Web 上の画像知識を用いた単語の視覚性に関する研究　(継続)

研究代表者　柳　井　啓　司　　電気通信大学情報工学科准教授
研究分担者　コーバス・バーナード　　米国アリゾナ大学コンピュータサイエンス学科准教授

1 はじめに

計算機によって，デジタル画像にその画像が表す内容に対応する単語を付与するのが画像認識の目的の 1 つであるが，その際に用いる単語を選ぶ方法は，従来は人手による選択で，定量的な指標は存在しなかった．単語には，例えば「ライオン」「山」というように視覚的特徴に直接対応しているものと，「動物」「乗り物」の様に様々な動物，乗り物が存在するために視覚的特徴と単語概念を直接結び付けることが困難であるものが存在し，直接的に画像認識に用いる単語は前者の様な視覚的特徴と結び付きの強いものであることが望まれる．従来の研究では認識できる単語の数が 10 個から多くても 100 個程度であったので，人手による選択以外の方法はまったく検討されなかったが，今後，実用的な認識を実現するためには，数千個，数万個オーダーでの単語を画像に自動的に付与することが求められる．そのためには，認識すべき単語を自動的に選び出す方法が必要である．そこで，昨年度，本研究では単語が表わす「概念」がどの程度，視覚的特徴を持ち合わせているかを定量的に評価する指標「画像領域エントロピー」を提案し，実験では，150 個の形容詞に関連した画像を Google Image Search を用いて各形容詞につき 300 枚，合計 45,000 枚収集し，それぞれの形容詞について画像領域エントロピーを求め，どの単語が画像認識に適しているか調査した．その際に，単語によっては必ずしも提案した「画像領域エントロピー」によってその単語の視覚性が評価されていないと言う問題点があった．たとえば，色を表す形容詞は，視覚性が高いことが予想されるにもかかわらず，視覚性が低いと判定されることもあった．

そこで，継続研究として，本年度は，Web から収集した画像からノイズを取り除く研究を重点的に行った．本論文では，Web 画像収集において，最新の画像認識手法を用いてノイズ画像を除去する方法について述べる．具体的には，我々が従来から研究している Web 画像収集システムに，局所特徴量のヒストグラムによって画像を表現する bag-of-keypoints アプローチを導入した結果について本論文では報告する．我々のシステムでは，(1)キーワードに関係する Web ページの HTML を収集・解析し，キーワードを表す可能性が高い画像のみを収集し，(2)HTML 解析によって評価の高かった画像を初期学習画像として一般物体認識手法を適用することによって画像選別を行う．

画像を局所特徴の集合として表現する bag-of-keypoints 表現[14]が近年，一般画像認識の研究において注目されている．Bag-of-keypoints では，一枚の画像から数百 数千個の局所特徴量を抽出し，ベクトル量子化によって出現特徴の分布をヒストグラム化して，そのヒストグラムを画像の特徴ベクトルとして利用する．ヒストグラムであるので，局所特徴の位置情報は利用しないが，物体認識においては従来の方法に比べて高い識別能力があることが示されている．一方，Bag-of-keypoints を改良する方法として，bag-of-keypoints に特徴分布の大まかな位置情報を埋め込む spatial pyramid matching[16]が提案されている．

そこで，bag-of-keypoins[14]及び spatial pyramid matching[16]を，Web 画像収集システム[12,13]に導入し，その効果について検証した結果を本論文では述べる．実験では，spatial pyramid matching を利用した結果が最もよい結果を示し，従来の領域表現とガウス混合分布による確率生成モデルを用いた結果よりも，10 単語による実験で平均 10 ポイント以上精度が向上した．最終的には，HTML 解析による仮学習画像の推定と，一般物体認識手法による画像認識によって，Web 画像検索に比べて平均 20 ポイント以上も精度が改善することが示された．

本論文の構成について簡単に述べる．2 章では関連研究および本研究の背景，3 章では Web 画像収集の処理の流れについて述べ，4 章では bag-of-keypoints および spatial pyramid matching について説明する．5 章では画像収集実験を行い，従来手法との比較を行う．最後に 6 章で全体のまとめを行う．

2 背景

2-1 画像アノテーション

近年，一般的な画像に対して，その画像が表す内容に対応する単語を付ける画像アノテーションの研究が盛んに行われている．古くは，画像をブロック部分領域に機械的に分割してそれぞれの部分領域の特徴量と名詞単語の関連付けを行った Photobook[1]の研究から始まり，近年では数個の単語が付与された画像を学習データとして，画像と単語の対応を学習する方法が多く研究されている．その代表的な研究が，森らの研究[2,3]および，本論文の第２著者のグループによる研究[4,5,6]である．これらの研究は，1980 年代に盛んに行われた領域分割とラベリングによる画像認識[7]とは，学習データから確率モデルを自動学習する点で大きく異なっている．

