける結果である．図 7.7に，CCD2 を各特徴量に適用した場合のスコアを重ねて示す．

7.3.2 大規模コーパスの定性的効果

ここでは，定性的な例をいくつか挙げながら，多数の学習サンプルを用いることの効果を述べる．提案手法はノンパラメトリックな認識手法であり，クエリ画像の近傍学習サンプルの持つ教師ラベルを重みづけてアノテーション結果とするものである．従って，適切な学習サンプルが近傍サンプルとしてヒットするほどアノテーションの精度は向上する．3 つのテスト画像について，学習データセットの大きさを 10 万枚，160 万枚，1200 万枚と変化させた際のアノテーション結果を図 7.8,7.9,7.10に示す．それぞれの場合について上位 10 単語のアノテーションと，最近傍の 25 画像を示している．ただし，画像のみからは類推困難な場所や時間に関するアノテーションは除外している．

図 7.8のステンドグラスの例が最も典型的である．10 万枚のデータセットでは，近傍 25 サンプルの中に一つしかステンドグラスの画像は含まれておらず，視覚的に類似しているチームスポーツの画像が多く含まれていることが分かる．その結果，アノテーションも“football”など実際の内容とは異なる単語が出力されている．学習サンプルを 160 万枚まで増やすと近傍サンプルの質は幾分改善されるものの，アノテーション結果の改善までには至っていない．これに対し，全データをもちいると 25 個の近傍サンプルが全てステンドグラスになっており，アノテーション結果も大幅に改善されていることが分かる．

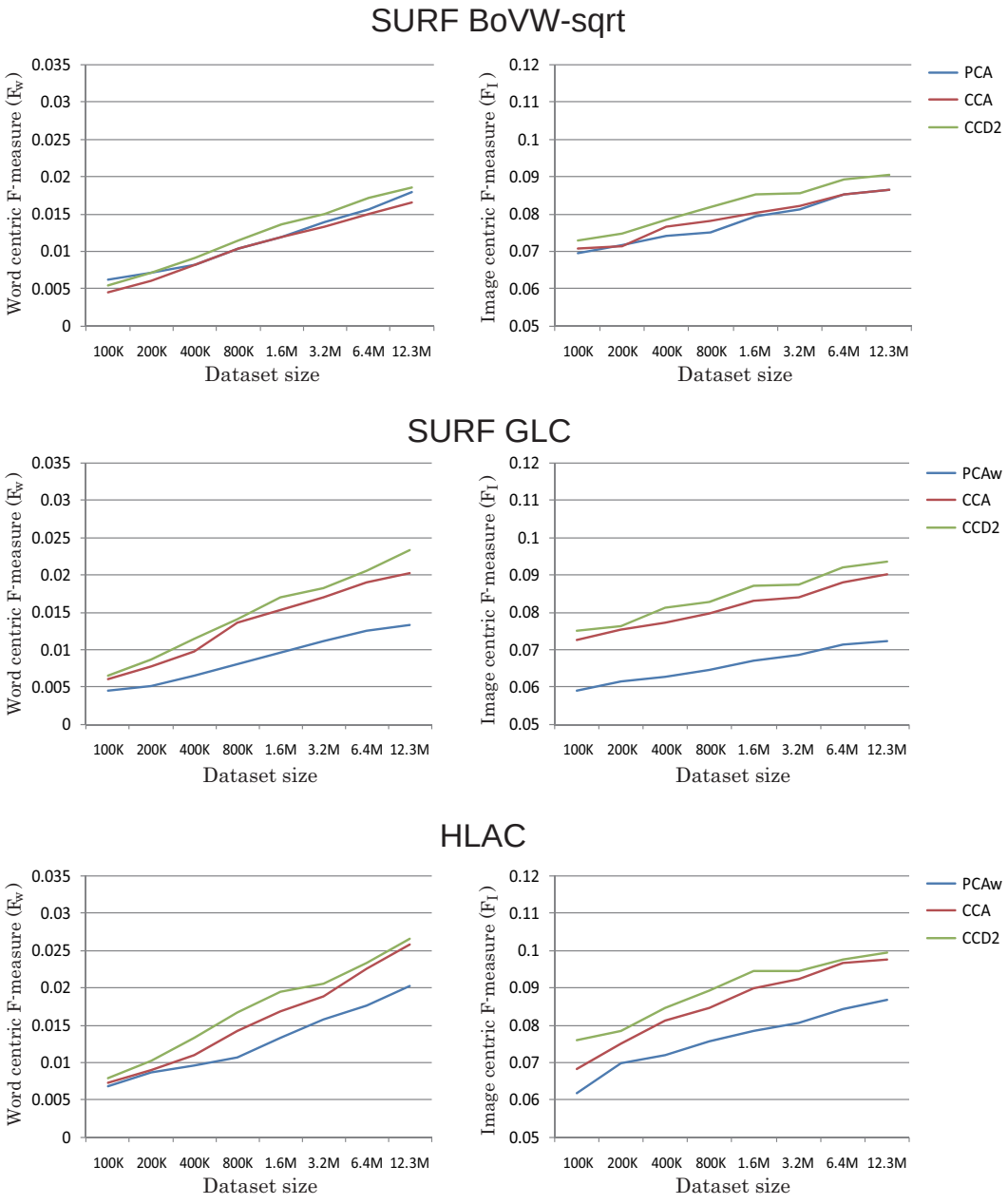


Figure 7.5: Annotation performance of SURF BoVW-sqrt, SURF GLC, and HLAC features (<12.3M samples).

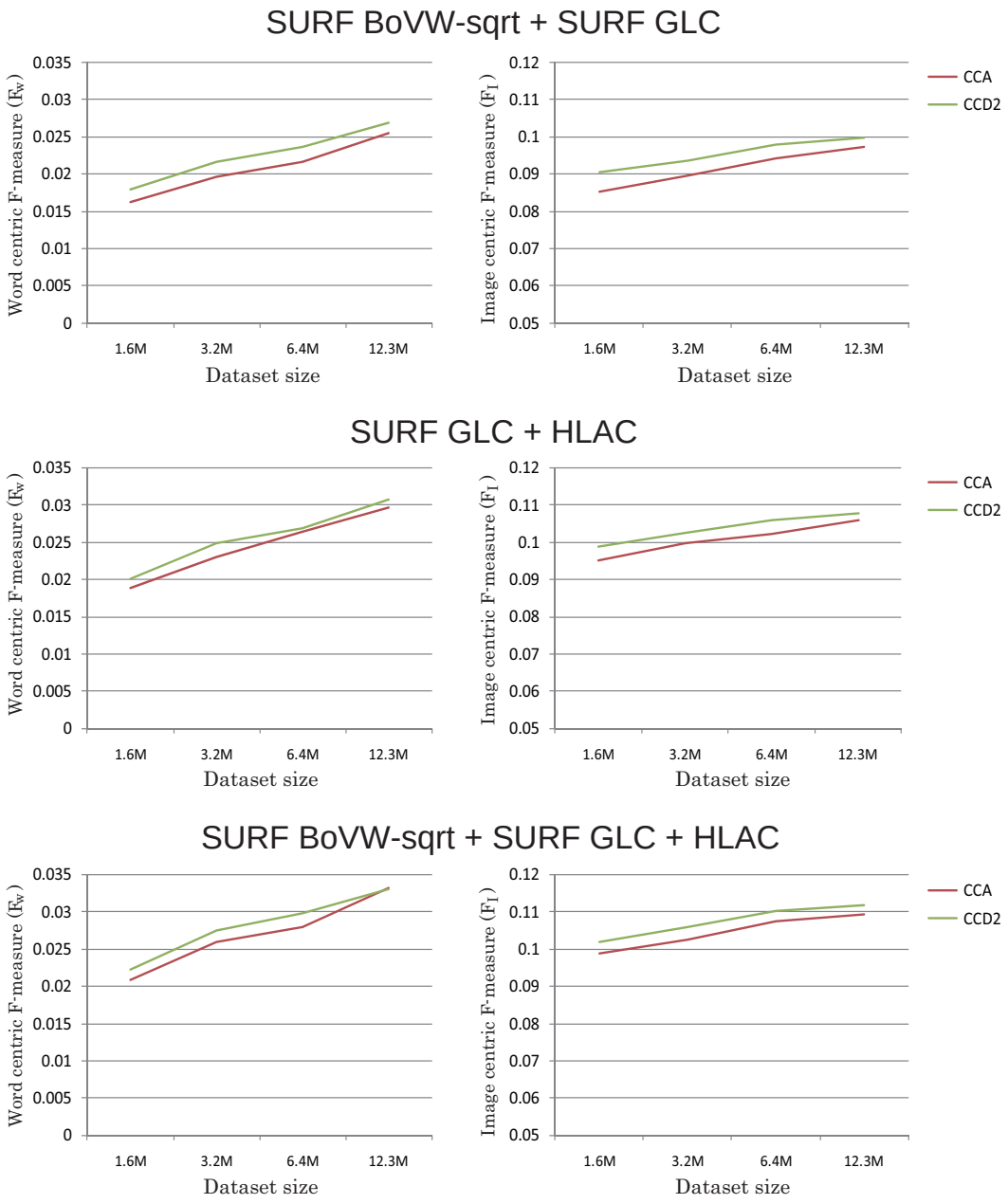


Figure 7.6: Annotation performance of combinations of SURF BoVW-sqrt, SURF GLC, and HLAC features (<12.3M samples).

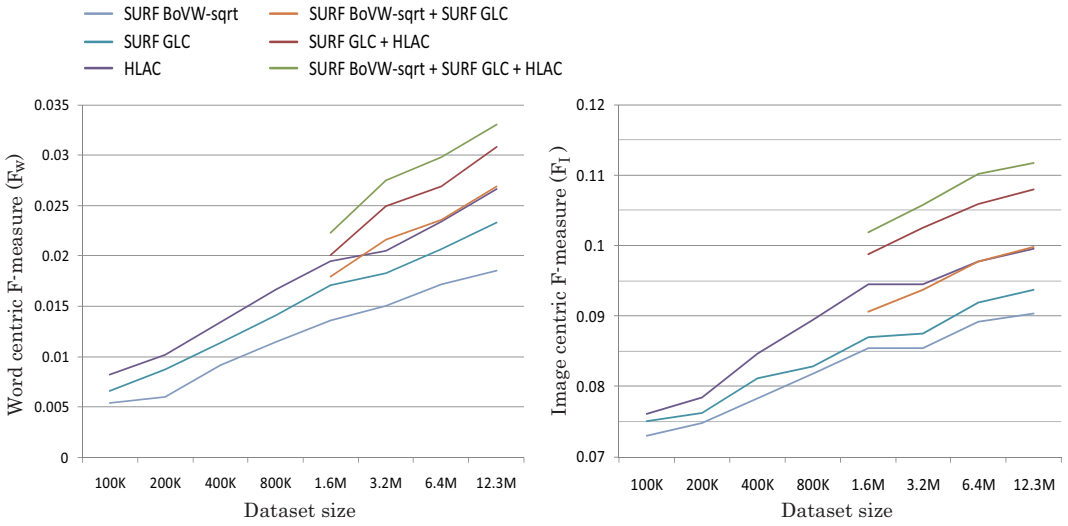


Figure 7.7: Comparison of annotation performance with CCD2 (<12.3M samples).

同様の結果は，図 7.9, 図 7.10からもみてとれる．図 7.9のイルカの画像は，サンプルが少ない時には海の風景画像に誤認されているが，サンプルが増えるにつれて飼育されているイルカの画像であると認識できるようになっていることが分かる．図 7.10は，space mountain というディズニーランドのアトラクション（ジェットコースター）の画像である．これも，10 万枚のサンプルでは全く認識できていないが，全サンプルを用いるとディズニーランドであることまで認識できるようになっている．このように，大規模な学習コーパスを用いることで，クエリと意味的に類似したサンプルが近傍サンプルとしてヒットするようになり，アノテーションの精度は安定に向上する．これは，semantic gap の問題が徐々に緩和されていることを示唆するものであるといえる．

Query image

	Annotation results	Nearest training samples
100K	football soccer varsity girls boys travel party family school high	
1.6M	football soccer festival college futbol park people cycling marchingband vacation	
12.3M	church stainedglass football bath city vacation travel cathedral window glass	

Figure 7.8: (1/3) Example of annotation using a varying number of training samples. Correct annotations are written in red. For each case, the 25 nearest images are shown.

Query image

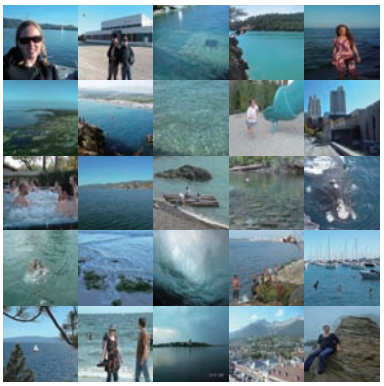
	Annotation results	Nearest training samples
100K	vacation beach travel trip nature sea island dale dalemorton mftmon	
1.6M	pool seaworld vacation swimming aquarium travel beach trip water sea	
12.3M	seaworld dolphin water sept zoo animals vacation beach orangecountry trip	

Figure 7.9: (2/3) Example of annotation using a varying number of training samples. Correct annotations are written in red. For each case, the 25 nearest images are shown.

Query image

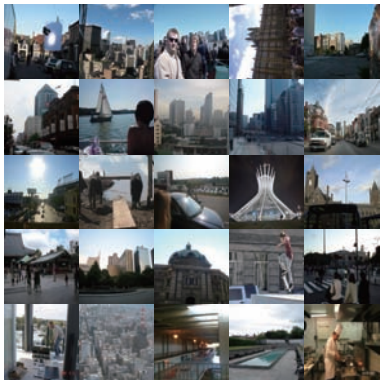
	Annotation results	Nearest training samples
100K	city trip buildings travel nikon road van birthday house car	
1.6M	travel rollercoaster freizeitpark vergngungspark achterbahn amusementparc vacation park city architecture	
12.3M	vacation festival travel trip disney city roadtrip waltdisneyworld rollercoaster matsuri	

Figure 7.10: (3/3) Example of annotation using a varying number of training samples. Correct annotations are written in red. For each case, the 25 nearest images are shown.

Chapter 8

結論と展望

8.1 結論

汎用性の高い一般画像認識を実現するためには，大量の事例データからの学習が鍵となる．しかしながら，従来の手法は学習サンプル数に対するスケーラビリティを欠いていたため，大規模な画像データセットを用いて学習・認識を行うことは著しく困難であった．そこで本研究では，学習サンプル数に対しスケーラブルかつ高精度な一般画像認識（画像アノテーション）アルゴリズムの開発に取り組みこれを実現した．これを可能としたのは，本研究で提案したサンプル間距離計量の統計的学習手法と，画像特徴量の抽出手法である．これらは互いに密接に影響するプロセスであり，両者の相性を考慮しそれぞれを設計することが極めて重要である．最終的に，開発した画像アノテーション手法を 1,200 万枚の画像データセットへ適用し，その有効性を示した．以下，本研究で得られた知見をまとめる．

サンプル間の判別的距離計量学習に基づく画像アノテーション (4 章 , 5 章)

一枚の画像に複数単語のラベルづけを行う画像アノテーションにおいては，ノンパラメトリックに事例サンプルを用いる識別則が有効である．これは，サンプル数が増えるほど認識精度が向上することが保証されているため，大規模学習コーパスの効果を得やすいアプローチと言える．しかしながら，以下の 2 点が大きな問題となる．

- Semantic gap を緩和したサンプル間の距離計量をどのように定義するか．
- サンプルの特徴次元数をいかにして圧縮するか．

これらの問題を解決するためには統計的機械学習の手法を用いる必要があるが，大規模コーパスを前提とする場合，学習サンプル数に対し線形オーダでの学

習の実現が望ましい．そこで本研究では，バイモーダルの線形次元圧縮手法である正準相関分析 (CCA) に着目した．古典的な CCA 自体はあくまで次元圧縮のみを行うものであり，サンプル間距離計量に関する知見は与えない．そこで，本研究では CCA の確率構造 (PCCA) を利用することで，理論的に最適なサンプル間距離計量の導出を行い，これを Canonical Contextual Distance (CCD) と名付けた．実験により，CCD に基づく画像アノテーション手法は相対的に少ない計算コストで学習・認識が可能であり，かつ先行研究と遜色ない認識精度を達成できることが示された．提案手法は，以下の特長を有する．

- サンプル数に対し線形オーダの計算コストで学習が可能．
- 解析解が求まるため，学習時に反復的にデータアクセスする必要が生じない．
- 認識において，サンプル間距離計算のコストが相対的に小さい．

CCA と同様の次元圧縮手法として，PLS, MLR などの利用も考えられ，これらの手法から得られる部分空間におけるユークリッド距離を距離計量として利用することも可能である．いくつかの画像特徴量においては，PLS などが CCA (CCD) よりも優れた認識精度を示す場合が見られた．これは，PLS が計算的に安定な手法であるため，非線形性を持つ画像特徴量に対しては CCA よりも安定に次元圧縮が行えるためであると考えられる．しかし，このような場合は線形手法をそのまま適用すること自体が不適当であり，どの手法も元の画像特徴空間において直接非線形な距離計量を用いる場合に比べて著しく認識精度が低下する．

一方，カーネル法を用い画像特徴量を線形化した場合，ないし画像特徴量がもともと線形な性質を有している場合は，CCD が常に最高の認識精度となった．これは，線形の前提のもとでは，CCD が一般的に最も優れた手法であることを示唆している．このように，CCD を有効に利用するためには入力となる画像特徴量に関して注意が必要であることも判明した．これに関しては次に述べる．

画像特徴量の抽出手法 (6 章)

CCD は画像特徴空間に対して線形性を仮定している．すなわち，内積が画像特徴の類似度を適切に定義していることを前提としている．しかしながら，現存する多くの画像特徴にはこの前提が成り立たず，線形手法である CCD を適用すると著しい性能低下につながる場合がある．このような場合，通常はカーネル法の利用により陰に問題解決が図られるが，十分な精度を得るためには多くの基底サンプルをカーネル化に用いる必要がある．結果として，学習アルゴリズムのスケーラビリティは著しく損なわれ，大規模コーパスへの適用は不可能となる．これは，既存の一般画像認識手法がスケーラビリティを欠いていた理由に他ならない．

この問題を解決するためには，線形性の仮定が成り立つように画像特徴を設計することが必要である．本研究では，局所特徴分布を単一のガウシアンによってモデル化する global Gaussian approach を提案した．また，ガウシアンを情報幾

何の手法により近似的にコーディングした大域的特徴ベクトルである generalized local correlation (GLC) の開発を行った．

Global Gaussian approach は，従来注目されなかった局所特徴分布の低次な統計的情報を活用することを目的としたものであり，3 つのシーン認識のベンチマークにおいて最高の認識精度を達成した．その特性は以下のとおりである．

- 任意の局所特徴記述子を利用可能である．
- 画像ごとに固有の表現である．
- 線形近似を行った場合でも，一般的な bag-of-visual-words と同程度の高い認識精度を得る．
- さらに，bag-of-visual-words とは相補的な関係にあるため，両者を併用することでさらに認識精度が向上する．

最終的な特徴ベクトルである GLC は，ガウシアンが為す多様体の座標系をとるものであり，以下の特性を持つ．

- CCD など，特徴空間のアフィン変換に対する不変性を有する線形手法に直接適用可能である．
- 一般的な bag-of-visual-words よりも高速に抽出可能である．

特に一つ目の特性により，CCD に基づく画像アノテーション手法にとって理想的な特徴表現になっているといえる．

なお，古典的な HLAC 特徴は GLC の特殊なケースであると解釈できる．HLAC 特徴は経験的に線形手法との相性のよさが知られているが，その理由も GLC の数理的導出過程より明らかとなった．

大規模画像コーパスの効果 (7 章)

CCD と GLC を利用した画像アノテーション手法を 1,200 万枚の画像データセットへ適用し，以下の重要な知見を得た．

- 学習サンプル数を増やすほど，認識可能な語彙数が増え，個々の画像に対するラベルづけの精度も向上する．
- CCD は他の次元圧縮手法と比較して常に高い認識精度を得る．
- 異なる多くの画像特徴量を用いるほど，認識精度が向上する．特に，GLC に基づく画像特徴が有効である．

