Web 上のジオタグ付き画像を用いた世界各地の文化的差異の発見

Discovering Cultural Differences using Geotagged Images on the Web

柳井 啓司
Keiji Yanai

電気通信大学 情報工学科
Department of Computer Science, The University of Electro-Communications

We propose a novel method to detect cultural differences over the world automatically by using a large amount of geotagged images on the photo sharing Web sites such as Flickr. We employ the state-of-the-art object recognition technique developed in the research community of computer vision to mine representative photos of the given concept for representative local regions from a large-scale unorganized collection of consumer-generated geotagged photos. The results help us understand how objects, scenes or events corresponding to the same given concept are visually different depending on local regions over the world.

1. はじめに

近年，Web 上には一般の人がアップロードしたメディアデータ (consumer-generated media, CGM) が増大してきている．特に Flickr や Panoramio などの写真共有サイトは代表的な CGM サイトであり，一般ユーザから大量の写真を集めて，それを誰もがアクセス可能な形で一般のユーザに提供している．アップロード時には写真だけでなくて，写真に付随してタグや写真の説明文などのメタデータも付加することができる．近年，そうした CGM データのタグとして，テキスト情報に加えて，緯度経度で表される「ジオタグ」と呼ばれる位置情報データが注目されている．「ジオタグ」を写真に付与することによって，写真を撮影位置で検索・分類することが可能となり，また，地図上に写真の撮影位置を表示することも可能となる．「ジオタグ」は一般に写真の場合，その写真が撮影された位置の緯度経度の座標値によって表現される．

正確なジオタグを得るには写真撮影時にカメラと一緒に GPS デバイスを携帯することが必要であるが，GPS 内蔵携帯電話で撮影する場合を除いて，GPS による位置情報をジオタグとして写真と一緒に CGM サイトにアップロードすることは手間が掛るため一般的ではない．そのため，デジタルカメラ，GPS デバイス共に 10 年以上前から存在するものの，ごく数年前まではジオタグ付き写真 (以下，ジオタグ写真と記す) は Web 上にほとんど存在しておらず，ジオタグ写真を大量に集めることが極めて困難であった．そうした状況は，2006 年に Flickr が地図ベースのジオタギングインターフェースを公開してから一変した．ユーザはアップロードした写真に対して，その撮影位置を画面に表示される地図上で指定することによって，簡単にジオタグ付与をすることが可能となった．[Kennedy 08] によると，2008 年の初めで 4 千万枚のジオタグ写真が Flickr で公開されており，毎月約 10 万枚のペースで増加し続けているとのことである．

大量のジオタグ写真を用いることにより，ジオタグが元々持っている位置による検索・分類，地図上へのマッピングという特性に加えて，ある特定地域の写真の傾向を分析したり，さらにテキストタグと組み合わせることによって，特定種類の写真の位置分布を分析することも可能になるなど，様々な新しいデータの利用法が考えられる．そこで本研究では，大量のジオタグ画像データの新しい利用方法の一つとして，特定の物体やシーンに関する世界規模の文化的差異の発見という新しい研究課題を提案する．具体的には，Flickr から特定のシーンや物体名に関連したテキストタグを持つジオタグ画像データをそれぞれ 2000 枚収集し，対応するシーンや物体を表す世界各地の代表的な画像を自動的に選び出すことを行う．

連絡先: 柳井 啓司 電気通信大学情報工学科 〒 182-8585
東京都調布市調布ヶ丘 1-5-1 E-mail: yanai@cs.uec.ac.jp

図 1: 指定されたコンセプトに関係する画像を最初にテキストタグサーチによって 2000 枚ずつ収集．次に，ノイズ画像の除去を行い，地域のクラスタリングを行った後に，各地域毎に代表的画像を選び出す．

2. 方法の概要

提案手法は大きく分けて 3 つのステップからなる．Flickr からテキストタグサーチで与えられたキーワードに関する 2000 枚のジオタグ画像を収集した後，(1) ノイズ除去, (2) 代表地域の推定，(3) 各代表地域の代表画像の選択 を行う．

最初のステップではノイズ画像の除去を行う．Flickr のテキストタグサーチによる画像検索の結果は Web 画像検索エンジンの結果に比べると精度は高いものの依然としてノイズ画像は含まれている．そこで，すべての画像を bag-of-features(BoF) 表現 [Csurka 04] に変換し，それら 2000 個の BoF ベクトルに対して k-means 法で $k = 100$ としてクラスタリングを行う．生成された各クラスタのメンバ数が 10 以下のクラスタを除去し，さらにクラスタ内メンバ間の類似度の平均値の上位 40 クラスタを残して，それ以外をノイズ画像として除去する．

2 番目のステップでは，非ノイズ画像として残った画像のジオタグの緯度経度ベクトルをクラスタリングすることによって，代表的な地域を選出する．実験では，k-means 法を用いて $k = 5$ として，与えられたコンセプトに対応する画像の代表的地域を推定した．

最後のステップでは，各代表地域について，代表画像の選出を行う．本研究では，Probabilistic Latent Semantic Analysis

(PLSA) [Hofmann 01] を用いて，ステップ 1 で得られた BoF ベクトルを，代表地域毎にさらにトピックベクトルに変換する．そして，そのトピックベクトルをクラスタリングし，最もメンバ数の大きなクラスタを代表像集合をする．本研究では，PLSA のトピック数は 20, その後のクラスタリングには再度 k-means を用いて $k = 5$ として実験を実施した．

ここでの提案手法は，Web から収集した画像から代表画像を選出する [Raguram 08] で提案された手法をジオタグ画像向けに改良したものであり，大量の画像を最初に収集し，クラスタリングを繰り返して少数の精度の高い代表画像集合を得るという考えに基づいている．なお，方法の詳細は [Yanai 09] を参照して欲しい．