森らの研究[2,3]では，百科辞典中の画像と説明文から画像の部分領域と単語の対応を自動的に学習する．この研究では，１つの画像に複数個の単語を持たせて，学習画像の部分領域を特徴量に関してベクトル量子化の方法によってクラスタリングし，各クラスタについて各単語の出現確率を予め求めておく．そして，テスト画像の各部分領域について，最も近いクラスタの単語出現確率の平均値の上位の単語がテスト画像の関連単語ということとしている．しかし，百科事典の扱う対象があまりにも広範囲に渡っているために，良好な精度が得られているとは言い難い．同じ手法を Web から収集したテキストと画像に対して行った研究[8]もある．

本論文の第２著者のグループによる研究[4,5,6]では，予め画像に数個のキーワードが付けられている Corel 画像データベースを用いて，画像へのアノテーションを行った．森らの研究とは異なり，単純なブロック分割ではなく，領域分割アルゴリズムを用いて領域分割し，領域毎の特徴量を利用して画像全体と単語の対応付けを行った．さらに，文献[5]では，画像と単語の対応のみで，領域と単語の対応付けがされていない学習データを用いて，領域分割された各画像領域と単語の対応付けを統計的に推定する手法を提案した．文単位で対応付けがされている２か国語で書かれた大量の文書(対訳コーパス)のみから，事前に辞書も文法の知識なしに確率モデルによって辞書と文法を自動的に学習し機械翻訳を行う統計的機械翻訳[9]の手法を画像に応用して，画像領域と単語の自動対応付けを実現した．領域分割によって画像から切り出したすべての領域を一方の言語で書かれた文，画像に付けられた複数の単語をもう一方の言語で書かれた文とみなし，単語が付与された画像を大量に用意することによって，確率モデル(image translation model)を学習し，画像の部分領域へのアノテーションを実現した．

2-2 単語選択手法の必要性

前節で紹介した研究に共通することは，アノテーションに用いる単語が名詞のみであり，それらは人手で選択されたか，もしくは学習に用いる画像データベースに最初から付けられていた単語を単にそのまま用いていたということである．人手による単語の選択は，単語の種類を多くする際の大きな問題である．我々は現在，単語の種類を増やす方向で研究を進めており，以下の２つの点から画像認識やアノテーションに適した単語を選ぶ方法を必要としている．

(1)名詞に加えて形容詞についてもアノテーションに用いる．

(2)World Wide Web から学習画像データを自動収集する．

(1)の形容詞の導入について，我々は文献[5,6]の translation model に形容詞を導入することを検討中である．例えば，学習データ中の「赤いボール」「赤いりんご」「赤い自動車」などから「ボール」「りんご」「自動車」に加えて「赤い」についても学習し，学習データにない名詞と「赤い」の組合せも認識することの実現を目指している．「赤い」は明らかに画像特徴と結び付きが強く，画像から認識可能であるが，例えば，「固い」「面白い」などは画像から直接認識可能であるかどうかは自明ではない．そこで，数多くある形容詞のうち，どの形容詞が画像から認識可能であるかを知るための客観的な指標が必要である．形容詞を用いた画像認識の一種として，90 年代初めに人間の感性に関連した形容詞(感性語)による画像検索[10,11]が研究されていたが，それらの研究では画像特徴と関係が深いと考えられる形容詞が予め人手によって選ばれていた．

(2)の World Wide Web から学習画像の自動収集では，Web からの画像収集の方法および画像認識への応用[12,13]の方法によって，一般物体認識のための大規模な学習画像データベースの構築することを研究中である．人手によるものと違ってノイズデータも含む不完全な形での学習データではあるが，Web から画像を収集することによって，どのような単語に関連する画像も収集可能であるという特徴がある．今後は 1000 語以上の単語についてデータベースを構築する予定で，どのような単語に対応する画像を Web から集めるべきか

を決める方法が必要である．

(1)や(2)のようなことを行う場合，実際に形容詞を導入して認識を行った結果の認識精度や，Web から画像を収集しそれを用いて学習した結果の認識精度を用いて，その単語の「視覚性」，つまり画像特徴との結び付きの強さを評価することは可能である，しかしながら，認識結果の評価を行う場合，正解データが必要である．Web 画像収集の精度は良くても 8 割程度というのが現状で，結果を評価するには人手による結果のチェック以外に方法はない．例えば，2000 単語について認識率を求める場合，それぞれの単語の認識結果の画像を 50 枚ずつサンプリングして評価するとなると，合計で 10 万枚の画像について人手による結果のチェックを行う必要がある．これを行うことは不可能ではないものの多大なコストが掛かる．

本研究の目的である単語が表す「概念」の「視覚性」の定量的評価はそれ自体が最終目的ではなく，あくまでも(1)や(2)の前処理であるので，「低コスト」であることが第一条件である．そのためには，最初に評価したい単語のリストを用意すれば，後の処理はすべて自動に行われることが望ましい．つまり，正解データを必要とせずに，単語の視覚性を評価することができる方法が必要である．

