論　文

画像クエリによる情報検索システム構築のための階層型オブジェクトカテゴリ認識手法

非会員　皆川　卓也*　**　　正　員　斎藤　英雄*

Hierarchical Object Category Recognition Technique for Image Based Search System

Takuya Minagawa*　**, Non-member,　Hideo Saito*, Member

In this paper, we present an object category recognition method for an information search system which is queried by camera of mobile phone or by servers of internet services. In such a system, processing speed is an important requirement.

To improve processing speed, the hierarchical object category recognition technique proposed by Serre[23] is modified using Haar-Like features, vector quantization of feature models, and reduction of processing area. In addition, by retaining the information of each feature's position, it compensates the accuracy which is a little reduced in exchange of processing speed. We implemented this method to web server, and proved this system can work in practical processing time.　Through the experiment for Caltech-101 image database and natural scene category images, we also confirm the accuracy of our approach.

キーワード：オブジェクトカテゴリ認識，情報検索システム

Keywords：Object Category Recognition, Information Search System

1.　はじめに

情報検索の技術分野では，従来のテキストによる検索手段とは別に，画像による検索の研究・実用化が進んでいる。これはデータベースに対するクエリーとして画像の特徴量を用いるもので，テキストによる部分一致や全文一致検索とは違い，言語化しにくい「見ため」そのものをキーに検索が行えるというメリットがある。

これらの検索手法では，画像クエリーの解釈のためにその用途に応じた様々な画像認識技術が利用されている。例えば，類似画像検索と呼ばれる技術は，色や構図などの情報を画像の特徴量として用いることで人間の主観的に「なんとなく似た画像」を探す検索手法であるが[20][26]，これらはファッションのショッピングサイトで，商品検索を行う際に利用されるなど，主観的見ためが重要な分野に応用されている[1][22]。

また，データベースに登録した物体画像を検索するオブジェクト認識と呼ばれる技術の多くは，局所特徴量を検索キーとして用いており，クエリー画像が歪んだり，回転していたり，一部が隠れたり変更されていても登録画像を正確に検索できる[16]。そのためこれらの技術は，ネット上の大量のデータから特定のロゴやランドマーク，絵画などの画像が使われているサイトを探すのに利用されたり[8][9]，広告などを携帯カメラで撮影することでキャンペーンに誘導するシステムなどに応用されている[10]。他にも顔認識/認証[11][21]や文字認識[4]を利用した検索など，様々な認識技術が画像クエリーによる検索サービスに利用されている。これらの技術とサービスの関係を Table 1 に示した。

本研究ではこのような，画像をキーとする情報検索システムに利用されることを想定したオブジェクトカテゴリ認識手法を提案する。オブジェクトカテゴリ認識は画像に写

Table 1.　Image Based Search System.

技術	検索キー	用途	事例
類似画像検索	色・構図など	主観的に似た画像の検索	ショッピングサイトの商品検索
オブジェクト認識	局所特徴量	DB に登録した物体画像の検索	広告キャンペーン
顔認識	顔	人物検索/類似人物検索	写真分類/診断系エンタメサービス
文字認識	文字	画像のタグ付け	DB 中の画像の自動インデクシング

*　慶應義塾大学理工学部
〒223-0061　神奈川県横浜市港北区日吉 3-14-1
Keio University Faculty of Science and Engineering.,
3-14-1 Hiyoshi, Kouhoku-ku, Yokohama-shi, Kanagawa 223-0061
**　ジェイマジック株式会社
〒105-0011　東京都港区芝公園 3-4-30　32 芝公園ビル 10F
J-Magic Inc.,
32Shibakouen Bld. 10F, 3-4-30 Shibakouen, Minato-ku, Tokyo 105-0011

© 2008 The Institute of Electrical Engineers of Japan.

っている対象がどのカテゴリ（バイク，顔，自動車など）に属するかを判定する技術で，検索向けに様々な応用が可能となる。例えばインターネット上の写真共有サイトやソーシャルネットワークサービス(以下 SNS)において，同じカテゴリの写真をアップしている友人を教えるなどのコミュニケーション手段や，大量の画像に対してカテゴリ名のタグを付与するなどの検索・分類手段，閲覧している画像に適した広告を出すなどの収益化の手段などに応用が考えられる。

オブジェクトカテゴリ認識手法として，近年特に良く用いられているのが，Bags of Features と呼ばれる画像内での局所特徴の数をヒストグラム化して認識する手法で，特徴間の位置関係を無視しても高い認識率を得ることができる[3] [27]。一方で局所特徴間の位置関係をベイズモデルであらわすことでオブジェクトを認識する手法も存在する[5] [7]。また，Serre らは脳の視覚システムを元にした階層型のオブジェクト認識アルゴリズムを提案した[23]。

これらの技術を検索目的に利用する場合，いくつか考慮すべき課題がある。例えばインターネットの世界では 8 秒ルールと言われるように，結果の表示時間があまりにかかってしまうと，ユーザは処理を中断して他のサイトへ移ってしまう傾向がある[28]。現在はブロードバンド回線の普及により，ユーザが表示を待てる時間は更に短くなったと言われており，通信によるオーバーヘッドも含めると認識技術に求められる処理時間は 1 秒程度と言える。

求められる認識精度や認識すべきカテゴリ数は，用途によって様々なパターンがありえる。例えば，花や動物を撮影してその名前や性質を調べるような辞書的な用途であれば，高い精度と多くのカテゴリの判別が必要になるが，一方で SNS において，好みの画像カテゴリを多くアップしている趣味の同じ人を教えてくれるといったコミュニケーション用途の場合，「顔」，「車」，「犬」，「風景」といった具合に絞ったカテゴリを判別することができれば良く，また精度に対する許容度も上がる。　本研究では，これらの条件を満たすような実用的な検索サーバーシステムの構築を目指す。ここではオブジェクトカテゴリ認識アルゴリズムとして，Serreの提案した階層型の認識手法を採用する[23]。Serre の手法は，比較的単純な構造ながら高い認識率を持ち，また将来的には Bileshi らが道路のシーン解析への応用で示したように[2]，より高度なコンテクスト解析への拡張が期待できる。一方で，ガボールフィルタの処理と 1000 個近くの形状特徴の計算に時間がかかるため，検索システムとして利用するには処理時間の面で課題がある。