前述のように，学習コストがサンプル数に対し線形オーダであるという条件のもとでCCDは理論的に最も優れた距離計量学習手法である．学習サンプル数が10万枚，20万枚の小規模なデータセットにおいてはPLSなど他手法がCCDを上回る場合があるが，サンプルが増えると認識精度は逆転し，CCDが優位となる．これは，サンプルが増えることでCCDの核であるCCAの計算が安定するためであると考えられる．このことは，CCDは大規模なデータセットにおいてこそ真の力を発揮する手法であることを示唆している．さらに，GLCはCCDが必要とする，特徴空間の線形性を近似的に保証している画像特徴であり，両者の組み合わせが最もよい結果となるのは妥当な結果である．これらは，大規模画像コーパスを用いた一般画像認識および本研究で開発した手法の有効性を裏付ける事実である．

8.2 解決すべき課題

本研究では，従来困難であった大規模な画像アノテーションのための道具立てを整えた．しかしながら，実用的な認識性能を持つ画像アノテーションシステムの実現のためには課題がいくつか存在する．

良質な大規模画像コーパスの作成

本研究では手法の開発に焦点をあてたが，大規模かつ良質な画像コーパスの作成そのものも重要な課題であり，一つの研究分野となりつつある．近年ではcrowd sourcingの利用により，人海戦術で大規模なコーパスを作成することが可能となりつつあるが，不特定多数の人間が作業に関わるため，良質かつ一貫したラベリングを行うことは容易ではない．また，一般画像認識においては，そもそもどのようなグラウンドツルースを作成すれば有用であるか明らかではないことも問題である．現在はWordNetなどの言語シソーラスを用いてトップダウンに設計を行うものが多いが，今後はより一般画像認識に適したグラウンドツルースを考える必要があろう．

インタラクティブな学習の実現

あらかじめ用意する大規模なコーパスにより獲得できるのは一般的な知識であり，世界の全てをカバーできるものではない．例えば，ローカルな環境にのみ存在する事物や，新しく発生した概念については対応できない．また，既に知っている概念であっても，学習した事例と著しく異なるアピアランスをしている場合は認識が困難である．このような場合に，これらに関する知識を逐次的に獲得することが重要となる．

そのためには，まずこれらの事物を未知なものであると判断できなければならない．これは，既知の事例からの汎化を行う統計学習とは背反する概念であり，

両者の両立は本質的に難しい課題である．また，未知と判断した事物について，適切にユーザに問い合わせ教示を受けるフレームワークを構築する必要がある．

8.3 本研究の発展

領域ラベリング手法への応用

本研究では，画像全体への複数ラベリングである画像アノテーションを最重要の課題であると考えこれに取り組んだ．将来的には，この上に画像領域認識（物体検出）のアルゴリズムを統合することが期待される．まず，画像アノテーションによる画像全体の認識結果から，そのおおまかなシーンを推定できる．そのシーンについて出現する可能性の高い物体について検出を行うことにより全体として効率のよいシステムが実現できると期待できる．また，さらに発展的な課題として，物体検出と画像アノテーションの同時最適化を行い互いに精度を向上させることも考えられる．

静止画像以外のリソースの利用

一般画像認識は基本的に静止画像のみを入力とするものであるが，場合によっては他の情報を統合することが合理的である．例えば，現在のスマートフォンではGPS 情報や慣性センサの値を容易に取得できるため，これらを用いて認識対象を絞り込むことが可能である．ロボットなどでは音声や動画像などさらに多様な情報を統合することで，より実用性の高いアプリケーションが実現できると期待できる．このような複数の情報（モダリティ）の統合には，マルチモーダル CCA [99] の利用が有効であると考えられる．

実世界画像認識への発展

個人アルバムや Web の画像を対象に認識を行いたいのであれば，Web 上の画像をサンプルとして学習すれば十分である．実際，現在一般画像認識に用いられる画像サンプルの多くは Web 上から収集されたものである．しかしながら，Web 上の画像のほとんどは人間により撮影され，何らかの目的のためにアップロードされたものである．言い換えれば，Web 上の画像はもともとはっきりした意味を保証されたものであるといえる．これに対し，我々人間やロボットが日々目にする画像は必ずしも意味を持たない雑多なものであり，性質が大きく異なる．このような実世界画像を適切に認識することが，一般画像認識の真のゴールであるべきである．その実現には，例えばライフログなどにより取得される，より目的に近い画像サンプルを利用した学習が必要であると考えられる．

Appendix A: 画像アノテーションおよびリトリーバルの評価プロトコル

本論文では，先行研究において定められた評価プロトコル [51; 92] に従い，画像アノテーション・リトリーバルの性能評価を行う．ここでは，その詳細について述べる．

A.1. アノテーションの評価指標

画像認識システムは，一つのテスト画像につき 5 単語ずつアノテーションを行う．システムの出力した単語はあらかじめ設定されているグラウンドツルースと一致した場合に正解とされる．ある単語 w_i について，システムが正しくラベルづけしたテスト画像数を a, 実際に w_i をグラウンドツルースとして持つテスト画像数を b, 正解・不正解に関わらずシステムが w_i をしてラベル付けしたテスト画像数を c とする．単語ごとの Recall, Precision は以下のように定義される．

$$\text{Recall}(w_i) = a/b, \tag{1}$$

$$\text{Precision}(w_i) = a/c. \tag{2}$$

これらの全単語平均をそれぞれ Mean-Recall (MR)，Mean-Precision (MP) とする．一般に MR と MP はトレードオフの関係にあるため，調和平均（F-measure）を統一的な評価指標として用いる [92]．

$$\text{F-measure} = \frac{2 \times \text{MR} \times \text{MP}}{\text{MR} + \text{MP}}. \tag{3}$$

なお，共通の評価プロトコルとして，先行研究では上位 5 単語を認識結果として出力するが，評価指標の値は出力単語数によって変化する．参考として図 1に，Corel5K に提案手法を適用し，出力する単語数を変化させた場合の Mean Recall, Mean Precision, F-measure の値を示す．基本的には，出力する単語数を増やすほど Recall が上がり，Precision が下がるが，グラウンドツルースの数に対して出力する単語数が少なすぎる場合は選択できる単語の種類が少なくなるため，Recall, Precision とも著しく低い値となる．Corel5K では，1 サンプルあたりのグラウン

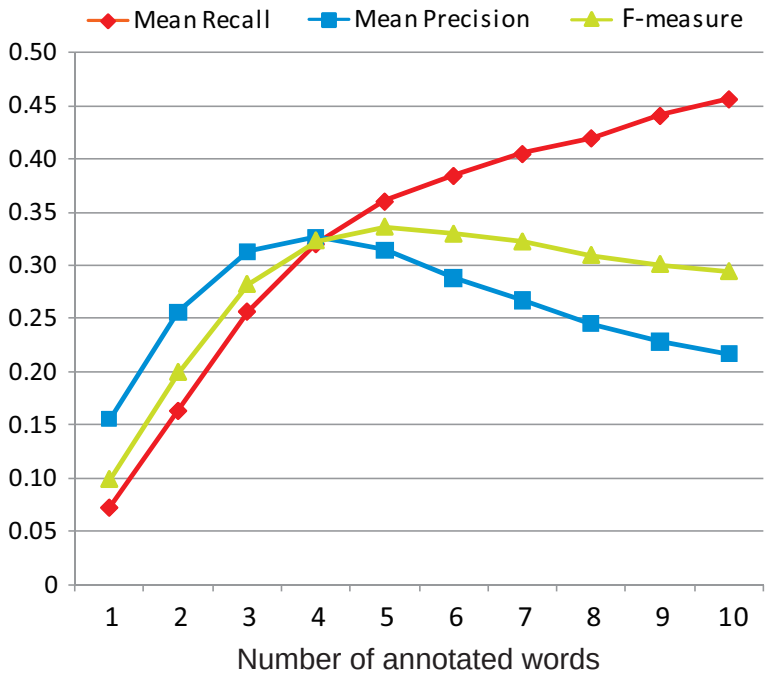


Figure 1: Annotation scores for the Corel5K dataset with varying numbers of output words. The proposed method (linear) + HLAC feature is used.

ドツルースの単語数の平均は 3.4 単語であるため，Precision は 4 単語出力から減少に転じるが，Recall は増加を続ける．実用的には，必要となるアノテーション数，辞書単語数を考慮して設計することが望ましい．また，一枚の画像が持つグラウンドツルースの数は異なるため，MR や MP の理論的上限値は多くの場合 1 にはならないことにも注意されたい．

これらに加え，リコールがゼロでない（すなわち，一度でもアノテーションに成功した）単語数も評価する $(N+)$.

A.2. リトリーバルの評価指標

リトリーバルでは，テスト単語それぞれについて全ての候補画像のランク付けを行い，Mean Average Precision (MAP) を用いて評価する．リトリーバル対象となる候補画像数を N_t とする．あるクエリ単語 w に関する Average Precision は，以下のように定義される．

$$\mathrm{AP}(w) = \frac{1}{\sum_{i=1}^{N_t} y_i^w} \sum_{i=1}^{N_t} \frac{y_i^w}{i} \sum_{k=1}^{i} y_k^w. \tag{4}$$

Figure 2: Illustration of “car” retrieval results. Correct images are ranked 2nd, 5th, and 7th, respectively.

ただし y_i^w は，i 番目にランキングされた画像がクエリ w に関連すれば 1，そうでなければ 0 をとるフラグである．例えば，10 枚の画像セットに対し，”car” をクエリとしてリトリーバルを行う（図 2）．実際に”car” に関連する画像が 3 枚存在するとする．システムが，これらの画像をそれぞれ 2 位，5 位，7 位にランキングした場合，AP(car)$= \frac{1}{3}\left(\frac{1}{2}+\frac{2}{5}+\frac{3}{7}\right) = 0.44$ となる．

MAP は AP の平均値であり，情報検索分野において標準的に用いられる性能指標の一つである．評価には，全辞書単語における平均 (MAP) と，アノテーションにおいて recall がゼロでなかった単語についてのみの平均 (MAP R+) の 2 通りを用いる．

Appendix B: カーネル主成分分析

B.1. 標準的な実装

まず，全学習サンプルをカーネル化に用いる標準的なカーネル主成分分析 (KPCA) [166] について述べる．学習サンプル数を N とする．元の特徴空間上の点 $\boldsymbol{x}$ を，高次元の空間へ射影する非線形写像 $\phi(\boldsymbol{x})$ を考える．一般に，このような写像 ϕ は陽には与えられず，カーネル関数を用いて内積のみが定義される．すなわち，$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$ を用いることにより，実際には高次元への射影を行うことなく，元の特徴空間において内積計算を行うことができる．

高次元空間における共分散行列を C とすると，定義より

$$C = \frac{1}{N} \sum_{i=1}^{N} \left(\phi(\boldsymbol{x}_i) - \frac{1}{N} \sum_{j=1}^{N} \phi(\boldsymbol{x}_j) \right) \left(\phi(\boldsymbol{x}_i) - \frac{1}{N} \sum_{j=1}^{N} \phi(\boldsymbol{x}_j) \right)^T, \tag{5}$$

となる．この高次元空間における主成分分析の解は，次の固有値問題によって得られる．

$$C\boldsymbol{v} = \lambda \boldsymbol{v}. \tag{6}$$

ここで，C の定義より，式 6は次のように書きかえられる．

$$\frac{1}{N} \sum_{i=1}^{N} \left(\phi(\boldsymbol{x}_i) - \frac{1}{N} \sum_{j=1}^{N} \phi(\boldsymbol{x}_j) \right) \left(\left(\phi(\boldsymbol{x}_i) - \frac{1}{N} \sum_{j=1}^{N} \phi(\boldsymbol{x}_j) \right)^T \boldsymbol{v} \right) = \lambda \boldsymbol{v}, \tag{7}$$

すなわち，固有ベクトル $\boldsymbol{v}$ は学習サンプルの射影 $\phi(\boldsymbol{x}_i)$ の線形結合として表せることが分かる．

$$\begin{aligned} \boldsymbol{v} &= \sum_{i=1}^{N} \alpha_i \left(\phi(\boldsymbol{x}_i) - \frac{1}{N} \sum_{j=1}^{N} \phi(\boldsymbol{x}_j) \right) && (8) \\ &= (\Phi - \Phi 1_N)\boldsymbol{\alpha}. && (9) \end{aligned}$$

ただし，

$$\Phi = \left(\phi(\boldsymbol{x}_1) \quad \phi(\boldsymbol{x}_2) \quad \ldots \quad \phi(\boldsymbol{x}_N) \right), \tag{10}$$

であり，$1_N \in \Re^{N\times N}$ は，要素が全て $1/N$ なる行列である．

これを式 6に戻すと，次のようになる．

$$\frac{1}{N}(\Phi - \Phi 1_N)(\Phi - \Phi 1_N)^T(\Phi - \Phi 1_N)\boldsymbol{\alpha} = \lambda(\Phi - \Phi 1_N)\boldsymbol{\alpha}. \tag{11}$$

両片に左から $(\Phi - \Phi 1_N)^T$ をかけると，

$$\frac{1}{N}(\Phi-\Phi 1_N)^T(\Phi-\Phi 1_N)(\Phi-\Phi 1_N)^T(\Phi-\Phi 1_N)\boldsymbol{\alpha} = \lambda(\Phi-\Phi 1_N)^T(\Phi-\Phi 1_N)\boldsymbol{\alpha}. \tag{12}$$

ここで，カーネルトリックにより以下のような置き換えが可能である．

$$\begin{aligned}(\Phi - \Phi 1_N)^T(\Phi - \Phi 1_N) &= \Phi^T\Phi - \Phi^T\Phi 1_N - 1_N\Phi^T\Phi + 1_N\Phi^T\Phi 1_N &(13)\\ &= K - K1_N - 1_N K + 1_N K 1_N, &(14)\end{aligned}$$

ただし，$K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ である．このように，カーネル関数を用いることでグラム行列 $\tilde{K} = K - K1_N - 1_N K + 1_N K 1_N$ を求めることができる．最終的に，求めるべき固有値問題は

$$\tilde{K}^2\boldsymbol{\alpha} = \lambda N\tilde{K}\boldsymbol{\alpha}, \tag{15}$$

となり，両辺から共通因子 $\tilde{K}$ を取り除くと，

$$\tilde{K}\boldsymbol{\alpha} = \lambda N\boldsymbol{\alpha}, \tag{16}$$

となる．固有ベクトルに必要な規格化条件は，$\boldsymbol{v}^T\boldsymbol{v} = 1$ であり，式 9, 式 16を用いて書き直すと，

$$\begin{aligned}1 &= \boldsymbol{\alpha}^T(\Phi - \Phi 1_N)^T(\Phi - \Phi 1_N)\boldsymbol{\alpha} &(17)\\ &= \boldsymbol{\alpha}^T\tilde{K}\boldsymbol{\alpha} &(18)\\ &= \lambda N\boldsymbol{\alpha}^T\boldsymbol{\alpha}. &(19)\end{aligned}$$

となる．

あるサンプル点 $\boldsymbol{x}_s$ の KPCA による射影は，

$$\begin{aligned}\boldsymbol{v}^T\left(\phi(\boldsymbol{x}_s) - \frac{1}{N}\sum_{j=1}^{N}\phi(\boldsymbol{x}_j)\right) &= \boldsymbol{\alpha}^T(\Phi - \Phi 1_N)^T(\phi(\boldsymbol{x}_s) - \Phi 1_N^c) &(20)\\ &= \boldsymbol{\alpha}^T(K_s - 1_N K_s - K1_N^c + 1_N K 1_N), &(21)\end{aligned}$$

となる．ここで，$1_N^c \in \Re^N$ は，全要素が $1/N$ の列ベクトルである．また，K_s は $\boldsymbol{x}_s$ のカーネルベースベクトルであり，

$$K_s = (k(\boldsymbol{x}_s, \boldsymbol{x}_1)\ \ k(\boldsymbol{x}_s, \boldsymbol{x}_2)\ \ \ldots\ \ k(\boldsymbol{x}_s, \boldsymbol{x}_N))^T. \tag{22}$$

B.2. 少数の基底サンプルを用いた実装

KPCA において中心となる考え方は，高次元空間における学習サンプルの線形結合により固有ベクトルを表現することである．これらのサンプルを基底サンプルと呼ぶことにする．一般に，より多くの基底サンプルを用いることにより，漸近的に真の解へ収束することが知られている．しかしながら，基底サンプル数次元の固有値問題を解く必要が生じるため，大規模なデータにおいて全学習サンプルを用いカーネル化を行うことは現実的でない．

ここでは，効率よくカーネル化を行う方法として，少数の学習サンプルを基底サンプルとして用いた実装を試みる．カーネル化に用いる基底サンプル数を n_K とする．式 9と同様に，n_K 個のサンプルの線形結合により固有ベクトルを表す．ここでのインデックスは，学習サンプル全体ではなく基底サンプルについて示していることに注意されたい．

$$\begin{aligned} \boldsymbol{v} &= \sum_{m=1}^{n} \beta_m \phi(\boldsymbol{x}_m) && (23) \\ &= \Phi_B \boldsymbol{\beta}. && (24) \end{aligned}$$

式 9では高次元空間におけるサンプル平均を差し引いているが，これは単なるオフセットに過ぎず，ここでは実装上のメリットがないため割愛する．

また，Φ_B は n_K 個の基底サンプルを列ベクトルにとり並べた行列であり，

$$\Phi_B = \left(\phi(\boldsymbol{x}_1) \quad \phi(\boldsymbol{x}_2) \quad \ldots \quad \phi(\boldsymbol{x}_n)\right), \tag{25}$$

である．

これを固有値問題の式に代入すると，

$$\frac{1}{N}(\Phi - \Phi 1_N)(\Phi - \Phi 1_N)^T \Phi_B \boldsymbol{\beta} = \lambda \Phi_B \boldsymbol{\beta}. \tag{26}$$

両辺に左から Φ_B^T をかけると，

$$\frac{1}{N}\Phi_B^T(\Phi - \Phi 1_N)(\Phi - \Phi 1_N)^T \Phi_B \boldsymbol{\beta} = \lambda \Phi_B^T \Phi_B \boldsymbol{\beta}. \tag{27}$$

となる．

ここで，カーネルトリックにより以下のような置き換えが可能である．

$$\begin{aligned} \Phi_B^T(\Phi - \Phi 1_N) &= \Phi_B^T \Phi - \Phi_B^T \Phi 1_N && (28) \\ &= K' - K' 1_N, && (29) \\ \Phi_B^T \Phi_B &= K_B. && (30) \end{aligned}$$

ここで $K' \in \Re^{n \times N}$ は学習サンプルのカーネルベースを並べた行列，$K_B \in \Re^{n \times n}$ は基底サンプルのグラム行列を示す．

最終的に，以下の一般化固有値問題として定式化できる．

$$(K' - K'1_N)(K' - K'1_N)^T \boldsymbol{\beta} = \lambda N K_B \boldsymbol{\beta}. \tag{31}$$

固有ベクトルに必要な規格化条件は次のようになる．

$$\boldsymbol{\beta}^T K_B \boldsymbol{\beta} = 1. \tag{32}$$

Appendix C: HLAC特徴の詳細

ここでは，本研究で用いるカラー高次局所自己相関特徴 (Color-HLAC 特徴)[97] の説明を行う．これは，グレースケール画像において定義される大域的画像特徴量である高次局所自己相関特徴 (HLAC 特徴)[149] をカラー画像へ拡張したものである．画像ごとに固有の表現であるため，bag-of-visual-words [40] のようにタスクに応じた前処理（ベクトル量子化）を必要としない．従って，システムが扱う問題の規模・質が変化した際も再計算が不要である．また，高速に抽出可能であるため，本研究で目的とする大規模なシステムにふさわしいものと言える．更に，特徴が位置不変性・加法性を有するため，画像中の物体の位置・数が未知である弱ラベリング問題を扱うのに適している．

Color-HLAC 特徴は，画像の局所的なピクセルパターンを全体で積分して得られる特徴である．図 3に高々1 次までの自己相関を考慮した場合の Color-HLAC 特徴のマスクパターンを示す．本論文では，高々2 次までの自己相関を用いる．また，スケール変化に対する耐性を向上させるため，元画像と 1/2 サイズに縮小した画像の両方から特徴抽出を行う．さらに，それぞれのサイズについて，前処理として Sobel フィルタによりエッジ抽出を行った画像からも特徴抽出を行う．これは，照明変動などの影響に頑健にすることが目的である．元のサイズの通常画像から抽出した Color-HLAC 特徴を $\boldsymbol{x}_{o_1}$，エッジ画像から抽出した Color-HLAC 特徴を $\boldsymbol{x}_{e_1}$ と表す．また，1/2 サイズに縮小後に抽出したものをそれぞれ $\boldsymbol{x}_{o_{1/2}}$, $\boldsymbol{x}_{e_{1/2}}$ と表す．これらを並べ，$\boldsymbol{x} = (\boldsymbol{x}_{o_1}^T, \boldsymbol{x}_{o_{1/2}}^T, \boldsymbol{x}_{e_1}^T, \boldsymbol{x}_{e_{1/2}}^T)^T$ を最終的な画像特徴とする．合計で，$739 \times 4 = 2956$ 次元のベクトルとなる．本論文で特にことわりなく HLAC 特徴と書く時は，このベクトルをさすものとする．

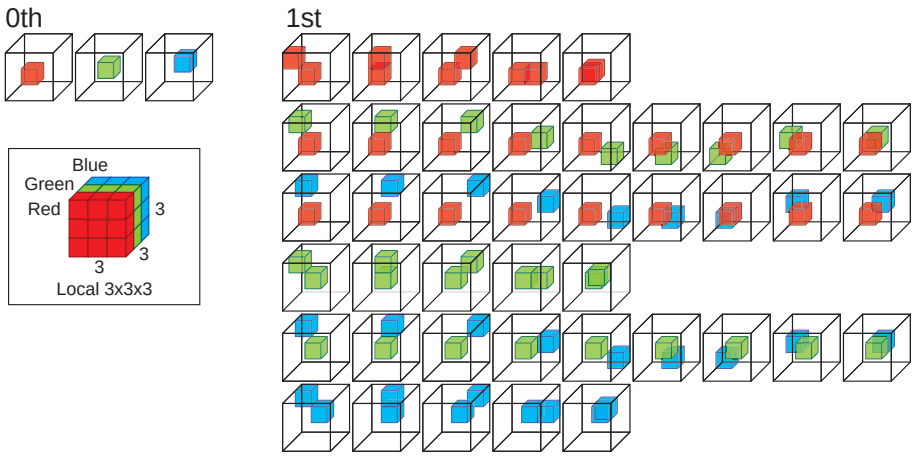


Figure 3: Mask patterns of at most the first order Color-HLAC features.