3 Web 画像収集の流れ

3-1 概要

本章では，我々の Web 画像収集システム ImageCollector[12,13]の処理の流れについて述べる．

ImageCollector は，与えられたキーワードに関係する画像を Web から数百枚程度収集することを目的とする．具体的には，処理は２つのフェーズからなり，テキスト解析および画像収集フェーズ，画像特徴量に基づく画像選択フェーズから成る．(1)最初にテキスト解析フェーズとして，与えられたキーワードに関係する Web ページの URL を Google などのテキスト検索エンジンから獲得し，それらすべての Web ページの HTML ファイルを収集する．そして，HTML ファイルから画像ファイルの URL を抽出し，それを HTML 解析によって A，B，C の３つにランク付けする．A はキーワードとの関係が強く，B は中程度，C は無関係とみなして，A，B にランク付けされた画像のみを実際に収集する．(2)次に画像解析フェーズとして，収集したすべての A，B 画像から画像 1 枚ごとに画像特徴を抽出し，A ランクの画像を仮の正例学習画像として利用して，一般物体認識の手法を用いてキーワードに関係する画像のみを選別する．一般物体認識の手法は教師あり学習であるのが一般的であるので，HTML 解析によって評価の高かった画像を正例学習データとして利用して，B 画像に加えて学習データとして利用した A 画像自体も認識対象画像として画像認識を行う．このように，A 画像も認識することによって，ノイズの除去が実現でき，A 画像自体の精度も向上する．なお，負例学習画像には予めランダムなキーワードで収集してある画像を利用することとする．

初めに Image Collector と同様であるテキスト処理フェーズの説明を簡単に行い，次に bag-of-keypoints を用いた画像処理フェーズの説明を行う．

3-2 画像収集と HTML 解析

ここは従来システムの Image Collector と同じあるので，簡単に説明する．

最初に収集した画像に関係するキーワードをシステムに与える．キーワードには，システムが検索エンジンに渡す検索エンジン用キーワードと HTML 文書の解析時に用いる画像分類用キーワードの２種類がある．

(1)既存の商用テキスト検索エンジンを利用し，ユーザの与えたキーワードに関係する Web ページの URL(Universal Resource Locator)を集める．検索エンジン用キーワードを検索エンジンに送る．

(2)集めた URL が示す Web ページにアクセスして，各 Web ページの HTML 文書を獲得する．

(3)各 HTML 文書に対して HTML タグに基づく解析し，HTML 文書からリンクされている画像ファイルとキーワードとの関係の強さについての評価を，画像ファイルへのリンクタグやタグ周辺のテキストにキーワードがどの程度含まれているかなど調べることによって行い，評価の高いものから順に A,B,C にランク分けする．ここでは，画像分類用キーワードを用いる．

(4)A ランク,B ランクに該当した画像を Web から収集し，それぞれ A 群画像，B 群画像と呼ぶ．C ランクに該当した画像はキーワードと無関係の画像と見なして収集しない．

ここで用いられている HTML タグを利用したキーワードと画像ファイルの関係の強度に関する評価方法は，テキスト検索エンジンにおいてキーワードと HTML 文書との関係の強度の評価に用いられている方法と類似した方法であり，Web 画像検索システムにおいては一般的に用いられてい

る手法である[11].

3-3 Bag-of-Keypoints および SVM による画像選別

画像解析フェーズでは，bag-of-keypoints[14]による特徴ベクトルを画像から生成し，それを SVM に入力して画像選別を行う.

Bag-of-keypoints では局所特徴量のヒストグラムを画像の特徴量として利用する.Bag-of-keypoints アプローチにおいては，分類は SVM(Support Vector Machine)を使うのが一般的であり，本研究においても SVM を分類器として利用する．SVM に代表される教師あり機械学習の手法では学習データが必要であるが，本研究においては対応可能なキーワードを限定せず，処理途中でのユーザのフィードバックも想定しないために確実な学習画像を用意することはできない．そこで，HTML 解析によって高い評価であった少数の画像を，正しくキーワードに対応する画像であると仮定し，正例の初期学習画像として利用する．一方，負例には予めランダムなキーワードで収集してある画像を利用することとする.

SVM は特徴ベクトルを高次元に写像して，高次元空間上で線形分離する超平面を求める判別手法による学習法であるが，実際には特徴ベクトルを高次元空間に写像する代りに，高次元空間での内積を与えるカーネル関数を用意すればよいという「カーネルトリック」と呼ばれる性質がある.カーネル関数は，一般にベクトル同士の類似度を表す関数であり，SVM では問題に応じたカーネル関数を利用することが可能である．そこで，bag-of-keypoints において大まかな位置情報を埋め込むためのカーネル関数 spatial pyramid kernel[16]が提案されている.

本研究では，通常の線形カーネルに加えて，spatial pyramid matching に基づくカーネル関数である spatial pyramid kernel も利用して，Web 画像の選別を行う.

4 Bag-of-Keypoints と Spatial pyramid matching

4-1 Bag-of-Keypoints

Bag-of-keypoints では，一枚の画像から数百から数千個の局所特徴量を抽出し，ベクトル量子化によって出現特徴の分布をヒストグラム化して，そのヒストグラムを画像の特徴ベクトルとして利用する.ヒストグラムであるので局所特徴の位置情報は利用しないという特徴があり，これは統計的言語処理において語順を無視して文章を単語の集合と考える bag-of-words と同様の考え方であるということで，bag-of-keypoints，もしくは bag-of-features と呼ばれている.

