# 位置情報とラベルの共起性に着目したランドマークアノテーション
## ～クリッカブル・リアルワールドの実現に向けて～

島田　敬士† Vincent CHARVILLAT†† 長原　一† 谷口 倫一郎†

† 九州大学 大学院システム情報科学研究院
〒 819–0395 福岡市西区元岡 744 番地
†† IRIT-ENSEEIHT
2 rue Camichel, B.P. 7122, F-31071 Toulouse Cedex 7, France
E-mail: †{atsushi,nagahara,rin}@limu.ait.kyushu-u.ac.jp, ††Vincent.Charvillat@enseeiht.fr

**あらまし**　Web 上で公開されている大規模画像データベースの画像とその画像に付与されているラベル情報を利用して，ラベルが道の画像に対して適切なラベル付けを行う画像アノテーションに関する研究が盛んに取り組まれている．本稿では，画像アノテーションのなかでも，ランドマークアノテーションと呼ばれる世界の有名建築物やモニュメントなどに対してアノテーションを実現する手法について提案する．ランドマークアノテーションの場合，位置情報を有効活用する手法が一般的であるが，画像がどのような場所で撮影されたものであるかという観点での利用がこれまでの手法であった．これに対して本研究では，位置とラベルの共起性にも着目したランドマークアノテーション手法を提案する．実験を通して，画像，ラベル，位置の 3 種類の情報間の関連性に基づくランドマークアノテーションの有効性を確認できた．

**キーワード**　画像アノテーション，撮影位置情報，クリッカブル・リアルワールド

## 1. はじめに

Flickr や Picasa をはじめ，Web 上に大量の画像が公開されるようになってきている．そのような画像共有サイトから取得した大量の画像とその画像に付与されたラベルを利用して，ラベルが未知の画像に対して適切なラベルの付与を行う画像アノテーションの研究が盛んに行われている [1]～[4]．画像特徴とラベル特徴の関係を学習することで，未知の画像に対してのラベル付けを行う手法が主流であるが，近年では，画像とラベルの組だけではなく，Wikipedia や Google Image Search の情報を連携して性能を向上させる手法 [5]～[7] が提案されている．それらの手法では画像に付与された位置情報を利用して学習サンプルの収集法を工夫している．本稿では，画像が撮影された位置情報を，先行研究のように学習サンプル収集のために利用するだけではなく，画像アノテーションの精度を向上させるために利用する方法を提案する．本稿では画像アノテーション問題を，Image-based Labeling と Label-based Localization という画像特徴からラベルを推定する問題と，ラベルから位置情報を推定する問題として定式化し，これらの組合せによって画像アノテーションを行う．

## 2. 位置情報を利用した画像アノテーション

位置情報を利用した画像アノテーションは，我々がこれまでに提案しているクリッカブル・リアルワールドと名付けられたモバイル端末による情報検索サービスの枠組み [8], [9] で利用される．詳細は文献 [8], [9] に譲るが，クリッカブル・リアルワールドでは，モバイル端末のカメラを利用して実世界中の対象を撮影することを情報検索のトリガとしている．すなわち，システム側は撮影された対象を識別し，情報検索のためのキーワードを推定する必要がある．学習サンプルには公開画像データベースを利用し，モバイル端末からは画像のみではなく，撮影位置が GPS で獲得できる環境を想定している．このような条件で，画像アノテーションの問題を定式化する．画像，ラベル，モバイル端末の位置をそれぞれ $\boldsymbol{x}, w, g$ をすると，ベイズの定理から推定すべきラベル $w$ に関する条件付確率は，次式で表される．

$$P(w|\boldsymbol{x},g) = \frac{P(w)P(\boldsymbol{x},g|w)}{P(\boldsymbol{x},g)} \tag{1}$$

$$\propto P(w)P(\boldsymbol{x}|w)P(g|\boldsymbol{x},w) \tag{2}$$

$\boldsymbol{x}$ と $g$ の独立性を仮定することで naive bayes により，

$$P(w|\boldsymbol{x},g) \propto P(w)P(\boldsymbol{x}|w)P(g|w) \tag{3}$$

と変形でき，さらに右辺の第 1 項と第 2 項にベイズの定理から，

$$P(w|\boldsymbol{x},g) \propto P(\boldsymbol{x})P(w|\boldsymbol{x})P(g|w) \tag{4}$$

