ジオタグ画像認識における位置情報の利用法の検討と分析

八重樫恵太† 丸山 拓馬† 柳井 啓司†

† 電気通信大学大学院 電気通信学研究科 情報工学専攻 〒 182-8585 東京都調布市調布ヶ丘 1–5–1
E-mail: †{yaegas-k,maruya-t,yanai}@mm.cs.uec.ac.jp

あらまし デジタル写真の認識対象とその撮影位置には密接な関係がある．本稿では，ジオタグ画像認識において，位置情報とそれらに関連する航空写真やテキストなどの情報を特徴量として使用することを提案する．撮影位置に対応する航空写真と，周辺情報のテキストおよび時間情報付加的な特徴量として用いることで，地理的なコンテキストを反映した画像認識が可能であるかどうかを画像カテゴリー分類における分類精度を比較検討することで検証する．また，それぞれの特徴量の有効性を Multiple Kernel Learning(MKL) による重みを以て推定する．結果として，28 種類の分類カテゴリーについて，周辺情報のテキストと画像特徴の組み合わせの平均適合率の平均値が，画像特徴のみの場合と比較して 77.71%から 80.87%へ向上した．

キーワード ジオタグ，位置情報付き写真，航空写真，一般画像認識，機械学習，マルチカーネル学習

A Study on Methods of Utilizing Geolocation Information for Geotagged Image Recognition

Keita YAEGASHI†, Takuma MARUYAMA†, and Keiji YANAI†

† Department of Computer Science, The University of Electro-Communications
Chōfugaoka 1–5–1, Chōfu-shi, Tokyo, 182–8585 Japan
E-mail: †{yaegas-k,maruya-t,yanai}@mm.cs.uec.ac.jp

Abstract Scenes and objects represented in photos have causal relationship to the places where they are taken. In this paper, we propose using geo-informations such as aerial photos and location-related informations as feature for geotagged image recognition. We verify the possibility for reflecting location contexts in image recognition, by evaluating not only recognition rates, but feature fusion weights estimated by Multiple Kernel Learning (MKL). As a result, the mean average precision for 28 categories increased up to 80.87% by the proposed method, compared with 77.71% by the baseline.

Key words geotag, geotagged photos, aerial photos, general image recognition, machine learning, Multiple Kernel Learning

1. はじめに

今日では，デジタルカメラの普及により，撮影位置の情報を持つ位置情報付き写真は WWW(World Wide Web) 上に大量に存在している．また，Flickr などの写真共有を行うソーシャルサイトの普及により，個人が多くの位置情報付き写真を収集，整理し，共有することが容易になった．一方で，Google Maps などのオンラインマッピングサービスも普及し，地図情報の検索機能と共に日々高度化している．地図情報から，デジタル写真や航空写真，Google Street View に代表されるような周辺写真情報へアクセスすることはネットユーザーにとってもはや当然のこととなった．

利用者は自宅から地図を利用した位置の検索だけでなく，航空写真の閲覧など高度なサービスを利用できるようになった．また，写真の収集に関しても，位置情報を積極的に活用しようとする試みが行われてきた．その例として Flickr では，写真に位置情報を付加して投稿し，地図 (Flickr Map) を用いて写真を検索，整理できるようになっている．

大量の写真情報の普及にもかかわらず，それらを自動で整理・分類し，ユーザーの手間を省くことは未だ困難な課題であり続けている．単純な画像分類へのソリューションとして，現状ではタグ (内容を表現する複数の単語の集合) やタイトルなど，テキストベースのメタデータが主流になっている．写真を分類するための一般画像認識の基礎技術が高度化する中で，認識精度の向上を図るにあたり，写真と関連する多様な情報を，特に実世界と関連の深い情報をいかに効率的に組み合わせるかが求められる．

これらの課題について，現在のところ，位置情報付き写真の一般画像認識において写真の撮影位置に対応する航空写真を付加的な画像特徴量として利用する研究が行われている [9], [10]．それらの研究においては，画像認識において，位置情報の有効性が高い認識カテゴリと，そうでないカテゴリを明確に区別し，デジタル写真と実世界情報との対応付けにおける有効な手段を提案することを目標として位置付けられている．

一般に，認識対象と位置は密接な関係があり，例えば，海岸は海と陸の境目にしかなく，海や陸の真ん中には存在し得ない．しかしながら，海岸の写真の認識において位置情報を役立てるには，世界中の海岸の位置を学習データに持っておかないといけない．これには，膨大なデータを用意する必要がある．そこで，我々は，写真の撮影位置の地理的な状況を表す情報源として，航空写真に加えて，位置の周辺を記述するテキスト情報に注目している．地図に掲載されているような，駅や建物の名前などのテキスト情報のことである．写真の画像特徴量に合わせて，航空写真から抽出した画像特徴量とテキストに由来する特徴量を認識に用いることで，写真の撮影場所の地理的なコンテキストを反映した認識が可能となると考えている．

本研究では，分類カテゴリによって，どの程度位置情報由来の特徴量の有効性に違いがあるか分析を行う．認識実験にあたり，Flickr から収集した位置情報付き写真と，それぞれの位置情報に対応する複数の縮尺 (レベル) の航空写真，そして周辺テキストから抽出した特徴量を用いる．各特徴量の有効性を明確に判定するにあたり，機械学習の段階においてマルチカーネル学習 (MKL) を用いて，画像から抽出した特徴量と，位置情報に基づく航空写真，周辺テキスト，緯度経度の各特徴量，さらに撮影時間についての特徴量の認識における重要度の重みを推定し，本稿で追加した周辺テキスト，緯度経度，撮影時間についての特徴の有効性を評価する．

本稿は以下のように構成される．まず実験の全体的手順と本研究の方針について第 3. 節で触れる．後述するように，実験には画像収集，特徴抽出，機械学習の手順を踏むが，特徴抽出で適用する手法の詳細は，第 4. 節で説明する．また，機械学習に用いる手法については第 6. 節で述べる．実験方法と評価方法，結果を第 7. 節で考察し，第 8. 節で結論付ける．

2. 関 連 研 究

我々の扱う位置テキストと航空写真を伴う画像認識について，それぞれ先に研究がおこなわれている．

位置テキスト情報を画像認識に取り入れた例として，Dhiraj ら [6] の研究が挙げられる．彼らの研究では，位置情報から GIS データベースの逆ジオコーディングで地名などを求め，オッズ比に基づく確率的手法で分類し，これと画像特徴量との統合には Naive fusion と Confidence-based fusion の 2 通りの手法を用いている．

Luo ら [7] は Google Earth から手作業で取得した 853 枚の比較的詳細な航空写真を用いた (以下この研究を Third Eye と呼ぶ)．画像は Flickr から収集した 981 枚の写真と，被験者に旅行させることで収集した 720 枚の写真を使用している．画像特徴量は SIFT に基づく bag of visual words 表現と，HSV 空間によるカラーヒストグラムを単純ベクトル結合したものであり，画像の分類には SVM を，航空写真の分類に AdaBoost を用いている．画像と航空写真の認識精度を融合するにあたっては，Qi ら [8] の手法に基づく SVM のメタ分類器を用いて行っている．実験の結果，多くの分類カテゴリで，画像よりも航空写真の精度が良いことを示している．