物体認識においては，bag-of-keypoints が従来の方法に比べて高い記述能力があることが示されている．局所特徴量の表現には SIFT 記述子(Scale Invariant Feature Transform descriptor)[15]が用いられるのが一般的である.

画像から bag-of-keypoints のヒストグラムベクトルを生成する手順は以下の様になっている.

(1) 各画像について局所特徴量を得るための局所パッチを数百から数千個程度抽出する.
(2) SIFT記述子によって局所特徴量を抽出.
(3) 学習画像から抽出した局所特徴量ベクトルに対して，k-means法によるクラスタリングを実行し，ベクトル量子化におけるコードブックを作成．コードブックの各要素をcodewordもしくはvisual wordと呼ぶこともある.
(4)各画像について，局所特徴量ベクトルを最も近いcodewordに分類し,codewordsのヒストグラムを作成．それを画像を表すbag-of-keypointsベクトルとする.

なお，SIFT は画像からの特徴点抽出も含んだ手法であるが，物体認識においては，SIFT の特徴点抽出は利用せずに，画像の中身と無関係に予め決められた格子点の周りから SIFT 記述子によって局所特徴量を抽出することも行われる．こうした局所特徴量の抽出は，SIFT の特徴点抽出やその他の特徴オペレータを用いる``sparse''表現に対して，``dense''表現と呼ばれている.
本研究に置いては，特徴点は，SIFT の特徴点抽出によって抽出する``sparse''表現，x,y 方向それぞれ 10 ピクセル毎に 4,8,12,16 ピクセルの 4 通りのスケールで抽出する``dense''表現(以下，``grid dense''と呼ぶ．)，ランダムな位置，スケールで 3000 個程度特徴点を抽出する``dense''表現(以下，``random dense''と呼ぶ．)の合計 3 種類を試した.

4-2 Spatial Pyramid Matching

Bag-of-keypoints は局所特徴の分布をヒストグラム化する方法であるので，局所特徴の位置情報はすべて捨てられてしまっている．そこで，spatial pyramid matching では，画像を数段階にグリッド分割し，グリッド毎に bag-of-words ヒストグラムを作成し，対応する位置同士の bag-of-words ヒストグラムの類似度を計算して,その線形和を画像間の距離とする，大まかな位置情報を考慮した画像間の類似度を計算法を提案している．図に示すように，レベル０では画像全体を１つのグリッド，レベル１では 22 のグリッド，レベル２では 44 のグリッドに画像をそれぞれ分割し,それぞれのグリッド毎に bag-of-keypoints ヒストグラムを作成する．図では 3 段階のピラミッドで合計 21 個の bag-of-keypoints ヒストグラムが作成される．画像全体をレベル０として，縦横それぞれ２倍ずつグリッドを増やしていって,L を最大レベルとすると，類似度は以下の式で定義される．

$$k(X,Y) = \frac{1}{2^L} I_0(X,Y) + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I_l(X,Y) \quad (1)$$

I0(X,Y)はレベル０におけるすべての bag-of-keypoints ヒストグラムのヒストグラムインターセクションである．１枚の画像を表す特徴ベクトルの次元は元の次元の４の１乗倍になってしまうが，特徴ベクトルの多くの成分が 0 であるスパースなベクトルで,しかも類似度計算が簡単なヒストグラムインターセクションであるために計算時間に著しく増大することは起らない．実験では，この類似度の式を SVM のカーネル関数として利用して，画像の選別を行った．

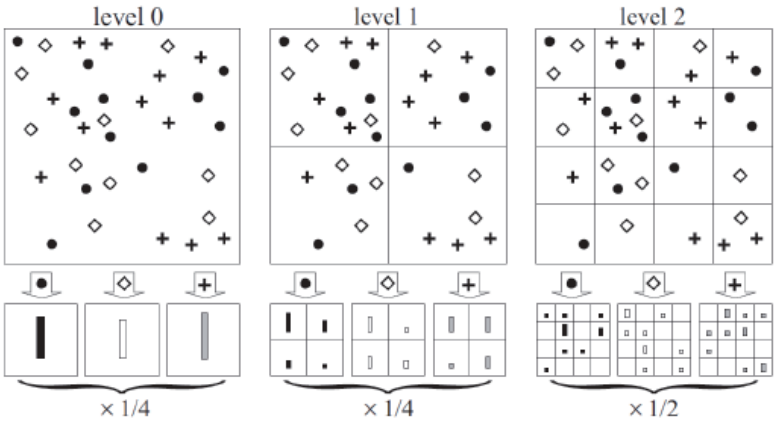

図 1　Spatial pyramid matching.（[16]より引用）

5 実験

10 種類のキーワード（夕暮れ，山，ラーメン，滝，海岸，花，ライオン，りんご，赤ちゃん，ノートパソコン)について，Web からの画像収集実験を行った．表１に，まずは比較のために，Google Image Search を用いて検索した場合の 1 から 500 位までの適合率，および Web からキーワードによって Web 検索エンジンを利用して収集した画像の結果(RAW)，さらに２つの従来手法による結果である，カラーヒストグラムを特徴量とした画像検索によって画像の選別を行った結果(CBIR)と,領域分割した画像の領域毎の特徴量の分布をガウス混合分布による確率生成モデルによって表現して画像選別をした結果(GMM)を示す．RAW は収集フェーズまでの結果を示している．この結果と最終結果の差分が画像選択フェーズによる画像選別の結果あると言える．

