ランドマーク検出のための Web 画像群からの共通画像特徴獲得

～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 実世界中に存在する有名建築物や観光名所などのランドマーク特徴を，Web で公開されている大規模画像データベース内の画像から自動獲得する方法を提案する．ランドマークは位置に依存して撮影されやすいため，画像データベース内の画像に付与されている画像の撮影位置情報を積極的に利用する．提案手法では，まず位置毎に荒いクラスタを作成し，その中で，画像の大局的特徴を局所的な特徴をクラスタリングすることで，ランドマーク特徴を自動検出している．このようにして検出されたランドマーク特徴は，クリッカブル・リアルワールドと呼ばれるモバイル端末を利用した実世界情報検索サービスにおいて，ランドマーク検出のために利用される．実験では，自動検出された画像特徴を利用してランドマークを検出できることが確認できた．

キーワード 対象検出，対象追跡，撮影位置情報，クリッカブル・リアルワールド

1. はじめに

近年，Flickr や Picasa などの画像共有サイトの画像を利用した画像アノテーション・リトリーバル [1]～[5] や物体認識 [6] に関する研究が盛んに行われている．画像共有サイトには，世界中のユーザから投稿された数多くのラベル付画像が公開されているため，これらの研究を行ううえで非常に有益な情報源である．さらに近年では，画像が撮影された位置情報も同時に取得できるようになってきており，学習サンプルの収集を工夫したり [7], [8]，認識性能を向上させたりする [9] ために利活用されるようになってきている．本稿では，画像共有サイトの位置情報付画像から同一のランドマークを表す画像の共通特徴を抽出し，それらを利用して撮影シーンからランドマークを自動検出する手法を提案する．

本研究の特長は，

（1） 画像共有サイトの位置情報付画像を利用することで，手作業で検出対象（ランドマーク）に関する情報を準備することなく，対象の画像特徴を獲得することができる．

（2） 獲得された画像特徴を利用して，シーン内のランドマークを検出することができる．

という点が挙げられる．一般に対象検出には，その対象の画像やラベル情報などを事前知識としてシステムに与え，学習ベースの手法で対象検出を行うケースが多い．それに対して本研究では，手作業による事前知識の準備は一切不要である．代わりに，本研究では，画像共有サイトの画像を利用して事前知識を自動獲得している．その際に位置情報を有効活用している点も本研究の特色である．文献 [8] でも，画像共有サイトの画像から得られる画像の共通特徴を利用した対象領域の検出方法について検討されているが，その手法では計算時間についての考察がなされていない．文献 [8] の手法を吟味する限りでは，画像の局所特徴を大量に利用した検出手法を採用している点から，計算時間がかかると推察できる．本研究では，次節で述べるフレームワークにおいて対象検出を行うことが目的であるため，処理の高速化は必要不可欠である．そのため，文献 [8] の手法よりもより厳選された画像特徴を獲得し，それを利用した高速な対象抽出法を本稿では提案する．また，連続した画像フレームで安定した対象検出を行うための手法についても検討を行ったため，本稿で報告する．

以降の本稿の構成は次のようになる．まず，2. 節で，画像共通サイトの画像を利用して本稿で提案する対象検出技術を利用する背景について述べる．次に，3. 節では，画像共有サイトの画像から同一対象（ランドマーク）の画像特徴を自動抽出する手法について述べる．画像の共通特徴を利用した対象検出については 4. 節で述べ，5. 節で実験結果を報告する．

2. クリッカブル・リアルワールド

位置情報付画像から得られる同一対象（ランドマーク）の画像特徴を利用した対象検出技術は，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