本研究では一般物体認識を視野に入れ，多量の画像と航空写真を扱う．また，テキスト情報は Yahoo!ローカルサーチを用いて抽出する．特徴量の融合にあたり，Multiple Kernel Learning(MKL) を導入する．

3. 手順と方針

我々の認識実験は全体的に，図 1 に示す要領で行われる．全体の流れとしては，Flickr より収集した位置情報付き画像から特徴抽出したものを，機械学習することで認識精度を検証するものである．

機械学習の段階において，各特徴量がどのように有効に利用されているかどうかを考察する必要がある．本研究では，第 6. 2 節で説明する MKL-SVM を用いて，認識精度のほかに本稿で追加した周辺テキスト，緯度経度，撮影時間についての特徴の有効性を MKL によって推定された重みに基づいて評価する．各手順において用いる手法の詳細については後述する．

4. 画像からの特徴抽出

本節では，画像から認識に必要な特徴を抽出する方法について述べる．画像の特徴を記述する手法としては，本実験では，特徴抽出のために局所特徴の一種である SIFT 特徴を用いる．また，この局所特徴を簡潔に記述するために Bag of Keypoints 手法を用いてデータをベクトル量子化する．画像からは BoK と HSV 色特徴，航空写真からは BoK のみを抽出する．

4. 1 SIFT 特徴

SIFT(Scale Invariant Feature Transform) [2] とは，David Lowe によって提案された特徴点とそれに付随する特徴ベクトルの抽出法である．特徴点周りの局所画

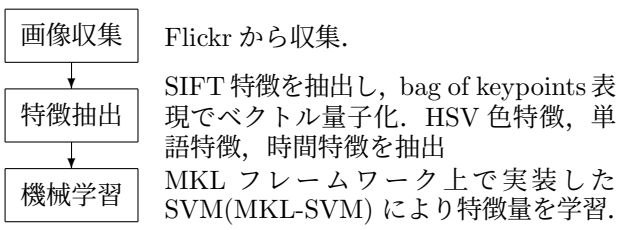

図 1　研究の全体的手順

像パターンを勾配方向ヒストグラムから成る 128 次元特徴ベクトルで表現する．特徴点の抽出は，自動で行う方法 (ガウシアン差分による選定，特定の物体同士の対応点検出に有用) と，手動で指定する方法 (主に格子状あるいはランダム点指定) がある．本実験では，特定の物体を認識することに固執しないので，10 ピクセルの格子状に点を抽出する方法を採用する．

4.2 Bag of keypoints 表現

Bag of keypoints モデル [1] とは，画像を局所特徴の集合と捉えた手法である．局所特徴の特徴ベクトルをベクトル量子化し，visual words と呼ばれる特徴ベクトルを生成する．それらの集合をコードブックと呼び，それを記述子として画像の特徴ベクトルを生成する．これにより画像を visual words の集合 (bag) として表現する．ベクトル量子化は，ユークリッド距離を尺度とし，k-Means 法を用いて行う．本研究では，処理速度や結果における精度の差異を考慮した上で，クラスタ数を 1000 に固定した．

4.3 HSV 色特徴

本研究では，配色の偏りに配慮するため，位置情報付き画像を 5×5 に分割し，各面について $4^3 = 64$ 次元を，画像 1 枚について計 1600 次元の HSV カラーヒストグラム (総和を 1 で正規化) を抽出する．

5. メタデータからの特徴抽出

5.1 地理テキストの特徴

本研究では，位置情報の近傍を周辺検索することで得られるテキスト情報を利用する．具体的には，Yahoo! ローカルサーチ API に撮影位置の緯度経度を与え，その周囲 500m 以内にあるランドマーク等に関する情報 (建物・施設・公園など地図上に掲載され得る情報) を取得する．テキスト情報を，機械学習への入力に対応する特徴ベクトルに変換するにあたり，茶筅 (chasen) [11] で単語に分割し，抽出した全名詞の出現頻度の上位 2000 語に対してヒストグラム (総和を 1 で正規化) を作成し，地理テキスト特徴とする．

5.2 時間・季節情報の利用

本研究では，メタデータとして撮影時刻に基づく時間情報も用いる．特徴量を得るにあたり，時間帯の類似度を判別できる形式とするため，時刻情報をヒストグラムに準ずる表現に変換する．具体的には，月について 12 ビンと時間について 24 ビンを用意し，該当する月・時刻とその隣接にそれぞれ 0.5，0.25 を投票し，ヒストグラムに準ずる表現とする．

5.3 緯 度 経 度

位置情報から抽出しうる特徴のみならず，メタデータとして緯度と経度の 2 次元ベクトルを直接学習・分類に用いる．測位系は Flickr で使用されている WGS84 を想定する．

5.4 航 空 写 真

航空写真を利用するにあたり，それぞれの画像に対応する位置情報を地図サイトで表示し，スクリーンキャプチャしたものを特徴抽出に利用する．特徴抽出の方法はデータセットの画像と同様である．写真の内容によっては，より詳細な縮尺を持つ航空写真の方が検証に際し有効であると考えられるが，撮影位置によっては，詳細な航空写真を入手できないこともある．広範なデータセットに対応するため，図 2 に示すような 4 種類のスケールを採用する．図 2 に示すように，航空写真は撮影位置が中央にくるような 1 辺 256 ピクセルの正方形に加工したものを利用する．

6. 学習と分類

本節では，機械学習で用いる手法の詳細を説明する．認識精度の検証に当たっては，その基本としてサポートベクタマシン (SVM) を用いる．これに加えて画像と航空写真の有効性について判定するために，マルチカーネル学習を導入する．

6.1 サポートベクタマシン

サポートベクタマシン (SVM) はニューロンのモデルとして最も単純な線形しきい素子を用いて，2 クラスのパターン識別器を構成する手法である．カーネル学習法と組み合わせると非線形の識別器になる．本実験では，カーネル関数として以下の式で表現されるヒストグラム特徴に有効であると言われている χ^2RBF カーネルを用いるが，緯度経度に関してのみ，ヒストグラム特徴でないため通常の RBF カーネルを用いる．

$$K(\boldsymbol{x},\ \boldsymbol{x}') = \exp\left(\sum_{i=1}^{D} (x_i - x_i')^2 / (x_i + x_i')\right) \qquad (1)$$

6.2 マルチカーネル学習

本研究では，特徴を統合して画像を認識するために，複数の特徴量のカーネルを線形結合することにより統合カーネルを作成し，それをサポートベクターマシン (SVM) に適用して特徴統合による画像認識を実現する．最適なカーネル (カーネルを重みつきで線形結合したカーネル) のサブカーネルに対する重み j を学習する．これはマルチカーネル学習 (Multiple Kernel Learning, MKL) [3] 問題と呼ばれ，統合カーネルは以下で定式化される．

$$K_{\text{combined}}(\boldsymbol{x},\ \boldsymbol{x}') = \sum_{j=1}^{K} \beta_j k_j(\boldsymbol{x},\ \boldsymbol{x}')$$

$$\beta_j \geqq 0,\ \sum_{j=1}^{K} \beta_j = 1 \qquad (2)$$

最近の研究では，この MKL 問題を凸面最適化問題として効果的に解く方法が提案されている [4]．マルチカーネル学習は SVM のみを前提としたものではないが，SVM のフレームワークで解く方法が一般的で，MKL-SVM と