Appendix D: Flickr12Mにおける実験データ

ここでは，7.2 節の実験結果の一部をまとめる．Flickr12M のサブセットにおいて，各画像特徴量を用いた際のアノテーション精度を以下にそれぞれ示す．

- 図 4, 5: Tiny image
- 図 6, 7: RGB color histogram
- 図 8, 9: GIST
- 図 10, 11: HLAC
- 図 12, 13: SURF GLC
- 図 14, 15: SURF BoVW
- 図 16, 17: SURF BoVW-sqrt
- 図 18, 19: RGB-SURF GLC

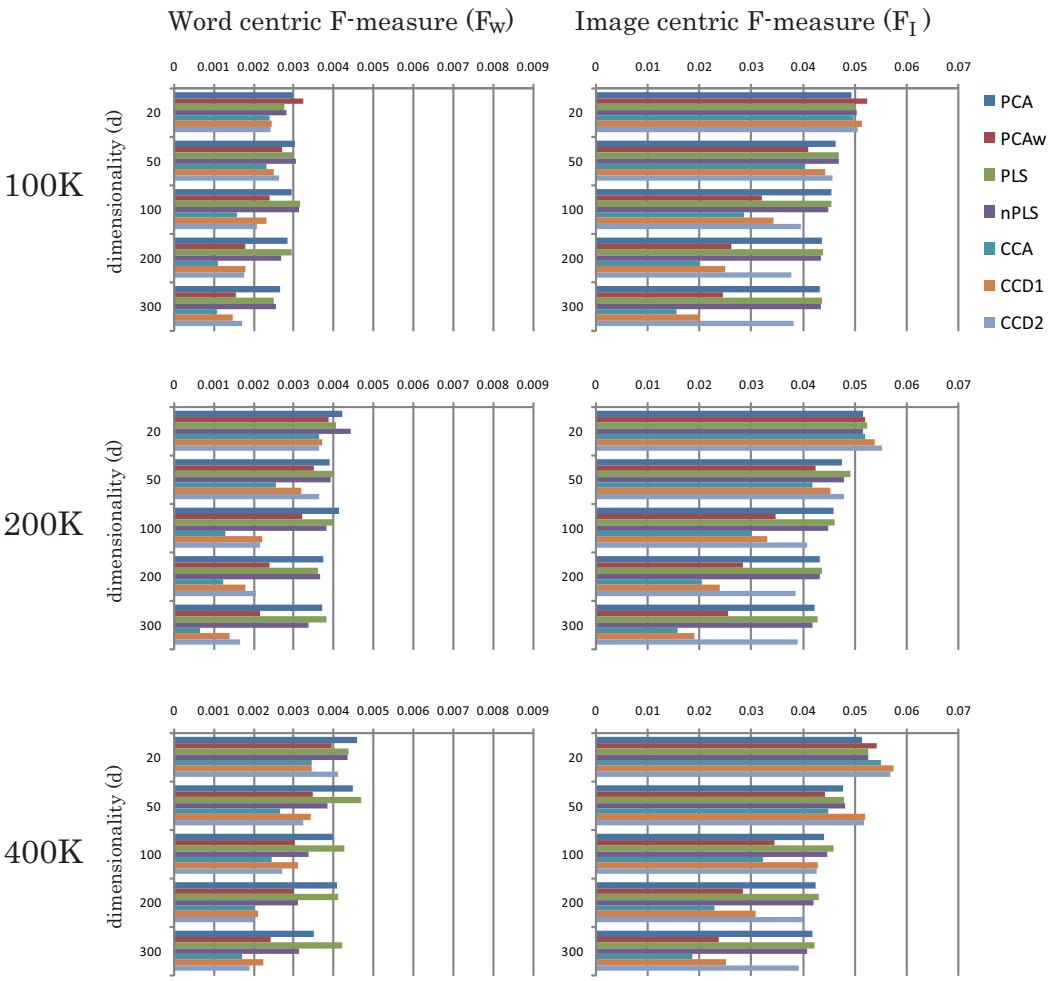


Figure 4: F-measures of **Tiny image** features for the 100K, 200K, and 400K subsets.

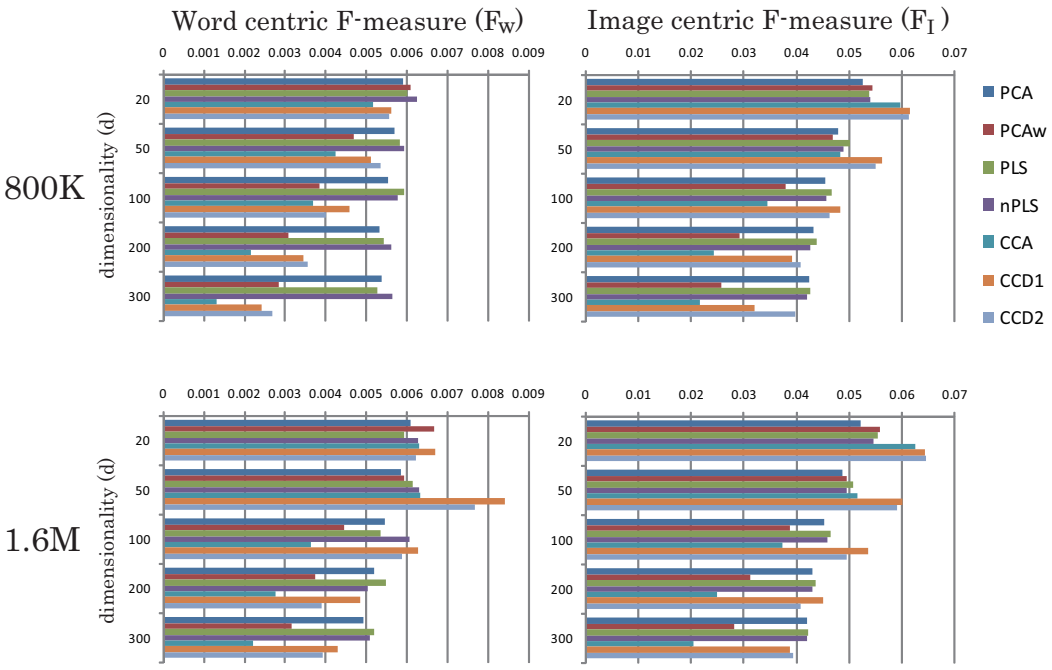


Figure 5: F-measures of **Tiny image** features for the 800K and 1.6M subsets.

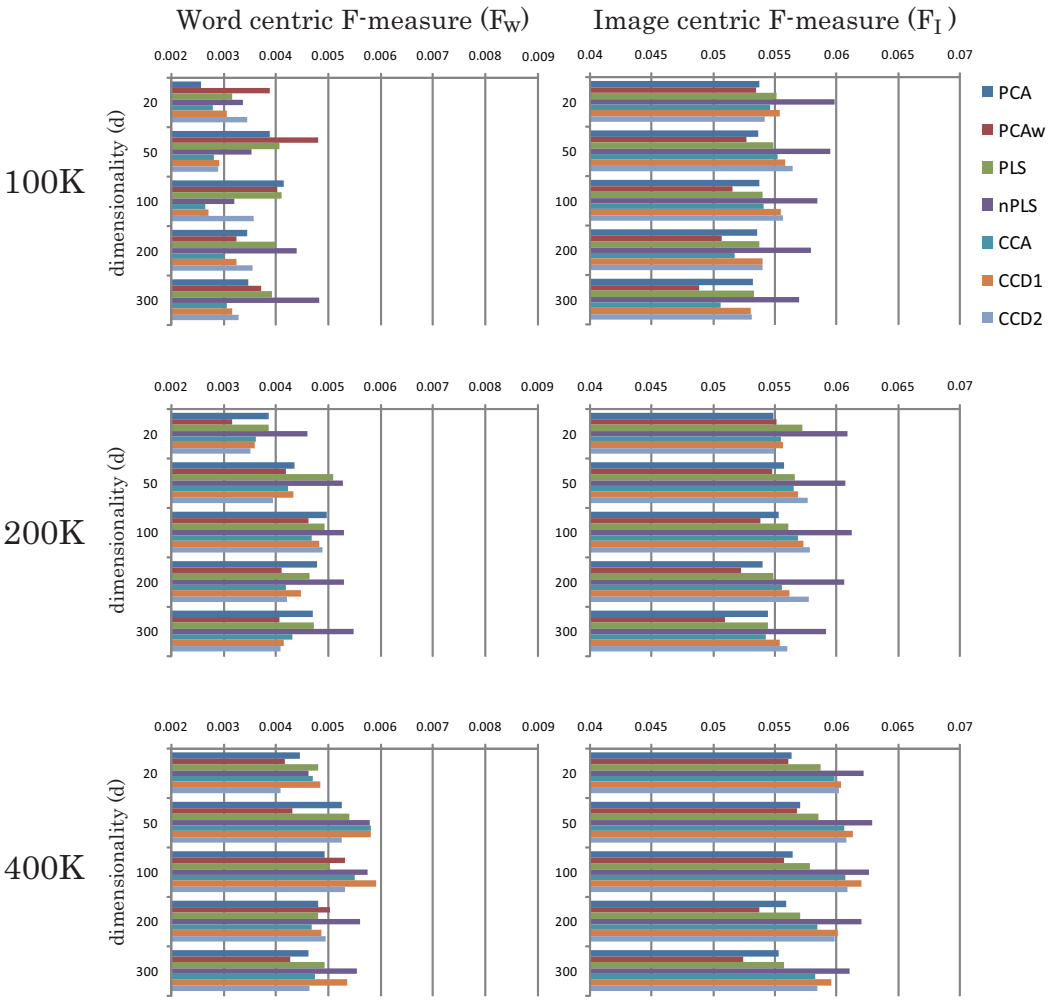


Figure 6: F-measures of the **RGB color histogram** for the 100K, 200K, and 400K subsets.

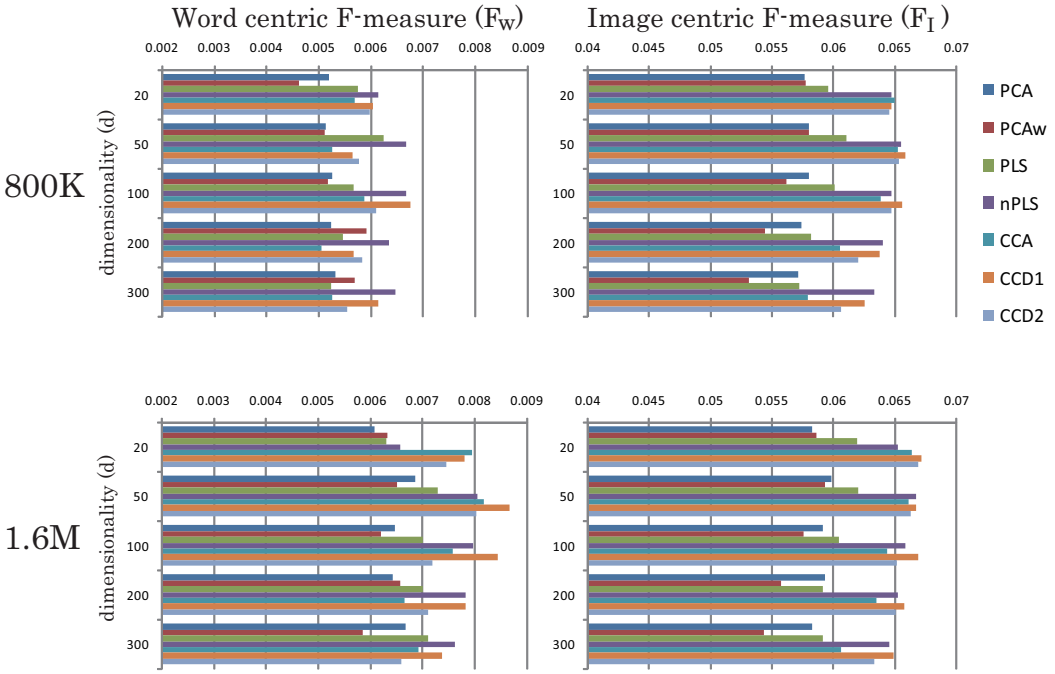


Figure 7: F-measures of the **RGB color histogram** for the 800K and 1.6M subsets.

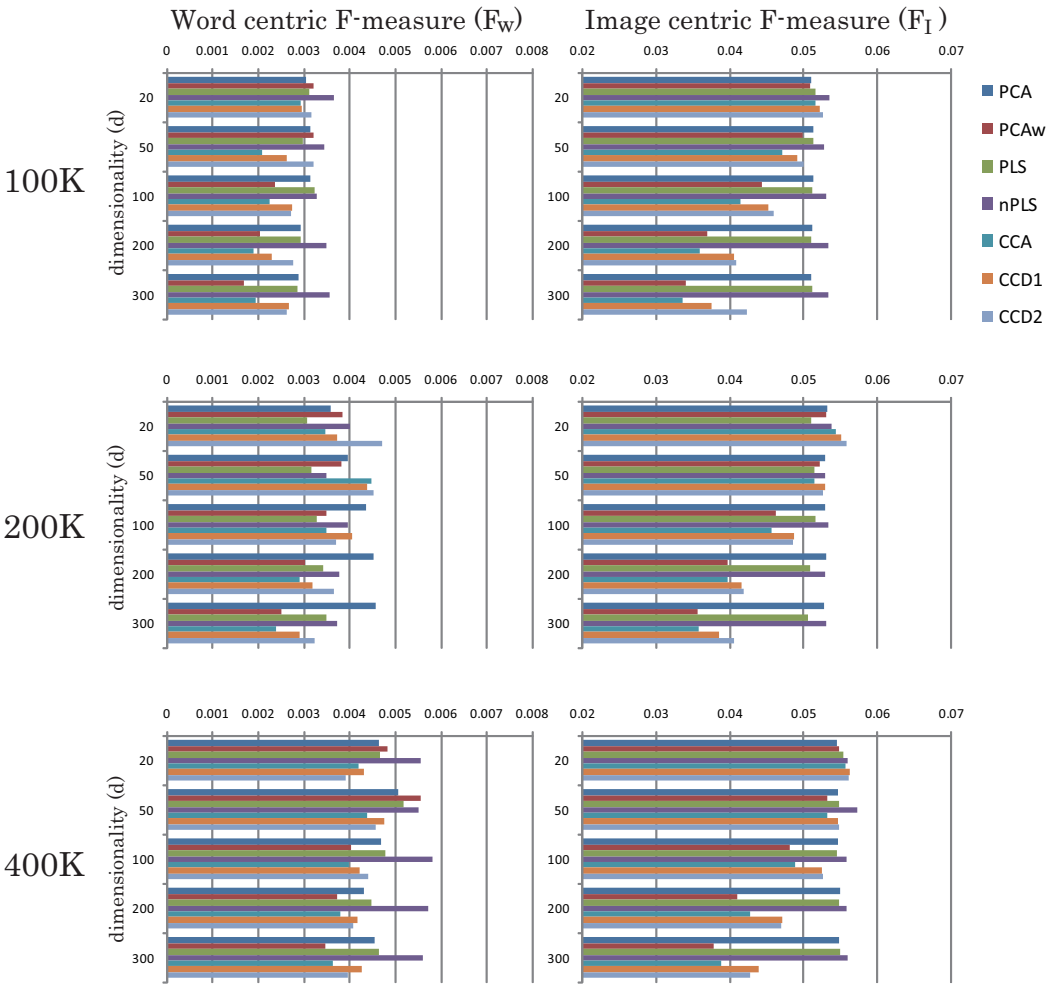


Figure 8: F-measures of **GIST** features for the 100K, 200K, and 400K subsets.

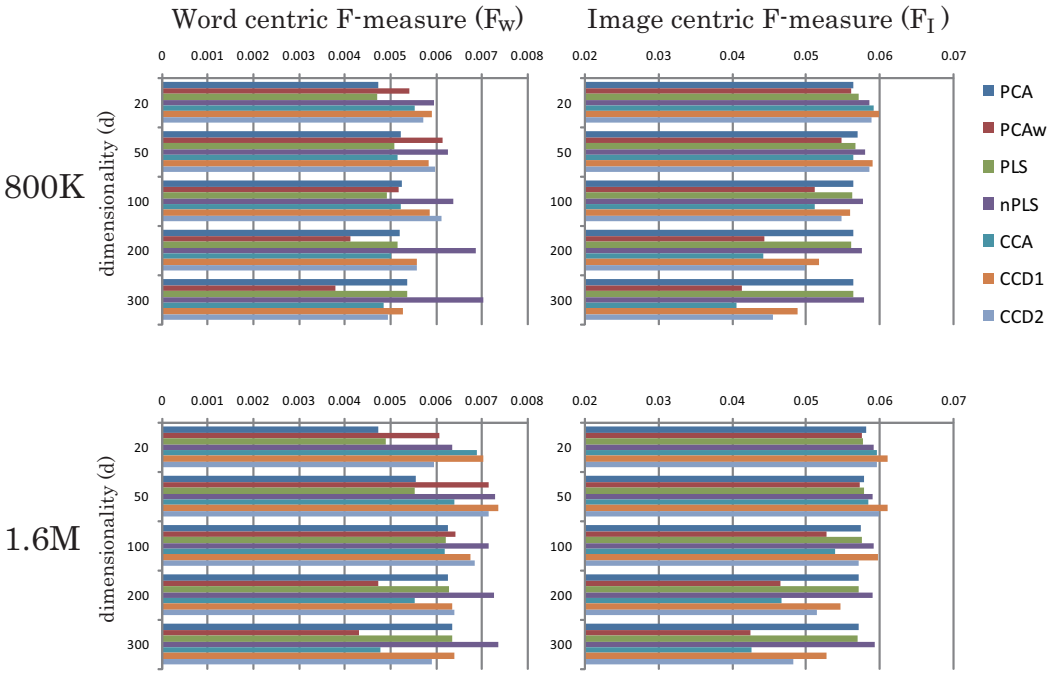


Figure 9: F-measures of **GIST** features for the 800K and 1.6M subsets.

