情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2012-CVIM-181 No.5
2012/3/15

大規模データを用いた一般物体・シーン認識の潮流と理論

原　田　達　也[†1,†2]

近年，web 上の大規模な画像データを用いた画像認識が注目されている．本論文でははじめに大規模データを用いた物体・シーン認識のトレンドについて概観する．次に，最先端の高精度物体・シーン認識システムのパイプラインと各構成要素を説明し，スケーラビリティを維持するための工夫点を紹介する．特に，大規模な画像データを活用するにはスケーラビリティを維持するために線形識別機を用いることが一般的である．線形の識別機であっても十分な識別能力を発揮するためには画像特徴をどのように表現するかが鍵となり，その詳細を解説する．

Trend and Theory in Large-Scale Object and Scene Recognition

Tatsuya Harada[†1,†2]

Recent year, the use of a large-scale image dataset collected from Web becomes a trend in the generic object and scene recognition. In this paper, I take a brief look at the large-scale object and scene recognition, and introduce a pipeline of the state-of-the art methods. Since scalability is a crucial issue in the large-scale object recognition, I mainly focus on devices to keep scalability and recognition accuracy. Especially, in order to keep scalability, a linear classifier is commonly utilized. To obtain high recognition accuracy, an image representation, which generates a feature vector from local descriptors, is a key technique. Therefore, I expound details of image representations, and give a unified view of those methods.

†1 東京大学
The University of Tokyo
†2 JST さきがけ
JST PRESTO

1. 物体・シーン認識とは

物体・シーン認識は画像をシステムに入力し，画像に適切なラベルを付与する過程のことを指す．物体とシーンは本来別個の概念であるが，コンピュータビジョンでは物体認識と言えばシーン認識も含んでいる場合が多い．そのため本論文でも物体認識と言えばシーン認識も含むことにする．物体認識は大きく分けて**特定物体認識（specific object recognition）**と**一般物体認識（generic object recognition）**に分類可能である．特定物体認識とは，データベースに認識対象とする物体の画像をすでに持つことを前提として，入力画像に写る物体とデータベース内の画像を照合することであり，一般物体認識はデータベースに存在しない入力画像のカテゴリを予測することである．これらの分類とは別に**画像アノテーション（image annotation）**という言葉が用いられるが，狭義としては複数ラベルが付与された画像データセットから，入力画像に複数のラベルを付与することであり，広義としては特定物体認識，一般物体認識を含む広い概念である．これらは全て最終的目標は同一でありながら，利用される手法は異なる場合が多い．

しかしながら物体認識の処理の流れは学習時と識別時のどちらの場合でもデータ→特徴抽出→識別機の順に処理されていくのが一般的である．この処理が行われる際に生じる現象を理解するために，まず data processing 定理を説明する．Data processing 定理では w をワールドの状態，d を収集したデータ，r はそのデータを処理したデータとする．これらの変数は $w \to d \to r$ のようなマルコフ連鎖の関係にあるとすると，これらの同時確率は $P(w,d,r) = P(w)P(d|w)P(r|d)$ となる．この時，D が伝達する W に関する平均情報量 $I(W;D)$ は，R が伝達する W に関する平均情報量 $I(W;R)$ よりも必ず大きくなる（$I(W;D) \geq I(W;R)$）．この定理の意味するところは，データ処理はデータの持つ情報を破壊させるのみで，決して増やすことがないということである．つまり，ゴミを入力として宝が生まれることは決してなく，W, D, R の順でいかに重要な情報を保持しているかが鍵となることが分かる．Data processing 定理の観点で物体認識のパイプラインを眺めると，認識性能を向上させるためにはデータ（知識）が最も重要であり，次に特徴抽出，最後に識別機となる．以下の章では，この重要度の順に説明していく．

2. 知識の収集

知識の収集方法は主に以下の 3 つに分類される．

(1)　Web 上の膨大な情報を利用

(2) 人に尋ねる
(3) 対象とは別の知識を活用

本論文では，Web 上の膨大な情報を利用した例を詳しく述べ，人に尋ねる例や対象とは別の知識を活用する例は近年の代表例を紹介するにとどめる．

2.1 インターネット上の膨大な情報を利用

Web 上には Flickr[⋆1] に代表される膨大なラベル付き画像データや Google Image Search のような画像検索が存在するため，これらを活用した画像知識構築が盛んに行われている．ここでは代表例として TinyImages，ImageNet，ARISTA，Visual Synset を紹介する．

TinyImages[⋆2][30] は 8,000 万枚の 32×32 の低解像度画像から構成される．これらの画像は WordNet[19] に含まれる全てのカテゴリを Flickr や Google などの画像検索エンジンで検索し収集している．カテゴリ数は 75,062 である．WordNet を利用することでカテゴリに偏りのない画像を収集可能である．大規模な画像データセットを扱うには画像の次元圧縮が必要だが，このデータセットでは単純に解像度を低下させることで実現している．画像の低解像度表現はストレージへの負荷を少なくするだけではなく，識別において重要な情報を失っていないことが実験的に調べられている．