表 1　カテゴリの種類と選定基準

カテゴリ	選定基準
ディズニーリゾート	ディズニーリゾートの敷地内で撮影されている．建造物や人物など
東京タワー	東京タワーが目立つように映っている．東京タワーの内部からの眺めは含まない．
橋	橋の構成要素が目立つように映っている．橋の上からの眺めは含まない．
神社	神社の境内や鳥居などの建造物が明確に映っている．
建物	高層建築または建物の全体が目立つように映っている．撮影位置から建物までの距離は遠すぎない．
城	城の建築が明確に映っている．
鉄道	鉄道車両が至近でまたは目立つように映っている．
公園	公園内の風景．遊具など公園特有の設備が映っている．
庭園	庭園内で撮影されている．水流や木々など庭園を象徴する人工構成要素が明確に映っている．
風景	周囲に遮るものがなく，遠方の物体が多く明確に映っている．
湖畔	湖畔で撮影されている．水面が目立つように映っている．
川	川岸で撮影されている．水面が目立つように映っている．
海岸	海岸で撮影されている．水面または砂浜が目立つように映っている．
像・モニュメント	仏像・銅像・モニュメントなどの不動の物体が明確に映っている．
自動車	自動車が明確に映っている．自動車からの眺めは含まない．
自転車	自転車が明確に映っている．自転車が小さいものや，自転車からの眺めは含まない．
落書き	落書きが目立つように映っている．
自動販売機	1 台または複数台の自動販売機が明確に映っている．
紅葉	紅葉が目立つように映っている．
桜・花見	花の咲いている桜の木が目立つように映っている．または，花見の情景が映っている
夕日	夕日が目立つように映っている．
コスプレ	派手な衣装をまとい 1 人または複数人で正面を向き顔・上半身・全体のいずれかが明確に映っている．
祭	神輿，提灯，縁日など祭りを演出する構成要素が十分に存在する．
猫	猫が目立つように映っている．
鳥	鳥が目立つように映っている．
花	花が目立つように映っているか，写真のほとんどが花で占められている．
ラーメン	食べられる状態にあるラーメンが明確に映っている．
寿司	食べられる状態にある 1 個または複数個の寿司が明確に映っている．

呼ばれることもある．本研究では，SHOGUN [5] ツールキットを用いて実装した MKL-SVM を使用して実験を行う．

7. 実験方法

実験は，各画像から抽出した特徴量をサポートベクタマシン (SVM) で学習させ，分類結果により精度を判定することにより行う．学習と分類は正解画像と不正解画像の 2 クラスで行う．正解画像については，画像の内容に応じていくつかのカテゴリのデータセット用意し，不正解画像は正解画像に用いないデータをランダムに選択することで構成される．また，SVM のフレームワークの下で，マルチカーネル学習で種類ごとの重みを推定する (MKL-SVM) ことにより，画像と航空写真のどちらがどれだけ有用であるかを判断する．

7.1 データセット

後述する複数のカテゴリを準備し，カテゴリ毎の精度検証を行う．Flickr では大量の画像を収集可能であり，位置情報付き写真は少なくとも 40 万枚は存在する．ただし，それらの写真は Flickr ユーザーの主観的な観点から撮影されたものであり，あるカテゴリにおいてそのカテゴリの写真そのものを描写しているとは限らない．例えばタワーに関する写真はタワーそのもののほかにタワーからの眺め，電車に関する写真は，電車自体のほかに車窓や駅舎などが含まれることがある．

データソースから得られる実験データセットを，より客観的なものに近付けるべく，我々は Flickr 画像を用いて認識の精度検証を行うにあたり，適切な画像を手動で選定することによりデータセットを作成する．

正解データセットについては，表 1 に示すような 28 種類のカテゴリを準備した．これらのデータは，タグなどのメタデータである程度分類したものの中から，カテゴリごとに写真の外観，位置情報ともに適切なものを手動で日本国内で撮影された 200 枚を選定した．正解画像の一例として，図 3 を挙げることができる．

カテゴリの選定に当たっては，Flickr で多く投稿される写真の内容の偏りを考慮しつつ，主に次のような 8 つの観点 (ジャンル) から選択した．また，それによる認識精度に関する仮説を示す．

位置に特有なランドマーク　ディズニーリゾート，東京タワー　[2 カテゴリ]

必然的に位置情報が偏るので，航空写真の特徴量に類似の外観ものが集中するので，航空写真の特徴量に依存する形で認識精度に改善がみられると考えられる．

狭い範囲の地理構成物　橋，神社，建物，城，鉄道　[5 カテゴリ]

地理的要素のうち，特定の物体を対象に描写している．航空写真がある程度詳細なものであれば，これらは把握可能であり，それらの特徴量によって認識精度に若干の改善がみられると考えられる．

広い範囲の地理構成物　公園，庭園，風景　[3 カテゴリ]

地理的要素のうち，撮影対象が広範囲で，画像中の物体も多岐にわたる．公園や庭園は，余ほど詳細なものでない限りは航空写真との相関は考えにくい．風景については，航空写真と関連することも考えられるが，撮影位置と視界の位置が異なるので，航空写真による精度はさほど変化しないと考えられる．

地形　湖畔，川，海岸　[3 カテゴリ]

位置情報が特定の箇所に偏在する．色特徴，SIFT 特徴共に航空写真との相関が考えられる．

屋外の人工物　像・モニュメント，自動車，自転車，落書き，自動販売機　[5 カテゴリ]

主に屋外の物体であるが，航空写真から自明なものを確認することはほぼ不可能である．位置情報・時間情報ともに位置情報による精度変化は考えにくい．

時期依存的要素　紅葉，桜・花見，夕日，コスプレ，祭　[5 カテゴリ]

周期的な季節・時刻を伴うものと，人為的なイベントに関係する．夕日，桜・花見，紅葉については時系列的な相関と色特徴が精度の変化に影響を与えると考えられる．位置情報による変化は考えにくいであろう．コスプレと祭については，そのイベントの性質上時間と共に位置情報に大いに依存する可能性が高い．