そこで，本研究では Serre の提案した階層型オブジェクトカテゴリ認識アルゴリズムに対してベクトル量子化や Haar-Like 特徴，処理画像領域の削減を用いて処理を高速化し，検索用サーバーシステム構築のためのオブジェクトカテゴリ認識手法の提案を行う。

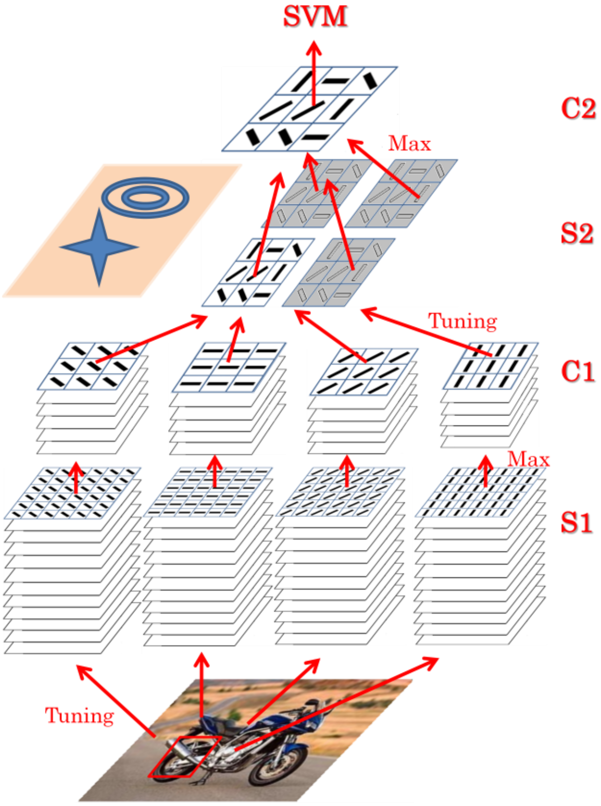


Fig. 1. Base method.

2.　ベース手法

Serre の提案したカテゴリ認識手法（以下，ベース手法）は Fig.1 のような階層構造型のアーキテクチャをもち，それぞれ特定の信号に対して選択性を持つ S 層と，位置やサイズに対しての不変性を持たせる C 層が交互に組み合わさっている[23]。

すなわち入力画像は S 層の処理を経てある特定のパターンを持つ信号が抽出され，C 層で位置やサイズが違うがパターンが同じ信号を統合して，一つの出力とする。これにより，人間が持つ信号に対する選択性や，位置やサイズのずれや歪みに対する不変性の両方を獲得することができる。

〈2･1〉　S1 層　　入力画像はまずグレースケールに変換され，S1 層への入力となる。グレースケール画像は，以下の式で表せる 2 次元ガボールフィルタにかけられる。

$$F(x,y) = \exp\left(-\frac{(x_0^2+\gamma^2 y_0^2)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda}x_0\right)$$

$$x_0 = x\cos\theta + y\sin\theta \text{ and } y_0 = -x\sin\theta + y\cos\theta \quad \cdots(1)$$

これはアスペクト比γ，方向成分θ，波長λを持ち，中心から離れるとσの割合で減衰する局所性を持ったフィルタである。

S1 層では画像に対し，16 種類のスケール帯（λおよびσ）と 4 種類の方向($\theta = 0^o, 45^o, 90^o, 135^o$)を持った，計 64 個のフィルタをかける。すなわち S1 層は画像平面上の各位置に 64 種類のユニットを持ち，それぞれのタイプのユニットが，ある特定の方向性を持つ 2 次元信号に対して選択性を持つ

ことになる。この各ユニットの出力が，次の C1 層のユニットへの入力となる。

〈2･2〉　C1 層　　C1 層は S1 層と同じく 4 種類の方向に選択性をもつユニットを各位置に持つ。各 C1 ユニットは，$N_S \times N_S$近傍の S1 ユニットの出力を 2 つのスケール帯から受け取り，その最大値を出力することで統合する。すなわちこの層では，各方向の信号毎に位置のズレやサイズの違いを吸収している。

従って，S1 の 16 のスケール帯は C1 では 8 つへと統合され，また 1 つの C1 ユニットは入力画像の$N_S \times N_S$ピクセル領域からの信号に対して反応を示す。N_Sの値はスケール帯によって違い，S1 のガボールフィルタの波長λが大きいものほど大きくなる。また隣り合った C1 ユニットがカバーする入力画像の領域は，ΔSだけオーバーラップするように設定する。

〈2･3〉　S2 層　　S2 層は，事前に学習されたある特定の形状に反応する層である。（Fig.2 参照）

各 S2 ユニットは対応するスケール帯の近傍 C1 ユニットのうち，4 方向全てからの出力を受けとり，事前に学習した画像の一部の領域（パッチ）の形状とのユークリッド距離を元に出力を決定する。例えば，ある S2 ユニットが n 近傍の C1 層からの出力を受け取る場合，1 つの S2 ユニットへの入力は n×n×4 次元のベクトルとなる。各 S2 ユニットの反応 r は以下の式で表すことができる。

$$r = \exp(-\beta \|\boldsymbol{X} - \boldsymbol{P_i}\|^2) \quad \cdots (2)$$

X は C1 ユニットからの入力ベクトル，$\boldsymbol{P_i}$は事前に学習された N 個のパッチのうちの一つ（N～1000），βは反応の鋭さを表す係数である。

従って，S2 層の 8 つのスケール帯の各位置で，上記の反応 r をパッチ数 N 個分計算を行う必要がある。尚，N 個のパッチ$\boldsymbol{P_i}$の学習方法は，以下の通りである。

1. 学習画像から，S1 層及び C1 層の計算を行う。
2. ランダムに S2 層のユニットを一つ選ぶ。
3. そのユニットへの C1 層からの入力ベクトルを学習されたパッチとする。