ImageNet[⋆3][36] は WordNet の階層的構造を利用した大規模な画像のオントロジである．2012 年 2 月 19 日現在，14,197,122 枚の画像と 21,841 カテゴリが収集されている．1 カテゴリあたり 500 から 1000 枚の画像が含まれており，TinyImages と異なり画像の質が統制されている点，高解像度の画像（400×350 程度）を扱っている点で異なる．画像検索エンジンの検索精度は約 10%であるという知見[30] から，各カテゴリの単語と WordNet の上位単語やその単語に対応する複数の言語を画像検索エンジンに入力して各カテゴリあたり 1 万枚程度（目標収集画像数の 10 倍）の画像を収集する．その画像群から Amazon Mechanical Turk（AMT）[⋆4]を通じ，人力でカテゴリに属する適切な画像を選別して質の高いデータセットを構築している．

ARISTA[35] は 20 億枚の画像を有し，論文で情報が公開されている画像データセットの中で最大である．この論文で，数百万枚のデータセットは Web 画像の一部を表現しているに過ぎず，人の生活になじみの深い絵画，有名人，映画や商品のカテゴリが欠落している点や，WordNet を基盤としたデータセットは，これらのカテゴリを含まないために日常的に利用される画像を反映できていない点を指摘している．オントロジとして Open Dictionary Project[⋆5] の方が Web 検索を反映していると述べている．さらに，従来ノイズとして扱ってきたほぼ類似な画像群（Near Duplicate Image, NDI）のアノテーションにおける重要性を指摘してる点は興味深い．人々が関心を集める対象は NDI を持ちやすく，NDI とそれに付随するタグが大量に得られればその画像に関する有益なタグのみ抽出可能となり，関心を集める画像のアノテーション性能を向上させることができる．

Visual Synset[⋆6][32] は現在公開されているラベルが付与されたデータセットの中で最大である．2 億枚の画像と 30 万のラベルが付与されている．このデータセットでは見た目が似ていてかつ意味的に関連する Visual Synset と呼ばれる画像群で構成されている．各 Visual Synset には複数のラベルが重み付きで付与されている．データの構築には，始めに画像をクラスタリングし Visual Synset を構成する．次に各クラスタに属する画像に付与されたラベル群に対して TF-IDF によって重みを計算し，Visual Synset のラベル群と各重みを決定している．Visual Synset により定義された概念クラスは単一ラベルで定義された概念クラスよりも容易に学習可能であることを示している．

TinyImages と ARISTA の主張は，大規模な画像データセットがあれば複雑な機械学習手法を利用せずともセマンティックギャップ[29] を回避できる点にある．セマンティックギャップとは画像から得られる特徴と画像に映し出されている意味との間に存在するギャップであり，長い間解決されていない問題である．TinyImages や ARISTA のアプローチは大規模な画像があれば画像特徴間の類似度がそのまま画像間の意味類似度に近づいていくというアイデアに基づく．一方，ImageNet は高品質な画像のオントロジ作成が目的である．このために WordNet とクラウドソーシングによるラベルのクリーニングを利用しているが，固定されたオントロジとクラウドソーシングのコストが問題となる．これに対して Visual Synset ではこれらを問題を回避し自動的にデータセットが構築可能であると主張している．

2.2 人に尋ねる

高品位のラベル付きデータセットの構築を行う最も確実な方法は，ラベルが不明なデータを人に尋ねて解決してもらうことである．近年のクラウドソーシングの発展により画像へのラベル付与や誤ったラベルのクリーニングが比較的容易に行えるようになってきたが，依

⋆1 http://www.flickr.com/
⋆2 http://groups.csail.mit.edu/vision/TinyImages/
⋆3 http://www.image-net.org/
⋆4 https://www.mturk.com/mturk/welcome
⋆5 http://www.dmoz.org/
⋆6 http://cpl.cc.gatech.edu/projects/VisualSynset/

然として人手によるラベル付与はコストが高くつくために，クラウドソーシングするデータを極力減らしながらも高品位なデータセットや識別機を構成する枠組みが望まれている．最も効率の良い学習を可能とするデータを選択し，それを人に尋ねながら識別機を学習していく手法は**能動学習（active learning）**と呼ばれている．

Vijayanarasimhan らの研究[33] では，能動学習とクラウドソーシングを融合して自動的に物体検出器を学習する枠組みを提案している．この枠組みの中で高速かつ高性能な物体検出手法と高速かつ適切なクラウドソーシング対象画像の選択手法が示されている．特に後者のクラウドソーシングする画像選択において Hyperplane-hashing[13] が利用されいている．クラウドソーシングするべき画像は識別機の境界面の近くに存在すると考えられる．能動学習では決定境界の超平面を表現する重みは逐次的に更新されるために，まじめに行うと超平面を表現する重みと全てのデータとの内積を更新されるたびに計算する必要がありコストが非常に高い．Hyperplane-hashing では決定境界のパラメータを入力として，決定境界の近いサンプル群がなるべく同じ hash 値になるようにすることで，逐次的に更新される識別機であっても高速にきわどいデータ群を選択できるようになっている．