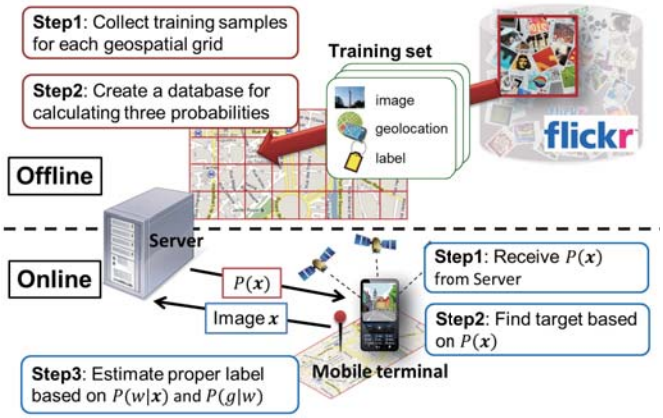

図 1　クリッカブル・リアルワールドのフレームワーク

(Clickable Real World (CRW)) 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 [10], [11] で利用される．CRW では，モバイル端末のカメラを利用して実世界中の対象（ランドマーク）を撮影することを情報検索のトリガとして，撮影対象の名称（ラベル）を推定して結果をユーザに返す．その際に，どのランドマークがクリック可能かという情報をモバイル端末上に提示することはユーザにとって有益な情報になり得る．その実現のために，画像共有サイト（Flickr）で公開されている画像，ラベル，位置情報を利用する．次節では，CRW 実現に向けた問題の定式化を行い，その後，本稿で提案する対象検出技術が CRW 全体の枠組みのどの部分で利用されるかについて説明する．

2.1　問題の定式化

モバイル端末には GPS が搭載されており，ユーザの位置情報は GPS を利用して獲得できる環境を想定する．また，画像共有サイトから得られる学習画像にもラベル情報のみならず，その画像が撮影された位置情報も利用できることを想定する．この条件下で，画像を $\boldsymbol{x}$，画像に付与すべきラベルを w，対応する位置情報を g とすると，画像と位置情報を利用したラベル推定問題 $P(w|\boldsymbol{x}, g)$ は，次の式で定義される．

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

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

ナイーブベイズにより，$\boldsymbol{x}$ と g の独立性を仮定すれば，数式は次のように変形できる．

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

さらに右辺の第 1 項と第 2 項に対してベイズの定理を適用して，最終的に次の式を得る．

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

画像と位置情報を利用したラベル推定問題は，3 つの確率モデルにより定式化されることがわかる．本研究では，右辺第 1 項 $P(\boldsymbol{x})$ を Image Prior，第 2 項 $P(w|\boldsymbol{x})$ を Image-based Labeling，第 3 項 $P(g|w)$ を Label-based Localization を呼んでいる．次節で述べるように，Image-based Labeling と Label-based Localization の組合せにより，ラベルが未知の画像に対するラベル付けを行う．その詳細については，文献 [9] を参照されたい．本稿で重要なのは，右辺第 1 項 $P(\boldsymbol{x})$ の Image Prior であり，このモデルを画像共有サイトの画像を利用して生成することが本稿における研究の目的である．また，生成されたモデルにより，対象検出ならびに追跡が行われる．

2.2　Image Prior の役割

図 1 に，CRW の全体の処理の流れを示す．処理はオフラインとオンラインに分けられる．オフライン処理は，次の 2 ステップで構成される．

（1） 地表を緯度経度に基づいてグリッドに分割する．サーバは各グリッドの中心の緯度経度をクエリとして Flickr に送り，ラベルと位置情報が付与された画像を収集する．

（2） 収集された学習サンプルは，上記の 3 つの確率モデルを計算するために利用される．Image Prior を得るためには，同一対象を撮影された画像を選定する必要がある．その中からさらに画像間の共通特徴を厳選することで，対応するグリッド内で撮影される対象（ランドマーク）の画像特徴を獲得する．本稿ではこのようにして得られる画像の共通特徴を “ランドマーク特徴” と呼ぶことにする．ランドマーク特徴は，画像共有サイトに投稿された画像のコンセンサスによって得られる特徴とも言える．Image-based Labeling と Label-based Localization のためのモデル生成 [9] についての説明はここでは省略する．

オンライン処理では，モバイル端末とサーバ間で情報のやりとりがなされる．

（1） ユーザがモバイル端末のアプリケーションを起動すると，アプリケーションはサーバから Image Prior $P(\boldsymbol{x})$ を取得する．