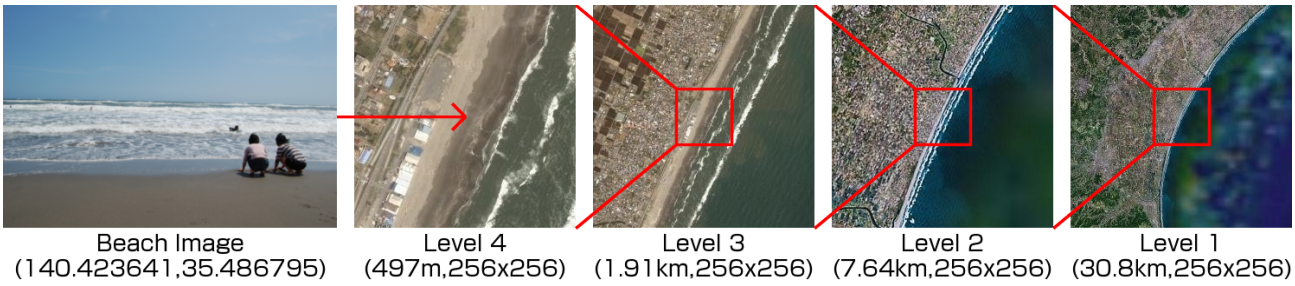

図 2　位置情報付き写真と航空写真の対応付け

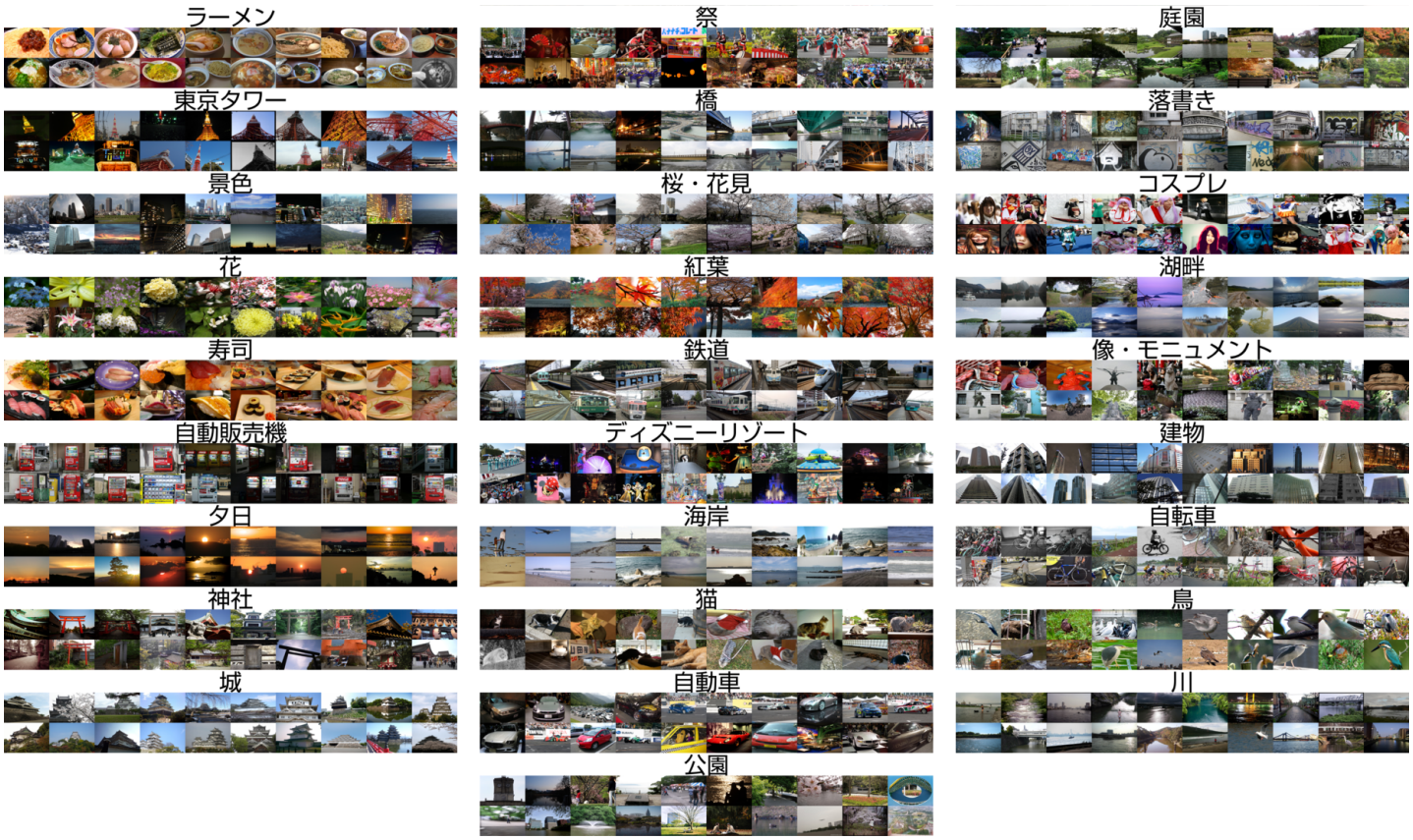

図 3　実験に用いる 28 種類の正解データセットの例.

天然の物体　猫，鳥，花　[3 カテゴリ]

動植物．航空写真・時間特徴共に何ら関連性は持たない．画像特徴で十分な精度が期待できるが，位置情報・時間情報ともに位置情報による精度変化は考えにくい．

食べ物　ラーメン，寿司　[2 カテゴリ]

位置情報に関しては若干関連があるかもしれない．画像特徴で十分な精度が期待できるが，位置情報・時間情報ともに位置情報による精度変化は考えにくい．

不正解データセットは，Flickr から収集した画像データの中から，正解データに用いられていないものをランダムに選定することで 200 枚準備した．

いずれのデータセットも，一般性を保証するため写真の撮影者が偏向しないように選定した．画像は縦と横のうち長い方の画素が 500 ピクセルになるようにリサイズしてある．

カテゴリごとのデータセットの枚数を決定するにあたり，各カテゴリにおいて Flickr から得られる適切な画像の枚数の平均を考慮した．正確な実験結果を保証するためには，一般に大量のデータセットを構築する必要があるが，本研究においては身近なデータからデータセットを構築することを重視した上で，今後万一データソースを見直す場合は認識精度の比較検討と共に考慮することになる．

7.2　特徴量の種類と組み合わせ

画像の bag of keypoints 表現 (以下，結果記述にあたり BoK とも記す)，航空写真の bag of keypoints 表現 (4 レベルでの 4 種類)，緯度経度，周辺情報に関するヒストグラム，撮影時刻に関するヒストグラム表現，HSV カ

表 2　MKL に与える特徴量の組み合わせ.

組合せ表記	BoK	HSV	航空写真 (A)	緯度経度 (G)	位置テキスト (T)	時刻 (D)
Vis	○	○	-	-	-	-
Vis+G	○	○	-	○	-	-
Vis+T	○	○	-	-	○	-
BoK+A	○	-	○	-	-	-
Vis+A	○	○	○	-	-	-
Vis+A+T	○	○	○	-	○	-
Vis+G+T+D	○	○	-	○	○	○
Vis+A+T+D	○	○	○	-	○	○
All	○	○	○	○	○	○

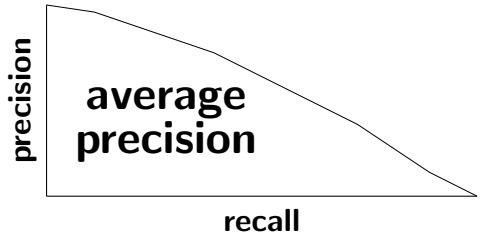

図 4　平均適合率.