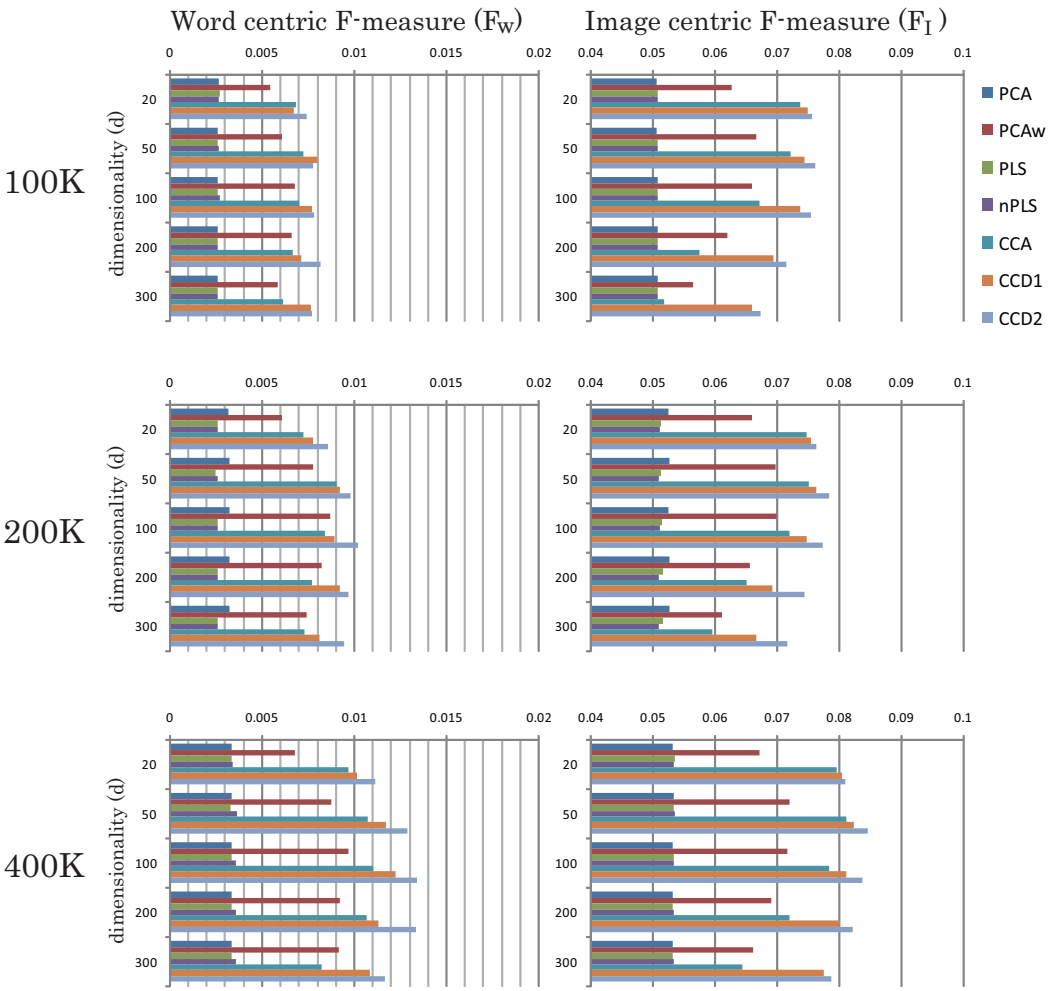


Figure 10: F-measures of **HLAC** features for the 100K, 200K, and 400K subsets.

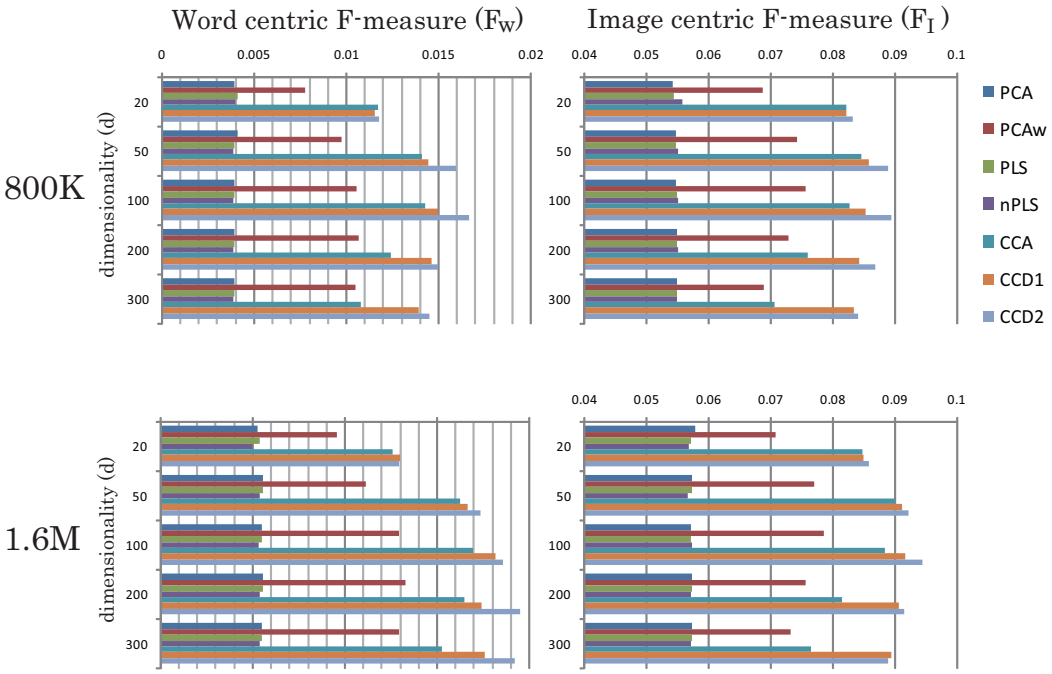


Figure 11: F-measures of **HLAC** features for the 800K and 1.6M subsets.

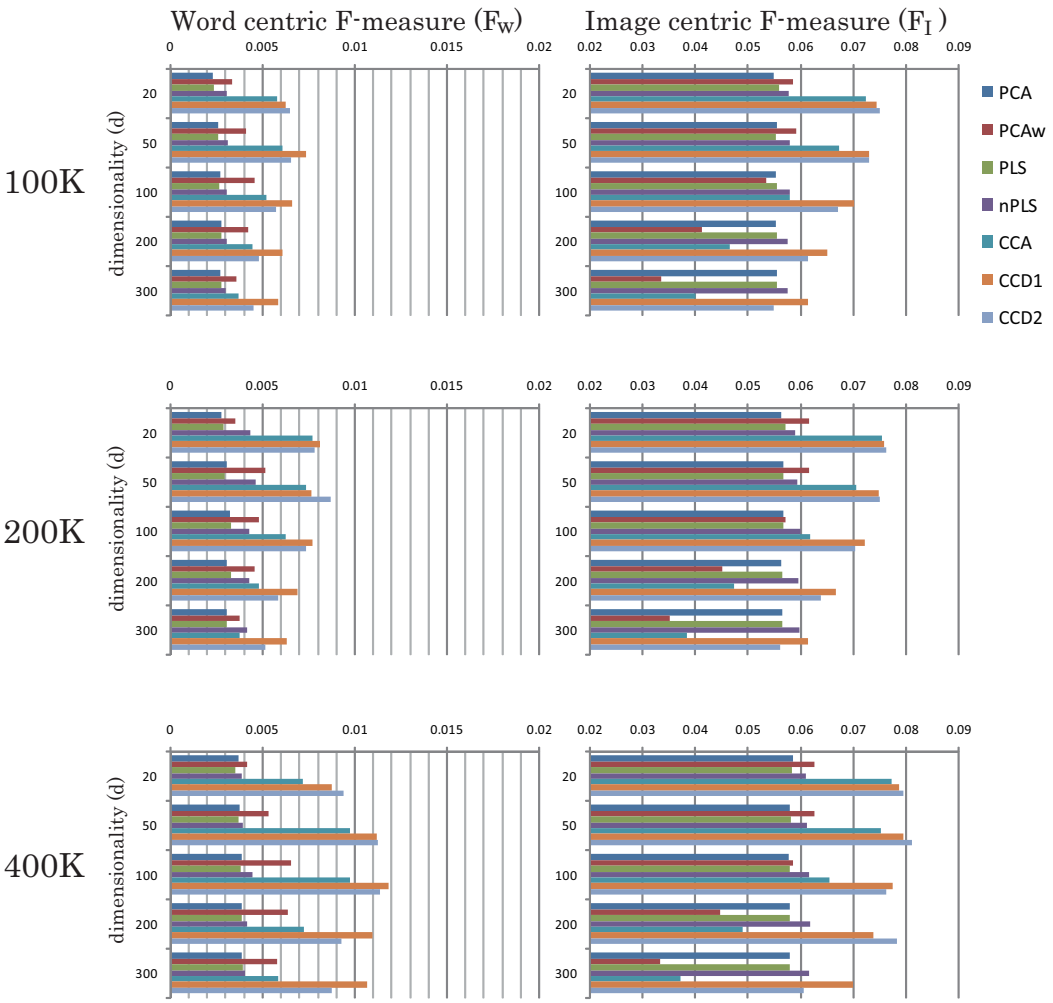


Figure 12: F-measures of **SURF GLC** features for the 100K, 200K, and 400K subsets.

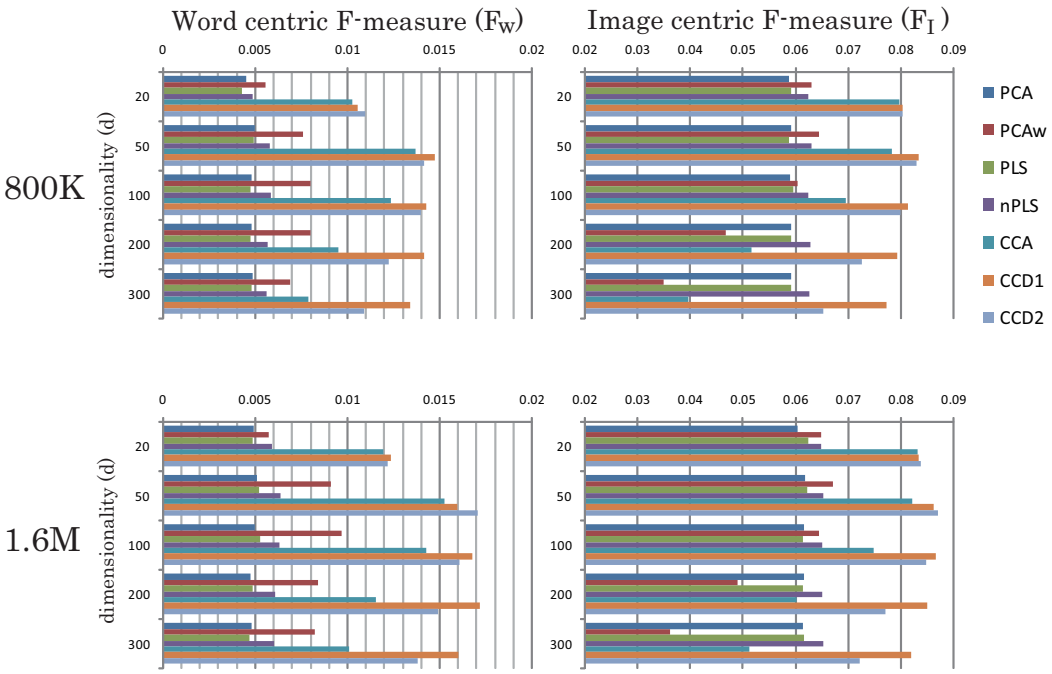


Figure 13: F-measures of **SURF GLC** features for the 800K and 1.6M subsets.

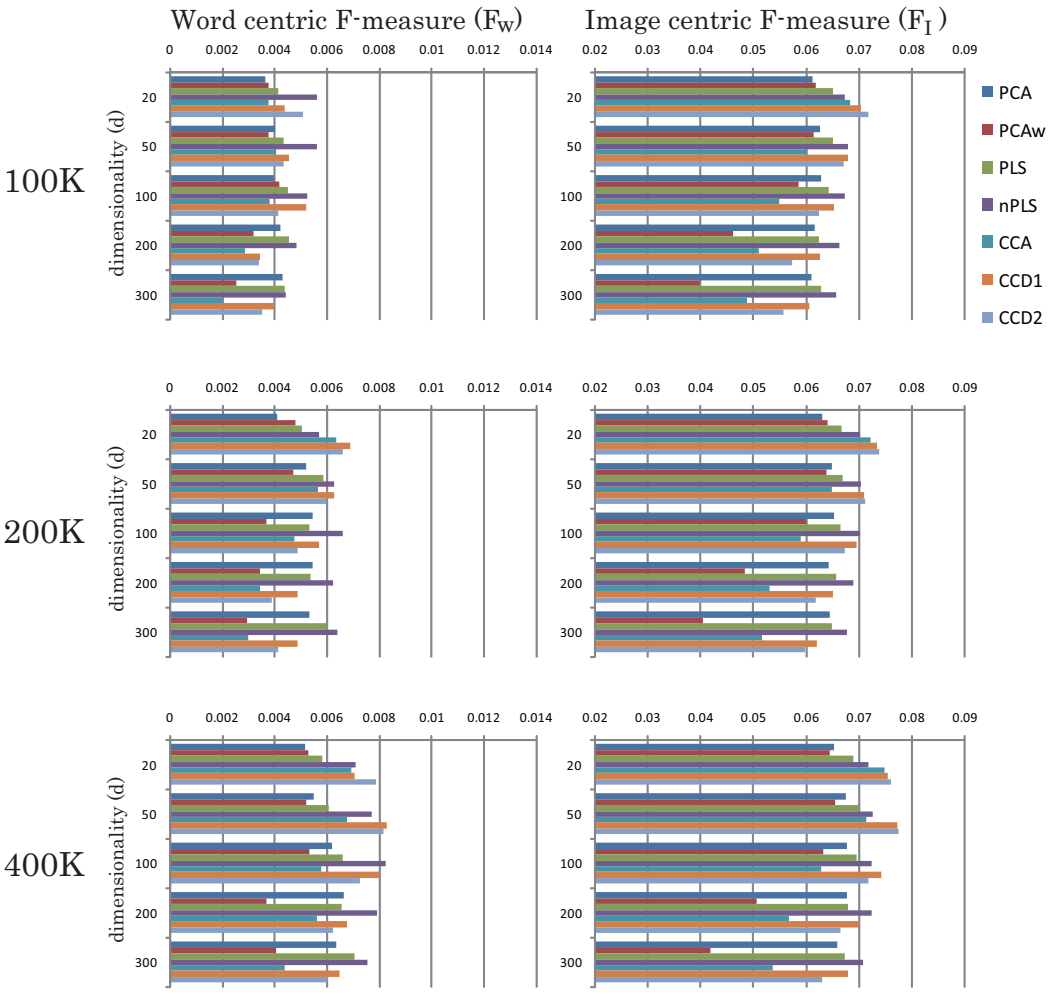


Figure 14: F-measures of **BoVW** features for the 100K, 200K, and 400K subsets.

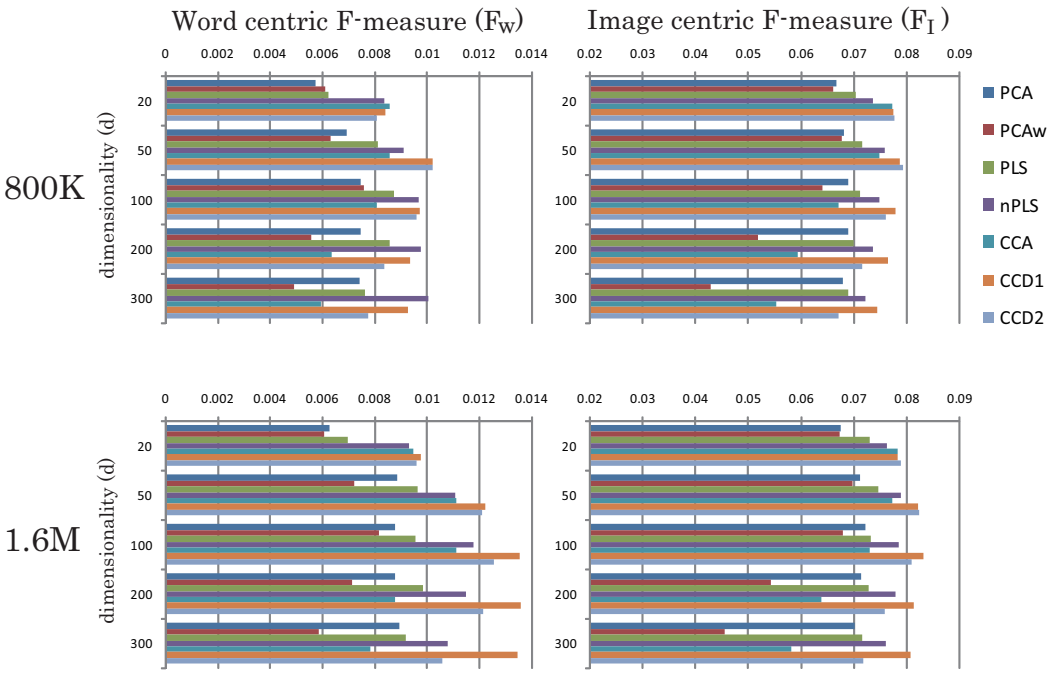


Figure 15: F-measures of **BoVW** features for the 800K and 1.6M subsets.

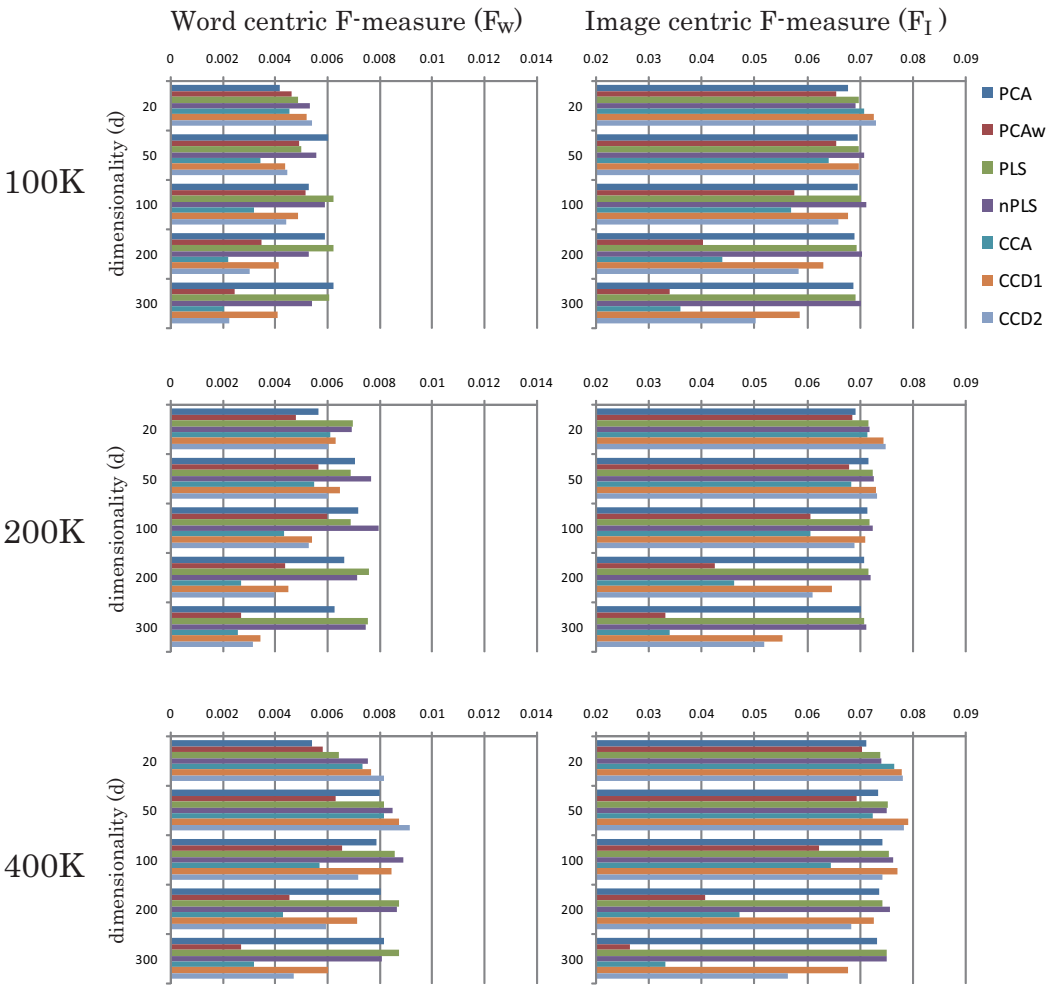


Figure 16: F-measures of **BoVW-sqrt** features for the 100K, 200K, and 400K subsets.