（2） ユーザは，モバイル端末で実世界を眺望する．Image Prior $P(\boldsymbol{x})$ によりクリック可能な対象（ランドマーク）が見つかれば，端末上にクリック可能であることを表すマークが重畳表示される．ユーザは情報獲得のために，そのマークが描かれたランドマークをクリック，すなわち撮影することができる．

（3） ユーザが対象をクリックした場合，その画像がサーバに送られて Image-based Labeling$P(w|\boldsymbol{x})$ と Label-based Localization$P(g|w)$ によってラベル推定がなされ，その結果がユーザに返る．

このように，3 つのモデルのうちのひとつ $P(\boldsymbol{x})$ はユーザのモバイル端末上で利用される．Image Prior は単にモバイル端末で対象を検出するためだけに役に立つのではなく，ユーザに対してクリック可能な対象を提示する

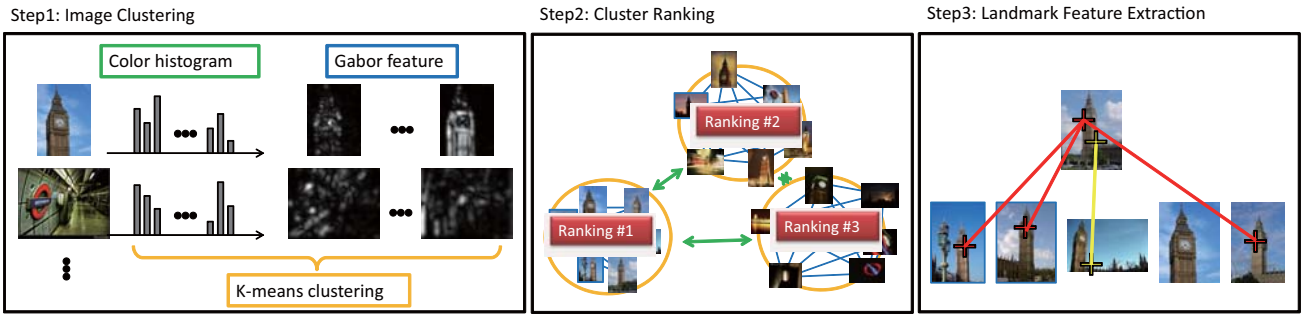

図 2　ランドマーク特徴抽出の流れ

ためにも役立つ．

3.　ランドマーク特徴の抽出

本節では，学習サンプルから Image Prior $P(\boldsymbol{x})$ すなわちランドマーク特徴を抽出する手法について述べる．画像共有サイトから収集された学習サンプルには，対象であるランドマークが撮影された画像の他にもランドマークとは関係のない画像も多く含まれる．そのような画像群から，対象のランドマークが撮影された画像を選定し，それらの画像から画像の共通特徴を抽出することでランドマーク特徴を獲得する．本研究では，ランドマーク画像は多くの画像投稿者によって撮影されることを想定して，ランドマーク特徴を獲得する手法を提案する．ランドマーク特徴の抽出の流れは図 2 に示すように 3 段階の処理で行われる．ランドマーク特徴は収集された学習サンプルの画像の少なくとも $R\%$ の画像に含まれているとする．つまり，閾値 R を設定することで，学習サンプルから抽出された画像特徴が他の画像からも得られる共通性の高い画像特徴であるかどうかを決定できる．しかし実際にはこの閾値はシーン，すなわち場所によって大きく異なるものであり，一意に決定することは不可能である．そこでまず，画像の大局的な特徴を利用して画像をクラスタリングする．次に，画像クラスタに対して，画像の一貫性に基づいてランキングを行う．高いランクにランクインされたクラスタは，同一の対象が多く撮影されているとし，そのクラスタから共通の画像特徴を獲得するという処理により，閾値 R を収集画像全体に対する割合ではなく，クラスタ内での画像の割合として利用することにする．以下では，これらの処理について詳しく述べる．