表 1 中の結果はそれぞれ実際に A 画像，B 画像として収集した枚数と，括弧内はその主観評価による適合率((キーワードに適合している画像の枚数)/(収集枚数))，さらに，A,B 合計の枚数と適合率を示す．選択画像については，適合率に加えて，その後に，再現率((選択画像中の適合画像の枚数)/(収集した画像中の適合画像の枚数))も示してある．なお，実際には 29944 枚すべての収集画像について主観評価を行うことは困難であるので，それぞれの結果について約 500 枚を無作為抽出して作られたグランドツールスデータを元に適合率，再現率を求めた．

収集フェーズでは，1 キーワードにつき 5000 程度の URL を複数のテキスト検索エンジンを用いて集めた．なお，通常検索エンジンは 1 キーワードにつき 1000 個までの URL しか返さないので，実際には検索語拡張(query expansion)の手法を用いて，検索キーワードのペアを自動生成して収集をした．収集フェーズ終了時での収集画像枚数は平均 2994.4 枚で，1500 枚から 5837 枚までキーワードによってかなりばらつきがあった．適合率は 43.6%から 79.2%までとおおむね４　８割程度であった．この適合率を上げるのが，収集フェーズの後の，画像選択フェーズの役目である．本研究での新規の実験は後半の選択手法に関してのみなので，この収集フェーズでの結果に対してフィルタリングを行ったことになる．

表２に通常の bag-of-keypoints と SVM(BOK+SVM)による結果を示す．局所パッチの切り出し方法は，SIFT による sparse，縦横 10 ピクセル毎のグリッドによる grid dense（g-dense と省略する．），3000 点のランダム抽出による random dense（r-dense)の 3 通りで行い，それぞれコードブックサイズ 500 (cb=500)と 1000（cb=1000)についての合計６通りについて実験した．

10 単語の平均については，g-dense（cb=1000)の場合，適合率が 82.4%で最も良い結果であったものの，どの組み合わせも 80%台前半で大きな違いは見られなかった．r-dense，g-dense はランダム，グリッドの違いがあるもののどちらもサンプリングする局所画像パッチを画像内容に無関係に機械的に決めているため本質的には同じであるため，結果が同程度であるのは予想の範囲内であるが，SIFT による画像パッチ抽出の sparse よりも平均的に dense の方が良い結果であるのは注目すべき結果である．特に，「ラーメン」「海岸」で dense の方が良い結果が得られている．

表３に Spatial Pyramid Kernel を利用した結果(BOK+SPK)を示す．BOK＋SVM と同様の 6 種類の局所パッチの切り出し法とコードブックサイズの組合せで実験を行った．g-dense（cb=500)の時が適合率 84.0%で最も良い結果が得られた．6 種類の結果すべてで，BOK+SVM の結果を僅かではあるが上回って，性能向上が見られた．

図２に RAW, CBIR, GMM, BOK+SVM, BOK+SPK の比較のグラフを示す．BOK＋SVM, BOK+SPK はそれぞれ最も平均の適合率の値が良かった組合せについて示している．RAW に比べるとすべての単語について適合率の向上がみられ，特に「夕暮れ」(図３)では，35 ポイント近くも上昇している．従来手法である CBIR, GMM との比較では，10 単語の平均でほぼ同じ再現率の適合率がそれぞれ最大 18 ポイント，10.5 ポイント向上している．個々の単語については，GMM が比較的有効であった「夕暮れ」「山」などの「風景」画像に加えて，GMM がほとんど適合率を改善すること出来なかった「リンゴ」「ライオン」などの「物体」画像を表す単語に関しても大幅に適合率を上昇させている．ただし，仮の正例学習画像として用いた RAW の A 画像の精度が 50%台であった「赤ちゃん」「ノート PC」(図４)に関しては，学習画像の適合率が低くノイズを多く含んでいたため，適合率は 10 ポイント程度しか上昇させることは出来なかった．

表 1　10 個のキーワードに関する Goggle Image Search の結果 (1-500 位の画像の適合率 (%))，HTML 解析直後の収集画像 (A 群,B 群, A と B の合計) に関するデータ (収集枚数, () 内は適合率 (%))，従来手法による結果 (A 群,B 群,A と B の合計それぞれについての収集枚数, () 内は適合率, 再現率 (%))．