ラーヒストグラムの，計 9 種類を扱う．これらを，MKL に与えるにあたり表 2 のように組み合わせる．ただし，航空写真を用いる際は 4 レベルをすべて統合するので，まとめて 1 つの欄に表記している．

7.3 評　　価

認識実験に際しては，後述のように写真の内容を予め各カテゴリに分類し，カテゴリごとに，画像と航空写真をペアにしたものを MKL-SVM を用いて学習・分類・重みの推定を行う．

学習と分類はクロスバリデーションで行う．すなわち，データセットを 5 分割し，1 つを分類用に，残り 4 つを学習用に充てることを，分類用に相当する各 fold について 5 通り繰り返す．

認識精度は各分類結果における SVM の出力値に対して，以下のように平均適合率を計算することで評価する．SVM による出力値に基づいて，テストデータをソートする．ソートしたデータを最初から順番に読み込み，Positive のデータが出現した時点で，それまでの読み込んだデータの数 m_i と，Positive データの出現頻度 i を記録する．ここで $p_i = \frac{i}{m_i}$ とおく．最後までテストデータを読み込んだときのすべての Positive データの出現数を n とすると，平均適合率 (Avarage Precision) は，

$$P = \frac{1}{n}\sum_{i=1}^{n} p_i \qquad (3)$$

で計算される．これは precision-recall グラフの面積に相当する (図 4).

7.4 実 験 結 果

画像と航空写真を MKL-SVM で分類した結果について示す．ここでは，画像と航空写真の 2 組について MKL-SVM で分類，重み推定を行った．結果については，5 回のクロスバリデーションの平均値を示す.

平均適合率の結果を表 3 に示す．表 3 において，

表 4　平均適合率におけるベースライン (BoK あるいは Vis) との差分 (gain).

記号	算出方法	概要
Geo	(Vis+G)−(Vis)	緯度経度の値による性能向上
Text	(Vis+T)−(Vis)	位置テキストによる性能向上
A(BoK)	(BoK+A)−(BoK)	航空写真による性能向上 (BoK のみの差分)
A(Vis)	(Vis+A)−(Vis)	航空写真による性能向上
AT	(Vis+A+T)−(Vis)	航空写真と位置テキストによる性能向上
GTD	(Vis+G+T+D)−(Vis)	航空写真以外の位置情報と時間による性能向上
ATD	(Vis+A+T+D)−(Vis)	緯度経度以外の位置情報と時間による性能向上
All	(All)−(Vis)	特徴統合全体による性能向上

ジャンル	Geo	Text	A(BoK)	A(Vis)	AT	GTD	ATD	All
ランドマーク	2.97	6.69	10.05	6.56	6.60	6.57	6.67	6.57
屋外の人工物	0.61	2.89	3.77	2.49	2.75	2.65	2.71	2.51
狭い地理構成物	1.03	3.19	3.18	3.02	3.30	2.76	2.95	2.85
広い地理構成物	0.98	4.05	2.99	3.77	4.35	3.72	3.84	3.62
時期依存的要素	0.41	1.69	2.50	1.53	1.61	1.46	1.72	1.48
食べ物	0.19	0.84	1.21	0.77	0.69	0.71	0.80	0.68
地形	0.78	3.44	3.73	3.42	3.53	3.18	3.47	3.27
天然の物体	1.02	4.04	3.71	3.79	3.88	3.80	4.11	3.73
総計	**0.89**	**3.16**	**3.61**	**2.96**	**3.15**	**2.89**	**3.08**	**2.88**

見出しは表 2 の組合せ表記を頭文字に因んで省略 (VGTD=Vis+G+T+D など) して表記する．ベースラインである画像 BoK 単体の精度，HSV 単体の精度，BoK+HSV を MKL で統合した精度を左端に示す.

特徴統合による精度の変化について，まず精度のみに着目して考察する．平均適合率におけるベースライン (BoK) との差分 (gain) を表 4 に示す．平均のみに着目すると，単純に緯度経度を使う場合 gain は 0.89 でほとんどないが，位置テキストや航空写真に変換することで，表 4 の ATD に関しては 3.08 に上昇することを示唆しており，これにより提案手法の有効性が示される．また，HSV を統合せず，元画像と航空写真のみを統合すると平均適合率に関する上昇は 3.61 と最も高い.

ジャンルごとの平均に着目すると，位置に固有なランドマークが ATD について+6.67 で位置情報が有効で，緯度経度でも+3.77 の gain である．緯度経度のみの gain は固有ランドマーク以外では 1%程度かそれより低いことから，固有ランドマーク以外では基本的には有効ではない．一方，天然の物体，広い範囲の地理構成物，地形は，+4.04〜4.05 で位置情報に由来する特徴が有効であるが，緯度経度は 1 前後でほとんど有効ではない．このことからも緯度経度を位置テキストや航空写真に変換する本手法の有効性が確認できる.

位置情報が単なる付加的情報ではなく，適切な重みにより分類されていることを確認するために，表 6 に全特徴の組み合わせで MKL を行った時の重みを示す (一部の特徴量の組み合わせによる重みは紙面の都合で割愛する)．また，いかなる位置テキストが実験に影響を及ぼしたかを示すため，位置テキストの重みが高かった 6 カテゴリと，そうでない 6 カテゴリについて，出現頻度の高い上位 20 語の単語を表 5 に示す.

位置テキストにおいて頻繁に出現する単語は，概ね地理的には均一に分散していると思われる構成物の名称が多い．ただし，位置テキストの重みが大きいカテゴリに関しては，特定の地名を表すいくつかの固有名詞 (有明，舞浜，姫路，六本木など) において高い頻度を記録している．このテキストの偏在が学習において大きく影響し

表 3　MKL-SVM における平均適合率の計算結果．各カテゴリごとと，各ジャンルごとの結果をそれぞれ示す．スペースの関係で見出しは表 2 の組合せ表記を頭文字に因んで省略 (VGTD=Vis+G+T+D など) して表記する．