このようにして N 個の学習パッチの S2 への入力が，そのまま特徴形状として学習される。

〈2･4〉　C2 層　　C2 層では，S2 層での出力を全て統合して，最終的な学習ベクトルとする。すなわち N 個の形状特徴それぞれについて，8 つの全スケール帯と各位置について，信号の最大値を取得する。これにより，C2 層の出力は最終的に N 個の要素（各形状特徴に対する信号の最大値）をもつベクトルとなる。

〈2･5〉　学習　　得られた N 個の要素をもつベクトルを元に，カテゴリの学習及び認識を行う。ここでは判別のアルゴリズムとして線形 Support Vector Machine(24)を用いた。

3. 提案システム

〈2･5〉　システム構成　　本システムの処理構成を Fig.3 に示す。まずユーザは，Web やメール経由で検索クエリーとしたい画像をシステムに送る。システムでは送付されてきた画像をオブジェクトカテゴリ認識プログラムへ送り，その画像がどのカテゴリに属するかを得る。次にシステムは，得られた結果を元に情報をデータベースやインターネット上から検索してユーザに返す。

本システムでは多クラス分類ではなく 2 クラス分類を行う判別器を，判別したいカテゴリの数だけ準備するものとし，認識の結果複数の候補が見つかった場合は，それら全ての情報を検索してユーザへ情報提示する。これは，複数の候補のうち正確に 1 つを求めて検索を行うよりも，候補をある程度絞らずに提示する方が，検索システムとして柔軟性が高いと考えるためである。また実運用を考えた場合，多クラスの分類器は，識別したいクラスが増える毎に，全クラスのカテゴリを学習し直さねばならないため，運用のコストがかかる。

このようなシステムを考えた場合，実用的にシステム全体で数秒程度の応答時間で結果が得られることが望ましい。

〈2･5〉　認識アルゴリズム　　認識アルゴリズムの処理速度は，その実装する言語によっても大きく作用される。本研究ではまず、2 章で述べたベースアルゴリズムを C/C++及び OpenCV を用いて実装を行った。Table 2 は，以下の条件で処理した際の処理速度内訳である。

測定条件：

CPU: Intel Core Duo 1.8GHz

RAM: 2G

画像サイズ: QVGA(240×320)

Fig. 2.　Examples of training image for S2 model feature

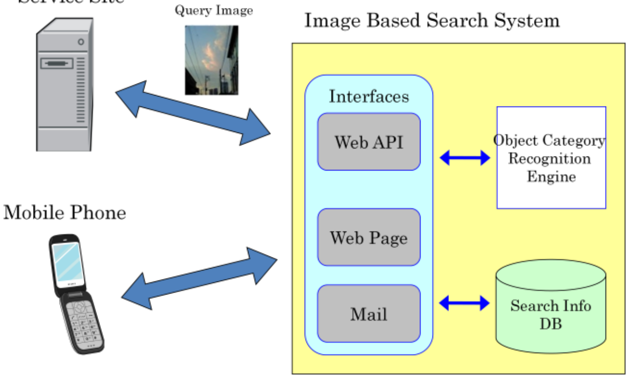


Fig. 3.　Search system overview

Table 2. Process time.

Process	Base Method		Our Method	
	mean(sec)	SD	mean(sec)	SD
S1	1.23	0,009	0.31	0.043
C1	0.24	0.006	0.27	0.025
S2	11.04	0.037	0.47	0.044
C2	0.05	0.005	0.05	0.007
SVM	0.00	0.000	0.00	0.003
Total	12.56	0.040	1.11	0.053

ベース手法と本手法を用いて，QVGA 画像 50 枚に対して計測した処理速度の平均(mean)と標準偏差(SD)。

ベース手法は，10 秒以上の処理時間となっており，検索に耐えうる処理時間とはなっていない。そこで，本研究では検索に耐えうる処理速度を実現するために，以下の 4 点について変更を行った。

1. ベクトル量子化による形状特徴数 N の削減
2. C1 の極値を用いた S2 の計算領域削減
3. Gabor-Wavelet の Haar-Like 特徴による近似
4. C2 の統合の抑制による構図の維持

Table 2 からも明らかなように S2 の処理に最も時間がかかっている。そこで，上記 1,2 で S2 の処理時間を，3 で S1 の処理時間を削減する。4 は特徴間の位置関係の情報を保持することによる精度の改善である。

（1） ベクトル量子化による特徴数の削減　S2 層では入力画像の全ての領域において，与えられた入力信号と事前に学習させた全ての形状特徴との距離を計算するため，形状特徴の数が処理時間に大きく影響する。2.3 節で述べたように，ベース手法では学習用画像のランダムな領域から C1 の出力を求め，それを形状特徴としてそのまま学習している。しかしながら，この手法だと似たような特徴が重複して学習される可能性がある。

そこで計算コストを削減するため，このような似た特徴をまとめることで，特徴数を減らしたい。本研究では，ベクトル量子化の代表的手法である LBG アルゴリズム[15]を用いて，クラスタリングによる特徴の削減を実施した。ベース手法では N=1000 の形状特徴を用いていたが，本手法では，N=200 まで特徴の削減を行った。

（2） C1 の極値を用いた S2 の計算　Bags of Features をはじめとする従来の局所特徴量に基づいた認識手法では，DoG(Difference of Gaussian)や Harris オペレータなどの特徴点検出オペレータを用いて，画像領域のうちエッジなどの特徴ある部分を注目点として見つけ，その周辺領域の特徴量を取得することで認識を行っている[16][12]。一方ベース手法では，入力画像の全領域に対して，S1，C1，S2，C2 の計算を行っている。そこで本手法にもこの特徴点の考え方を導入することで，認識率の低下を最小限に抑えながら計算量の削減を行いたい。特に Table 2 で示した通り，S2 における処理時間が最も大きな比率を占めているため，S2 の計算を特徴点上に限定することを検討する。