2.3 対象とは別の知識を活用

認識したい対象の知識がない場合に，他の認識対象の知識を活用することが考えられる．このような枠組みは**転移学習（transfer learning）**，**知識転移（knowledge transfer）**，**ドメイン適合（domain adaptation）**と呼ばれている．転移学習は見たことのない物体の認識に良く用いられるが，見たことのないカテゴリを学習，認識する手法は**ゼロショット学習（zero-shot learning）**と呼ばれている．転移学習は幅広い概念であり統一的な理解は難しいが，物体認識では Rohrbach らの論文[25] においてゼロショット学習のための知識転移手法を系統立てて比較している．この中で 1) 階層構造で表現された各カテゴリ間の関係を活用するもの，2) アトリビュートを活用するもの，3) 階層構造ではなく物体間の直接的な関係性を活用するもの，を知識転移の手法としてあげている．ここで**アトリビュート（attribute）**[10] とは物体カテゴリ間で共有される人間が理解可能な属性のことを言う．アトリビュートの適応先としては知識転移のみならず，物体識別を補助する中間特徴として利用されている．実際にアトリビュートを中間特徴として利用すると物体識別性能が向上することが知られている[31]．また，画像検索の特徴として利用しても検索性能が向上することが報告されている[8]．面白い応用例として，アトリビュートを利用して画像の美しさや魅力を推定する研究がある[7]．

3. 画 像 特 徴

3.1 画像特徴とは

ここでの画像特徴は，1 枚の画像を代表する一つの特徴ベクトルのことを指す．画像特徴ベクトルを取得するプロセスは次の通りである．

(1) 特徴点検出
(2) 特徴点回りの記述：**局所記述子（local descriptor）**の獲得
(3) 特徴空間における局所記述子群のモデル化
(4) 局所記述子群のモデルの代表ベクトルの抽出

上記 (4) のステップで得られる代表ベクトルを画像特徴と呼ぶ．また，局所記述子から代表ベクトルを得るまでのプロセスをここでは**画像表現（image representation）**と呼ぶことにする．(1) の特徴点検出器としては Laplacian of Gaussian （LoG），FAST コーナー検出器[26] などが利用される．また，特別な検出器を利用せずに画像の等間隔なグリッド上の点を特徴点とする場合があり一般物体認識によく利用される．これを**密なサンプリング（dense sampling）**と呼ぶ．局所記述子として，形状情報を表現する輝度勾配ヒストグラムがよく利用される．例えば SIFT 記述子[18]，SURF 記述子[1]，HOG 記述子[5] がある．またテクスチャ情報を表現するものとして Local Binary Patterns（LBP）[21]，カラーヒストグラムなどがある．ここでは特徴点検出器や局所記述子については紹介のみにとどめ，局所記述子が与えられたとして，そこから一枚の画像を表現する特徴ベクトルの獲得過程に焦点を当て詳細に見ていくことにする．

3.2 Bag of Features

一般物体認識を行うには局所記述子同士の「堅い」比較で**類似度（Similarity）**評価を行うのではなく，画像の持つ局所記述子の統計量を画像特徴と見なして類似度評価を行えば「柔らかい」比較が可能となる．局所記述子群から，その統計量を計算する手法として **Bag of Features（BoF）**[4),28] が広く利用されている．BoF は文章特徴である **Bag of Words（BoW）**のアナロジーから生まれた特徴である．BoW は単語の並び順，文法などを考慮しない文書特徴であり，例えば文章中に出てきた単語のヒストグラムが利用される．BoF は訓練集合から代表的ないくつかの局所記述子を取り上げ，画像の中に代表的な局所記述子がいくつ出現するかヒストグラムで表現したものである．BoF は **Bag of Visual Words（BoVW）**とも呼ばれる．BoF の計算プロセスを以下に示す．

(1) 訓練画像群から各画像に対して局所記述子を抽出する．

(2) 全ての局所記述子から K 個の代表的な局所記述子を選択する．選択した代表的な局所記述子を**コードワード（codeword）**と呼び，選択されたコードワードの集合を**コードブック（codebook）**と呼ぶ．コードワードにそれぞれ $w_1, \ldots, w_K$ とラベルを付与する．コードブックは**辞書（dictionary）**とも呼ばれる．

(3) 全ての局所記述子をいずれかのコードワードに対応させる．この操作により，全ての局所記述子に $w_1, \ldots, w_K$ のラベルが付与される．

(4) 各画像において，コードワードに関するヒストグラムを計算する．つまり，ある画像に w_k とラベル付与された局所記述子の数をカウントする．コードワードのヒストグラムをその画像の特徴ベクトルとする．つまり特徴ベクトルの次元は K となる．

上記の (2)，(3) のステップで局所記述子群から代表的な局所記述子の選択，局所記述子群を代表的な局所記述子への割り当てを行っているが，BoF のフレームワークでは K-means を利用することが多い．BoF を利用したカテゴリ識別手法では**カーネル法（kernel method）**と**サポートベクトルマシン（Support Vector Machine, SVM）**の組み合わせが広く利用されている．カーネルとして，**カイ 2 乗カーネル（χ^2 kernel）**や**インターセクションカーネル（intersection kernel）**が BoF との相性がいいことが実験的に示されている．

3.3 混合ガウス分布による Bag of Features の改良

BoF においてコードワードを作成する目的はいくつかある．

- 類似した局所記述子を最近傍のコードワードに割り当てることで局所記述子の表現にある程度のロバストネスを持たせる．
- 全ての画像に対して，局所記述子を同じコードブックに適用することで同じ長さの特徴ベクトルを得る．
- 局所記述子間の幾何学的関係は視点依存であるので，局所記述子の相対的な位置関係を無視することで識別のロバストネスが向上できる．