となる．実世界を撮影される画像の事前確率は一定と仮定すると，最終的に，

$$P(w|\boldsymbol{x}, g) \propto P(w|\boldsymbol{x})P(g|w) \tag{5}$$

と変形することができる．本稿では，式（5）右辺第 1 項の $P(w|\boldsymbol{x})$ を Image-based Labeling，第 2 項 $P(g|w)$ を Label-based Localization と呼ぶことにする．従来の画像アノテーションでは，Image-based Labeling$P(w|\boldsymbol{x})$ によって画像に与えるラベルを推定する手法が多く提案されている [1]〜[4]．一方で，画像アノテーションに位置情報を利用した手法もあるが，学習サンプルを収集する際に利用する方法 [5]〜[7] が主流であり，本稿で提案するように，ラベルが付与された位置情報を画像アノテーションに利用する方法は提案されていない．提案手法では，ふたつの条件付確率の積として最終的にラベル $w$ の確率を計算する．Label-based Localization では画像特徴は利用されず，ラベル $w$ とモバイル端末の位置 $g$ との関係のみを利用して，ラベル付与の性能改善が期待される．以降の節では，Image-based Labeling と Label-based Localization について述べ，実験結果を報告する．

## 3. Image-based Labeling

Image-based Labeling は，Harada らが提案している確率正準相関分析 [10] による画像アノテーション手法 [2] を利用する．詳細については，文献 [2] に委ね，ここでは処理の概要を述べる．学習サンプルの各画像に対して，SURF [11] 特徴を抽出し，k-means クラスタリングを行うことで，画像特徴を Bag-of-Visual-Words 表現する．一方，ラベル特徴としては各ラベルの単語ヒストグラムを作成する．ひとつの画像には複数のラベルが付与されていることもあり，学習サンプル全体で付与されているラベルの種類を $W$ とすると，単語ヒストグラムはビン数 $W$ のバイナリ形式になる．以降では，画像特徴を $\boldsymbol{x}$，ラベル特徴を $\boldsymbol{y}$ で表す．確率的正準相関分析 [10] を利用して，画像特徴 $\boldsymbol{x}$ とラベル特徴 $\boldsymbol{y}$ を潜在確率変数 $\boldsymbol{z}$ によって間接的に結びつけるトピックモデルを生成する．各学習サンプル $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$ は潜在空間で正規分布 $p(\boldsymbol{z}|\boldsymbol{x}_i, \boldsymbol{y}_i)$ を構成し，画像とラベルの特徴が似たサンプルのなす分布同士が潜在空間では近くに配置される．Harada らはこれらをそれぞれ識別器と考え，各単語についてのベイズの最適識別器を構築している．

未知画像を $\boldsymbol{x}_{new}$，$l_i$ を分布 $p(\boldsymbol{z}|\boldsymbol{x}_i, \boldsymbol{y}_i)$ が未知画像 $\boldsymbol{x}_{new}$ の隠れ状態として選択される仮説を表すものとする．このとき，単語 $w$ に対するベイズの最適識別器による事後確率は，

$$P(w|\boldsymbol{x}_{new}) = \sum_{i=1}^{N} P(w|l_i)P(l_i|\boldsymbol{x}_{new}). \tag{6}$$

となる．一方，未知画像 $\boldsymbol{x}_{new}$ は，潜在空間で $P(\boldsymbol{z}|\boldsymbol{x}_{new})$ なる分布を構成する．この分布と，各学習サンプルが構成する分布との同時確率分布から，各仮説の事後確率 $P(l_i|\boldsymbol{x}_{new})$ は次式で定義される．

$$P(l_i|\boldsymbol{x}_{new}) = \frac{\int \sqrt{p(\boldsymbol{z}|\boldsymbol{x}_i, \boldsymbol{y}_i)p(\boldsymbol{z}|\boldsymbol{x}_{new})}dz}{\sum_{j=1}^{N} \int \sqrt{p(\boldsymbol{z}|\boldsymbol{x}_j, \boldsymbol{y}_j)p(\boldsymbol{z}|\boldsymbol{x}_{new})}dz} \tag{7}$$