ジャンル	カテゴリ	BoK	HSV	Vis	VG	VT	BA	VA	VAT	VGTD	VATD	All
狭い範囲の地理構成物	橋	69.78	63.87	69.78	71.95	74.76	74.32	75.43	**76.44**	74.65	74.83	75.12
	建物	78.58	74.35	79.92	80.64	**81.77**	79.80	81.41	81.26	81.35	80.90	81.04
	城	80.13	81.21	82.09	82.91	83.72	82.99	83.77	83.75	83.74	**83.79**	83.73
	神社	70.32	70.60	71.34	73.11	**77.28**	75.01	76.72	77.15	76.08	77.05	76.20
	鉄道	74.43	72.59	77.71	77.35	**79.25**	77.03	78.62	78.70	78.83	79.04	78.98
	平均	74.65	72.52	76.17	77.19	79.36	77.83	79.19	**79.46**	78.93	79.12	79.02
屋外の人工物	自転車	76.05	71.85	76.72	77.22	**79.76**	77.90	78.80	79.39	78.87	79.45	78.80
	自動車	70.80	70.74	75.72	76.18	**77.54**	74.84	76.54	76.72	77.01	77.05	76.70
	自動販売機	79.88	82.81	83.26	83.11	83.22	81.59	83.08	83.03	**83.39**	83.10	83.28
	像・モニュメント	66.45	65.94	67.78	68.43	72.81	70.78	72.94	**73.25**	72.34	72.27	72.49
	落書き	72.59	75.60	76.68	78.27	81.27	79.53	81.27	81.53	81.81	**81.82**	81.42
	平均	73.15	73.39	76.03	76.64	**78.92**	76.93	78.53	78.78	78.69	78.74	78.54
時期依存的要素	コスプレ	75.67	79.42	80.29	81.53	83.95	83.68	84.04	84.04	83.73	**84.09**	83.73
	紅葉	80.77	82.53	83.27	83.51	83.83	82.08	83.82	83.77	83.96	83.91	**83.97**
	祭	73.31	74.89	76.90	77.30	**80.19**	76.20	79.59	79.97	79.23	80.08	79.58
	桜・花見	80.22	79.70	82.13	82.29	**82.91**	80.56	82.67	82.77	82.80	82.88	82.55
	夕日	82.90	83.41	83.68	83.71	83.83	82.83	83.83	83.78	83.86	**83.90**	83.84
	平均	78.57	79.99	81.25	81.67	82.94	81.07	82.79	82.86	82.72	**82.97**	82.73
広い範囲の地理構成物	公園	70.42	69.05	71.24	71.95	78.29	74.99	77.63	**78.76**	77.76	77.53	77.33
	庭園	77.57	77.61	79.80	81.06	82.39	81.06	82.35	**82.72**	82.20	82.59	82.27
	風景	74.23	74.28	75.75	76.73	78.27	75.16	78.12	**78.37**	78.00	78.20	78.06
	平均	74.08	73.65	75.60	76.58	79.65	77.07	79.36	**79.95**	79.32	79.44	79.22
地形	海岸	81.02	80.06	81.70	81.99	**83.76**	83.50	83.54	83.61	83.72	83.62	83.55
	湖畔	78.27	76.25	79.41	80.53	**82.87**	81.42	82.42	82.67	82.67	82.78	82.39
	川	74.56	74.22	75.81	76.74	80.61	80.11	81.22	81.24	80.06	80.93	80.78
	平均	77.95	76.84	78.97	79.75	82.41	81.68	82.39	**82.51**	82.15	82.44	82.24
天然の物体	花	77.98	77.70	80.71	81.02	81.53	78.30	81.31	81.24	82.23	**82.65**	82.22
	鳥	69.75	71.32	73.00	74.02	**81.38**	78.89	80.89	81.21	80.36	81.20	80.26
	猫	67.96	67.24	71.72	73.43	**74.63**	69.62	74.60	**74.63**	74.23	73.93	74.14
	平均	71.89	72.09	75.14	76.16	79.18	75.60	78.93	79.03	78.94	**79.26**	78.87
位置に特有なランドマーク	ディズニーリゾート	68.37	70.98	73.20	78.41	84.16	84.14	84.14	**84.17**	84.10	84.16	84.09
	東京タワー	79.34	80.59	81.54	82.28	**83.95**	83.67	83.73	83.75	83.78	83.91	83.78
	平均	73.86	75.79	77.37	80.34	**84.05**	83.90	83.93	83.96	83.94	84.03	83.93
食べ物	ラーメン	82.59	80.85	83.03	83.09	**83.28**	82.77	83.18	83.11	83.00	83.10	82.97
	寿司	79.85	79.51	81.76	82.09	83.18	82.09	83.15	83.07	83.20	**83.28**	83.19
	平均	81.22	80.18	82.40	82.59	**83.23**	82.43	83.17	83.09	83.10	83.19	83.08
全体平均		75.49	75.33	77.71	78.60	**80.87**	79.10	80.67	80.86	80.61	80.79	80.59

ていると考えられる．

また，表 6 で比較すると，航空写真よりも位置テキストの方が重みが高い．このことからも位置情報付き画像の認識にはテキストが画像認識に寄与するうえで，地理的コンテキストを大きく反映したと考えることができる．

一方，緯度経度の重みは著しく少ないことから，緯度経度情報のみではその位置のコンテキストを記述するには十分でないことが言える．したがって，位置情報を，対応するコンテキストに変換することの有効性を見て取ることができる．ただし，表 4 において航空写真による gain が+2.96 であり，Text,AT,ATD に次いで多いことから，本実験よりも詳細な航空写真が利用可能であると仮定し，かつ地理構成物を認識対象にする場合，元画像の特徴量に航空写真特徴のみを統合した場合でも十分な精度向上を見込める可能性はあると考えられる．

時間情報の特徴は，1 日と 1 年の特定の周期に集中する場合に，認識に対して有効に作用する．本実験においては，カテゴリ「桜・花見」に対して高い重みを示したが,「夕日」や「紅葉」など，期間が長い場合やイベントの発生する時期に (同じ周期内であっても) ばらつきが目立つ場合，時間情報は有効性ではない．したがって，時間情報を画像認識に十分に適用するには，時間情報を補強するに足るメタデータの利用が必要になると思われる．

以上のことから，周辺テキストの特徴量は画像認識に有効であるが，緯度経度・時間情報はさほど有効ではないことがいえる．緯度経度・時間情報については追加のメタデータが必要であると考えられるが，位置情報を航空写真やテキスト情報などで補強することにより，地理的コンテキストを反映した上での認識への有効性が示された．

なお，Flickr から得られる写真は個人により撮影されているものである以上，位置情報の偏向については常に考慮しなければならない問題である．航空写真については地図サイト上での必要なリソースの利用可能性が実験方法を左右する．これらについてはデータの大量収集と効率的なデータ・カテゴリの選定に対して包括的に配慮することで，引き続き対応していきたいと考える．

8. おわりに

8.1 まとめ

本研究では，位置情報付き写真の一般画像認識を拡張するにあたり，写真の撮影位置に対応する航空写真と周辺テキストの情報を付加的な画像特徴量として利用する手法を導入した．認識実験においては各種特徴量の組み合わせによる認識精度の変化を検証するとともに，マルチカーネル学習 (MKL,Multiple Kernel Learning) を導入することで，特徴量の種類ごとの認識への関与を定量的に分析した．28 種類のカテゴリによる実験では，画像単体の場合と比較して 75%から 80%へ向上した．

緯度経度のみでは，特定のランドマークを除いて十分な精度向上は得られなかった．一方，位置情報に由来する特徴量は，緯度経度の拡張・補強し，認識精度の向上に有効に作用した．位置テキストは，その有効性につい