さらに BoF によりテキスト分類のテクニックをそのまま画像分類に適用できる．

しかしながら BoF には，識別性能がコードワードの選び方に依存する．また，局所記述子のヒストグラムを特徴空間における局所記述子の確率密度分布推定と考えると，ヒストグラムによる表現は粗い推定と言える．よって確率密度分布推定をより正確に行えば識別性能の向上につながると考えられる．そこで**混合ガウス分布（Gaussian Mixture Model, GMM）**を用いることで，BoF 表現を改善する試みが行われている[9]．

混合ガウス分布はガウス分布の線形重ね合わせで書ける．

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{x}), \tag{1}$$

ここで $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), p_k(\boldsymbol{x})$ は**混合要素（mixture component）**であり，平均 $\boldsymbol{\mu}_k$ と分散 $\boldsymbol{\Sigma}_k$ を持つ．π_k は混合係数である．混合ガウス分布のパラメータ集合を $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$，データ集合 $\mathcal{X} = \{\boldsymbol{x}_n \in R^D\}_{n=1}^{N}$ とすると，混合ガウス分布の**最尤法（maximum likelihood estimation）**によるパラメータ推定は次のように求められる．

$$\boldsymbol{\theta}_{ml} = \arg\max_{\theta} \log p(\mathcal{X}|\boldsymbol{\theta}) = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}). \tag{2}$$

この対数尤度関数を最大化するパラメータは閉形式の解析解で得られないために **EM アルゴリズム**を用いてパラメータを求める．混合ガウス分布のための EM アルゴリズムは以下の通りである．

E-step

$$\gamma_n(k) = p(k|\boldsymbol{x}_n, \boldsymbol{\theta}^{(t)}) = \frac{\pi_k p_k(\boldsymbol{x}_n)}{\sum_{j=1}^{K} \pi_j p_j(\boldsymbol{x}_n)}. \tag{3}$$

M-step

$$\pi_k^{(t+1)} = \frac{N_k}{N}, \tag{4}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_n(k) \boldsymbol{x}_n, \tag{5}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_n(k) (\boldsymbol{x}_n - \boldsymbol{\mu}_k^{(t+1)})(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{(t+1)})^\top, \tag{6}$$

ここで $N_k = \sum_{n=1}^{N} \gamma_n(k)$ であり，事後確率 $\gamma_n(k) = p(k|\boldsymbol{x}_n, \boldsymbol{\theta}^{(t)})$ は**負担率（responsibility）**とも見なすことができる．

混合ガウス分布を利用するメリットとして，BoF は局所記述子とコードワードへの距離が単なるユークリッド距離で計量されるが，混合ガウス分布を構成する各ガウス分布がそれぞれ共分散を持つために共分散を考慮した距離計量を利用できることがあげられる．また，BoF は局所記述子が一つのコードワードのみに割り当てられるが，混合ガウス分布では局所記述子と多くのコードワードとの関係を表現できるので，特徴空間における局所記述子の

位置に関する情報をエンコードできるメリットもある．しかしながら，デメリットして混合ガウス分布表現は BoF と比較してパラメータが多い．混合ガウス分布は $\mathcal{O}(K(D^2/2+D))$ のパラメータ数であるが，BoF は $\mathcal{O}(KD)$ ですむ．そのため，混合ガウス分布は訓練データに対して**過剰適合（overfitting）**する可能性があり，混合ガウス分布の学習時に**正則化（regularization）**を行う必要がある．そこで混合ガウス分布のパラメータに関する事前知識を導入し，**事後確率最大化（maximum a-posterior, MAP）**によりパラメータを求める手法が提案されている[9)23)]．

BoF は局所記述子がただ一つのコードワードと関連し，混合ガウス分布では局所記述子が全てのコードワードと関連した表現となっている．しかしながら BoF のように一つのコードワードへの割り当てでは量子化誤差が大きく，類似した局所記述子であっても量子化後の符号が異なる可能性がある．また，全てのコードワードと関連づける方法は多くの関連のないコードワードとの関連性も表現するためにコードワード数が増えた場合，局所記述子の顕著なパターンを捉えにくくなる．そこで，局所記述子を少数のコードワードのみで表現する**スパースコード化（sparse coding, SC）**[36)] が提案されている．これにより上記の問題を解決しつつ，ベクトル量子化よりも量子化誤差を低減可能である．

さらに，データを近傍に存在するいくつかの基準点の線形和で局所的に近似する手法が提案されている[37)]．この結果，得られた線形和の重みはデータの**局所座標符号化（local coordinate coding, LCC）**と呼ばれる．この論文中で，ある仮定の下では局所性がスパースネスよりも本質であると述べている．しかしながらスパース符号化と同じように，LCC も L1 ノルム最適化問題を解く必要があり，計算コストが高い問題を抱える．そこで，LCC の高速な実装と見なせる**局所制約線形符号化（locality-constrained linear coding, LLC）**[34)] が提案されている．スパース符号化ではコードワードが過剰であるためにスパース性を優先することで類似した局所記述子に対して全く異なるコードワードを選択する可能性があるが，局所制約線形符号化は類似した局所記述子には類似したコードを出力可能である．