ただし，分母は $\sum_{j=1}^{N} P(l_i|\boldsymbol{x}_{new}) = 1$ とするための正規化項である．最後に，$P(w|l_i)$ は

$$P(w|l_i) = \mu\delta_{w,l_i} + (1-\mu)\frac{\log(N/N_w)}{\log(N)} \tag{8}$$

で定義される．ここで，$N_w$ は，ラベル $w$ が付与されていた画像の枚数，$N$ は画像の総数である．また，$\delta_{w,l_i}$ は，学習サンプル $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ にラベル $w$ が付与されていれば 1，そうでなければ 0 となる．$\mu$ は，アルファブレンドの値であり，本稿では $\mu = 0.5$ とした．

## 4. Label-based Localization

### 4.1 処理の概要

Label-based Localization の処理の流れを図 1 に示す．まず，クエリ位置と学習サンプルのラベルが付与された位置間の距離を計算する．位置情報は緯度と経度により表されているため，2 点間の距離はヒュベニの距離計算式 [12] により獲得する．得られた距離を利用して，各ラベル $w$ に対してクエリ位置からの相対距離マップ（図 1 の Label map）を作成する．クエリ画像内のランドマークを表すラベルは，クエリ位置から近い場所で多く付与され，ランドマーク以外の地理情報などは，クエリ位置周辺のみではなく，クエリ位置から離れた場所でも付与されると期待できる．そのため，適切な閾値 $TH$ を設定し，その閾値より近い範囲で付与されたラベルの数を基に，確率 $P(g|w)$ を計算する．確率 $P(g|w)$ と距離の閾値 $TH$ の具体的な求め方については後述の節で述べる．

### 4.2 条件付確率 $P(g|w)$ の計算

Label-based Localization で得られる条件付確率 $P(g|w)$ は，ラベル $w$ が与えられた時に，そのラベルがどれだけクエリ位置で付与されているかを表したものである．厳密には，位置 $g$ に対する条件付き確率を計算する必要があるが，実際には撮影位置は撮影者によって多少変わるものであり，また，GPS の計測誤差等が原因で位置がずれてレベルが付与されていることも考えられる．そこで，クエリ位置を中心にある範囲内（実験では $1,000m$ とした）の学習サンプル内のラベル $w$ を利用して，クエリ位置周辺でそのラベルがどれだけ付与されやすいかという確率を計算することにする．収集した学習サンプルに含まれるラベル $w$ の数を $N_w$ とする．また，$d_i, (i = i, \ldots, N_w)$ をラベル $w$ が付与されたクエリ位置からの距離とする．このとき，条件付確率 $P(g|w)$ を，以下の式で定義する．

$$P(g|w) = \frac{\sum_{i=1}^{N_w} I(d_i, w)}{N_w} \tag{9}$$

$$I(d_i, w) = \begin{cases} 1 & \text{if } d_i < TH \text{ for label} w \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

閾値 $TH$ よりも近い範囲で，付与されているラベルの数が多いほど確率は高くなる．

### 4.3 閾値 $TH$ の獲得

ランドマークによっては近い位置から撮影する対象や遠くから撮影する対象があるため，式 (10) で示した閾値 $TH$ を全てのシーンで同じ値を利用してしまうと，所望の条件付確率 $P(g|w)$ が得られる場合とそうでない場合が生じてしまう．また，閾値 $TH$ そのものを経験則で決定することも非常に難しい．そこで，シーンに応じて適応的に閾値 $TH$ を推定する．ラベルの種類の総数を $W$，ラベル $w \in W$ が付与されたクエリ位置からの距離の平均（$d_i, (i = i, \ldots, N_w)$ の平均）を $M_w$ とする．このとき，閾値 $TH$ を次式で決定する．

$$TH = \min_{w \in W} M_w \quad (11)$$

ランドマークを示すラベルは他のラベルに比べて，“平均的” にクエリ位置から近い場所で付与されることが多いという仮定から，このように各ラベル $w$ の距離の平均 $M_w$ の最小値を閾値としている．紙面の都合で本稿の実験結果には掲載できないが，閾値 $TH$ を全てのシーンで固定した場合よりもこのようにシーンに応じて適応的に決定した場合の方が，所望のラベルに対する式 (9) の条件付確率が他のラベルに比べて高くなることが確認できている．