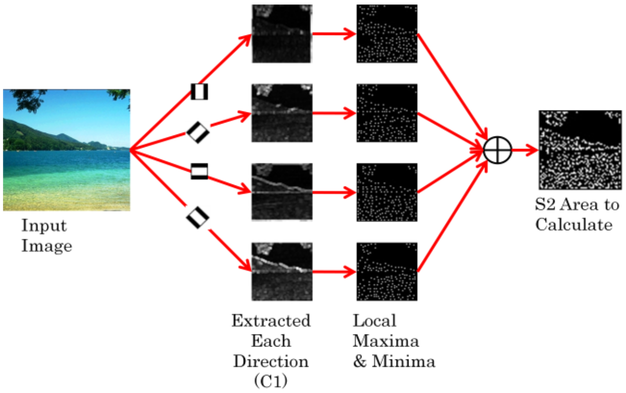
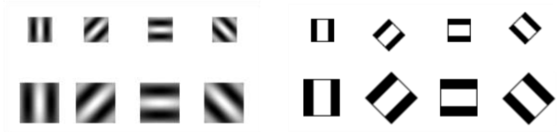
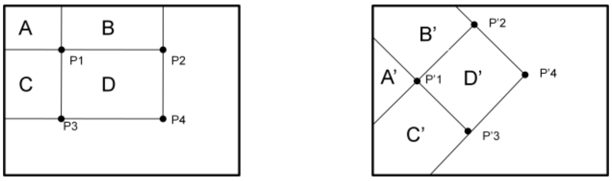


Fig. 4. S2 calculating area



(a) Gabor filter　(b) Haar-like feature

Fig. 5. Gabor filter and Haar-like feature.



(a)Normal　(b)Skew

Fig. 6. Integral image.

本手法では Harris オペレータ等の特徴点検出手法を新たに実装するのは避け，効率化のため C1 までの計算結果を利用することを考えた。C2 の計算では，S2 の出力結果のうち各々の形状特徴毎に最大値を出力する値のみが利用され，それ以外の情報は破棄される。そこで，近傍の S2 ユニット同士の出力は似通っていることが想定されるため，計算領域を近傍で間引いても結果に大きな影響を及ぼさないとみなし，C1 層の出力のうち，極大値，もしくは極小値を取るもの以外は，S2 の計算から除外する(Fig.4 参照)。これにより， S2 の処理速度を大幅に改善することが可能になる。

（3） Haar-Like 特徴による近似　S1 層では 16 スケール×4 方向，計 64 種類のガボールフィルタを用いて方向性をもつ画像特徴の計算を行っている。具体的には 7×7 ピクセルから 37×37 ピクセルまでのサイズの異なる複数のフィルタを画像に対して畳み込んでいるため，処理に時間がかかる。そこで本システムでは，ガボールフィルタを Haar-Like 特徴で近似することで処理の高速化を行う(Fig.5 参照)。

Haar-Like 特徴は Fig.5 (b)のように，2 値のマスクで白領域の画素値の総和から，黒領域の画素値の総和を引いた値

を出力する。すなわち，次の式(3)であらわすことができる。

$$F^{'}(x,y)=\begin{cases}-1 & -\frac{\lambda}{2}\leq x_0<-\frac{\lambda}{4}\cap-\frac{\lambda}{2}\leq y_0<-\frac{\lambda}{2},\\ 1 & -\frac{\lambda}{4}\leq x_0<\frac{\lambda}{4}\cap-\frac{\lambda}{2}\leq y_0<\frac{\lambda}{2},\\ -1 & \frac{\lambda}{4}\leq x_0<\frac{\lambda}{2}\cap-\frac{\lambda}{2}\leq y_0<-\frac{\lambda}{2},\\ 0 & \text{otherwize}\end{cases}\quad\cdots(3)$$

Viola らは，インテグラル画像という手法を用いて，Haar-Like 特徴を高速に計算する手法を提案し[25]，Lienhart らはその手法を更に拡張して斜め方向の線分特徴を導入した[14]。本手法では Lenhart らの手法に従う。

インテグラル画像は各画素の値 II(a,b)が，原画像 I に対して，次の式により変換したものである。

$$II(a,b)=\sum_{i=0,j=0}^{a,b}I(i,j)\quad\cdots(4)$$

すなわち Fig.6.(a)において，インテグラル画像上の P1 の値は入力画像の領域 A の全画素の総和として与えられる。同様に，斜め方向については Fig.6.(b)のように P’1 の値が入力画像の領域 A’の画素の総和を持つ。従って，入力画像における領域 D の画素の総和および領域 D’の画素の総和はインテグラル画像の画素値を用いて，式(5)，(6)のような単純な式で表わすことができる。

$$D=P4-P2-P3+P1\quad\cdots(5)$$

$$D^{'}=P^{'}4-P^{'}2-P^{'}3+P^{'}1\quad\cdots(6)$$

（4）特徴毎の位置情報の維持　ベース手法では，S2 からの出力を全て統合して，N 個の要素をもつベクトルとして各カテゴリの認識と学習を行った。すなわち各特徴間の位置関係の情報は失われてしまっている。Mutch らは同様の階層型アルゴリズムにおいて認識率を向上させるための手段として，特徴の位置情報を保持する手法を提案している[17]。これは，C 層において位置の不変性が確保されているため，有効な方法である。従って本手法では，C2 の出力を $N_{c2}\times N_{c2}$領域へ分割し，各領域毎に形状特徴数 N 個の出力を求めることで，構図の情報を保持した。尚，分割した各領域間は，比率r_{c2}(0~1)でオーバーラップさせる。

4. 実験

ここでは本手法を用いて，処理速度の評価や認識精度実験，および本手法の各改善点の精度および速度への影響について評価を行った。本実験で用いたパラメータを Table 3 に示す。尚，S2 で使用する形状特徴の学習は，インターネットから取得した画像 1200 枚を元に行った。