3.1　画像のクラスタリング

画像の大局的な特徴を利用して，撮影構図の類似する画像をクラスタリングする．大局的な画像特徴には，色分布を表す画像の色ヒストグラムと画像のテクスチャを表現するための Gabor 特徴を利用する．色ヒストグラムには RGB の 3 次元色空間を利用して，各色 4 階調に量子化した 64 次元の色ヒストグラムを生成する．Gabor 特徴は 4 スケール，6 オリエンテーションの 24 種類のフィルタを適用し，画像全体での特徴の平均値と分散を利用する．従って Gabor 特徴は，48 次元の特徴として表現される．色ヒストグラムと Gabor 特徴の計 112 次元ベクトルを k-means クラスタリングによりクラスタリングする．クラスタ数は，Bayesian Information Criterion (BIC) を利用して決定した．本稿で述べる実験では，約 10 のクラスタが生成された．

3.2　クラスタへのランク付け

k-means クラスタリングによって得られた各クラスタにランク付けを行う．ランク付けは画像の一貫性を基準に行われる．すなわち，上位にランキングされるクラスタは画像の一貫性が高いことになる．一貫性は次のように計算される．まず，各クラスタの代表ベクトル間の距離をクラスタ間距離 $d1$ として計算する．次に，各クラスタ内の全ての要素間の距離を計算し，その平均値をクラスタ内距離 $d2$ とする．最後にそれらの比である $d1/d2$ を計算する．画像の一貫性の高いクラスタは，クラスタ内距離が小さく，クラスタ間距離が大きくなるため，$d1/d2$ が大きいクラスタが画像の一貫性の高いクラスタであると言える．本研究では同一のランドマークが多くの画像投稿者によって撮影されていることを想定しているため，$d1/d2$ が高いクラスタに対象が撮影された画像が含まれることになる．一方で，$d1/d2$ が低いクラスタには，投稿者からの雑多な画像が含まれることになる．

3.3　特徴の選定

最上位にランキングされたクラスタの画像から画像の局所特徴 SURF [13] を抽出する．各局所特徴が複数の画像から共通して抽出される特徴であるかを検証するために，画像間での局所特徴のマッチングを行う．マッチングは全ての画像で抽出された局所特徴に対して行われ，クラスタ内の画像の $R\%$以上の画像でマッチしたと見なされた局所特徴をランドマーク特徴とする．実験では $R = 10$ とした．ここで選定したランドマーク特徴は Image Prior として後続の対象検出で利用される．

4.　ランドマーク特徴による対象検出

4.1　ランドマーク検出

モバイル端末で眺望しているシーンからのランドマー

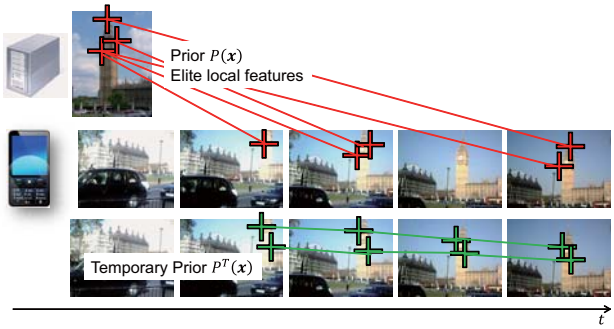

図 3　Temporary Prior による対象の追跡

ク検出はランドマーク特徴と類似した特徴を持つ対象を探すことにより実現される．特徴間の類似性は，局所特徴間の L2 ノルムにより検証される．高い類似性を持つ局所特徴がシーン内から検出された場合は，その局所特徴をクリック可能な対象の一部であると判断する．ここで，ランドマーク特徴を利用したランドマーク検出の利点と欠点について考える．ランドマーク特徴は，前節で述べたように，画像の一貫性を基準に厳選された画像クラスタ内の画像から，さらに厳選した局所特徴である．そのため，その数は画像全体から抽出される局所特徴の数と比較して，非常に少ない．その利点は，モバイル端末のように処理速度がサーバに比べて遅い端末でも特徴マッチングのコストを削減できる点にある．しかしその一方で，その数の少なさが故に，例えシーン内に検出すべき対象が写っていても毎フレームそれを検出できるとは限らないことである．CRW では，クリック可能なマークをモバイル端末上に重畳表示するため，そのような検出ミスが頻発することは，画像マッチングの観点からのみならずユーザの立場からも望ましいことではない．そこで，ランドマーク検出後に，安定して対象を検出し続けるために，対象を追跡する手法について次節では検討する．