## 5. 実 験 結 果

### 5.1 実験の条件

Flickr から収集した 87 のシーンについて，画像アノテーションの実験を行った．画像を収集する位置は，ユーザが写真を撮る場所を考慮して，手作業で緯度経度を決定し，その位置を中心とする半径 $1,000m$ の範囲で，画像とその画像に付与されているラベルと位置情報を獲得した．各シーンで平均1,000 枚の画像を収集したが，シーンによっては 200 枚程度しか画像が集まらないシーンもあった．収集した画像とラベル，位置情報を利用して，Image-based Labeling と Label-based Localization を行うための確率モデル $P(w|\boldsymbol{x})$，$P(g|w)$ を生成させた．画像アノテーション用のクエリ画像は，学習に利用した画像とは異なる画像を各シーンで用意した．

性能比較のために，確率的正準相関分析による手法 [2]（本稿では $P(w|\boldsymbol{x})$ のみを利用した場合に相当）と，位置情報とラベル情報のみを利用した手法（本稿では $P(g|w)$ のみを利用した場合に相当）による画像アノテーションも行った．評価方法としては，上位 5 つにランキングされたラベルに対して，手作業で正解と不正解のラベルを検証した．また，各ランキングを表 1 のようにスコアとして換算し，以下の Precision Score($Ps$) と Recall Score($Rs$) を評価した．

表 1　スコア換算表

| Ranking | #1 | #2 | #3 | #4 | #5 |
|---|---|---|---|---|---|
| Score | 5 | 4 | 3 | 2 | 1 |

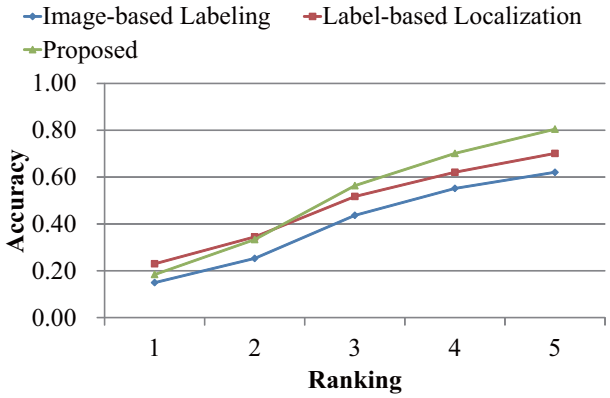



図 2　ランキングに対する正解ラベル再現性能．Ranking #3 の場合，正解ラベルが 3 位以内に入っている割合を示す．

$$Ps = \frac{Total\ Score}{\#\ of\ scenes\ given\ a\ correct\ label} \quad (12)$$

$$Rs = \frac{Total\ Score}{\#\ of\ scenes} \quad (13)$$

ここで，$\#\ of\ scenes = 87$ である．Precision Score($Ps$) では，正しいラベルが与えられたシーンのみを評価対象として，どれだけ高いランクにそのラベルがランキングされたかを評価している．一方で，Recall Score($Rs$) では，シーン全体（正しいラベルが与えられなかった場合も含む）で獲得されたスコアの平均値を評価している．ともに高い値ほど，良い結果が得られたことを意味する．ひとつのシーンに複数の正しいラベルが付与されている場合は，高いスコアの方を採用した．

### 5.2 実 験 結 果

まず，図 2 に，クエリ画像に付与されたラベルが上位 5 つにランキングされた割合を示す．縦軸はラベルを正しく付与できる割合（再現率）であり，横軸は再現率を許容するランキングである．例えば，第 3 位までを許容すれば，Image-based Labeling では，約 44%，Label-based Localization では約 52%，提案手法では約 56%の再現率が得られた．5 位までにランキングされる割合は，提案手法が最も高く，続いて Label-based Localization，Image-based Labeling の順であった．一方で，1 位にランキングされる割合に注目すると，Label-based Localization による結果が最も良い結果であった．画像情報を一切使わない手法であるにもかかわらず，このような結果が得られたことは非常に興味深い．