〈4･1〉 処理速度の測定

本手法の処理速度の平均と標準偏差を Table 2 に示す。ここでは，QVGA(240×320)の画像 50 枚に対して処理を行った。この表から，従来は 10 秒以上かかっていた処理が，1 秒程度で終了していることがわかる。尚ベース手法では，画像サイズが同一の場合処理のステップ数も同一となるため，処理時間の揺らぎはほとんどない。一方本手法では，C1 の出力から求めた注目点に対してのみ S2 の計算を行うよう変更を行ったため，入力画像によって処理時間が違うと予想される。実際，平均値で標準偏差を正規化した変動係数（標準偏差/平均）でベース手法が 0.003 のところを本手法では 0.048 と大きく違っている。しかしながら揺らぎはそれでもせいぜい±0.1 秒程度(信頼区間 95%)であり，ほとんどの画像において安定した処理速度を実現できた。

Table 3. Parameters of experiment

C1 layer			S1 layer
Scale band	Spatial pooling grid ($N_S\times N_S$)	Overlap ΔS	Haar filter size
1	8×8	4	3×3 5×5
2	10×10	5	7×7 11×11
3	12×12	6	14×14 17×17
4	14×14	7	20×20 22×22
5	16×16	8	25×25 28×28
6	18×18	9	31×31 34×34
7	20×20	10	36×36 39×39
8	22×22	11	42×42 45×45

C2 layer		S2 layer	
Output grid size ($N_{c2}\times N_{c2}$)	Overlap r_{c2}	Model feature size	Model feature number N
5×5	0.25	8	200

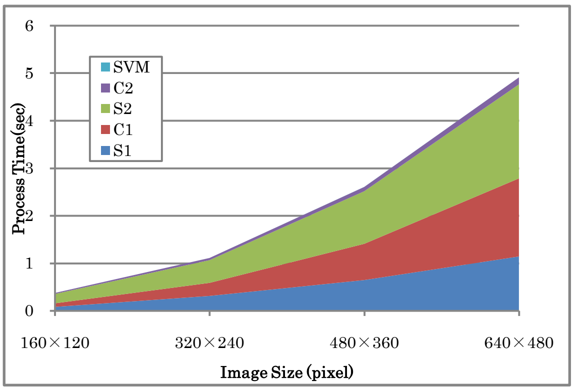


Fig. 7. Process time and image size

画像サイズと処理速度の関係は Fig.7 に示した通り，画像の面積にほぼ比例する。また認識対象となるカテゴリ数を増やした場合，C2 の処理までは共通でカテゴリの数だけ SVM による判別処理を繰り返すことになるので，カテゴリ数に比例して SVM の処理時間が増加する。しかしながら Table 2 に示した通り，全体の処理速度に対して，線形 SVM での処理時間は非常に小さく，12 カテゴリに増やしてもせ

いぜい 6msec 程度であった。

〈4･2〉 精度実験

(１) オブジェクトカテゴリ　ベース手法との比較のため，Caltech101 データベース(5) の画像を元に提案手法の精度評価を行う。Caltech101 データベースは，最近のオブジェクトカテゴリ認識の研究において標準的に用いられている画像データで，101 種類のカテゴリと背景カテゴリの計 102 種類から構成されている。具体例を Fig.8 に示した。

各オブジェクトカテゴリについて，正解画像 80 枚，非正解画像 100 枚（Caltech101 の背景データから選択）を元に学習を行った。また実験は，学習に用いなかった正解画像 200 枚と同じく学習に用いなかった非正解画像 200 枚を用いた。ただし，Car(Side)に関しては，画像枚数が十分でなかったために, 正解画像および非正解画像を 43 枚ずつで行った。結果を Table 4 に示す。また参考としてベース手法の認識率および局所特徴量を用いた他の認識手法として Fergus の Constellation Model による手法(7)の精度も示した。このように，認識率を大きく落とすことなく処理速度の改善を行うことができた。

(２) シーンカテゴリ　画像共有サイトや SNS においては，顔などの単体オブジェクト画像だけでなく，風景（山，海岸，街，etc）などのシーン画像も多くアップロードされる。そこで，本手法をシーンカテゴリに対しても実験を行った。実験画像は Oliva & Torralba の８カテゴリの風景画像を用いた(19)。シーンカテゴリ画像例を Fig.9 に示す。

学習画像は，正解画像 80 枚，非正解画像 100 枚で行った。非正解画像は Caltech101 の背景カテゴリの画像 30 枚，正解カテゴリ以外のカテゴリ画像を 10 枚ずつ用いた．実験は，正解画像 100 枚，非正解画像 100 枚（背景カテゴリ 30 枚，その他各カテゴリから 10 枚ずつ）を用いた。Table 5 に示した通り，それぞれ 8 割近くの認識率を得ることができた。これは例えばユーザのアップした画像に対し，タグをつけたり広告を返すなどの用途に関しては実用に耐えうる認識率であると言える。

シーンカテゴリ認識の分野でも，このような局所特徴量を用いたアプローチが近年多く用いられている。例えば Fei-Fei らは局所特徴量と中間オブジェクトクラスを用いることでシーンカテゴリ分類を行う手法を提案し (6)，また Lazebnik らは Bags of Features の考え方に空間的な情報を持たせることでシーンカテゴリだけでなく単体オブジェクトに対しても同一の手法で認識を行った(13)。これらの場合，局所特徴量で表されるシーン上のテクスチャとそれらの空間的な位置関係が認識に利用されている。本手法の場合も同様に，局所特徴量としての S2 の形状特徴がテクスチャ情報として認識に有効に働いているものと思われる。

〈4･3〉 改善点評価　本研究で提案した 4 つの改善点について，その精度及び速度における影響を調査する。

(１) 精度・速度比較　Table 6 に本手法と各改善点を未実装にした場合の精度比較を，Table 7 に処理速度比較を示した。それぞれベクトル量子化を行なわずランダムに取得した形状特徴を用いた場合(Without VQ), 注目点による S2 の計算領域を削減しなかった場合(Full area calc of S2)，そして Haar-Like 特徴ではなく Gabor フィルタを用いた場合(Gabor Filter)を記述してある。ただし，処理速度に関してベ

Table 4. Recognition rate of Caltech-101(%).