	Goo.	raw images			CBIR	region-based representation + GMM		
concepts	prec.	A	B	A+B	A+B	A	B	A+B
夕暮れ	79.8	790 (67)	710 (44)	1500 (55.3)	828 (62.2, 62.1)	387 (96, 72)	249 (83, 68)	**636 (91.0, 70.2)**
山	48.8	1950 (88)	3887 (71)	5837 (79.2)	3423 (82.6, 61.2)	1237 (93, 65)	2273 (85, 65)	**3510 (89.0, 65.0)**
ラーメン	65.2	901 (78)	1695 (55)	2596 (66.6)	1492 (71.0, 61.3)	453 (85, 49)	813 (69, 59)	**1266 (77.0, 53.2)**
滝	72.4	2065 (71)	2584 (70)	4649 (70.3)	3281 (71.4, 71.7)	1569 (77, 73)	1935 (76, 77)	**3504 (76.8, 74.6)**
海岸	63.2	768 (69)	1155 (62)	1923 (65.5)	1128 (67.3, 60.3)	414 (74, 61)	569 (73, 64)	**983 (73.3, 62.5)**
花	65.6	576 (72)	1418 (67)	1994 (69.6)	952 (79.3, 54.4)	282 (77, 50)	476 (66, 33)	**758 (71.9, 41.0)**
ライオン	44.0	511 (87)	1548 (49)	2059 (66.0)	967 (71.0, 50.5)	234 (88, 59)	477 (55, 48)	**711 (69.4, 53.6)**
リンゴ	47.6	1141 (78)	2137 (59)	3278 (64.3)	1495 (68.8, 48.8)	499 (79, 45)	753 (60, 33)	**1252 (67.2, 37.7)**
赤ちゃん	39.4	1833 (56)	1738 (53)	3571 (54.5)	1831 (55.1, 51.8)	801 (66, 49)	537 (60, 41)	**1338 (63.9, 45.9)**
ノート PC	60.2	781 (57)	1756 (32)	2537 (43.6)	1290 (46.9, 54.6)	295 (68, 46)	572 (44, 50)	**867 (56.0, 47.6)**
TOTAL/AVG.	58.6	11316 (72)	18628 (56)	29944 (62.2)	16687 (66.0, 57.7)	6171 (80, 57)	8654 (67, 54)	**14825 (73.5, 55.1)**

表 2　Bag-of-keypoints と SVM(BOK+SVM) による画像収集結果．() 内は適合率と RAW に対する再現率 (%)．

単語	sparse (500) A+B	sparse (1000) A+B	r-dense (500) A+B	r-dense (1000) A+B	g-dense (500) A+B	g-dense (1000) A	g-dense (1000) B	g-dense (1000) A+B
夕暮れ	546 (92.9,60.3)	563 (93.3,62.5)	642 (90.9,69.4)	652 (91.5,70.9)	703 (90.4,75.5)	462 (93,81)	235 (86,64)	703 (91.6,76.5)
山	2656 (94.2,66.5)	2762 (93.7,68.7)	2343 (93.6,58.3)	2778 (93.7,69.1)	2471 (93.0,61.0)	1317 (94,72)	1303 (91,57)	2593 (92.7,63.9)
ラーメン	996 (84.1,51.2)	1020 (86.9,54.2)	780 (93.2,44.5)	809 (92.0,45.5)	878 (92.0,49.4)	437 (88,55)	542 (91,53)	984 (90.3,54.3)
滝	3285 (81.9,82.2)	3261 (82.3,82.0)	3127 (82.6,78.9)	3211 (81.4,79.8)	3141 (83.1,79.7)	1442 (84,83)	1497 (85,70)	3015 (84.8,78.1)
海岸	699 (87.0,48.9)	701 (86.5,48.7)	733 (91.2,53.7)	737 (91.0,53.9)	756 (92.1,55.9)	368 (90,63)	393 (94,52)	786 (92.1,58.2)
花	695 (86.1,43.8)	707 (86.7,45.0)	716 (87.7,46.1)	711 (87.4,45.6)	779 (86.6,49.4)	278 (91,61)	401 (87,36)	718 (89.4,47.1)
ライオン	754 (85.5,53.6)	790 (85.5,56.1)	724 (85.3,51.4)	719 (86.7,51.8)	724 (80.3,48.4)	324 (91,66)	349 (62,28)	706 (79.3,46.6)
リンゴ	961 (82.7,36.9)	970 (85.3,38.5)	1180 (82.3,45.1)	1089 (86.0,43.6)	1257 (87.5,51.2)	566 (92,58)	691 (87,47)	1279 (89.9,53.5)
赤ちゃん	1961 (53.9,54.3)	2042 (55.7,58.4)	2073 (57.3,61.0)	2114 (59.1,64.2)	1412 (58.7,42.6)	1083 (58,61)	502 (67,37)	1710 (60.7,53.3)
ノート PC	1183 (54.3,63.8)	1213 (55.0,66.3)	1111 (51.5,56.9)	1090 (52.9,57.3)	740 (50.6,37.2)	358 (60,48)	322 (42,24)	696 (52.9,36.6)
合計/平均	13000 (80.3,56.2)	13297 (81.1,58.0)	12872 (81.6,56.5)	13153 (82.2,58.2)	12559 (81.4,55.0)	6290 (84,65)	6206 (79,47)	12812 (82.4,56.8)