次に，ランキングをスコア換算した結果を，図 3 に示す．図 3 では，正しいラベルが付与されたシーンの

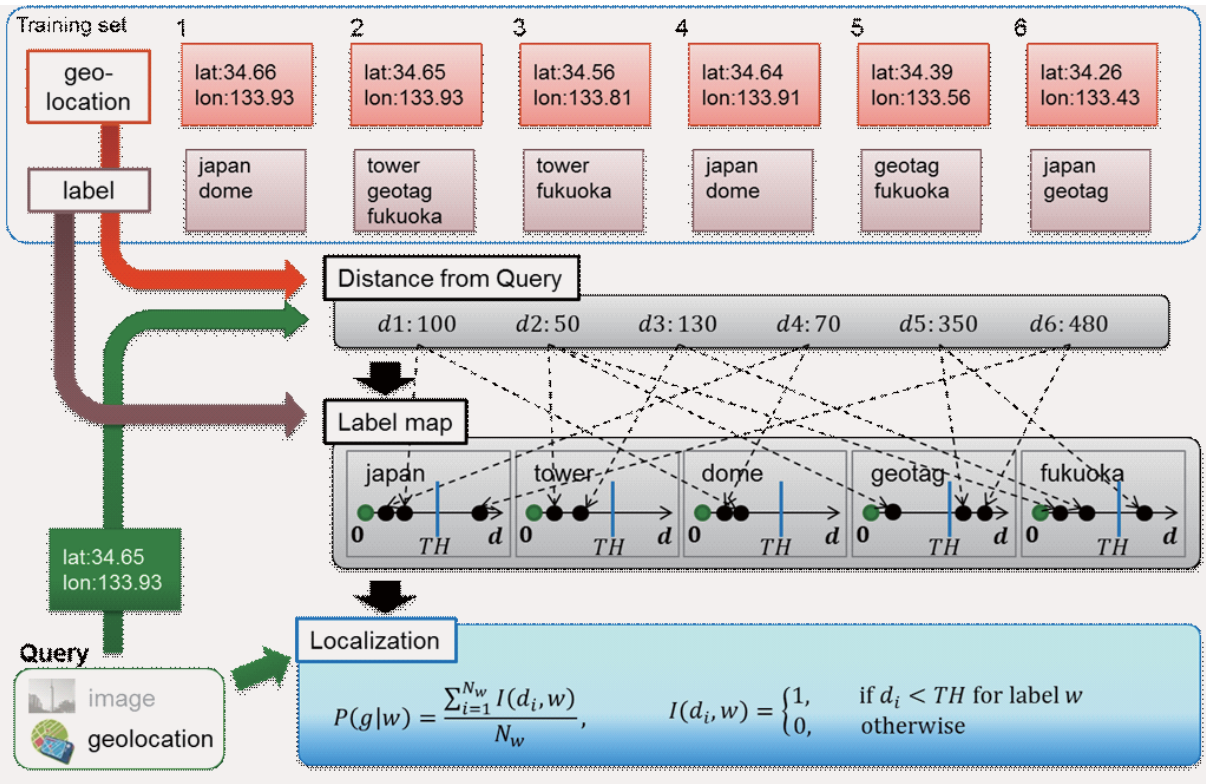



図 1　Label-based Localization の処理の流れ

表 2　各手法の Precision Score と Recall Score

| | Image-based Labeling | Label-based Localization | Proposed Method |
|---|---|---|---|
| Total Score | 175 | 210 | 225 |
| # of scenes given a correct label | 54 (62.07%) | 61 (70.11%) | 70 (80.46%) |
| Precision Score | 3.24 | 3.44 | 3.21 |
| Recall Score | 2.01 | 2.41 | 2.59 |

みを利用して，スコアの割合を表示している．ここでも Label-based Localization がスコア '5' を獲得している割合が最も高かった．表 2 には，各シーンで得られたスコアの合計，正しいラベルが付与されたシーンの数，の Precision Score($Ps$)，Recall Score($Rs$) をそれぞれ示している．図 3 の結果とも関連するが，Label-based Localization が最も高い Precision Score を獲得した．一方で，Recall Score は提案手法が最も高かった．Image-based Labeling と比べると平均で約 0.6 ポイントの差が開いていることから，画像のみではなく位置情報も利用した画像アノテーションは，非常に効果的であることがわかる．

最後に，表 3 と表 4 に画像アノテーションの成功例と失敗例を示す．表中の太字で示したラベルが，正解ラベルである．成功例では，提案手法が比較手法と同等以上の良い結果を得ることができている．一方で失敗例では，提案手法の方が，比較手法のいずれかよりも悪い結果が得られている．その原因を解析したところ，主に次のふたつの要因により悪い結果が得られていた．