Category	Our approach	Base method	Fergus(7)
Cars(Side)	98.8	99.8	88.5
Faces	91.5	98.1	94
Airplanes	95.3	94.9	90.2
Motorcycles	96.3	97.4	92.5

Fig. 8. Example of Caltech-101 image set.

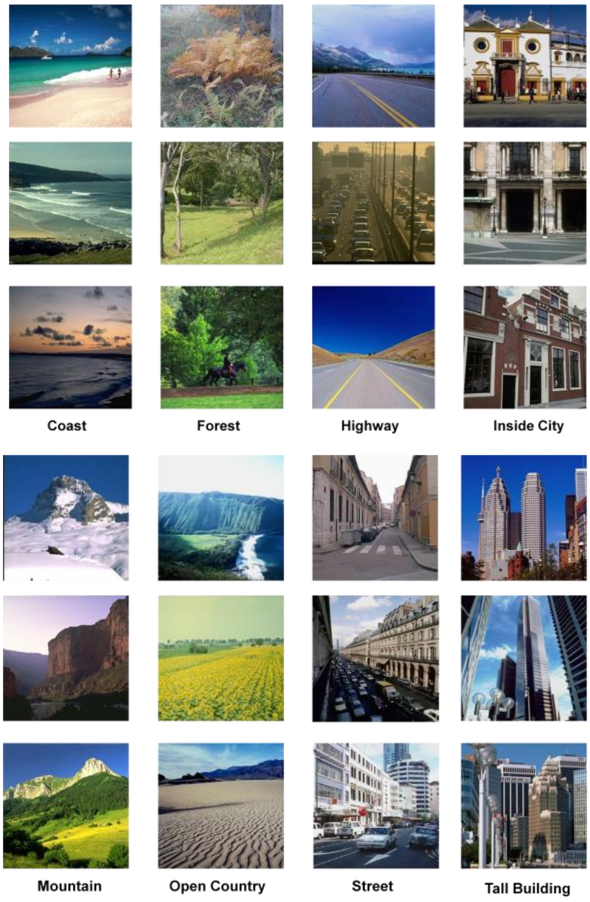


Fig. 9. Example of scene category image set.

Table 5.　Result of scene category image.

Category	Our approach (%)
Coast	83.5
Forest	86.5
Highway	83.5
Inside city	78.5
mountain	80.5
Open country	77.0
Street	86.0
Tall building	88.0

Table 6.　Evaluation of our improvement in recognition rate.(%)

Category	Our approach	Without VQ	Full area calc of S2	Gabor Filter
Cars(Side)	98.8	97.7	98.8	100.0
Faces	91.5	86.3	91.5	93.5
Airplanes	95.3	96.3	96.5	95.3
Motorcycles	96.3	95.3	96.3	98.5
Average	95.5	93.7	96.1	97.3

Table 7.　Evaluation of our improvement in process time. (sec)

Process	Our approach	Full area calc of S2	Gabor Filter
S1	0.31	0.31	1.23
C1	0.27	0.30	0.23
S2	0.48	1.81	0.41
C2	0.05	0.04	0.05
SVM	0.00	0.00	0.00
Total	1.11	2.46	1.92

クトル量子化の有無は影響しないため Table 7 からは除いている。尚，量子化のクラスタリング数と速度・精度の関係については後述する。

前述の通り同じ特徴数であれば，ベクトル量子化を用いた方が似たような特徴を一つにまとめる分，より広範な特徴をカバーでき，それにより認識率の向上が期待できる。一方で細かい差異についての判別能力を落とす可能性も存在し，Table 6 において Airplane での認識率が 1%低下しているのはこれが原因と考えられる。しかしながら全体では約 2%の精度向上を達成できた。S2 の計算領域の削減では，精度は全体でわずかに 0.6%程度の低下がみられたのに対し，S2 における処理速度が 1/4 以下まで大幅に改善することに成功している。また Gabor フィルタの Haar-Like 特徴への置き換えも約 2%程度の認識率の低下があったものの，S１の処理速度を 1/3 以下にまで抑えることができた。

(２)　量子化クラス数の検討　　ベクトル量子化のクラスタリング数（S2 の特徴数）と精度・速度との関係を Fig.10 に示した。図は，S1，C1，S2，C2，SVM の処理の積み上げグラフであり，ここに示されている通り，クラスタリング数は S2 の処理時間にほぼ比例する形で影響を与えている。一方で精度は緩やかな影響しか受けず，今回使用した

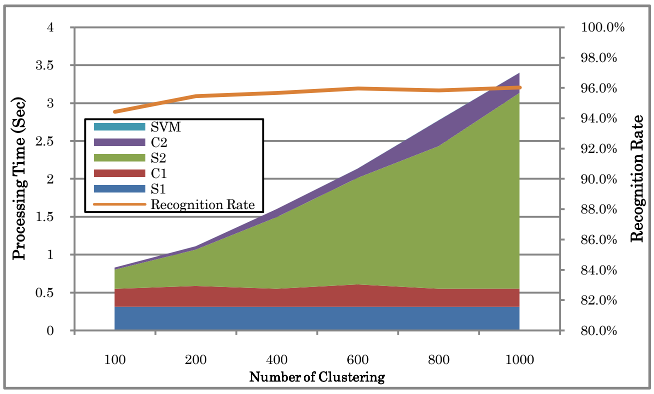


Fig. 10.　Process time and recognition rate for each number of clustering

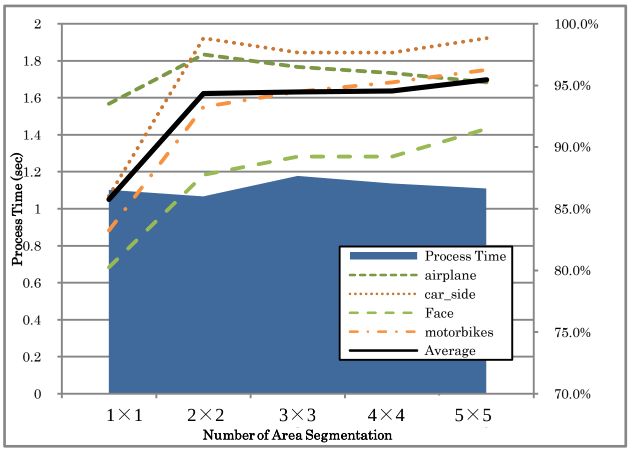


Fig. 11.　Process time and recognition rate for each number of C2 area segmentation