4.2　ランドマーク追跡

上記のランドマーク検出が不安定に行われることを回避するために，ランドマーク検出後にその対象を追跡する手法を提案する．ランドマーク検出に失敗する要因は，ランドマーク特徴の数が少ないことが挙げられるため，一旦ランドマーク特徴にマッチする画像の局所特徴が検出された際には，一時的に利用可能な局所特徴を新たに導入して安定した対象検出と追跡を実現させる．一時的に利用可能な局所特徴を Temporary Prior と本稿では呼ぶことにする．Temporary Prior $P^T(\boldsymbol{x})$ には，ランドマーク特徴すなわち Image Prior とマッチした画像の局所特徴の周辺で抽出された局所特徴（図 3 の緑色の十字マーク）を利用する．Temporary Prior $P^T(\boldsymbol{x})$ はフレーム間での特徴のマッチングのみに利用され，新規の対象検出には利用されない．すなわち，対象検出・追跡のトリガはあくまで Image Prior $P(\boldsymbol{x})$ によって行われることになる．実装上は，追跡には Kalman Filter を利用している．Kalman Filter で推定する真の状態ベクトルを $\theta_t = [x_t, y_t]^T$ とし，毎フレーム獲得される観測ベクトル z_t を，$P(\boldsymbol{x})$ と $P^T(\boldsymbol{x})$ にマッチした局所特徴の座標の平均値として計算している．Kalman Filter の事後確率 $P(\theta_t|z_t)$ が閾値を超えた場合に，クリック可能なマークを画面に表示するようにしている．

表 1　ランドマーク特徴抽出過程の結果

	Big Ben	Louvre	Byodoin
# of clusters	8	9	9
# of elements	26	47	40
# of elite features	14	9	8

5.　実 験 結 果

5.1　実験の条件

Flickr からランドマーク周辺の学習サンプルを収集した．実験では，ビッグベン，ルーブル美術館，平等院の 3 カ所のシーンを利用した．各シーンで，ランドマーク周辺の緯度経度を手作業で調べ，その位置情報をクエリとして Flickr からクエリ位置から 1km 以内で撮影された位置情報付画像を約 1,000 枚収集した．学習サンプルから Image Prior $P(\boldsymbol{x})$ であるランドマーク特徴を抽出した．

検証用のシーンには，YouTube からランドマークが撮影されているビデオを収集して利用した．各シーンのビデオは約 30 秒で構成され，対象となるランドマーク以外にも他の建物や車，草木などの自然物が含まれている．これらの検証用ビデオを利用して，Image Prior のみを利用した場合の対象検出と，Temporary Prior を併用した場合の対象検出の性能について主観的な評価を行った．

5.2　ランドマークの検出と追跡結果

各シーンに対応する学習サンプルからランドマーク特徴を抽出する過程で得られた画像のクラスタ数，最上位クラスタの要素数，ならびに最終的に得られたランドマーク特徴数を表 1 に示す．各シーンで約 10 のクラスタが生成されていることがわかる．また，各シーンで最上位にランキングされたクラスタ内には，26 ～ 47 の要素（画像）が含まれていた．ビッグベンのシーンについて最上位クラスタに含まれていた画像を図 4 に示す．撮影構図が類似した画像が多く含まれていることが確認できる．一方で，図 5 は下位にランキングされたクラスタに含まれていた画像である．図 4 に比べて，撮影構図にばらつきがあり，撮影対象も異なるものが含まれていることがわかる．最後に，最上位クラスタに含まれていた画像から共通する局所特徴を選定した結果，各シーンで 10 前後のランドマーク特徴が抽出された．