まず 1 番目のケースでは，Image-based Labeling に失敗し，Label-based Localization に成功してた．提案手法は，Image-based Labeling と Label-based Localization の組合せ手法であるため，いずれかが極端に悪い結果である場合，それに影響され提案手法も結果が悪くなってしまう．Image-based Labeling に失敗する原因は，ランドマークを外部から写した画像よりもランドマーク内，あるいはランドマーク内から外の景色を写した画像が多く含まれているためであった．このような場合，画像特徴とラベル特徴の共起関係を学習することができないため，画像あのテー四がうまくいかなくなる．この問題を解決するためには，収集した学習サンプル内で，ランドマークを撮影した画像のみを画像間の特徴の類似性などから抽出するといった前処理を今後検討する必要がある．一方で，Label-based Localization では，画像特徴は一切使わないため，ランドマークの外部から撮られた画像であっても内部から撮られた画像であっても，ランドマーク名がラベルとして画像に付与されていれば，良い結果が得られていた．この点は非常に興味深い結果であった．

次に 2 番目のケースでは，Image-based Labeling に成功し，Label-based Localization に失敗していた．この原因は，位置情報が付与された学習サンプルが極端に少ないことであった．学習サンプルが少なすぎると，クエリ位置周辺（$TH$）で付与されるラベルと周辺から離れた位置（$TH$ 外）で付与されるラベルの数に大きな差が生じないため，Label-based Localization に失敗してしまう．撮影位置情報が付与された画像の枚数はシーンに

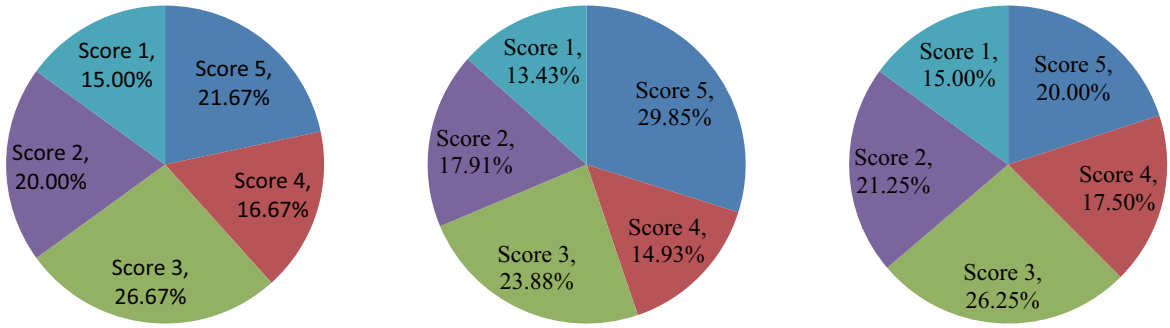



Image-based Labeling　　Label-based Localization　　Proposed

図 3　各手法の獲得スコア分布

よって異なるため，今後位置情報付の画像枚数が増えれば結果の改善が期待できる．Image-based Labeling は少ない学習サンプルであっても，ランドマークが撮影されていて，そのランドマーク名が付与されていればラベル推定に成功していた．

## 6. おわりに

本稿では，位置情報を利用した画像アノテーション手法について提案した．従来の画像アノテーション手法では，画像特徴とラベル特徴を使用した手法が主流であったが，提案手法ではその手法に加え，学習サンプルのラベルが付与されている位置情報を積極的に利用する方法を取り入れた．画像アノテーション問題を，画像特徴からラベルを推定する Image-based Labeling と，ラベルから位置情報を推定する Label-based Localization として定式化し，これらの組合せにより画像アノテーションの性能向上を図った．実験結果から従来法よりも良い結果が得られることが確認できた．今後の課題は，実験シーンの追加，収集した学習サンプル内の不要画像除去，他の Web 情報（Wikipedia や Google Image Search など）との連携による精度向上の検討などが挙げられる．

## 文　献


[1] Yunpeng Li, David J. Crandall, and Daniel P. Huttenlocher. Landmark classification in large-scale image collections. In *International Conference on Computer Vision (ICCV)*, pp. 1957–1964, 2009.

[2] Tatsuya Harada, Hideki Nakayama, and Yasuo Kuniyoshi. Image annotation and retrieval based on efficient learning of contextual latent space. In *IEEE International Conference on Multimedia and Expo*, pp. 858–861, 2009.

[3] 木村昭悟, 中野拓帆, 亀岡弘和, 杉山将, 前田英作, 坂野鋭. SSCDE: 画像認識検索のための半教師付正準密度推定法. 画像の認識・理解シンポジウム (MIRU2010), pp. 1396–1403, 2010.