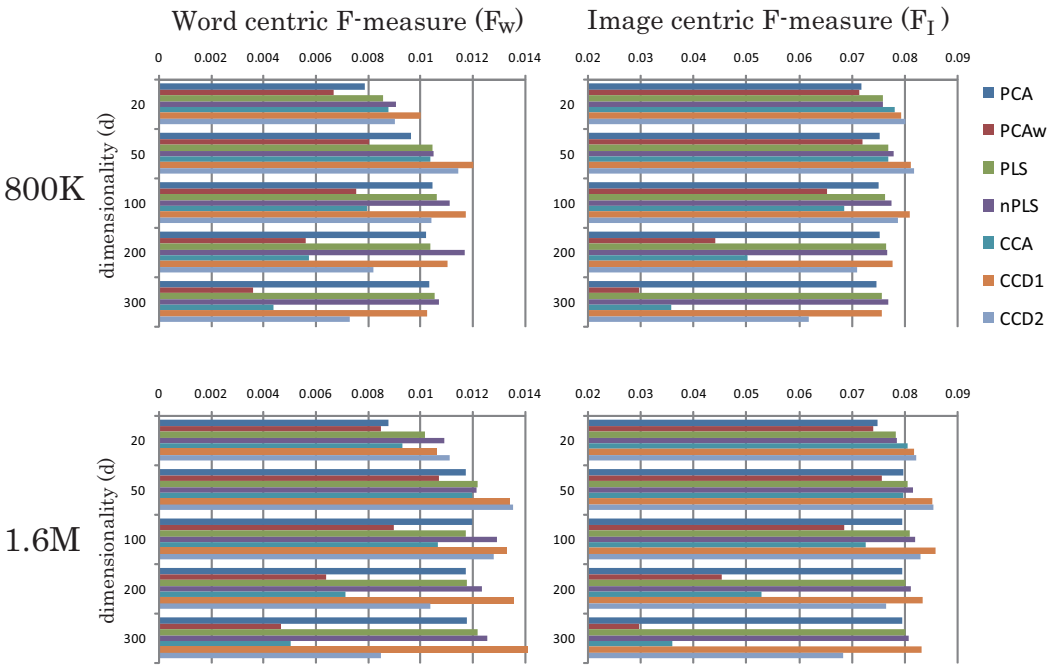


Figure 17: F-measures of **BoVW-sqrt** features for the 800K and 1.6M subsets.

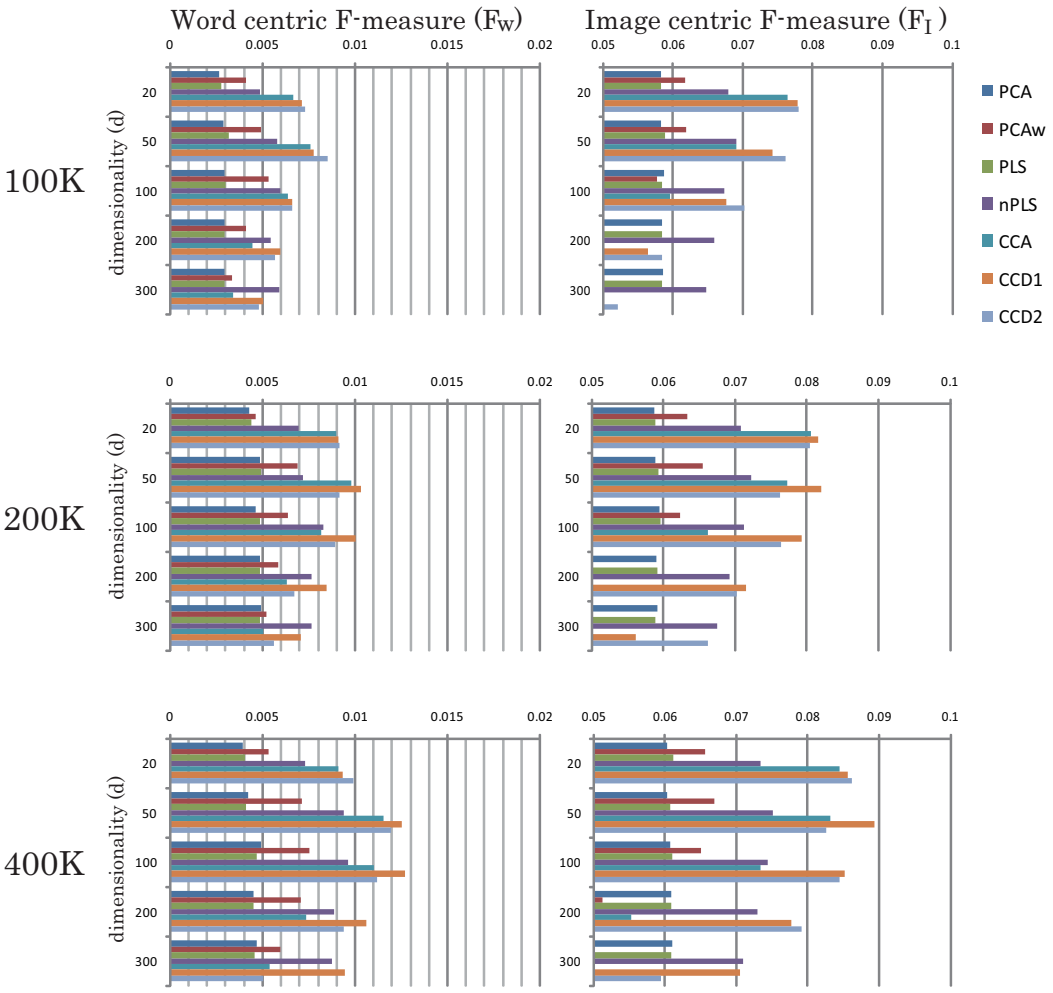


Figure 18: F-measures of **RGB-SURF GLC** features for the 100K, 200K, and 400K subsets.

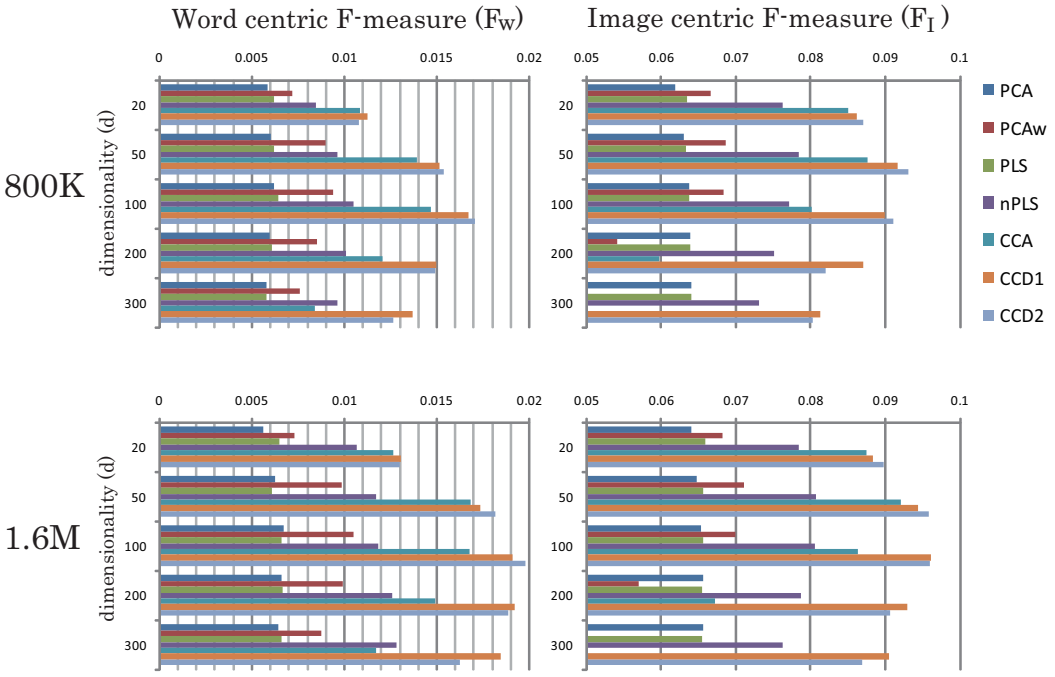


Figure 19: F-measures of **RGB-SURF GLC** features for the 800K and 1.6M subsets.

Appendix E: ハッシングに基づくアノテーションの高速化

本論文で提案する CCD と標準的なハッシング手法を組み合わせることにより，アノテーションの更なる高速化を行う．各画像サンプルは小さなバイナリコードによって表現されるため，数千万枚のサンプルを PC 一台のメモリ上に格納することが可能となる．本付録では，まず contents-based image retrieval (CBIR) の文脈で近似最近傍探索手法の開発を行う．CBIR は，ノンパラメトリックな画像アノテーションにおいても核となるプロセスである．最後に，開発した手法を画像アノテーションへ応用する．

E.1. 概要

CBIR は長い間研究が続けられてきており，近年では商用レベルへ近づきつつある．しかしながら，数百万枚・数千万枚の大規模な画像データベースから適切な画像を検索することは依然として困難な課題である．CBIR の難しさは主に二つの問題に起因する．第一に，画像特徴量は一般的に高次元であることである．大規模な問題においては，単純な線形探索は計算量・メモリ使用量とも膨大になり，実行困難である．従って，大量の高次元データを扱うためには，何らかの効率的な検索アルゴリズムが必要である．しかしながら，一般に高次元空間においては，どのような手法を用いても短時間で近傍探索を行うことは非常に困難である．これは，“次元の呪い”として知られている問題でる．

第二の問題は，low-level な画像特徴と高次の意味の間には隔たりがあるという，いわゆる semantic gap の問題である (2.2.2 節)．基本的に，最近傍探索の高速化に関する研究は，unsupervised な状況を考える．すなわち，元の画像特徴空間におけるユークリッド距離を近似したコンパクトな表現の導出を行う．これらの方法は，near-duplicated image など，視覚的に類似した画像を検索するのに有効である一方，画像間の意味的な近さを測ることには必ずしも向いていない．Semantic gap を緩和するためには，テキストなど画像と対になるモダリティを機械学習の枠組みで利用する，supervised なセッティングを考えることが有効である．しかしながら，一般にこのような手法の学習コストは大きく，本研究が考慮するような大規模なデータベースにおいて利用することは容易ではない．

本付録では，本論文で提案する CCD の枠組みを応用することで，効率よく supervised な近似最近傍探索を行う．CCD2 (4.2.3 節) では，以下のようなマルチモーダルなセットアップを扱う．画像特徴量 $\boldsymbol{x} \in \mathcal{R}^p$ とテキスト特徴 $\boldsymbol{y} \in \mathcal{R}^q$ を，N 個のサンプルからそれぞれ抽出し，検索用のデータベース $\mathcal{T} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$ を得る．クエリとなる入力画像 $\boldsymbol{x}_Q$ に最も類似したサンプルをデータベースの中から検索することを考える．すなわち，

$$\mathrm{NN}(\boldsymbol{x}_Q) = \arg\min_{\boldsymbol{x} \in \mathcal{T}} D(\boldsymbol{x}_Q, \{\boldsymbol{x}, \boldsymbol{y}\}), \tag{33}$$

ここで, D は 2 つのインスタンスの間の距離を表す．これは，近年の Web データの指数的な増加を踏まえれば，適切なセットアップであるといえる．例えば，携帯電話で撮影した画像をクエリとし，Web ドキュメント内の類似画像を検索することなどが可能となる．

検索のための距離計算や，メモリ使用量を削減するために，画像・テキスト両方の類似度を埋め込んだ，小さなバイナリコードを学習する．D は，コード間のハミング距離によって計算される．データベース中の各サンプルは数バイトから数十バイト程度で表現されるため，線形探索を用いても高速な検索が可能となる．Flickr12M を用いた実験により，大規模データベースにおける提案手法の有効性を示す．

E.2. 関連研究

初期の研究では，厳密な最近傍探索の高速化が試みられてきた．Bentley により発案された kd-tree [12] をはじめとする木探索アルゴリズムはその代表例であり，二分探索による枝刈りを利用することにより，低次元のデータにおける高速な厳密最近傍探索を実現した．しかしながら，このような二分探索に基づくアルゴリズムは，高次元のデータにおいては有効に機能せず，線形探索と同程度の計算コストを要することがその後の研究により明らかとなった [111; 205]．高次元空間における厳密最近傍探索の高速化は現在に至るまで未解決の課題である．

このため，近年では近似最近傍探索のアプローチが多数を占めるようになってきている．これは，厳密な最近傍でなくとも，十分に高い確率で最近傍であると予想されるサンプルが検索できればよしとする枠組みであり，検索精度と計算コストの現実的なトレードオフを実現することを目指すものである．この考え方は，Indyk らによる Euclidean locality-sensitive hashing (E2LSH) [42; 86] において初めて提案された．E2LSH は，ランダム射影を用い，類似するサンプルが高い確率で衝突するようなハッシュ関数を多数構築する．確率的なアプローチをとることにより，理論的に保証された形で検索精度と速度のトレードオフを議論することが可能であることを示した．E2LSH はサンプル間類似度がユークリッド距離で測られることを前提としているが，LSH はその後任意のマハラノビス距離や非線形距離を扱えるように拡張が行われている [100; 101]．

LSH は様々な高次元データにおいて高速な近似最近傍探索を実現し，コンピュータビジョンにおいても応用が多く行われている．しかしながら，問題がいくつか存在する．標準的な LSH で得られるものはあくまでハッシュ関数群であり，クエリがヒットしたバケットにサンプルが存在しない場合，隣接バケットを探索する別の方策が必要となる．さらに，バケット内に候補サンプルが存在する場合でも，最終的な最近傍の決定が必要なタスクの場合には元のベクトルを用いた距離計算が必要となる．従って，任意の入力に対し高速に近似最近傍を決定するためには，元の空間におけるサンプル点を全てメモリ上に展開しておく必要が生じる．これは，我々が考慮するスケールの問題においては非現実的である．このため，機械学習の手法を用い，ハッシュ関数の生成と特徴ベクトルの圧縮を同時に行うアプローチが注目を浴びている [164; 183; 208]．これは，元の空間におけるサンプル間距離（ユークリッド距離）を，ハミング距離により近似するバイナリコードへの射影を学習するものである．十分低次元のバイナリコードへ圧縮を行えば，これを直接ハッシュ値として用いることが可能である．また，ハミング距離がサンプル間類似度を保存しているため，比較的少量のメモリで多くのサンプルを保持することができ，類似度計算も高速に行うことが可能である．例えば，spectral hashing [208] は，一様分布のグラフラプラシアンのスペクトル分解を利用し，教師なし学習の枠組みでバイナリコードの学習を行う．この場合，ハッシュ関数は解析解により得られるため，極めて高速にハミングコードの学習を行うことができる．前提として，データの分布に一様分布を仮定するという強い制約を置いており，実装もデータの主成分軸を段階的に分割していくシンプルなものであるが，経験的に少ないビット数で非常によい検索精度が得られることが示されている．

学習に基づくハッシングの枠組みでは，サンプルに与えられる教師情報 (ラベル）を自然な形で利用することができる．例えば，AdaBoost [63] を利用した BoostSSC [169] が先駆的な例として挙げられる．また，Torralba ら [183] は，情報検索の分野で開発された restricted Boltzmann machine (RBM) [79; 164] を画像検索へ応用し，LSH や BoostSSC よりも精度よく GIST 特徴 [148] の圧縮が行えることを示した．RBM におけるハッシュの作成では，教師付き・教師なし両方の枠組みをとることが可能である．教師付き学習の際に必要となるバックプロパゲーションには，neighborhood components analysis (NCA) [68] の目的関数を用いている．しかしながら，RBM の学習コストは非常に高いことが指摘されており，動的なデータベースへの適用は困難であるといえる．バイナリコード学習によるハッシングのアプローチはその後も盛んに研究されており，半教師付き学習 [119] や，コードのオンライン学習 [200] を行うものが提案されている．また，Jégou ら [90] は，product quantization によりバイナリコードの学習を行い，メモリ使用量・検索精度のトレードオフにおいて spectral hashing を上回る性能を得られることを報告している．

バイナリコード学習の他にも，近似最近傍探索の手法は多く提案されている．例えば，optimized kd-tree [175] や hierarchical k-means [146] などは，データの分布をヒューリスティックに利用することで，古典的な手法の応用により LSH を上回る探索精度を得ている．また，一般的に画像特徴量は非常に高次元となる場合

が多いため，画像データの効率のよい表現方法自体が重要な課題となる [49; 89]. 近年の研究では，画像表現と後段の近似最近傍探索の両方を考慮した設計により，bag-of-visual-words による画像検索を大幅に高速化・省メモリ化した例が報告されている [91].

E.3. 提案手法

CCD の枠組みにより，画像とテキストは既に，semantic gap を緩和した低次元のユークリッド空間（潜在空間）に埋め込まれている．さらに潜在空間上で，ユークリッド空間において定義される標準的なハッシング手法を利用し，バイナリコードの学習を行う．これは，PCCA におけるトピックレベルの類似度を近似したコードと解釈できる．検索時には，クエリは同様にバイナリコードへ変換され，ハミング距離により高速に類似度計算が行われる．さらに，コードを 30 ビット程度まで圧縮できれば，これを直接ハッシュテーブルとして用いることが可能となるため，データ数に関わらず類似サンプルを極めて短時間で検索することができる [183].

4.2.3 節で述べたように，潜在空間における KL ダイバージェンスは式 4.19と式 4.20によって定義される $\boldsymbol{r}$ のユークリッド距離によって計算できる．ここでは，$\boldsymbol{r}$ をトピック特徴と呼ぶことにする．$\boldsymbol{r}$ に以下で述べるハッシング手法を適用し，c ビットのバイナリコードへ変換する．トピック特徴は定義より zero mean となっていることに注意されたい．

Simple Binarization

ベースラインとして，トピック特徴の各次元を正負により二値化しこれをハッシュ関数として用いる場合を考える．この場合，コードのビット数 c は CCD の次元数 d となる．

Locality Sensitive Hashing (LSH)

LSH のハッシュ関数は，特徴空間のランダム射影によって学習される．バイナリコードの学習を行う場合，次のようなハッシュ関数により各ビットをコーディングする．

$$h(\boldsymbol{r}) = sign(\boldsymbol{w}^T \boldsymbol{r} + b), \tag{34}$$

ここで，$\boldsymbol{w}$ は各要素が p-stable distribution （ここではガウス分布）から独立にサンプリングされたベクトルである．また，b は一様乱数により得られるオフセットである．このようなハッシュ関数をランダムに c 個生成し，c ビットのバイナリコードを得る．なお，本研究では経験的に $b = 0$ と固定した場合に最もよい性能が得られたため，これに固定する．また，超球状のサンプルに対しては $b = 0$ とした場合に最もバランスのとれたハッシュ関数が得られることが指摘されている [35; 119].