3.4 フィッシャーベクトル

局所記述子の混合ガウス分布を用いた正確な確率密度分布推定による BoF の改良を述べた．混合ガウス分布は**生成モデル（generative model）**と見なせるが，生成モデルを判別的なアプローチに適応可能なより洗練された手法があれば識別性能の改善につながるはずである．**フィッシャーカーネル（Fisher kernel）**[12)] は**生成的アプローチ（generative approach）**と**判別的アプローチ（discriminative approach）**を結合させる強力な枠組みである．フィッシャーカーネルでは，まず局所記述子を生成する確率密度分布から導出される勾配ベクトルを計算し，画像を表現する一つの特徴ベクトルとする．そしてこの特徴ベクトルを分類機に入力する．

BoF と比較してフィッシャーカーネルを利用するメリットは，コードブックサイズが同じであればフィッシャーカーネルの方がより要素数の多い特徴ベクトルが得られる点にある．つまり，特徴ベクトルの表現する情報が多いため計算コストの高いカーネル法を利用して高次元空間へ射影する必要がなく，線形識別機でも十分な識別性能を出すことが可能となる．

ここで，u_θ をあらゆる画像の内容を表現する**確率密度関数（probability density function, pdf）**とし $\boldsymbol{\theta}$ を確率密度関数のパラメータとする．局所記述子群を $\mathcal{X}$ とすると，このデータを次に示す勾配ベクトルで表現する．

$$G_\theta^{\mathcal{X}} = \frac{1}{N}\nabla_\theta \log u_\theta(\mathcal{X}|\boldsymbol{\theta}). \tag{7}$$

対数尤度の勾配はデータに最も適合するように確率密度関数のパラメータが修正すべき方向を表現している．また異なるデータサイズの $\mathcal{X}$ をパラメータ数に依存した決まった長さの特徴ベクトルに変換する．

この勾配ベクトルは様々な識別機に利用できるが，内積を利用する識別機ではベクトルを適切な計量を用いて正規化する必要がある．この正規化には**フィッシャー情報行列（Fisher information matrix）**が利用できる．

$$F_\theta = E_X[\nabla_\theta \log u_\theta(\mathcal{X}|\boldsymbol{\theta})\nabla_\theta \log u_\theta(\mathcal{X}|\boldsymbol{\theta})^\top]. \tag{8}$$

フィッシャー情報行列を用いて正規化された勾配ベクトルは次のように与えられる．

$$\mathcal{G}_\theta^X = F_\theta^{-1/2}\nabla_\theta \log u_\theta(\mathcal{X}|\boldsymbol{\theta}). \tag{9}$$

このようにしてできた画像の特徴ベクトルを局所記述子群 $\mathcal{X}$ の**フィッシャーベクトル（Fisher vector）**と呼ぶ[22)]．計算コストの観点からフィッシャー情報行列を単位行列と近似する場合もあるが，[22)] では対角行列として近似している．

フィッシャーカーネルをコードブックに適応するにあたり，局所記述子の特徴空間における確率密度分布を混合ガウス分布で表現する．一枚の画像から得られる局所記述子の集合を $\mathcal{X}$ とする．$\gamma_n(k)$ を式 (3) で示した局所記述子 $\boldsymbol{x}_n$ が k 番目のコンポーネントから生成される確率とする．この時，対数尤度 $\mathcal{L}(\mathcal{X}|\boldsymbol{\theta}) = \log u_\theta(\mathcal{X}|\boldsymbol{\theta})$ の微分は以下のようになる．

$$\frac{\partial \mathcal{L}(\mathcal{X}|\boldsymbol{\theta})}{\partial \pi_k} = \sum_{n=1}^{N} \left[\frac{\gamma_n(k)}{\pi_k} - \frac{\gamma_n(1)}{\pi_1} \right], \tag{10}$$

$$\frac{\partial \mathcal{L}(\mathcal{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k^d} = \sum_{n=1}^{N} \gamma_n(k) \left[\frac{\boldsymbol{x}_n^d - \boldsymbol{\mu}_k^d}{(\boldsymbol{\sigma}_k^d)^2} \right], \tag{11}$$

$$\frac{\partial \mathcal{L}(\mathcal{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}_k^d} = \sum_{n=1}^{N} \gamma_n(k) \left[\frac{(\boldsymbol{x}_n^d - \boldsymbol{\mu}_k^d)^2}{(\boldsymbol{\sigma}_k^d)^3} - \frac{1}{\boldsymbol{\sigma}_k^d} \right], \tag{12}$$

ここで，ベクトルの上付き文字 d はベクトルの d 番目の要素を示す．また混合ガウス分布の共分散行列は対角行列（$\boldsymbol{\sigma}_k^2 = \mathrm{diag}(\boldsymbol{\Sigma}_k)$）と仮定している．

フィッシャー情報行列を対角行列と仮定し，$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \pi_k}$，$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \mu_k^d}$，$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \sigma_k^d}$ のそれぞれに対応するフィッシャー情報行列の要素を f_{π_k}，$f_{\mu_k^d}$，$f_{\sigma_k^d}$ とすると，これらは次に示すように閉じた解として近似的に求められる．

$$f_{\pi_k} = N\left(\frac{1}{\pi_k} + \frac{1}{\pi_1}\right), \quad f_{\mu_k^d} = \frac{N\pi_k}{(\boldsymbol{\sigma}_k^d)^2}, \quad f_{\sigma_k^d} = \frac{2N\pi_k}{(\boldsymbol{\sigma}_k^d)^2}. \tag{13}$$