表 3　Spatial pyramid matching (BOK+SPK) による画像収集結果．() 内は適合率と RAW に対する再現率 (%)．

単語	sparse (500) A+B	sparse (1000) A+B	r-dense (500) A+B	r-dense (1000) A+B	g-dense (500) A	g-dense (500) B	g-dense (500) A+B	g-dense (1000) A+B
夕暮れ	644 (91.2,69.8)	667 (90.1,71.4)	637 (92.8,70.3)	521 (93.1,57.6)	479 (91,83)	219 (85,60)	709 (90.1,75.9)	707 (90.6,76.2)
山	1997 (94.4,50.1)	2721 (92.5,66.8)	2603 (94.1,65.1)	2242 (94.8,56.4)	1271 (94,70)	1112 (95,52)	2344 (95.2,59.3)	2496 (94.9,62.9)
ラーメン	958 (84.2,49.4)	999 (82.8,50.6)	696 (92.8,39.5)	620 (93.5,35.5)	388 (94,52)	457 (92,45)	852 (93.5,48.8)	978 (89.2,53.4)
滝	3261 (83.8,83.4)	3309 (82.7,83.6)	4645 (70.5,100.0)	2635 (85.9,69.1)	1538 (82,86)	1404 (85,66)	3073 (83.8,78.6)	2974 (85.1,77.2)
海岸	788 (84.3,53.4)	780 (82.9,51.9)	722 (92.5,53.7)	533 (92.1,39.4)	391 (90,66)	369 (95,49)	798 (92.0,59.0)	806 (92.8,60.1)
花	696 (90.0,46.0)	676 (90.2,44.7)	591 (88.9,38.5)	1882 (72.5,100.0)	321 (90,69)	344 (87,31)	730 (88.9,47.6)	659 (89.5,43.2)
ライオン	721 (86.1,51.6)	714 (87.5,51.9)	610 (86.5,43.9)	1952 (61.6,100.0)	343 (93,72)	208 (71,19)	614 (87.2,44.5)	596 (84.4,41.8)
リンゴ	1321 (86.5,53.2)	1400 (84.7,55.1)	1179 (87.9,48.2)	850 (88.2,34.9)	651 (88,64)	595 (86,40)	1307 (88.0,53.5)	1271 (88.2,52.2)
赤ちゃん	1554 (64.3,51.3)	1615 (65.7,54.5)	1649 (66.1,56.0)	1004 (69.8,36.0)	1003 (61,60)	354 (70,27)	1502 (63.1,48.7)	1653 (61.9,52.6)
ノート PC	822 (61.6,50.3)	784 (57.5,44.8)	743 (52.4,38.7)	550 (55.9,30.5)	394 (64,57)	269 (45,21)	690 (58.1,39.8)	646 (55.2,35.5)
合計/平均	12553 (82.6,55.8)	13093 (81.7,57.6)	12479 (82.4,55.4)	12874 (80.7,55.9)	6535 (85,68)	5295 (81,41)	12291 (84.0,55.6)	12399 (83.2,55.5)

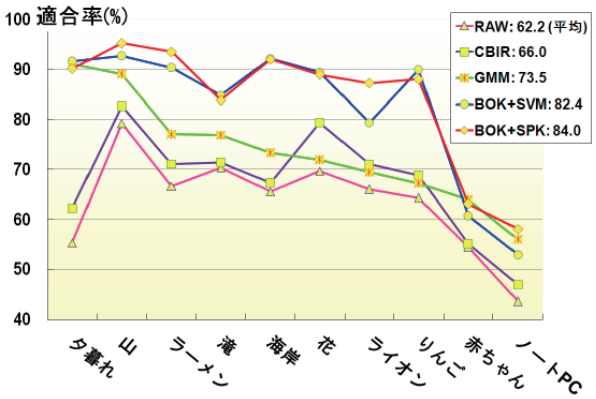

図 2　収集結果の比較．

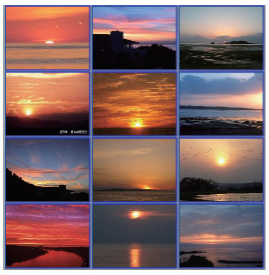

図 3　“夕暮れ”．

図 4　“ノート PC”．

5．おわりに

本論文では，「Web 上の画像知識を用いた単語の視覚性に関する研究」の精度向上のための，bag-of-keypoints アプローチによる一般物体認識手法を用いた Web 画像収集システムについて述べた．10 種類のキーワードについて実験した結果，最大 82.4%の平均適合率を得られた．特に，領域分割とガウス混合分布による確率モデルを用いた方法でよい結果が得られなかった「ライオン」「リンゴ」などの「物体」を表す単語についても大幅な適合率の向上が確認された．さらに，bag-of-keypoints に大

まかな位置情報を導入する spatial pyramid matching を導入した結果，平均適合率は 84.0%まで上昇した．このことから，HTML 解析による仮学習画像の推定と，bag-of-keypoints アプローチによる一般物体認識手法による画像認識によって，不完全な学習画像データしか得られない Web 画像収集においても bag-of-keypoints と SVM の組合せが有効であることが示された．

今後の課題としては，HTML 解析の精度向上や query expansion の改良による仮学習画像の精度向上，1 クラス SVM の利用による学習画像からのノイズ除去などが挙げられる．さらに最終的には，単語の視覚性分析を bag-of-keypoints を用いて行い，人間の感覚に合った「画像エントロピー」の提案を目指している．

【参考文献】

[1] Minka, T. P. and Picard, R. W.: Vision Texture for Annotation, *ACM/Springer Journal of Multimedia Systems*, Vol. 3, pp. 3-14 (1995).