Spectral Hashing (SH)

N 個のサンプルの類似度行列 (affinity matrix) を $W \in \Re^{N\times N}$ とする．ただし，$W(i,j) = \exp(-||\boldsymbol{r}_i - \boldsymbol{r}_j||^2/\epsilon^2)$ とする．ビット列へ圧縮されたサンプルを $\{\boldsymbol{h}_i\}_{i=1}^N$ とする．ここで，$\boldsymbol{h}$ は c ビットのバイナリベクトルである．$\boldsymbol{h}$ のハミング距離が，$\boldsymbol{r}$ のユークリッド距離をできるだけ近似するように学習を行いたい．この問題は，以下のように定式化できる．

$$\begin{aligned} \text{minimize:} \quad & \sum_{ij} W_{ij}||\boldsymbol{h}_i - \boldsymbol{h}_j||^2, \\ \text{subject to:} \quad & \boldsymbol{h}_i \in \{-1,1\}^c, \quad \sum_i \boldsymbol{h}_i = 0, \\ & \frac{1}{N}\sum_i \boldsymbol{h}_i \boldsymbol{h}_i^T = I. \end{aligned} \tag{35}$$

上述の問題は，$c = 1$ の場合であっても NP 困難であり [208]，$c > 1$ の一般の場合は更に困難となる．しかしながら，拘束条件を一部緩和することで，グラフスペクトル分析により簡単に解を得ることができる．

もちろん，実際のデータは一様分布に従うことは考えにくいが，[208] ではあらかじめ主成分空間に射影を行い無相関化を行うだけで，実験的によい性能が得られることを報告している．PCA と CCA は近い関係にあり，無相関の前提はトピック特徴にも成り立つため，提案手法では直接トピック特徴のハッシングを行う．

E.4. 画像検索実験

本実験では，二つのデータセットを用いる．一つ目は，LabelMe [163] である．LabelMe の画像は，人手により画像中の物体が切り出されており，それぞれ物体の名称が与えられている．ここでは，物体のラベルのみ用い，位置情報は無視する．本研究では，一般に公開されているデータのうち，60,000 枚を検索対象のサンプル，1,191 枚をクエリとして用いる．二つ目は，7.1 節で用いた Flickr12M である．これは，Flickr からダウンロードした画像とソーシャルタグで構成される大規模データセットであり，1230 万枚の画像と 4,130 種類の単語からなる．ここでは，5,000 枚のテスト画像をクエリに用いる．

テキスト特徴 ($\boldsymbol{y}$) としては，単語ヒストグラムを用いる．CCD の実装では，実験的に最もよい d（潜在空間の次元数）を決定する．

LabelMe 画像検索

ここでは，画像特徴量として GIST 特徴 [148] を用いる．グラウンドツルースの最近傍は，ラベルヒストグラム間の χ^2 距離によって定義する．各検索アルゴリ

ズムを用い，60,000 枚のデータから検索された上位 n 画像のうち，真の最近傍 50 サンプルがどれだけ含まれるか (recall) によってその性能を評価数する．

図 20に，検索された上位 5000 画像における recall の値を，コードに用いるビット数を変化させながらプロットする．また，図 21に，検索画像数 n に対する recall の変化を示す．全体として，CCD に基づくハッシング手法は unsupervised な手法を大きく上回る検索精度を示しており，数十ビット程度のコードで元の GIST と同程度のスコアとなっている．特に，CCD+LSH はビット数の増加に伴い一貫して精度が向上している．図 22に，いくつかのクエリ画像と検索された近傍画像の例を示す．ビット数が増えるにつれて，提案手法は見た目が類似している画像のみならず，意味的に類似している画像を検索することが可能となる．

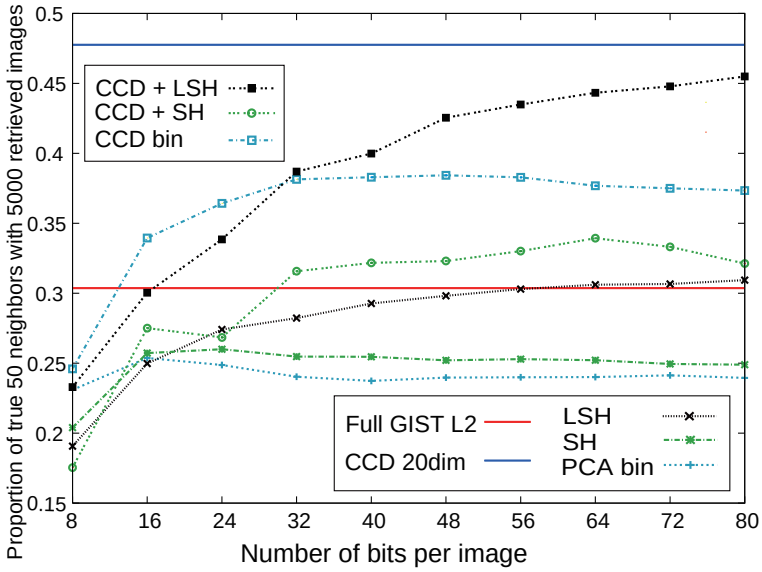


Figure 20: Retrieval performance with a varying number of bits for the LabelMe dataset.

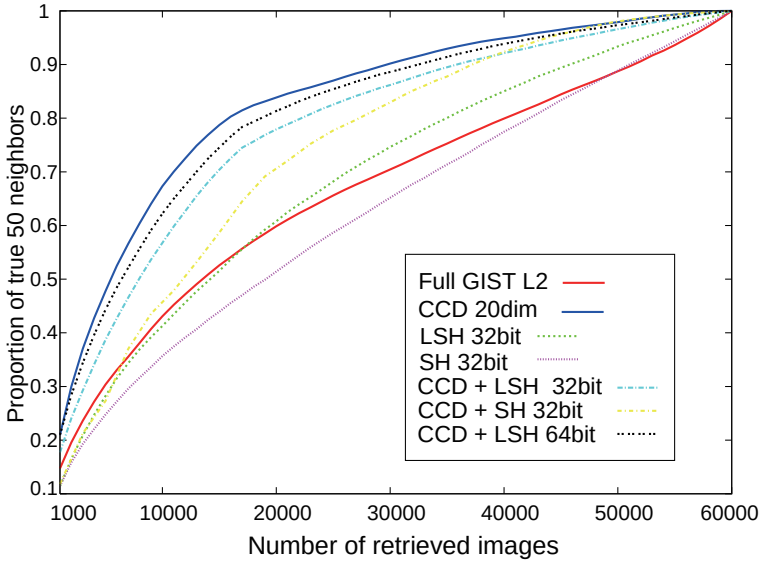


Figure 21: Retrieval performance as a function of retrieved images for the LabelMe dataset.

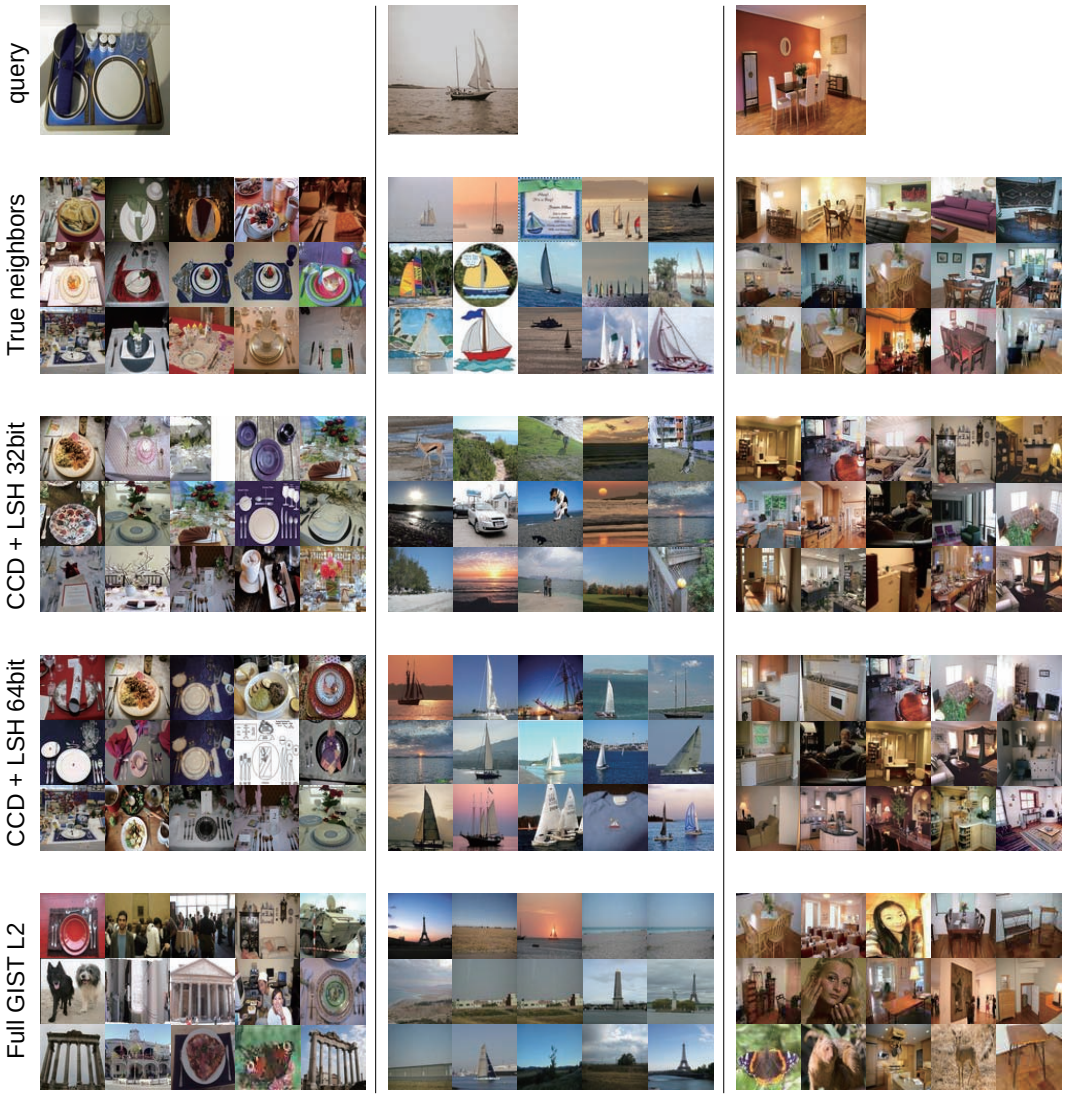


Figure 22: Examples of retrieved images (15 neighbors) for the LabelMe dataset.

Flickr 画像検索

Flickr12M では単語ヒストグラムは 2 値であるため，真の最近傍を以下のように定義する．まず，データベース中の各画像を，クエリ画像と共通する単語数の降順に並び替える．さらに，一致しない単語数の昇順に並び替え，クエリ画像に対するランクとする．画像特徴としては，標準的な bag-of-visual-words (BoVW) [40] を用いる．実装には k-means 法を用い，1000 個の visual word を生成する．BoVW の検索性能の評価には，χ^2 距離を用いる．なお，[155] より，BoVW の Bhattacharyya カーネルは，BoVW の各要素の平方根をとったベクトル (BoVW-sqrt) の線形カーネルと等価であることが示されている．これは，BoVW-sqrt が線形手法にとってより適した表現になっていることを示唆している．そこで，本実験でも BoVW-sqrt をコード学習手法に用いる．図 23，図 24に結果を示す．Flickr12M においても，CCD に基づくハッシング手法は unsupervised な手法を上回るスコアを示している．しかしながら，この大規模なデータセットにおいては，LabelMe の場合よりも多くのビットを必要とすることが分かる．CCD+SH はビット数が少ない場合に相対的に優位であるが，その性能は早い段階で頭打ちとなり，元の BoVW には及ばない．一方，CCD+LSH はビット数に対して安定に性能を向上させており，128 ビット程度で元の BoVW と同程度の検索精度となっている．図 25に定性的な例を示す．

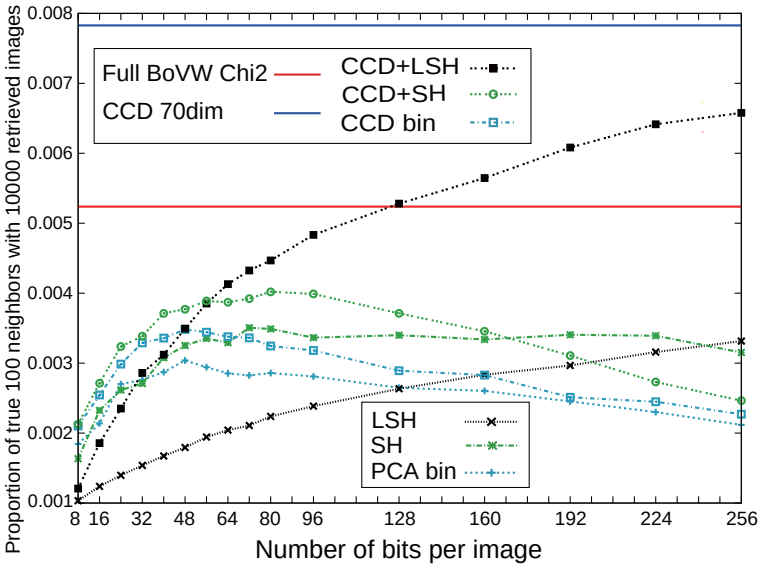


Figure 23: Retrieval performance with a varying number of bits for the Flickr12M dataset.

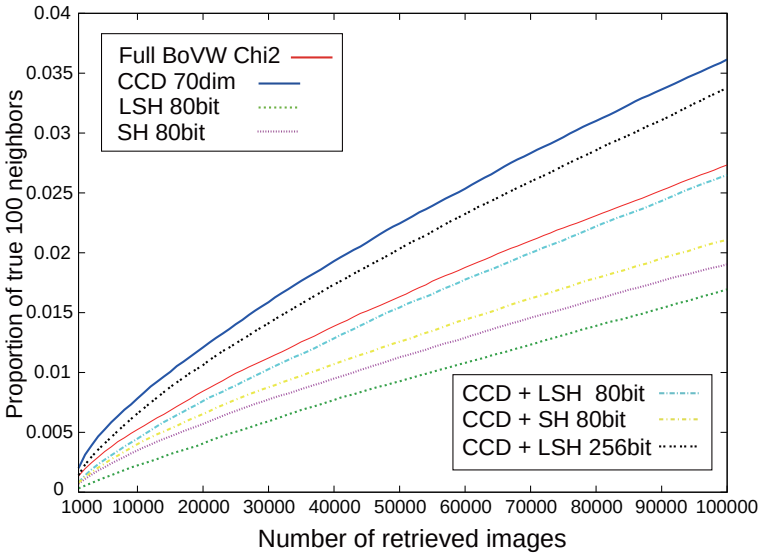


Figure 24: Retrieval performance as a function of retrieved images for the Flickr12M dataset.

Figure 25: Examples of retrieved images (15 neighbors) for the Flickr12M dataset.

Table 1: Retrieval time per image for Flickr12M (s) using a single CPU.

Full BoVW Chi2	219
CCD 70dim	6.2
32 bit code	0.16
80 bit code	0.21
256 bit code	0.44

Table 2: Computation time for training with the Flickr12M dataset using an 8-core desktop machine.

PCA	31m
CCD	4h 31m
SH	5h 7m
LSH	14m
CCD+SH	5h 42m
CCD+LSH	4h 34m

計算時間

表 1に，単一の CPU (3.20GHz) を用い，Flickr12M から最近傍サンプルを検索する際の計算時間を示す．ここでは，クエリの画像特徴抽出にかかる時間は含まない．また，コーディングの計算時間も検索時間と比較して微小であるため無視する．小さなコードを用いることにより，単純な線形探索であるにも関わらず，1,200 万枚の画像を 0.5 秒未満で検索することが可能である．表 2に，8CPU の PC を用いた際の，Flickr12M における各手法の学習時間を示す．CCD に基づく手法は unsupervised な手法に比べやや長い計算時間を要するものの，単一の PC 上で数時間のうちに学習を終えることが可能である．

E.5. 画像アノテーション実験

次に，開発した検索手法を k 最近傍法による画像アノテーションへ応用し，Flickr12M を用いてその性能を検証する．ここでは，7 章と同じ実験セットアップに従う．画像特徴量としては，HLAC, SURF GLC, SURF BoVW-sqrt を直列に結合したものを用いる．これは，7.3 節において最も良い認識精度を示した特徴量である．これを，“All features” と表記する．CCD の次元数は $d = 200$ とする．

図 26，図 27に，コードのビット数 c に対するアノテーション精度を示す．ベースラインとして，通常の CCD をいくつかの画像特徴量に適用した場合のスコア

も示している．検索実験の結果と同様に，ビット数が少ない場合に SH は LSH よりも相対的に優位である．しかしながら，SH の性能は 256 ビットを境に低下している．一方，LSH はビット数の増加に伴い安定に性能を向上させ，元の CCD に近づいていることが分かる．全体的には，256~512 ビット程度のコードがアノテーション精度と計算コストのよいトレードオフを実現しているといえる．これらのコードを用いると，元の CCD の 85~95%程度のアノテーション精度を実現し，かつ Flickr12M の全データを 375~750MB 程度に納めることができる．

なお，元の CCD においても，次元数 d と学習データ数を調整する事によりトレードオフを操作することがある程度可能である．そこで，CCD のアノテーション精度を，d と学習データセットの大きさを変えながら検証する．図 28，図 29に，使用するメモリ量とアノテーション精度の関係を示す．また，ハッシング手法を利用した際の結果を重ねて表示する．実験結果が示すように，ハッシングを応用した提案手法はより優れたトレードオフを実現しているといえる．

このように，ハッシング手法の応用は CBIR のみならず，アノテーションにおいても有効である．

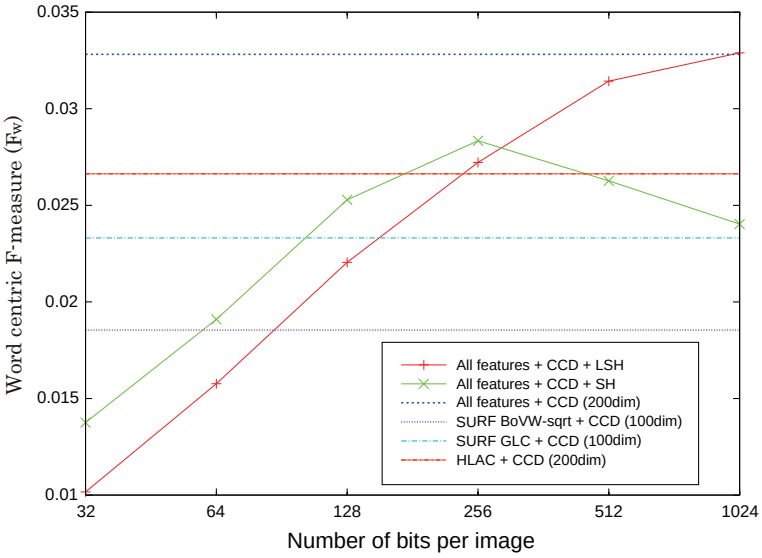


Figure 26: Annotation scores (F_W) with a varying number of bits for the full Flickr12M dataset.

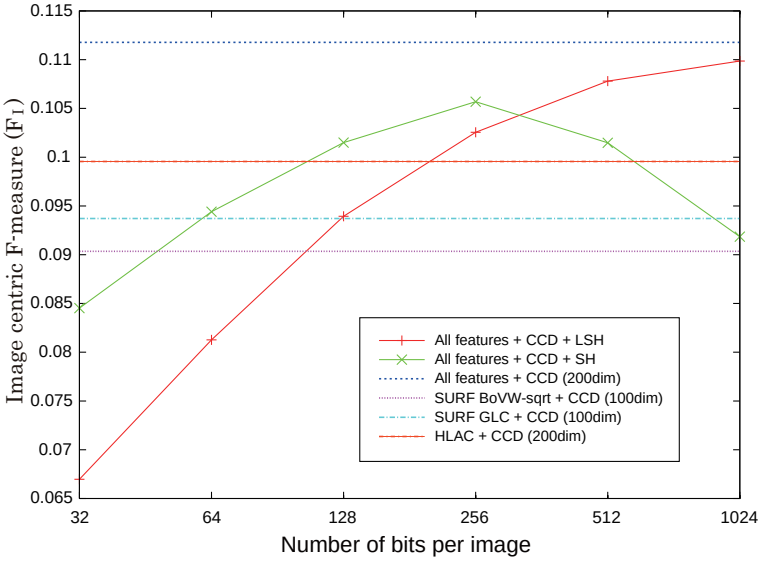


Figure 27: Annotation scores (F_I) with a varying number of bits for the full Flickr12M dataset.

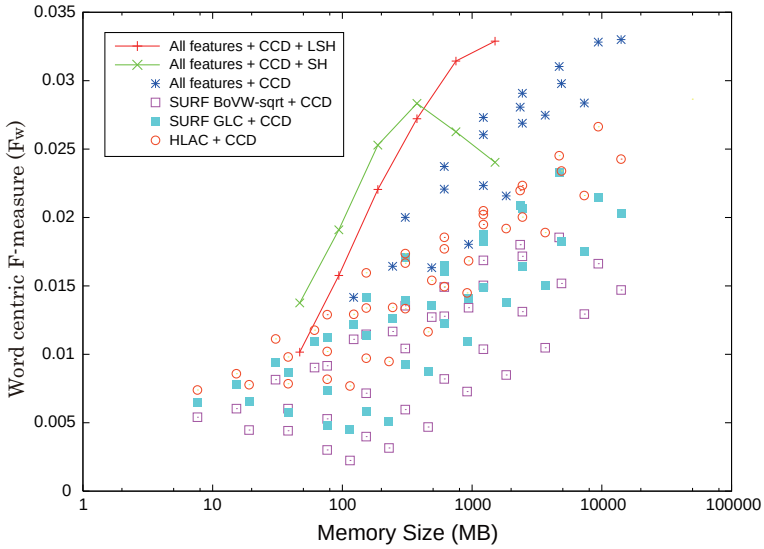


Figure 28: Annotation scores (F_W) with a varying amount of memory (MB).