BoF や混合ガウス分布による特徴ベクトルは負担率を用いて

$$\boldsymbol{f} = \frac{1}{N}\sum_{n=1}^{N}[\gamma_n(1), \cdots, \gamma_n(K)]^\top \in R^K. \tag{14}$$

と表現できるが，これは混合比が一定の仮定を設けると式 (10) に示すフィッシャーベクトルの 0 次の統計量と同じとなる．一方フィッシャーベクトルは 0 次だけではなく平均（1 次），分散（2 次）の統計量を考慮している．コードブックのサイズを K とすると，BoF は K 次元のベクトルとなるがフィッシャーベクトルは $(2d+1)K-1$ 次元となる．つまりフィッシャーベクトルは小さなコードブックサイズで豊かな表現が可能となる．

3.5 フィッシャーベクトルの改良

フィッシャーベクトルは BoF と比較して画像を豊かに表現しているにも関わらず，そのまま画像識別に利用しても BoF とさほど性能に差がない．そこで[24] ではフィッシャーベクトルの改良を提案している．

ここで一枚の画像から得られた局所記述子群 $\mathcal{X} = \{\boldsymbol{x}_n\}_{n=1}^N$ は確率密度分布を $p(\boldsymbol{x})$ に従っているとする．十分大きな N のとき大数の法則から式 (7) は以下のように近似できる．

$$G_\theta^{\mathcal{X}} \approx \nabla_\theta \int_x p(\boldsymbol{x}) \log u_\theta(\boldsymbol{x}) d\boldsymbol{x}. \tag{15}$$

確率密度分布 $p(\boldsymbol{x})$ を画像に依存しない分布 $u_\theta(\boldsymbol{x})$ と画像に特定の分布 $q(\boldsymbol{x})$ に分解する．

$$p(\boldsymbol{x}) = \omega q(\boldsymbol{x}) + (1-\omega)u_\theta(\boldsymbol{x}), \tag{16}$$

ここで ω は 0 から 1 の間の値を取るパラメータである．式 (16) を式 (15) に代入する．

$$G_\theta^{\mathcal{X}} \approx \omega\nabla_\theta \int_x q(\boldsymbol{x}) \log u_\theta(\boldsymbol{x}) d\boldsymbol{x} + (1-\omega)\nabla_\theta \int_x u_\theta(\boldsymbol{x}) \log u_\theta(\boldsymbol{x}) d\boldsymbol{x}. \tag{17}$$

パラメータ $\boldsymbol{\theta}$ は最尤法によって求められているとすると式 (17) の右辺第二項はゼロと見なせるので，結局，

$$G_\theta^{\mathcal{X}} \approx \omega\nabla_\theta \int_x q(\boldsymbol{x}) \log u_\theta(\boldsymbol{x}) d\boldsymbol{x}, \tag{18}$$

となるため，フィッシャーベクトルを利用すると画像に依存しない多くの部分を無視することができる．しかしながら ω の値の大小により $G_\theta^{\mathcal{X}}$ の値が変化する．これは前景と背景の割合によって $G_\theta^{\mathcal{X}}$ の値が変化することを意味する．そのために[24] では $G_\theta^{\mathcal{X}}$ の **L2 正規化（L2 normalization）**によって ω の影響を排除する手法を提案している．

また，混合ガウス分布の混合数 K を増加させるとフィッシャーベクトルがスパースになる現象が観測されている．これは混合数の増加により局所記述子が複数のコードワードと接近するため，大きな負担率 $\gamma_n(k)$ を持つ局所記述子が少なくなることによる．その結果，フィッシャーベクトルの要素はゼロ近くの頻度が高くなる．L2 正規化により得られたベクトルの内積は L2 距離と同じであるが，スパースなベクトルに L2 距離を適用しても高い識別性能が得られないことが知られている[20]．このときの対処方法として，カーネル法の利用が考えられるが，一般にカーネル法は計算コストが高く大規模データへの適応は難しい．[24] では**パワー正規化（power normalization）**によりスパースネスを緩和することで内積による類似度を維持する手法を提案している．具体的にはフィッシャーベクトルの各要素 z に $f(z) = \mathrm{sign}(z)|z|^\alpha$ を適用する．ここで α は正規化のためのパラメータである．正規化の順序はパワー正規化を行った後に L2 正規化を適用している．

三番目のフィッシャーベクトルの改良点として**空間ピラミッド（spatial pyramid）**[16] の適応を行っている．空間ピラミッドは画像を一定間隔のグリッドに分割し，分割されたセル内で画像特徴を求める．分割の粒度（レベル）を変えて得られた全てのセルの画像特徴をつなげて一つの特徴ベクトルとする手法である．[24] では画像を 1×1，2×2，3×1 の合計 8 個のセルに分割し，8 個のフィッシャーベクトルを計算している．

フィッシャーベクトルと関連の深い画像表現として **VLAD (vector of locally aggregated descriptors)**[15] や**スーパーベクトル符号化（super vector coding）**[38] がある．

フィッシャーベクトルの平均の関する項のみを利用すれば VLAD，負担率と平均の項を利用するとスーパーベクトル符号化とほぼ等価となる．2011 年度 TRECVID の Semantic indexing でトップになった手法は，GMM の高速な MAP 適合によって得られる **GMM supervectors**[3] を利用してる[11]．GMM supervectors はフィッシャーベクトルの平均成分とほぼ等価である．