Caltech101 の 4 カテゴリでは特徴数が 600 を超えて以降，ほぼ飽和している。そこで，本手法では N=200 として実装を行うことで精度の低下を最小限にしつつ全体の処理速度を抑えた。

(３)　特徴毎の位置情報の維持　　C2 の出力を 1×1, 2×2, 3×3，4×4，5×5 と領域を分割していったときの認識率と処理速度の変化を求めた。Fig.11 より領域分割数を多くすれば，全体的に精度の向上が見られることが見て取れる。また，その際処理速度の変化はほとんどない。すなわち，特徴毎の位置情報をある程度残すことが有用であることを示しているが，これは Caltech101 のようにある程度オブジェクトが画像の中央にあるような対象に対する結果であることには注意が必要である。

〈4･4〉　**運用試験**　　本オブジェクトカテゴリ認識プログラムをサーバー上へ実装し，インターネットからアクセス可能な環境を整えることで，実運用に近い環境での試験を行った。試験に用いたサーバー環境を Table.8 に示す。

本試験では，Fig.13 のような Web 画面からフォームを使って画像を送信し，認識結果を得る。認識結果をクリックすると Google Image Search で送信した画像と同種の画像を

Table 8.　Environment of search system.

Server	IBM eServer x306
CPU	Intel Pentium4 3GHz
RAM	2GB
Web App Server	Apache Tomcat 6.0

Table 9.　Recognition rate of operational test.

Category	Our Approach	Base Method
Natural Scene	64.5%	63.3%
Dog & Cat	70.5%	77.3%

Fig. 12.　Example of training image (pet animal)

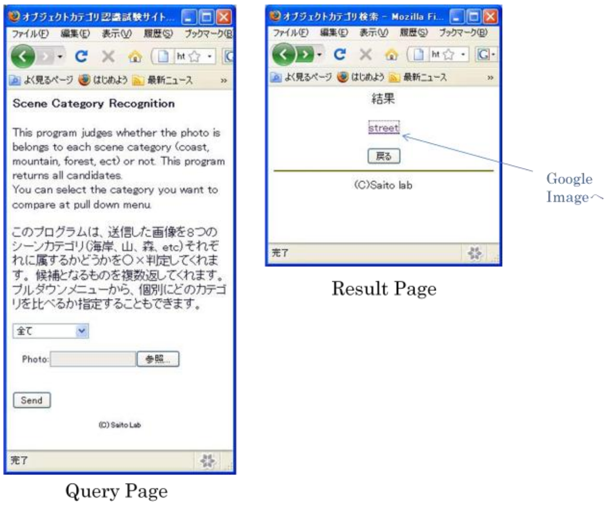


Fig. 13.　Search and result page.

検索できるようにした。本システムでは 2 クラス分類のオブジェクトカテゴリ認識を用いているため，ユーザに対し候補を複数提示する。

認識対象は，比較実験で用いた Oliva & Torralba の 8 カテゴリの風景画像[19] 及びインターネット上で収集した犬猫画像を学習させたものを用いた（Fig.12 参照）。どちらも写真共有サイトなどで多い画像カテゴリである。犬及び猫の判別機は犬画像 55 枚+猫画像 24 枚，非犬猫画像 100 枚で学習を行った。

このサイトを複数の被験者に対して公開して，試験を行ってもらった。試験数はシーンカテゴリが約 40 枚，ペット（犬猫）が約 60 枚で，その際の認識率は Table 9 の通りである。尚，シーンカテゴリに関しては，候補の中に正しいカテゴリが含まれていた場合に認識成功とした。尚，この実験で用いた画像をベース手法に対して適用した結果も Table 9 に合わせて示す。

どちらも比較実験の時と比べると認識率が低下している。シーンカテゴリに関しては，非風景画像を送った時に，8 つの判別機のうちのいずれかが反応してしまうことが多かったために低下した。ユーザに候補を提示するという目的であれば，このような誤認識が許容される場合もあるが，判別機が増えた時にユーザに提示される候補が増えすぎてしまう懸念がある。また，犬猫画像に関しては Caltech-101 の試験で用いた対象(車，顔，飛行機，オートバイ)のようなある程度形状固定のオブジェクトではなく，撮影アングルやポーズも様々なもののため，それが大きく響いていると考えられる。ベース手法と比較した場合，犬猫認識で約 7 %もの違いが出てしまったのも，ベース手法が特徴間の位置情報を破棄しているため，形状変化に強かったものと考察される。今後，これらの問題に対して更なる検討が必要である。

処理速度は，HTTP のリクエストを受信してから，結果を返信するまでの Java サーブレット内での処理時間を測定した（つまりネットワークの影響はない）。その結果，平均 0.78 秒で標準偏差 0.18 であった。従ってネットワークの影響を考えても，十分実用に耐えうる処理時間を実現したと言える。

5.　まとめと今後の課題

本論文では，画像クエリーによる情報検索システム構築のための階層型オブジェクトカテゴリ認識手法について提案した。本手法では Serre らの提案した認識手法[23]をベースに，ベクトル量子化や Haar-like 特徴の導入，処理領域の削減などを行うことで，認識性能を大きく落とすことなく検索システムとして運用に耐えうる処理速度を実現した。また，単体オブジェクトだけでなくシーンカテゴリに対しても同一のアルゴリズムで対応できることを示した。

例えば，車を撮影したら販売店のサービス情報が得られるなどのキャンペーン用途や，ブログや SNS などの画像に連動した占いを出すなどのように，対象や撮影方法をある程度限定できる場合や多少の誤認識が許容される（揺らぎが逆にエンターテイメントの要素になる）場合であれば，現在の認識精度でも十分実用には耐えうると考えられる。しかしながら，用途によっては撮影アングルやポーズなどの条件を限定するのは難しく，また運用試験で見たようにカテゴリが増えれば誤認識も増えるため，今後更なる実運用に耐えるための改良を進める必要がある。例えば，カテゴリ追加に伴う学習や運用の手間が少ない多クラス分類手