表 5　周辺テキストの抽出結果．ヒストグラムビンをカテゴリごとに平均し，対応する単語を降順で上位 20 語まで示す．9 種類の分類において位置テキストの重みが特に高かった 6 カテゴリと，特に低かった 6 カテゴリを示す．

位置テキストの重みが高いカテゴリ

コスプレ		鳥		ディズニーリゾート		城	
有明	0.0391	市立	0.0202	場	0.117	姫路	0.0324
ビル	0.0388	寺	0.0197	駐車	0.1074	市立	0.0252
東京	0.025	局	0.0197	東京	0.0649	局	0.0201
場	0.0223	郵便	0.0191	舞浜	0.0424	ビル	0.0199
原宿	0.0222	小学校	0.0189	リゾート	0.0416	学校	0.0173
センター	0.0181	ビル	0.0171	パーキング	0.0347	橋	0.0161
公園	0.0144	橋	0.015	ホテル	0.0321	郵便	0.0158
棟	0.013	線	0.0136	南葛西	0.024	寺	0.0149
幕張	0.0128	センター	0.0131	ベイ	0.0197	公園	0.0143
号	0.0124	院	0.0121	入口	0.0139	小学校	0.0133
明治	0.0119	公園	0.0117	東急	0.012	センター	0.0131
神宮	0.0099	幼稚園	0.0114	区立	0.0115	高等	0.0126
橋	0.0088	保育園	0.0111	シアター	0.0108	神社	0.0124
ホテル	0.0083	場	0.0094	公園	0.0105	幼稚園	0.0123
ホール	0.0081	学校	0.0088	ランド	0.01	会館	0.0114
東郷	0.0079	神社	0.0088	野毛	0.01	私立	0.0109
会館	0.0078	東京	0.0087	立体	0.0092	城	0.0107
通り	0.0077	中学校	0.0087	クリスタル	0.0089	大阪	0.0101
記念	0.0076	旭川	0.0078	保育園	0.0089	中学校	0.01
代々木	0.0071	館	0.007	小学校	0.0087	県立	0.0092

テキストが高いカテゴリ				テキストが低いカテゴリ			
神社		東京タワー		花		桜・花見	
ビル	0.034	ビル	0.1089	ビル	0.0326	ビル	0.0359
局	0.0187	大使館	0.0395	局	0.0186	局	0.0177
郵便	0.0182	会館	0.0255	郵便	0.0178	郵便	0.0171
院	0.0174	六本木	0.0211	寺	0.0163	寺	0.0151
寺	0.0169	神谷町	0.0199	小学校	0.0161	新宿	0.0136
小学校	0.0136	日本	0.0186	市立	0.0158	小学校	0.0133
神社	0.0132	麻布	0.0183	新宿	0.0136	東京	0.0124
市立	0.0131	聖	0.0183	幼稚園	0.0121	市立	0.0115
線	0.0122	虎ノ門	0.0181	東京	0.0117	センター	0.0113
幼稚園	0.0113	飯倉	0.018	センター	0.0116	院	0.0111
センター	0.0107	寺	0.017	線	0.0112	会館	0.0108
館	0.0104	タワー	0.0164	保育園	0.0111	学校	0.0108
会館	0.0099	館	0.014	学校	0.0103	幼稚園	0.0107
保育園	0.0094	泉	0.0111	公園	0.0098	保育園	0.0104
ホテル	0.0092	局	0.0109	院	0.0094	館	0.0084
公園	0.0086	郵便	0.0109	セブン	0.0092	公園	0.0082
学校	0.0085	芝	0.0101	イレブン	0.0091	中学校	0.0079
東京	0.0082	殿	0.0097	中学校	0.0088	区立	0.0078
場	0.0079	教会	0.0093	橋	0.0085	橋	0.0076
中学校	0.0078	振興	0.0092	会館	0.0082	神社	0.0075

位置テキストの重みが低いカテゴリ

自動販売機		紅葉		ラーメン		夕日	
ビル	0.0286	院	0.0274	ビル	0.0522	線	0.027
局	0.02	寺	0.0241	局	0.0173	公園	0.0187
郵便	0.0187	市立	0.0202	郵便	0.0169	市立	0.0175
保育園	0.0168	局	0.0166	市立	0.0117	小学校	0.0166
富山	0.0154	小学校	0.0163	センター	0.0115	局	0.0166
小学校	0.0141	郵便	0.0161	ホテル	0.0109	ビル	0.0165
区立	0.0132	線	0.0159	小学校	0.0109	郵便	0.0158
寺	0.0123	ビル	0.0142	寺	0.0106	センター	0.0141
センター	0.0122	蜷峨	0.0125	東京	0.0102	橋	0.0139
幼稚園	0.0122	保育園	0.0116	保育園	0.0099	寺	0.0127
ホテル	0.0114	京都	0.0103	幼稚園	0.0099	保育園	0.0124
会館	0.0108	堂	0.0101	公園	0.0095	横浜	0.0107
東京	0.0104	橋	0.0101	会館	0.0093	東京	0.0101
公園	0.01	神社	0.01	渋谷	0.0083	場	0.0094
神社	0.0099	幼稚園	0.0097	橋	0.0083	幼稚園	0.0092
橋	0.0093	館	0.0095	セブン	0.0081	セブン	0.0087
館	0.009	場	0.0089	イレブン	0.008	イレブン	0.0087
セブン	0.0086	病院	0.0088	京都	0.008	ホテル	0.0084
イレブン	0.0084	公園	0.0084	ローソン	0.0079	中学校	0.0077
病院	0.0081	学校	0.0084	銀行	0.0078	国道	0.0074

て航空写真特徴を凌駕した．時間情報は，特定の周期を捉えられない場合は有効に作用せず，画像認識への適切な導入にあたりこれを補完するメタデータが必要とされる．

8.2　今後の課題

位置情報による認識精度の向上をさらに検証し，役立てるには，大量かつ良質なデータの収集について改めて模索していく必要がある．本研究では国内画像で 28 カテゴリで 2 クラス分類による実験を行ったが，今後はデータ収集から再検討した上で，世界規模でより多くのカテゴリについて実験するほか，マルチクラス分類による実験も視野に入れる．

位置テキスト特徴をさらに有効なものにするには，単語の抽出方法については更に検討の余地がある．具体的には形態素解析に際して，地名に関する固有名詞の辞書を強化するほか，単語を分割しすぎている部分を辞書の補強で抑制し，最終的に地理的コンテキストの抽出に適切な状態にする必要がある．