[4] Tatsuya Harada, Hideki Nakayama, Yasuo Kuniyoshi, and Nobuyuki Otsu. Image annotation and retrieval for weakly labeled images using conceptual learning. *New Generation Computing*, Vol. 28, No. 3, pp. 277–298, 2010.

[5] T. Quack, B. Leibe, and L. Van Gool. World-scale mining of objects and events from community photo collections. In *ACM Conference on Image and Video Retrieval (CIVR'08)*, 2008.

[6] S. Gammeter, L. Bossard, T. Quack, and L. Van Gool. I know what you did last summer: object-level auto-annotation of holiday snaps. In *International Conference on Computer Vision (ICCV)*, pp. 614–621, October 2009.

[7] Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, and Hartmut Neven. Tour the world: building a web-scale landmark recognition engine. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, Miami, Florida, U.S.A, June, 2009.

[8] 島田 敬士, 大神 渉, 谷口 倫一郎. クリッカブル・リアルワールド：モバイル端末を利用した実世界インタラクション. CD-ROM Proc. of 映像メディア処理シンポジウム (IMPS2009), 2009.

[9] 島田 敬士, 大神 渉, 阿部 尚之, 谷口 倫一郎. クリッカブル・リアルワールド：実世界情報獲得のための新たな実世界インタラクション. インタラクション 2010, pp.21–24, 2009.

[10] Francis R. Bach and Michael I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical report, 2005.

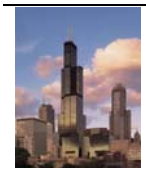
[11] H. Bay, A. Ess, T. Tuytelaars, and L.J. Van Gool. Speeded-up robust features (surf). Vol. 110, No. 3, pp. 346–359, June 2008.

[12] K. Hubeny. Isotherme koordinatensysteme und konforme abbildungen des rotationsellipsoids. In *Vienna: Osterreichischer Verein fur Vermressungswesen*, 1953.

表 3　画像アノテーション成功例．表中の太字は正しいラベルを示す．

| | Image-based Labeling | Label-based Localization | Proposed Method |
|---|---|---|---|
| | london<br>uk<br>england<br>**bigben**<br>westminster | clock<br>**bigben**<br>londres<br>housesofparliament<br>westminster | london<br>**bigben**<br>uk<br>england<br>westminster |
| | japan<br>fukuoka<br>geotagged<br>**yahoodome**<br>softbankhawks | softbankhawks<br>ホークス戦観戦<br>**yahoodome**<br>iphone<br>weatherbug | **yahoodome**<br>softbankhawks<br>ホークス戦観戦<br>fukuoka<br>geotagged |
| | washingtondc<br>dc<br>washington<br>**lincolnmemorial**<br>memorial | abrahamlincoln<br>lincoln<br>nationalmall<br>**lincolnmemorial**<br>night | washingtondc<br>**lincolnmemorial**<br>washington<br>dc<br>lincoln |
| | nyc<br>newyork<br>newyorkcity<br>manhattan<br>museum | met<br>vacation<br>statue<br>america<br>travel | nyc<br>art<br>newyork<br>**metropolitanmuseum**<br>newyorkcity |

表 4　画像アノテーション失敗例．表中の太字は正しいラベルを示す．また，表示が??となっている箇所は，文字エンコーディングの問題でもとの文字を表示できなかったもの．

| | Image-based Labeling | Label-based Localization | Proposed Method |
|---|---|---|---|
| | cologne<br>koln<br>germany<br>cathedral<br>dom | colognecathedral<br>**kolnerdom**<br>koln<br>cathedral<br>gothic | cologne<br>cathedral<br>koln<br>germany<br>**kolnerdom** |
| | rome<br>italy<br>roma<br>vatican<br>italia | basilica<br>church<br>**sanpietro**<br>vaticancity<br>cittadelvaticano | rome<br>italy<br>vatican<br>roma<br>vaticano |
|  | chicago<br>illinois<br>searstower<br>sears<br>tower | **willistower**<br>skyscraper<br>searstower<br>illinois<br>sears | chicago<br>illinois<br>searstower<br>sears<br>tower |
| | india<br>agra<br>**agrafort**<br>asia<br>uttarpradesh | ??????????<br>????<br>indian<br>nikon<br>geotagged | india<br>agra<br>mughal<br>uttarpradesh<br>**agrafort** |