4. 識 別 機

識別対象のクラスが増えると低次元の特徴ベクトルでは十分な識別性能が得られなくなるため特徴ベクトルの次元は増大させる必要がある．実際に Sánchez らの論文[27] において，高い識別率を維持するためには特徴ベクトルの高次元化が重要であることを実験的に示している．特徴ベクトルが高次元であっても破綻しない識別機が必要となるため，次元の呪いにかからないと言われている SVM が一般に用いられる．クラス数が増えた場合，通常の 2 クラス識別機である SVM を多クラスに拡張する必要があるが，ほとんどの場合 1-vs-all SVM が利用される．1-vs-all SVM は各クラス識別機の出力値の大小関係が正規化されていないために多クラス識別問題において適切である保証はないが，多くの物体認識研究で十分に高い性能が得られることが報告されている．また，1-vs-all SVM は他のクラス識別機との調整を取る必要がないため，各クラスの識別機の学習と識別を並列実行することが可能であり，大規模データへの適応が容易に行えるメリットがある．

大規模データを**一括学習（batch learning）**するにはメモリの問題や追加学習の困難さもあり利用しにくいため，**逐次学習（sequential learning）**が用いられる．物体認識では，SVM の逐次学習バージョンとして**確率的勾配降下法（stochastic gradient descent method，SGD method）**がよく利用される[2]．重み $\boldsymbol{w}$，バイアス b の線形識別機を $y = \boldsymbol{w}^\top \boldsymbol{x} + b$，ラベル付き訓練画像のペア $\{\boldsymbol{x}_t, y_t\}_{t=1}^T$ とすると SVM のコスト関数は次式のように表わされる．

$$L = \sum_{t=1}^{T} L(\boldsymbol{w}, b, \boldsymbol{x}_t, y_t) = \sum_{t=1}^{T} \frac{\lambda}{2} ||\boldsymbol{w}||^2 + \max\left[0, 1 - y_t(\boldsymbol{w}^\top \boldsymbol{x}_t + b)\right], \tag{19}$$

ここで λ は正則化パラメータである．このコスト関数に確率的勾配降下法を適用した時の重みとバイアスの更新式は次のようになる．

$$\boldsymbol{w}^t = \boldsymbol{w}^{t-1} - \eta \nabla_w L(\boldsymbol{w}, b, \boldsymbol{x}_t, y_t), \;\; b^t = b^{t-1} - \eta \nabla_b L(\boldsymbol{w}, b, \boldsymbol{x}_t, y_t). \tag{20}$$

標準の SGD では収束に時間がかかるために重みとバイアスに対して平均化のスキームを取り入れて収束の高速化が行われる[17]．

$$\bar{\boldsymbol{w}}^t = (1 - 1/t)\bar{\boldsymbol{w}}^{t-1} + \boldsymbol{w}^t/t, \;\; \bar{b}^t = (1 - 1/t)\bar{b}^{t-1} + b^t/t. \tag{21}$$

逐次学習により大規模データへの適応が可能になる．しかしながら，データが高次元かつ膨大になると上記の重みの更新式よりも，ハードディスクなどの補助記憶装置から主メモリへのデータ転送時間が学習速度のボトルネックとなる．2010 年度の大規模画像認識コンペティション（ILSVRC2010）[⋆1]でトップの成績を残した NEC laboratory America のチームは，Hadoop[⋆2]を利用して 1-vs-all SVM を並列計算させているが，メインメモリにロードした学習サンプルをなるべく多くの識別機の学習に同時利用することで，データのロード回数を減らし学習高速化を実現している[17]．ILSVRC2011[⋆3]でトップの成績を残した Xerox Europe のチームは，**直積量子化（product quantization, PQ）**[14] を用いてデータを圧縮し，全ての訓練データをメインメモリに蓄積し，学習時には圧縮されたデータを復号化して重みを更新することで学習の高速化を実現している[27]．

このように当然ではあるが，大規模データを利用した認識システムの構築には計算機システムのアーキテクチャも強く意識したアルゴリズム作りが大切となる．

5. ま と め

本論文では大規模データを用いた物体・シーン認識のトレンドについて概観した．次に，最先端の物体認識システムのパイプラインを説明し，データ，特徴抽出，識別機の順で高い性能が要求されることを述べ，各構成要素を説明した．特に大規模データを扱うには高い識別率を備えるだけではなく，スケーラビリティが重要となる．スケーラビリティを維持するために線形識別機を用いることが一般的であるが，線形識別機であっても十分な識別能力を発揮するためには画像特徴をどのように表現するかが鍵となり，その詳細を解説した．

参 考 文 献

1) Bay, H., Ess, A., Tuytelaars, T. and Gool, L.V.: SURF: Speeded Up Robust Features, *CVIU*, Vol.110, No.3, pp.346–359 (2008).
2) Bottou, L.: Large-Scale Machine Learning with Stochastic Gradient Descent, *COMPSTAT* (2010).
3) Campbell, W.M., Sturim, D.E. and Reynolds, D.A.: Support Vector Machines Us-