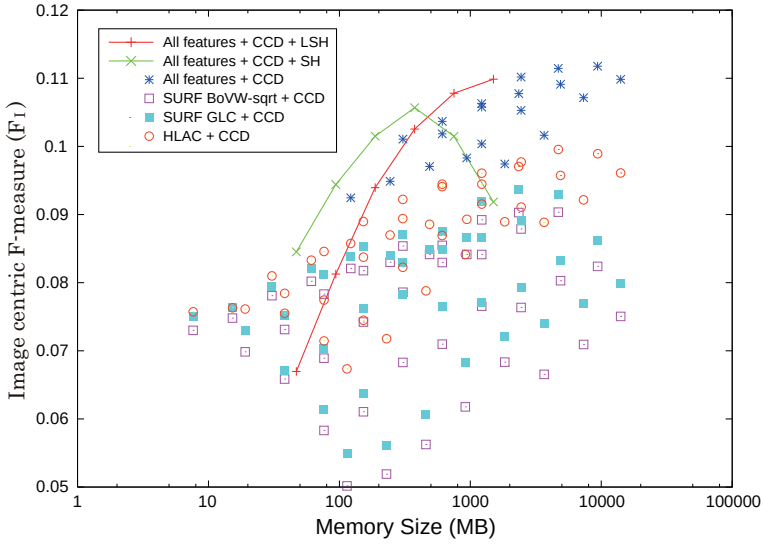


Figure 29: Annotation scores (F_I) with a varying amount of memory (MB).

References

[1] ImageCLEF home page. http://ir.shef.ac.uk/imageclef/. 15

[2] S. Akaho. The e-PCA and m-PCA: Dimension reduction of parameters by information geometry. In *Proceedings of International Joint Conference on Neural Networks*, 2004. 80

[3] S. Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, **8**, 1995. 77

[4] S. Amari and H. Nagaoka. *Methods of Information Geometry*. AMS and Oxford University Press, 2000. 73, 77

[5] F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, 2005. 44

[6] M. Banko and E. Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 26–33, 2001. 19

[7] K. Barnard, P. Duygulu, and D. Forsyth. Clustering art. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages II:434–439, 2001. 23, 30

[8] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, **3**:1107–1135, 2003. 30

[9] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *Proceedings of IEEE International Conference on Computer Vision*, pages II:408–415, 2001. 23, 30

[10] H. Bay, A. Ess, T. Tuytelaars, and L. van Gool. SURF: Speeded up robust features. *Computer Vision and Image Understanding*, **110**[3]:346–359, 2008. 8, 84

[11] P. R. Beaudet. Rotationally invariant image operators. In *Proceedings of the 4th International Joint Conference on Pattern Recognition*, 1978. 8

[12] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, **18**[9]:509–517, 1975. 102, 156

[13] A. Berg, J. Deng, and L. Fei-Fei. ImageNet large scale visual recognition challenge 2010. http://image-net.org/challenges/LSVRC/2010/index. 18

[14] T. L. Berg and D. A. Forsyth. Animals on the web. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006. 19

[15] A. Berger, V. D. Pietra, and S. D. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, **22**[1]:39–71, 1996. 24

[16] I. Biederman. Human image understanding: recent research and a theory. *Computer Vision, Graphics and Image Processing*, **32**[1]:29–73, 1985. 2, 7

[17] T. O. Binford. Spatial understanding: the successor system. In *Proceedings of IEEE conference on Systems and Control*, 1971. 7

[18] M. B. Blaschko and C. H. Lampert. Correlational spectral clustering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 38

[19] D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127–134, 2003. 30

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**:993–1022, 2003. 29

[21] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 74, 76

[22] M. Borga, T. Landelius, and H. Knutsson. A unified approach to PCA, PLS, MLR and CCA. Report LiTH-ISY-R-1992, ISY, 1997. 39

[23] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification via pLSA. In *Proceedings of European Conference on Computer Vision*, pages 517–530, 2006. 31

[24] A. Bosch, A. Zisserman, and X. Muñoz. Image classification using random forests and ferns. In *Proceedings of IEEE International Conference on Computer Vision*, 2007. 89

[25] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**[4]:712–727, 2008. 31, 84, 89, 92, 98, 101, 103, 104

[26] R. Brooks. Symbolic reasoning among 3D models and 2D images. *Artificial Intelligence Journal*, **17**:285–348, 1982. 7

[27] P. Brown, V. D. Pietra, S. D. Pietra, and R. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, **19**[2]:263–311, 1993. 23

[28] M. Calonder, V. Lepetit, and P. Fua. Keypoint signatures for fast learning and recognition. In *Proceedings of European Conference on Computer Vision*, 2008. 8

[29] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**[3]:394–410, 2007. ix, 26, 27, 33, 68

[30] G. Carneiro and N. Vasconcelos. A database centric view of semantic image annotation and retrieval. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005. 26

[31] G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 26

[32] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**[8]:1026–1038, 2002. 23

[33] C.-C. Chang and C.-J. Lin. *LIBSVM: A library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm. 82

[34] E. Chang, K. Goh, G. Sychay, and G. Wu. CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology*, **13**[1]:26–38, 2003. 28

[35] M. Charikar. Similarity estimation techniques from rounding algorithms. In *Proc. ACM Ann. Symp. Theory of Computing*, 2002. 158

[36] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. An online algorithm for large scale image similarity learning. In *Proceedings of Conf. Advances in Neural Information Processing Systems*, 2009. 35

[37] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of ACM International Conference on Image and Video Retrieval*, 2009. 53, 54, 55

[38] R. L. Cilibrasi and P. M. B. Vitanyi. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, **19**[3]:370–383, 2007. 19

[39] V. Clément and M. Thonnat. A knowledge-based approach to integration of image processing procedures. *Computer Vision, Graphics and Image Processing*, **57**[2]:166–184, 1993. 8

[40] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proceedings of ECCV Workshop on Statistical Learning in Computer Vision*, 2004. 9, 31, 32, 73, 75, 76, 135, 163

[41] C. Cusano, G. Ciocca, and R. Schettini. Image annotation using SVM. In *Proceedings of Internet Imaging IV*, **SPIE**, 2004. 27

[42] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of ACM Symposium on Computational Geometry*, pages 253–262, 2004. 102, 156

[43] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of International Conference on Machine Learning*, pages 209–216, 2007. 35

[44] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, **41**[6]:391–407, 1990. 29

[45] J. Deng, A. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *Proceedings of European Conference on Computer Vision*, 2010. 9, 18

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 10, 13, 18, 19

[47] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *Proceedings of IEEE International Conference on Computer Vision*, pages 229–236, 2009. 13

[48] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. In *Proceedings of Symposium on Discrete Algorithms*, 2006. 49

[49] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid. Evaluation of GIST descriptors for web-scale image search. In *Proceedings of ACM International Conference on Image and Video Retrieval*, 2009. 158

[50] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. In *Proceedings of Conf. Advances in Neural Information Processing Systems*, 1996. 67

[51] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of European Conference on Computer Vision*, pages 97–112, 2002. ix, 14, 15, 23, 32, 33, 53, 56, 68, 127

[52] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results. http://www.pascal-network.org/challenges/VOC/voc2009/. 17

[53] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) chellenge. *International Journal of Computer Vision*, **88**[2]:303–338, 2010. 17

[54] J. Fan, Y. Shen, N. Zhou, and Y. Gao. Harvesting large-scale weakly-tagged image databases from the web. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 19

[55] L. Fei-Fei, R. Fergus, and P. Perona. A Bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1134–1141, 2003. 9

[56] 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. In *Proceedings of IEEE CVPR Workshop on Generative-Model Based Vision*, 2004. ix, 15, 16, 91

[57] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**[4]:594–611, 2006. ix, 15, 16

[58] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 524–531, 2005. 30, 84, 91

[59] C. Fellbaum. *WordNet: An electronic lexical database*. Bradford Books, 1998. 10, 18, 19

[60] S. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli relevance models for image and video annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2004. ix, 24, 25, 33, 68

[61] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from Google's image search. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1816–1823, 2005. 18, 19

[62] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 264–271, 2003. 9

[63] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, **55**[1]:119–139, 1997. 157

[64] A. Frome, Y. Singer, and J. Malik. Image retrieval and classification using local distance functions. In *Proceedings of Conf. Advances in Neural Information Processing Systems*, 2006. 35

[65] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *Proceedings of IEEE International Conference on Computer Vision*, 2007. 35

[66] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *Proceedings of IEEE International Conference on Computer Vision*, pages 221–228, 2009. 17

[67] A. Globerson and S. Roweis. Metric learning by collapsing classes. In *Proceedings of Conf. Advances in Neural Information Processing Systems*, pages 451–458, 2006. 35

[68] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *Proceedings of Conf. Advances in Neural Information Processing Systems*, pages 513–520, 2005. 35, 157

[69] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, **8**:725–760, 2007. 103

[70] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. ix, 16

[71] G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 9

[72] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proceedings of IEEE International Conference on Computer Vision*, pages 309–316, 2009. 32, 33, 53, 55, 68, 69, 70, 71

[73] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? Metric learning approaches for face identification. In *Proceedings of IEEE International Conference on Computer Vision*, pages 498–505, 2009. 35

[74] A. Guzmán. Analysis of curved line drawings using context and global information. In B. Meltzer and D. Michie, editors, *Machine Intelligence 6*, pages 325–375. John Wiley & Sons, 1971. 7

[75] D. R. Hardoon, C. Saunders, S. Szedmak, and J. Shawe-Taylor. A correlation approach for automatic image annotation. In *Proceedings of*

International Conference on Advanced Data Mining and Applications, 2006. 29, 38, 52

[76] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 147–151, 1988. 8, 84

[77] X. He and P. Niyogi. Locality preserving projections. In *Proceedings of Conf. Advances in Neural Information Processing Systems*, 2003. 35

[78] N. Hervé and N. Boujemaa. Image annotation: which approach for realistic databases? In *Proceedings of ACM International Conference on Image and Video Retrieval*, 2007. 101, 103

[79] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Nature*, **313**[5786]:504–507, 2006. 157

[80] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, **42**:177–196, 2001. 29

[81] S. C. H. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2072–2078, 2006. 35

[82] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2137–2144, 2006. 9

[83] H. Hotelling. Relations between two sets of variates. *Biometrika*, **28**[3/4]:321–377, 1936. 38

[84] M. Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, **8**[2]:179–187, 1962. 7

[85] S. Ikeda, T. Tanaka, and S. Amari. Stochastic reasoning, free energy, and information geometry. *Neural Computation*, **16**:1779–1810, 2004. 77

[86] P. Indyk, R. Motwani, P. Raghavan, and S. Vempala. Locality-preserving hashing in multidimensional spaces. In *Proceedings of ACM Symposium on Theory of Computing*, pages 618–625, 1997. 156

[87] S. Ioffe. Probabilistic linear discriminant analysis. In *Proceedings of European Conference on Computer Vision*, pages 531–542, 2006. 82

[88] P. Jain, B. Kulis, I. S. Dhillon, and K. Grauman. Online metric learning and fast similarity search. In *Proceedings of Conf. Advances in Neural Information Processing Systems*, 2008. 35

[89] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometry consistency for large scale image search. In *Proceedings of European Conference on Computer Vision*, 2008. 158

[90] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. 157

[91] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 158

[92] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003. 24, 33, 68, 127

[93] J. Jeon and R. Manmatha. Using maximum entropy for automatic image annotation. In *Proceedings of International Conference on Image and Video Retrieval*, pages 24–32, 2004. 24, 33, 68

[94] Y. Jing, S. Baluja, and H. Rowley. Canonical image selection from the web. In *Proceedings of International Conference on Image and Video Retrieval*, 2007. ix, 11

[95] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proceedings of IEEE International Conference on Computer Vision*, 2005. 75

[96] F. Kang, R. Jin, and R. Sukthankar. Correlated label propagation with application to multi-label learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1719–1726, 2006. 32

[97] T. Kato, T. Kurita, N. Otsu, and K. Hirata. A sketch retrieval method for full color image database –query by visual example–. In *Proceedings of International Conference on Pattern Recognition*, **1**, pages 530–533, 1992. 135

[98] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation of local image descriptors. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, **2**, pages 506–513, 2004. 8, 99

[99] J. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, **58**[3]:433–451, 1971. 125

[100] B. Kulis and K. Grauman. Kernelized locality-sensitive hashing for scalable image search. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2130–2137, 2009. 156

[101] B. Kulis, P. Jain, and K. Grauman. Fast similarity search for learned metrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**[12]:2143–2157, 2009. 156

[102] S. Kumar, M. Mohri, and A. Talwalkar. Ensemble Nyström method. In *Proceedings of Conf. Advances in Neural Information Processing Systems*, 2010. 49

[103] A. Kutics, A. Nakagawa, and M. Nakajima. Image retrieval via connecting words to salient objects. In *Proceedings of IEEE International Conference on Image Processing*, 2003. 10

[104] L. Ladický, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical CRFs for object class image segmentation. In *Proceedings of IEEE International Conference on Computer Vision*, 2009. 14

[105] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of International Conference on Machine Learning*, 2001. 14, 101

[106] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: object localization by efficient subwindow search. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 14

[107] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, **5**:27–72, 2004. 29, 32, 67

[108] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002. 24

[109] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Proceedings of Conf. Advances in Neural Information Processing Systems*, 2003. 24, 29, 33, 42, 68

[110] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006. x, 9, 81, 82, 84, 85, 91, 101, 103

[111] D. T. Lee and C. K. Wong. Worst-case analysis for region and partial region searches in multidimensional binary search trees and balanced quad trees. *Acta Informatica*, **9**[1]:23–29, 1977. 156

[112] Y. J. Lee and K. Grauman. Object-graphs for context-aware category discovery. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 10

[113] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Proceedings of ECCV Workshop on Statistical Learning in Computer Vision*, 2004. 14

[114] L.-J. Li, G. Wang, and L. Fei-Fei. OPTIMOL: automatic online picture collection via incremental model learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 19

[115] Li-Jia Li and L. Fei-Fei. What, where and who? Classifying events by scene and object recognition. In *Proceedings of IEEE International Conference on Computer Vision*, 2007. x, 81, 82, 89

[116] M. Li, J. T. Kwok, and B.-L. Lu. Making large-scale Nyström approximation possible. In *Proceedings of International Conference on Machine Learning*, 2010. 49

[117] R. Lienhart, S. Romberg, and E. Hörster. Multilayer pLSA for multimodal image retrieval. In *Proceedings of International Conference on Image and Video Retrieval*, 2009. 31

[118] R. Lienhart and M. Slaney. PLSA on large scale image databases. In *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, **4**, pages 1217–1220, 2007. 31

[119] R.-S. Lin, D. A. Ross, and J. Yagnik. SPEC hashing: similarity preserving algorithm for entropy-based coding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 157, 158

[120] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma. Image annotation via graph learning. *Pattern Recognition*, **42**:218–228, 2009. 28, 33, 68

[121] J. Liu, M. Li, W.-Y. Ma, Q. Liu, and H. Lu. An adaptive graph model for automatic image annotation. In *Proceedings of ACM International Workshop on Multimedia Information Retrieval*, pages 61–70, 2006. 28, 33, 68

[122] J. Liu, B. Wang, M. Li, Z. Li, W.-Y. Ma, H. Lu, and S. Ma. Dual cross-media relevance model for image annotation. In *Proceedings of ACM International Conference on Multimedia*, pages 605–614, 2007. 19, 32, 33, 68

[123] N. Loeff and A. Farhadi. Scene discovery by matrix factorization. In *Proceedings of European Conference on Computer Vision*, **451-464**, 2008. 28, 33, 68

[124] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1150–1157, 1999. 8, 55, 74, 84, 92

[125] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, **60**[2]:91–110, 2004. 8

[126] Z. Lu, Horace H. S. Ip, and Q. He. Context-based multi-label image annotation. In *Proceedings of International Conference on Image and Video Retrieval*, 2009. 32, 33, 68

[127] S. Maji and A. C. Berg. Max-margin additive classifiers for detection. In *Proceedings of IEEE International Conference on Computer Vision*, pages 40–47, 2009. 87

[128] S. Maji and J. Malik. Object detection using a max-margin Hough transform. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 14

[129] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *Proceedings of European Conference on Computer Vision*, pages 316–329, 2008. ix, 15, 32, 33, 53, 54, 68, 69, 70, 71

[130] C. D. Manning and H. Schütze. *Foundation of Statistical Natural Language Processing*. The MIT Press, 1999. 9, 75

[131] D. Metzler and R. Manmatha. An inference network approach to image retrieval. In *Proceedings of International Conference on Image and Video Retrieval*, pages 42–50, 2004. 25, 33, 68

[132] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**[10]:1615–1630, 2005. 84

[133] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proceedings of ACM International Conference on Multimedia*, pages 275–278, 2003. 31

[134] F. Monay and D. Gatica-Perez. PLSA-based image auto-annotation: constraining the latent space. In *Proceedings of ACM International Conference on Multimedia*, pages 348–351, 2004. 31

[135] F. Monay and D. Gatica-Perez. Modeling semantic aspects for cross-media image indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**[10]:1802–1817, 2007. 31

[136] P. J. Moreno, P. P. Ho, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In *Proceedings of Conf. Advances in Neural Information Processing Systems*, 2003. 74, 76, 80

[137] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999. 23, 33, 68

[138] N. Morioka and S. Satoh. Building compact local pairwise codebook with joint feature space clustering. In *Proceedings of European Conference on Computer Vision*, pages 692–705, 2010. 89

[139] N. Morioka and S. Satoh. Learning directional local pairwise bases with sparse coding. In *Proceedings of British Machine Vision Conference*, 2010. 89, 102, 103

[140] H. Murase and S. K. Nayar. Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, **14**[9]:5–24, 1995. 8

[141] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi. Information geometry of U-boost and Bregman divergence. *Neural Computation*, **16**:1437–1481, 2004. 77

[142] H. Nakayama, T. Harada, and Y. Kuniyoshi. Canonical contextual distance for large-scale image annotation and retrieval. In *Proceedings of*

the 1st ACM International Workshop on Large-Scale Multimedia Mining and Retrieval, pages 3–10, 2009. 45

[143] H. Nakayama, T. Harada, and Y. Kuniyoshi. Evaluation of dimensionality reduction methods for image auto-annotation. In *Proceedings of British Machine Vision Conference*, 2010. 45

[144] H. Nakayama, T. Harada, and Y. Kuniyoshi. Global Gaussian approach for scene categorization using information geometry. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 76, 103

[145] H. Nakayama, T. Harada, Y. Kuniyoshi, and N. Otsu. High-performance image annotation and retrieval for weakly labeled images. In *Proceedings of Pacific-Rim Conference on Multimedia*, pages 601–610, 2008. 43

[146] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2161–2168, 2006. 157

[147] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *Proceedings of European Conference on Computer Vision*, pages 490–503, 2006. 84, 96, 102

[148] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, **42**[3]:145–175, 2001. 32, 55, 91, 109, 157, 159

[149] N. Otsu and T. Kurita. A new scheme for practical, flexible and intelligent vision systems. In *Proceedings of IAPR Workshop on Computer Vision*, pages 431–435, 1988. 79, 135

[150] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 663–658, 2004. 28

[151] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. GCap: Graph-based automatic image captioning. In *Proceedings of IEEE CVPR Workshop on Multimedia Data and Document Engineering*, 2004. 28

[152] A. Perina, M. Cristani, U. Castellani, V. Murino, and N. Jojic. A hybrid generative/discriminative classification framework based on

free-energy terms. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2058–2065, 2009. 101, 103

[153] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 74

[154] F. Perronnin, C. R. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. In *Proceedings of European Conference on Computer Vision*, pages 464–475, 2006. 75

[155] F. Perronnin, J. Sánchez, and Y. Liu. Large-scale image categorization with explicit data embedding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 49, 110, 163