図 6 に，ランドマーク検出結果を示す．比較のために，Temporary Prior $P^T(\boldsymbol{x})$ を利用しない場合，つまり

図 4　最上位にランキングされたクラスタの画像（全画像）

図 5　下位にランキングされたクラスタの画像（紙面の都合上一部画像のみを掲載）

Image Prior $P(\boldsymbol{x})$ のみを利用した場合の結果についても掲載している．各シーンの左列が Image Prior $P(\boldsymbol{x})$ のみを利用した場合の結果で，右列が Image Prior $P(\boldsymbol{x})$ と Temporary Prior $P^T(\boldsymbol{x})$ を利用した提案手法による結果である．Kalman Filter により対象の追跡が開始された際には，クリック可能なマークを青色のマークで示している．このマークが表示されている間が提案手法により対象であるランドマークが検出されているとみなす区間になる．Image Prior $P(\boldsymbol{x})$ とマッチする局所特徴が見つかってもすぐにクリック可能なマークが表示されないのは，Kalman Filter によって状態推定を行い，その事後確率 $P(\theta_t|z_t)$ が閾値を超えたときにだけ，そのマークを表示するためである．Temporary Prior $P^T(\boldsymbol{x})$ を利用した場合（右列）の方が，利用しない場合と比べて，高精度に対象を検出できていることがわかる．

処理時間を計測したところ，画像サイズ 320×240 の画像に対して，約 3 ～ 6fps であった．ランドマーク特徴数は全てのシーンでほぼ同数であることから，計算時間は各シーンで抽出される画像の局所特徴の数に影響を受けたと考えられる．本稿では，学習サンプルから抽出される画像特徴を厳選することで，Prior 側の特徴数を削減することには成功しているが，シーンから抽出される特徴点数については，十分にその数を絞り込むことは検討されていない．処理時間は，単純に局所特徴の数に比例することから，高速化のためにはシーンから検出される特徴数を制御する必要がある．シーンから抽出される局所特徴を絞り込む方法としては，文献 [14] で検討されているような注視度や顕著性の利用が有用であると考えられるため，本研究の今後の課題とする．

6. おわりに

本稿では，画像共有サイトから得られる位置情報付画像を利用して，多くの投稿者から共通して撮影されているランドマークの画像特徴を抽出する手法と，その抽出された特徴を利用してランドマークを検出する手法について提案を行った．複数の画像に共通して出現する画像の局所特徴をランドマーク特徴（Image Prior）として抽出した．ランドマーク特徴は非常に数の少ない厳選された特徴であるため，ランドマーク検出を行うシーンとの特徴マッチングにかかるコストを軽減することが可能になる．しかし，一方でその特徴数の少なさが故に，ランドマークを未検出してしまうという問題も生じた．そこでランドマーク検出後にその対象を安定して追跡するための Temporary Prior を導入した．Temporary Prior の効果により，対象の検出と追跡が安定して行えることが確認できた．今後の課題としては，ランドマーク検出を行うシーンから抽出される画像の局所特徴の数を制御することで処理時間の安定化を図ることが挙げられる．また，シーン内に対象が 2 つ以上存在する場合への対応なども挙げられる．

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] 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.

[6] Keita Yaegashi and Keiji Yanai. Geotagged Image Recognition by Combining Three Different Kinds of Geolocation Features. *ACCV2010*, 2010.

[7] 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.

[8] S. Gammeter, L. Bossard, T. Quack, and L. Van Gool.