⋆1 http://www.image-net.org/challenges/LSVRC/2010/
⋆2 http://hadoop.apache.org/
⋆3 http://www.image-net.org/challenges/LSVRC/2011/

ing GMM Supervectors for Speaker Verification, *IEEE Signal Processing Letters*, Vol.13, No.5, pp.308 – 311 (2006).
4) Csurka, G., Dance, C.R., Fan, L., Willamowski, J. and Bray, C.: Visual Categorization with Bags of Keypoints, *ECCV International Workshop on SLCV* (2004).
5) Dalal, N. and Triggs, B.: Histograms of Oriented Gradients for Human Detection, *CVPR* (2005).
6) Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database, *CVPR* (2009).
7) Dhar, S., Ordonez, V. and Berg, T.L.: High Level Describable Attributes for Predicting Aesthetics and Interestingness, *CVPR* (2011).
8) Douze, M., Ramisa, A. and Schmid, C.: Combining attributes and Fisher vectors for efficient image retrieval, *CVPR* (2011).
9) Farquhar, J., Szedmak, S., Meng, H. and Shawe-Taylor, J.: Improving "bag-of-keypoints" image categorisation: Generative Models and PDF-Kernels, Technical report, University of Southampton (2005).
10) Ferrari, V. and Zisserman, A.: Learning Visual Attributes, *NIPS* (2007).
11) Inoue, N. and Shinoda, K.: A Fast MAP Adaptation Technique for GMM-supervector-based Video Semantic Indexing, *ACM Multimedia* (2011).
12) Jaakkola, T. and Haussler, D.: Exploiting Generative Models in Discriminative Classifiers, *NIPS* (1998).
13) Jain, P., Vijayanarasimhan, S. and Grauman, K.: Hashing Hyperplane Queries to Near Points with Applications to Large-Scale Active Learning, *NIPS* (2010).
14) Jégou, H., Douze, M. and Schmid, C.: Product Quantization for Nearest Neighbor Search, *IEEE Trans. on PAMI*, Vol.33, pp.117–128 (2011).
15) Jégou, H., Douze, M., Schmid, C. and Pérez, P.: Aggregating local descriptors into a compact image representation, *CVPR* (2010).
16) Lazebnik, S., Schmid, C. and Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *CVPR* (2006).
17) Lin, Y., Lv, F., Zhu, S., Yang, M., Cour, T., Yu, K., Cao, L. and Huang, T.: Large-scale Image Classification: Fast Feature Extraction and SVM Training, *CVPR* (2011).
18) Lowe, D. G.: Distinctive image features from scale-invariant keypoints, *IJCV*, Vol.60, No.2, pp.91–110 (2004).
19) Miller, G.A.: WordNet: A Lexical Database for English, *Communications of the ACM*, Vol.38, No.11, pp.39–41 (1995).
20) Nistér, D. and Stewénius, H.: Scalable Recognition with a Vocabulary Tree, *CVPR* (2006).
21) Ojala, T., Pietikäinen, M. and Harwood, D.: Performance Evaluation of Texture Measures with Classification Based on Kullback Discrimination of Distributions, *ICPR* (1994).
22) Perronnin, F. and Dance, C.: Fisher Kernels on Visual Vocabularies for Image Categorization, *CVPR* (2007).
23) Perronnin, F., Dance, C., Csurka, G. and Bressan, M.: Adapted vocabularies for generic visual categorization, *ECCV* (2006).
24) Perronnin, F., Sánchez, J. and Mensink, T.: Improving the Fisher Kernel for Large-Scale Image Classification, *ECCV* (2010).
25) Rohrbach, M., Stark, M. and Schiele, B.: Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting, *CVPR* (2011).
26) Rosten, E. and Drummond, T.: Machine learning for high-speed corner detection, *ECCV* (2006).
27) Sánchez, J. and Perronnin, F.: High-Dimensional Signature Compression for Large-Scale Image Classification, *CVPR* (2011).
28) Sivic, J. and Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos, *ICCV* (2003).
29) Smeulders, A. W.M., Worring, M., Santini, S., Gupta, A. and Jain, R.: Content-based image retrieval at the end of the early years, *IEEE Trans. on PAMI*, Vol.22, No.12, pp.1349–1380 (2000).
30) Torralba, A., Fergus, R. and Freeman, W.T.: 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition, *IEEE Trans. on PAMI*, Vol.30, No.11, pp.1958–1970 (2008).
31) Torresani, L., Szummer, M. and Fitzgibbon, A.: Efficient Object Category Recognition Using Classemes, *ECCV* (2010).
32) Tsai, D., Jing, Y., Liu, Y. and A.Rowley, H.: Large-Scale Image Annotation using Visual Synset, *ICCV* (2011).
33) Vijayanarasimhan, S. and Grauman, K.: Large-Scale Live Active Learning: Training Object Detectors with Crawled Data and Crowds, *CVPR* (2011).
34) Wang, J., Yang, J., Yu, K., Lv, F., Huang, T. and Gong, Y.: Locality-constrained Linear Coding for Image Classification, *CVPR* (2010).
35) Wang, X.-J., Zhang, L., Liu, M., Li, Y. and Ma, W.-Y.: ARISTA - Image Search to Annotation on Billions of Web Photos, *CVPR* (2010).
36) Yang, J., Yu, K., Gong, Y. and Huang, T.: Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification, *CVPR* (2009).
37) Yu, K., Zhang, T. and Gong, Y.: Nonlinear Learning using Local Coordinate Coding, *NIPS* (2009).
38) Zhou, X., Yu, K., Zhang, T. and Huang, T.S.: Image Classification using Super-Vector Coding of Local Image Descriptors, *ECCV* (2010).