[2] Mori, Y., Takahashi, H. and Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words, *Proc. of First International Workshop on Multimedia Intelligent Storage and Retrieval Management* (1999).

[3] 森　靖英，高橋裕信，岡　隆一：単語群つき画像の分割クラスタリングによる未知画像からの関連単語推定，電子情報通信学会論文誌 D-II， Vol. J84-D-II, No. 4, pp. 649-658 (2001).

[4] Barnard, K. and Forsyth, D.: Learning the Semantics of Words and Pictures, *Proc. of IEEE International Conference on Computer Vision*, pp. 408-415 (2001).

[5] Duygulu, P., Barnard, K., Freitas, N. d. and Forsyth, D.: Object Recognition as Machine Translation: Learning a Lexicons for a Fixed Image Vocabulary, *Proc. of European Conference on Computer Vision*, pp. IV:97-112 (2002).

[6] Barnard, K., Duygulu, P., Freitas, N. d., Forsyth, D., Blei, D. and Jordan, M.: Matching Words and Pictures, *Journal of Machine Learning Research*, Vol. 3, pp. 1107-1135 (2003).

[7] Tenenbaum, J. M. and Barrow, H. G.: Experiments in Interpretation Guided Segmentation, *Artificial Intelligence*, Vol. 8, pp. 241-274 (1977).

[8] 森 靖英，高橋裕信，保科雅洋，野崎俊輔，岡 隆一：WWW 上の文書・画像混在データのクロスメディア検索，第 15 回情報統合研究会資料 SIG-CII-2001-MAR (2001).

[9] Brown, P. F., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Lafferty, J. D., Mercer, R. L. and Roossin, P. S.: A statistical approach to machine translation, *Computational Linguistic*, Vol. 16, No. 2, pp. 79-85 (1990).

[10] Kato, T., Kurita, T. and Shimogaki, H.: Multimedia interaction with image database systems, *ACM SIGCHI Bulletin*, Vol. 22, No. 1, pp. 52-54 (1990).

[11] 栗田多喜夫，加藤俊一，福田郁美，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， Vol. 33, No. 11, pp. 1373-1383 (1992).

[12] Yanai, K.: Generic Image Classification Using Visual Knowledge on the Web, *Proc. of ACM International Conference Multimedia*, pp. 67-76 (2003).

[13] 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，人工知能学会論文誌， Vol. 19, No. 5, pp. 429-439 (2004).

[14] Csurka, G., Bray, C., Dance, C. and Fan, L.: Visual categorization with bags of keypoints, *Proc. of ECCV Workshop on Statistical Learning in Computer Vision*, pp. 1-22 (2004).

[15] D. G. Lowe: “Distinctive image features from scaleinvariantkeypoints”， International Journal of Computer Vision, **60**, 2, pp. 91- 110 (2004).

[16] S. Lazebnik, C. Schmid and J. Ponce: “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories”, Proc. of IEEE Computer Vision and Pattern Recognition, pp. 2169- 2178 (2006).

〈発表資料〉

題名	掲載誌・学会名等	発表年月
確率的 Web 画像収集	人工知能学会誌	2007/01
一般物体認識のための単語概念の視覚性の分析	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	2007/02
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	2007/11
Image Collector III: A Web Image-Gathering System with Bag-of-Keypoints	Proc. of the Sixteenth International World Wide Web Conference	2007/05
Web Image Gathering with a Spatial Pyramid Kernel	Proc. of the IEEE Workshop on Multimedia Information Processing and Retrieval	2007/12
The News Flusher: a Photo News Clustering Browser	Proc. of the Pacific-Rim Conference on Multimedia	2007/12
Web Image Gathering with a Part-based Object Recognition Method	Proc. of the International Multimedia Modeling Conference	2008/01
Associating Faces and Names in Japanese Photo News Articles on the Web	Proc. of the 2008 IEEE International Symposium on Mining the Asian Web	2008/03
Automatic Web Image Selection with a Probabilistic Latent Topic Model	Proc. of the Seventeenth International World Wide Web Conference	2008/04
Web Image Selection by PLSA	Proc. of the International Conference on Multimedia and Expo	2008/06
Web Video Retrieval Based on the Earth Mover's Distance by Integrating Color, Motion and Sound Features	Proc. of the ICIP Workshop on Multimedia Information Retrieval	2008/10