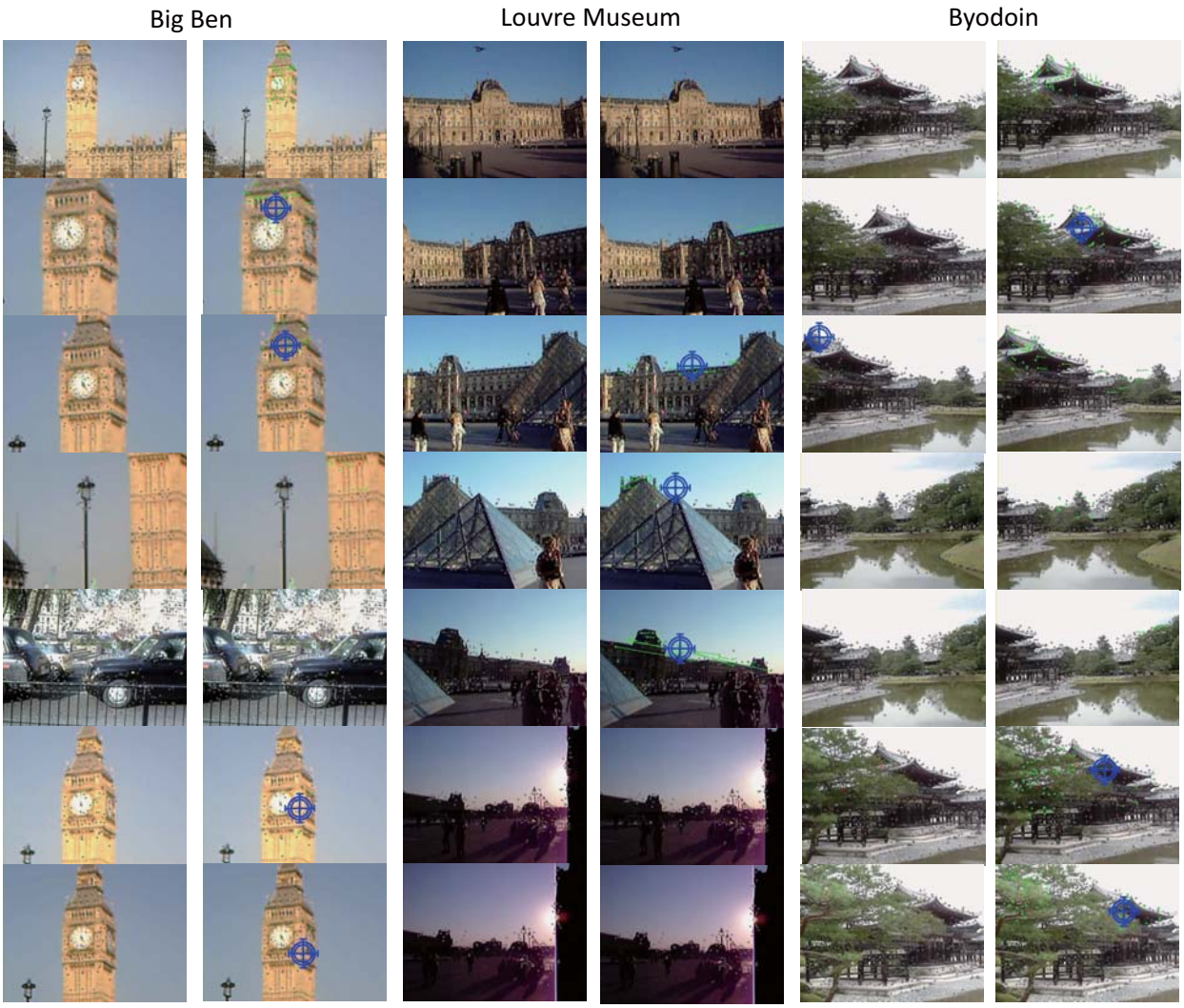

図 6　ランドマーク検出結果．各シーンの左列は，Image Prior のみを利用した場合の結果．右列は，Image Prior と Temporary Prior の両方を利用した結果．

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.

[9] 島田 敬士, Vincent Charvillat, 長原 一, 谷口 倫一郎. 撮影位置情報を利用した画像アノテーションに関する検討. IEICE-PRMU2010-113, pp.1–6, 2009.

[10] 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.

[11] 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.

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

[13] 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.

[14] 阿部 尚之, 大神 渉, 島田 敬士, 谷口 倫一郎. モバイル端末を利用した実世界インタラクションのための対象特定に関する検討. IEICE-PRMU2009-247, pp.85–90, 2010.