表 6　MKL による全種類の特徴の重み推定結果．

カテゴリ		画像情報（BoK,HSV）	航空写真（1,2,3,4）	緯度経度	位置テキスト	時刻
狭い地理構成物	橋	0.3843(0.2690,0.1153)	0.3408(0.0529,0.0557,0.0295,0.2028)	0.0089	0.2289	0.0371
	神社	0.4504(0.0999,0.3505)	0.0690(0.0012,0.0020,0.0001,0.0657)	0.0043	0.4307	0.0456
	建物	0.6147(0.3897,0.2251)	0.1207(0.0500,0.0425,0.0004,0.0278)	0.0455	0.2123	0.0068
	城	0.2441(0.0654,0.1787)	0.1094(0.0877,0.0041,0.0124,0.0053)	0.0521	0.5941	0.0002
	鉄道	0.6918(0.3016,0.3903)	0.1004(0.0212,0.0151,0.0004,0.0637)	0.0001	0.1869	0.0208
	平均	**0.4771(0.2251,0.2519)**	**0.1480(0.0426,0.0239,0.0085,0.0731)**	**0.0222**	**0.3306**	**0.0221**
屋外の人工物	像	0.2223(0.2156,0.0067)	0.1381(0.0587,0.0041,0.0464,0.0288)	0.4492	0.0336	0.1568
	自動車	0.6772(0.3290,0.3481)	0.0285(0.0065,0.0000,0.0001,0.0219)	0.0146	0.2678	0.0119
	自転車	0.5520(0.4410,0.1111)	0.0642(0.0076,0.0143,0.0003,0.0420)	0.0098	0.3391	0.0349
	落書き	0.2304(0.0682,0.1621)	0.2846(0.0141,0.1064,0.1043,0.0598)	0.0493	0.3728	0.0630
	自動販売機	0.8755(0.2851,0.5904)	0.0533(0.0064,0.0127,0.0094,0.0248)	0.0032	0.0085	0.0594
	平均	**0.5115(0.2678,0.2437)**	**0.1137(0.0187,0.0275,0.0321,0.0355)**	**0.1052**	**0.2044**	**0.0652**
時期依存的要素	紅葉	0.6158(0.2298,0.3861)	0.1490(0.0836,0.0297,0.0037,0.0318)	0.0066	0.0025	0.2261
	桜・花見	0.4437(0.3291,0.1147)	0.0402(0.0002,0.0019,0.0381,0.0001)	0.0016	0.0093	0.5051
	夕日	0.8096(0.4022,0.4074)	0.0635(0.0001,0.0004,0.0000,0.0631)	0.0000	0.0002	0.1267
	コスプレ	0.0485(0.0352,0.0133)	0.0036(0.0001,0.0001,0.0001,0.0033)	0.0126	0.9291	0.0062
	祭	0.5667(0.2347,0.3320)	0.1248(0.0012,0.0004,0.0005,0.1226)	0.0015	0.2154	0.0916
	平均	**0.4969(0.2462,0.2507)**	**0.0762(0.0170,0.0065,0.0085,0.0442)**	**0.0045**	**0.2313**	**0.1911**
広い地理構成物	公園	0.3971(0.1688,0.2283)	0.0881(0.0034,0.0101,0.0205,0.0541)	0.0169	0.4178	0.0800
	庭園	0.4740(0.1338,0.3402)	0.1227(0.0080,0.0033,0.0170,0.0944)	0.0269	0.3512	0.0252
	風景	0.5634(0.2895,0.2739)	0.0452(0.0000,0.0001,0.0078,0.0374)	0.0043	0.3587	0.0283
	平均	**0.4782(0.1974,0.2808)**	**0.0853(0.0038,0.0045,0.0151,0.0620)**	**0.0161**	**0.3759**	**0.0445**
地形	湖畔	0.3539(0.2184,0.1355)	0.2694(0.0749,0.0877,0.0150,0.0918)	0.0040	0.3435	0.0292
	川	0.3465(0.2492,0.0973)	0.3181(0.0063,0.0523,0.0614,0.1981)	0.0108	0.2447	0.0800
	海岸	0.1888(0.1816,0.0073)	0.5003(0.0108,0.1632,0.0013,0.3250)	0.0015	0.2934	0.0159
	平均	**0.2964(0.2164,0.0800)**	**0.3626(0.0307,0.1011,0.0259,0.2049)**	**0.0054**	**0.2938**	**0.0417**
天然の物体	猫	0.6929(0.2690,0.4239)	0.0718(0.0002,0.0032,0.0007,0.0677)	0.0100	0.2217	0.0036
	鳥	0.1869(0.0888,0.0980)	0.0296(0.0004,0.0039,0.0002,0.0252)	0.0000	0.7126	0.0708
	花	0.8196(0.3356,0.4840)	0.0063(0.0001,0.0007,0.0015,0.0041)	0.0179	0.0269	0.1293
	平均	**0.5665(0.2312,0.3353)**	**0.0359(0.0002,0.0026,0.0008,0.0323)**	**0.0093**	**0.3204**	**0.0679**
ランドマーク	ディズニー	0.0001(0.0001,0.0001)	0.3288(0.1460,0.1827,0.0001,0.0001)	0.0004	0.6706	0.0000
	東京タワー	0.1479(0.1474,0.0005)	0.4230(0.0000,0.0011,0.4217,0.0001)	0.0007	0.4280	0.0003
	平均	**0.0740(0.0737,0.0003)**	**0.3759(0.0730,0.0919,0.2109,0.0001)**	**0.0006**	**0.5493**	**0.0002**
食べ物	ラーメン	0.9317(0.4588,0.4729)	0.0563(0.0000,0.0000,0.0092,0.0470)	0.0018	0.0005	0.0097
	寿司	0.6789(0.2435,0.4354)	0.0021(0.0001,0.0005,0.0000,0.0016)	0.0043	0.2616	0.0532
	平均	**0.8053(0.3512,0.4541)**	**0.0292(0.0001,0.0002,0.0046,0.0243)**	**0.0030**	**0.1310**	**0.0314**
全体平均		**0.4717(0.2314,0.2403)**	**0.1411(0.0229,0.0285,0.0286,0.0611)**	**0.0271**	**0.2915**	**0.0685**

文　献

[1] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *Proc. of ECCV Workshop on Statistical Learning in Computer Vision*, pp. 59–74, 2004.

[2] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110, 2004.

[3] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, Vol. 5, pp. 27–72, 2004.

[4] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schaölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, Vol. 7, pp. 1531–1565, 2006.

[5] Shogun - A Large Scale Machine Learning Toolbox. http://www.shogun-toolbox.org/

[6] D. Joshi and J. Luo. Inferring generic activities and events from image content and bags of geo-tags. In *Proc. of ACM International Conference on Image and Video Retrieval*, 2008.

[7] J. Luo, J. Yu, D. Joshi, and W. Hao. Event recognition: Viewing the world with a third eye. In *Proc. of ACM International Conference Multimedia*, 2008.

[8] G. Qi, X. Hua, Y. Rui, J. Tang, T. Mei, and H. Zhang. Correlative multi-label video annotation. In *Proc. of ACM International Conference Multimedia*, pp. 17–26, 2007.

[9] K. Yaegashi and K. Yanai. Can Geotags Help Image Recognition?. *Proc. of the Pacific-Rim Symposium on Image and Video Technology*, pp. 361–373, 2009.

[10] 八重樫恵太, 柳井啓司. マルチカーネル学習を用いた画像特徴と航空写真特徴の重要度の推定, 電子情報通信学会パターン認識・メディア理解研究会, 2009.

[11] 茶筌. http://chasen.naist.jp/hiki/ChaSen/.