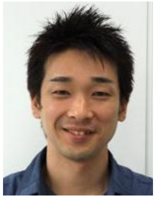
法の検討や，ペットなど身近な対象についてもより認識率を上げていく必要がある。また 1 枚の画像に必ず一つのオブジェクトが存在するとは限らないため，複数のオブジェクトが同時に存在する場合などにも対応できるアルゴリズムの検討が必要である。

（平成●●年●月●日受付，平成●●年●月●日再受付）

文　　献

（１） ALBERT: 「見つかる.jp」, http://mitsukaru.jp/, 2008/09/29 閲覧
（２） S. Bileschi, L. Wolf: "A Unified System For Object Detection, Texture Recognition, and Context Analysis Based on the Standard Model Feature Set", Proc. of British Machine Vision Conference, Sept 2005.
（３） G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray: "Visual Categorization with Bags of Keyponts", Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision, pp. 1-22, May 2004.
（４） Evernote: "EVERNOTE", http://evernote.com/, accessed on 2008-09-29
（５） L. Fei-Fei, R. Fergus, P. Perona: "Learning generative visual models from few training examples: an incremental Baysian approach tested on 101 object categories", Workshop on Generative-Model Based Vision, Computer Vision and Pattern Recognition, June 2004
（６） L. Fei-Fei, P. Perona: "A Bayesian Hierarchical Model for Learning Natural Scene Categories", Proc. of Conference on Computer Vision and Pattern Recognition, June 2005
（７） R. Fergus, P. Perona, A. Zisserman: "Object Class Recognition by Unsupervised Scale-Invariant Learning", Proc. of Conference on Computer Vision and Pattern Recognition, volume 2, pp.264-271, 2003
（８） idée: "TinEYE", http://tineye.com/, accessedon 2008-09-29
（９） Y. Jing, S. Baluja: "PageRank for Product Image Search", Proc. of International conference on World Wide Web, April 2008
（１０） J-Magic: "SAYL", http://www.j-magic.co.jp/j/sayl/, 2008/09/29 閲覧
（１１） J-Magic: 「顔ちぇき！～誰に似てる？～」, http://www.j-magic.co.jp/j/kaocheki/, 2008/09/29 閲覧
（１２） G Kim, C. Faloutsos, M. Hebert: "Unsupervised Modeling of Object Categories Using Link Analysis Techniques" Proc. of Conference on Computer Vision and Pattern Recognition, June 2008
（１３） S. Lazebnik, C. Schmid, J. Ponce: "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories", Proc. of Conference on Computer Vision and Pattern Recognition, June 2006
（１４） R. Lienhart, J. Maydt: "An extended set of Haar-like features for rapid object detection", Proc. of International Conference on Image Processing, Vol.1, pp.900-903, Sept 2002
（１５） Y. Linde, A. Buzo, R. Gray: "An algorithm for vector quantizer design", IEEE Trans.Commun, COM-28, pp.84-95, 1980
（１６） D. Lowe: "Object Recognition from Local Scale-Invariant Features", Proc. of the International Conference on Computer Vision, pp.1150-1157, Sept 1999
（１７） J .Mutch, D. Lowe: "Multiclass Object Recognition with Sparse, Localized Features", Proc. of Conference on Computer Vision and Pattern Recognition, June 2006
（１８） MyHeritage.com: "my Heritage", http://www.myheritage.com, accessed on 2008-12-29
（１９） A. Oliva, 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
（２０） G. Pass, R. Zabih: "Histogram Refinement for Content-Based Image Retrieval", IEEE Workshop on Applications of Computer Vision, 1996
（２１） Polar Rose: "Polar Rose", http://www.polarrose.com/, accessed on 2008-09-29
（２２） Riya: "Like.com", http://www.like.com, accessed on 2008-09-29
（２３） T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, T. Poggio: "Robust Object Recognition with Cortex-like Mechanisms", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.29, No.3, pp. 411-426, March 2007
（２４） 高木 幹雄, 下田 陽久(監修): 「新編 画像解析ハンドブック」, pp.209-216, 東京大学出版会, 2004
（２５） P. Viola, M. Johnes: "Rapid Object Detection using a Boosted Cascade of Simple Features", Proc of Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii USA, I-511- I-518 vol.1, June 2001
（２６） 横山 貴紀, 渡辺 俊典, 菅原 研, 上野 芳郎: 「フラクタル符号に基づく構造的な類似性抽出手法の提案」, 電子情報通信学会技術報告,PRMU2001-285, pp.105－112, 2002
（２７） H. Zhang, A. Berg, M. Maire, J. Malik: "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition", Proc of Conference on Computer Vision and Pattern Recognition, June 2006
（２８） Zona Research: "The Need for Speed II", Zona Market Bulletin, Issue05, April 2001

皆　川　卓　也



（非会員） 1974 年 11 月 10 日生。1997 年 3 月慶應義塾大学卒業。1999 年 3 月同大学院理工学研究科修士課程電気工学専攻修了。IT のシステムエンジニアを経て，現在ジェイマジック株式会社マジックラボ主任，及び慶應義塾大学大学院理工学研究科博士課程開放環境科学専攻に所属。

斎　藤　英　雄

（正員） 1965 年生。1987 年 3 月 慶應義塾大学理工学部電気工学科卒業。1992 年 3 月同大学院理工学研究科博士課程電気工学専攻修了。1992 年同大学助手。同講師，同助教授を経て，2006 年同教授。1997 年～1999 年米国カーネギーメロン大学ロボット工学研究所訪問研究員。2000 年～2003 年科学技術振興事業団さきがけ研究 21「情報と知」領域研究員兼務。博士（工学）。主としてコンピュータビジョンに関する研究に従事。電子情報通信学会，情報処理学会，映像情報メディア学会，計測自動制御学会，日本 VR 学会，IEEE 会員。