[156] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *Proceedings of European Conference on Computer Vision*, 2010. 73

[157] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman. Dataset issues in object recognition. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward category-level object recognition*, pages 29–48. Springer, 2006. 14

[158] J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors. *Toward category-level object recognition*. LNCS 4170. Springer, 2006. 1

[159] A. R. Pope. Model-based object recognition: a survey of recent research. Report TR-94-04 TR-94-04, University of British Columbia, Computer Science Department, 1994. 7

[160] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009. x, 81, 82, 89

[161] D. Ramanan and S. Baker. Local distance functions: a taxonomy, new algorithms and an evaluation. In *Proceedings of IEEE International Conference on Computer Vision*, pages 301–308, 2009. 35

[162] L. G. Roberts. Machine perception of three-dimensional solids. In J. Tippett, D. Berkowitz, L. Clapp, C. Koester, and A. Vanderburgh, editors, *Optical and Electro-optical Information processing*, pages 159–197. MIT Press, 1965. 7

[163] B. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, **77**[1-3]:157–173, 2008. 17, 159

[164] R. R. Salakhutdinov and G. E. Hinton. Semantic hashing. In *Proceedings of ACM SIGIR workshop on Information Retrieval and Applications of Graphical Models*, 2007. 157

[165] B. Schiele and J. L. Crowley. Recognition using multidimensional receptive field histograms. *Proceedings of European Conference on Computer Vision*, pages 610–619, 1996. 8

[166] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**[5]:1299–1319, 1998. 48, 131

[167] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *Proceedings of IEEE International Conference on Computer Vision*, 2007. 19

[168] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis. Human detection using partial least squares analysis. In *Proceedings of IEEE International Conference on Computer Vision*, pages 24–31, 2009. 37

[169] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. In *Proceedings of IEEE International Conference on Computer Vision*, pages 750–757, 2003. 157

[170] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**[8]:888–905, 2000. 23, 24

[171] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 75

[172] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proceedings of European Conference on Computer Vision*, 2006. 14

[173] L. Si, R. Jin, S. C. H. Hoi, and M. R. Lyu. Collaborative image retrieval via regularized metric learning. *Multimedia Systems*, **12**[1]:34–44, 2006. 35

[174] B. Siddiquie and A. Gupta. Beyond active noun tagging: modeling contextual interactions for multi-class active learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 10

[175] C. Silpa-Anan and R. Hartley. Optimized kd-trees for fast image descriptor matching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 157

[176] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **22**[12], 2000. 8, 11

[177] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. de Bona, A. Binder, C. Gehl, and V. Franc. The SHOGUN machine learning toolbox. *Journal of Machine Learning Research*, **11**:1799–1802, 2010. 67

[178] A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. In *Proceedings of CVPR workshop on Internet Vision*, 2008. 18

[179] A. Stein and M. Hebert. Incorporating background invariance into feature-based object recognition. In *Proceedings of IEEE Workshop on Applications of Computer Vision*, pages 37–44, 2005. 8

[180] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, **7**[1]:11–32, 1991. 8

[181] A. Talwalkar, S. Kumar, and H. Rowley. Large-scale manifold learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 49

[182] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**[11]:1958–1970, 2008. 10, 12, 13, 19, 109

[183] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 157, 158

[184] A. Torralba, K. Murphy, and W. Freeman. Using the forest to see the trees: a graphical model relating features, objects and scenes. In

Proceedings of Conf. Advances in Neural Information Processing Systems, 2003. 9

[185] M. Turk and A. Pentland. Eigenfaces for recognition. *Cognitive Neuroscience*, **3**[1]:71–96, 1991. 8

[186] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, **9**:187–222, 1991. 25

[187] T. Tuytelaars and C. Schmid. Vector quantizing feature space with a regular lattice. In *Proceedings of IEEE International Conference on Computer Vision*, 2007. 75

[188] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Proceedings of European Conference on Computer Vision*, pages 589–600, 2006. 75, 76

[189] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on Riemannian manifolds. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 75, 76

[190] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *Proceedings of European Conference on Computer Vision*, pages 334–348, 2006. 8, 55

[191] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. In *Proceedings of European Conference on Computer Vision*, pages 696–709, 2008. 75

[192] N. Vasconcelos, P. P. Ho, and P. J. Moreno. The Kullback-Leibler kernel as a framework for discriminant and localized representations for visual recognition. In *Proceedings of European Conference on Computer Vision*, 2004. 74

[193] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 48, 87

[194] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2001. 13

[195] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 319–326, 2004. 15, 17, 54

[196] L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 55–64, 2006. 17

[197] C. Wang, S. Yan, L. Zhang, and H.-J. Zhang. Multi-label sparse coding automatic image annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 32, 33, 68

[198] C. Wang, L. Zhang, and H.-J. Zhang. Learning to reduce the semantic gap in web image retrieval and annotation. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 355–362, 2008. 35

[199] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *Proceedings of IEEE International Conference on Computer Vision*, pages 537–544, 2009. 28

[200] J. Wang, S. Kumar, and S.-F. Chang. Semi-supervised hashing for scalable image retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 157

[201] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 73, 75, 102, 103

[202] X. J. Wang, L. Zhang, F. Jing, and W. Y. Ma. Annosearch: Image auto-annotation by search. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1483–1490, 2006. 19

[203] X.-J. Wang, L. Zhang, M. Liu, Y. Li, and W.-Y. Ma. ARISTA - image search to annotation on billions of web photos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 12, 13, 18, 19, 105

[204] Y. Wang and S. Gong. Conditional random field for natural scene categorization. In *Proceedings of British Machine Vision Conference*, 2007. 101, 103

[205] R. Weber, H.-J. Schek, and S. Blott. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *Proceedings of International Conference on Very Large DataBases*, pages 194–205, 1998. 156

[206] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *Proceedings of Conf. Advances in Neural Information Processing Systems*, pages 1473–1480, 2006. 35

[207] K. Q. Weinberger and L. K. Saul. Fast solvers and efficient implementations for distance metric learning. In *Proceedings of International Conference on Machine Learning*, pages 1160–1167, 2008. 35

[208] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *Proceedings of Conf. Advances in Neural Information Processing Systems*, 2008. 157, 159

[209] C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Proceedings of Conf. Advances in Neural Information Processing Systems*, pages 682–688, 2000. 49

[210] H. Wold. Partial least squares. In S. Kotz and N. Johnson, editors, *Encyclopedia of Statistical Sciences*, **6**, pages 581–591. John Wiley & Sons, 1985. 37

[211] J. Wu and J. M. Rehg. Beyond the Euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *Proceedings of IEEE International Conference on Computer Vision*, pages 630–637, 2009. 75, 84, 89, 103

[212] L. Wu, S. C. H. Hoi, R. Jin, J. Zhu, and N. Yu. Distance metric learning from uncertain side information with application to automated photo tagging. In *Proceedings of ACM International Conference on Multimedia*, pages 135–144, 2009. 35

[213] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li. Flickr distance. In *Proceedings of ACM International Conference on Multimedia*, pages 31–40, 2008. 18, 19

[214] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. SUN database: large-scale scene recognition from abbey to zoo. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 18, 89

[215] O. Yakhnenko and V. Honavar. Annotating images and image objects using a hierarchical Dirichlet process model. In *Proceedings of ACM SIGKDD workshop on Multimedia Data Mining*, 2008. 31

[216] O. Yakhnenko and V. Honavar. Multiple label prediction for image annotation with multiple kernel correlation models. In *Proceedings of IEEE CVPR workshop on Visual Context Learning*, 2009. 29, 38, 67

[217] J. Yang, Y. Li, Y. Tian, L. Duan, and W. Gao. Group-sensitive multiple kernel learning for object categorization. In *Proceedings of IEEE International Conference on Computer Vision*, pages 436–443, 2009. 17, 89

[218] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 73, 75, 76, 102, 103

[219] B. Yao, X. Yang, and S.-C. Zhu. Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks. In *Proceedings of CVPR workshop on Energy Minimization Methods*, pages 169–183, 2007. 18

[220] A. Yavlinsky, E. Schofield, and S. Rüger. Automated image annotation using global features and robust nonparametric density estimation. In *Proceedings of International Conferences on Image and Video Retrieval*, pages 507–517, 2005. 32, 33, 68

[221] J. Yuen, B. Russell, C. Liu, and A. Torralba. LabelMe video: building a video database with human annotations. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1451–1458, 2009. 17

[222] M. Zerroug and R. Nevatia. From an intensity image to 3-d segmented descriptions. In J. Ponce, M. Hebert, and A. Zisserman, editors, *Object Representation in Computer Vision II*, pages 11–24. 1996. 7

[223] H. Zhang, A. C. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, **2**, pages 2126–2136, 2006. 103

[224] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: An in-depth study. Technical Report RR-5497, INRIA, 2005. 95

[225] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, **73**[2]:213–238, 2007. 48

[226] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. N. Metaxas. Automatic image annotation using group sparsity. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 32, 33, 68, 69, 70, 71

[227] X. Zhou, N. Cui, Z. Li, F. Liang, and T. S. Huang. Hierarchical Gaussianization for image classification. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1971–1977, 2009. 74, 76, 84, 89, 102, 103

[228] X. Zhou, K. Yu, T. Zhang, and T. S. Huang. Image classification using super-vector coding of local image descriptors. In *Proceedings of European Conference on Computer Vision*, 2010. 73

[229] X. Zhou, X. Zhuang, H. Tang, M. Hasegawa-Johnson, and T. S. Huang. A novel Gaussianized vector representation for natural scene categorization. In *Proceedings of International Conference of Pattern Recognition*, 2008. 74

[230] 栗田多喜夫, 加藤俊一, 福田郁美, and 板倉あゆみ. 印象語による絵画データベースの検索. **情報処理学会論文誌**, **33**[11]:1373–1383, 1992. 29, 38, 52

[231] 岡部 孝弘, 近藤 雄飛, 木谷 クリス 真実, and 佐藤 洋一. カテゴリーの共起を考慮した回帰による複数物体認識. **電子情報通信学会論文誌** *D*, **J92-D**[8]:1115–1124, 2009. 29

[232] 中山英樹, 原田達也, and 國吉康夫. 大規模 web 画像のための画像アノテーション・リトリーバル手法-web 集合知からの自律的画像知識獲得へ向けて-. In **第** *12* **回画像の認識・理解シンポジウム** *(MIRU 2009)*, pages 55–62, 2009. 45

[233] 中山英樹, 原田達也, 國吉康夫, and 大津展之. 画像・単語間概念対応の確率構造学習を利用した超高速画像認識・検索方法. In **電子情報通信学会技術研究報告**, *PRMU2007-147*, pages 65–70, 2007. 43

[234] 柳井啓司. 一般画像自動分類の実現へ向けた world wide web からの画像知識の獲得. **人工知能学会誌**, **19**[5]:429–439, 2004. 18, 19

[235] 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・イメージメディア**, **48**[SIG16 (CVIM19)]:1–24, 2007. 1

Publications

Journal

1. 中山英樹, 原田達也, 國吉康夫, “大規模 Web 画像のための画像アノテーション・リトリーバル手法,” 電子情報通信学会論文誌, Vol.J93-D, No.8, pp.1267-1280, 2010.

2. Hideki Nakayama, Tatsuya Harada, and Yasuo Kuniyoshi, “Dense Sampling Low-Level Statistics of Local Features,” IEICE Transactions on Information and Systems, Vol.E93-D, No.7, pp.1727-1736, 2010.

3. Tatsuya Harada, Hideki Nakayama, Yasuo Kuniyoshi, and Nobuyuki Otsu, “Image Annotation and Retrieval for Weakly Labeled Images using Conceptual Learning,” New Generation Computing, Vol.28, No.3, pp.277-298, 2010.

4. 原田達也, 中山英樹, 國吉康夫, “AI Goggles: 追加学習機能を備えたウェアラブル画像アノテーション・リトリーバルシステム,” 電子情報通信学会論文誌, Vol.J93-D, No.6, pp.857-869, 2010.

Reviewed Conference

1. Hideki Nakayama, Tatsuya Harada, and Yasuo Kuniyoshi, “Evaluation of Dimensionality Reduction Methods for Image Auto-Annotation,” Proceedings of the 21st British Machine Vision Conference (BMVC 2010), Aberystwyth, United Kingdom, Sep., 2010.

2. Tatsuya Harada, Hideki Nakayama, and Yasuo Kuniyoshi, “Improving Local Descriptors by Embedding Global and Local Spatial Information,” Proceedings of the 11th European Conference on Computer Vision (ECCV 2010), Crete, Greece, Sep., 2010.

3. Hideki Nakayama, Tatsuya Harada, and Yasuo Kuniyoshi, "Global Gaussian Approach for Scene Categorization Using Information Geometry," Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010), San Francisco, USA, June, 2010.

4. Asako Kanezaki, Hideki Nakayama, Tatsuya Harada, and Yasuo Kuniyoshi, "High-speed 3D Object Recognition Using Additive Features in a Linear Subspace," Proceedings of the 2010 IEEE International Conference on Robotics and Automation (ICRA 2010), pp.3128-3134, Anchorage, USA, May, 2010.

5. Hideki Nakayama, Tatsuya Harada, and Yasuo Kuniyoshi, "Canonical Contextual Distance for Large-Scale Image Annotation and Retrieval," Proceedings of the 1st ACM International Workshop on Large-Scale Multimedia Mining and Retrieval (LS-MMRM 2009), pp.3-10, Beijing, China, Oct., 2009.

6. Hideki Nakayama, Tatsuya Harada, and Yasuo Kuniyoshi, "Dense Sampling Low-Level Statistics of Local Features," Proceedings of the 8th ACM International Conference on Image and Video Retrieval (CIVR 2009), Santorini, Greece, July, 2009.

7. Tatsuya Harada, Hideki Nakayama, and Yasuo Kuniyoshi, "Image Annotation and Retrieval Based on Efficient Learning of Contextual Latent Space," Proceedings of the 2009 IEEE International Conference on Multimedia & Expo (ICME 2009), pp.858-861, New York, USA, June, 2009.

8. Hideki Nakayama, Tatsuya Harada, and Yasuo Kuniyoshi, "AI Goggles: Real-time Description and Retrieval in the Real World with Online Learning," Proceedings of the 6th Canadian Conference on Computer and Robot Vision (CRV 2009), pp.184-191, Kelowna, Canada, May, 2009.

9. Hideki Nakayama, Tatsuya Harada, and Yasuo Kuniyoshi, "Scene Classification Using Generalized Local Correlation," Proceedings of the 11th IAPR Conference on Machine Vision Applications (MVA 2009), pp.195-198, Hiyoshi, Japan, May, 2009.

10. Hideki Nakayama, Tatsuya Harada, and Yasuo Kuniyoshi, "High-Performance Image Annotation and Retrieval for Weakly Labeled Images," Proceedings of the 2008 Pacific-Rim Conference on Multimedia (PCM 2008), LNCS 5353, pp.601-610, Tainan, Taiwan, Dec., 2008.

11. Rie Matsumoto, Hideki Nakayama, Tatsuya Harada, Yasuo Kuniyoshi, and Nobuyuki Otsu, "Journalist Robot: Robot System Making News Articles

from Real World," Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2007), pp.1234-1241, San Diego, USA, Oct., 2007.

Reviewed Domestic Conference

1. 中山英樹, 原田達也, 國吉康夫, "大規模 Web 画像のための画像アノテーション・リトリーバル手法 -Web 集合知からの自律的画像知識獲得へ向けて-," 第 12 回画像の認識・理解シンポジウム (MIRU 2009), pp.55-62, 松江, July, 2009.
2. 金崎朝子, 中山英樹, 原田達也, 國吉康夫, "部分空間法とカラー立体高次局所自己相関特徴を用いた高速三次元物体認識," 第 12 回画像の認識・理解シンポジウム (MIRU 2009), pp.103-110, 松江, July, 2009.
3. 中山英樹, 原田達也, 國吉康夫, "画像情報からリアルタイムに実世界記述・検索を行うサイバーゴーグル," 第 13 回ロボティクスシンポジア, pp.192-199, 高松, Mar., 2009.
4. 松本理恵, 中山英樹, 原田達也, 國吉康夫, "ジャーナリストロボット -実世界からの自律的ニュース探索と高性能な事象キーワードの記述-," 第 13 回ロボティクスシンポジア, pp.73-80, 高松, Mar., 2009.

Un-reviewed Domestic Conference

1. 牛久祥孝, 中山英樹, 原田達也, 國吉康夫, "Web 画像と文章の大域的特徴から得る潜在的意味に基づくデータ検索 -Web 上での一般画像認識実現への新たなアプローチを目指して-," 電子情報通信学会技術研究報告, PRMU2009-100, pp.45-50, 石川, Nov., 2009.
2. 原田達也, 松本理恵, 中山英樹, 國吉康夫, "ニュース性により記事生成を行うジャーナリストロボットの試み," 第 27 回ロボット学会学術講演会, pp.2G1-07, 横浜, Sep., 2009.
3. 原田達也, 中山英樹, 國吉康夫, "自らの視覚記憶を言葉で検索可能とする AI Goggles," 第 23 回人工知能学会全国大会, 高松, June., 2009.
4. 原田達也, 中山英樹, 國吉康夫, "超高速汎用的画像認識検索手法の開発と実世界応用," 第 4 回デジタルコンテンツシンポジウム, 千葉, June., 2008.
5. 中山英樹, 原田達也, 國吉康夫, "サイバーゴーグル: 画像情報からリアルタイムに実世界記述・検索を行うシステム," 情報処理学会第 70 回全国大会, pp.5-89 - 5-90, 筑波, Mar., 2008.

6. 松本理恵, 中山英樹, 原田達也, 國吉康夫, “ジャーナリストロボットシステム: 実世界からの自律的ニュース探索と事象の記述,” 情報処理学会第 70 回全国大会, pp.5-91 - 5-92, 筑波, Mar., 2008.

7. 原田達也, 中山英樹, 國吉康夫, 大津展之, “画像・単語列間の確率的な概念獲得による高速かつ高精度な汎用的画像認識・検索手法,” 情報処理学会第 70 回全国大会, pp.5-87 - 5-88, 筑波, Mar., 2008.

8. 中山英樹, 原田達也, 國吉康夫, 大津展之, “画像・単語間概念対応の確率構造学習を利用した超高速画像認識・検索方法,” 電子情報通信学会技術研究報告, PRMU2007-147, pp.65-70, 神戸, Dec., 2007.

9. 松本理恵, 中山英樹, 原田達也, 國吉康夫, “ジャーナリストロボット: 実世界からニュース記事を生成するロボットシステム,” ロボティクス・メカトロニクス講演会 2007 (ROBOMECH), 2A1-L02, 秋田, May, 2007.

10. 下畠康幸, 中山英樹, 原田達也, 大津展之, “カメラの動きに頑健な異常検出手法に基づく移動物体の検出,” ロボティクス・メカトロニクス講演会 2007 (ROBOMECH), 2P1-C08, 秋田, May, 2007.

Others

1. **(Competition)** Tatsuya Harada, Hideki Nakayama, Yoshitaka Ushiku, Yuya Yamashita, Jun Imura, and Yasuo Kuniyoshi, Got the 3rd place in the ImageNet Large Scale Visual Recognition Challenge 2010 (in conjunction with ECCV 2010), Crete, Greece, Sep., 2010.

2. **(Invited talk)** 中山英樹, “実世界指向画像認識・検索手法の開発とその応用,” 第 8 回情報科学技術フォーラム（FIT 2008）イベント企画：次世代を担う若い情報・システム研究者を迎えて, pp.11-12, 仙台, Sep., 2009.

Awards

1. 2008 年　日本機械学会三浦賞
2. 2008 年　PRMU 研究奨励賞
3. 2008 年　計測自動制御学会 SI 部門賞若手奨励賞
4. 2009 年　情報処理学会第 70 回全国大会大会奨励賞
5. 2009 年　MIRU 2009 シングルトラックオーラルセッション採択

Patents

1. 特徴量生成装置, 特徴量生成法および特徴量生成プログラム, ならびにクラス判別装置, クラス判別方法およびクラス判別プログラム, 特願 2009-121244, 2009.

2. 対応関係学習装置および方法ならびに対応関係学習用プログラム, アノテーション装置および方法ならびにアノテーション用プログラム, および, リトリーバル装置および方法ならびにリトリーバル用プログラム, 特願 2007-240272, 2007.