3. 実験結果

実験では，5 つの物体に関するキーワード “noodle”, “wedding cake”, “flower”, ”castle”, “car” と，2 つのシーンに関するキーワード “waterfall”, “beach” の合計 7 つの英語キーワードについて，FlickrAPI を用いてテキストタグサーチによって Flickr からそれぞれ 2000 枚のジオタグ画像を収集し，提案手法を適用した．

適合率を (キーワードに適合している画像の枚数)/(全画像枚数) とすると，7 種類のキーワードの平均の適合率 (各種類 100 枚のランダムサンプリングによる値．主観評価による．) は，ステップ 1 のノイズ除去の処理の前と後で比べると，43%から 80% へ上昇した．

次に，提案手法によって得られた，いくつかの代表地域に関する代表画像を示す．図 2，図 3 に，”noodle” 画像について自動的に選ばれた地域のうち，それぞれ日本とヨーロッパについての結果を示す．この結果から，日本の代表的な “noodle” はラーメン，ヨーロッパの代表的な “noodle” はスパゲティであることが分かる．今回の提案手法では，最も代表的な画像のみを選び出しているので，日本独自の “noodle” であるそば，うどんはほとんど選ばれておらず，Flickr 上に画像が多いラーメンが主に選ばれている．他には，東南アジア，アメリカ中東部，アメリカ西部が代表的地域として選ばれている．東南アジアは台湾や中国，タイの “noodle” が選ばれており，アメリカは中東部，西部ともにアジアとヨーロッパの “noodle” が混じった画像集合が代表画像として選ばれた．

図 4，図 5 は，それぞれアメリカ中部とヨーロッパの “wedding cake” 画像である．アメリカ中部の方が，ヨーロッパよりより高く派手なケーキが多いことが分かる．

他にも，“waterfall” 画像については，東アジア地域，ヨーロッパ，北アメリカ，南アメリカが主な地域として選択され，南アメリカではイグアスの滝に代表される巨大な力強い滝の写真が多い一方，東アジアでは細い美しい滝の写真が多いことが分かる．

ここに示した以外の結果は
http://img.cs.uec.ac.jp/yanai/ASRP/ で見ることができる．

4. おわりに

本研究では，大量のジオタグ画像データの新しい利用方法として，特定の物体やシーンに関する世界規模の文化的差異の発見という新しい研究課題を提案した．そのための第一歩として，本予稿では，Flickr から特定のシーンや物体名に関連したテキストタグを持つジオタグ画像データをそれぞれ 2000 枚収集し，対応するシーンや物体を表す世界各地の代表的な画像を自動的に選び出す実験を 7 種類のキーワードについて行った．その結果，例えば，“noodle” の場合，日本の代表的な”noodle” はラーメンである一方，ヨーロッパではスパゲティであることが大量のジオタグ画像データから明らかになった．今後は，地域毎の差異の大きな単語，差異があまりない単語を自動的に分類することを予定している．また，得られた結果を用いたアプリケーションも検討中である．

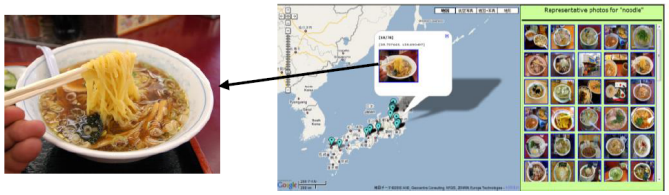

図 2: 日本の”Noodle” 画像．ラーメンが代表的．

図 3: ヨーロッパの”Noodle” 画像．スパゲティが大部分．

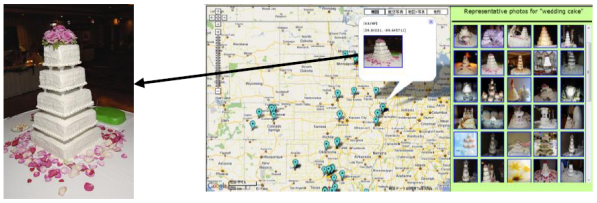

図 4: アメリカ中部の”Wedding cake” 画像．背が高い．5 段のもある．

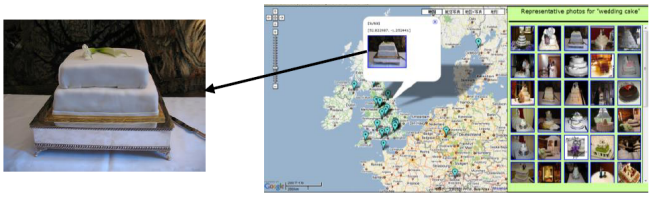

図 5: ヨーロッパの”Wedding cake” 画像．アメリカのより背が低く，シンプル．

参考文献

[Csurka 04] Csurka, G., Bray, C., Dance, C., and Fan, L.: Visual categorization with bags of keypoints, in *Proc. of ECCV Workshop on Statistical Learning in Computer Vision*, pp. 59–74 (2004)

[Hofmann 01] Hofmann, T.: Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*, Vol. 43, pp. 177–196 (2001)

[Kennedy 08] Kennedy, L. and Naaman, M.: Generating Diverse and Representative Image Search Results for Landmarks, in *Proc. of the International World Wide Web Conference*, pp. 297–306 (2008)

[Raguram 08] Raguram, R. and Lazebnik, S.: Computing Iconic Summaries of General Visual Concepts, in *Proc. of IEEE CVPR Workshop on Internet Vision* (2008)

[Yanai 09] Yanai, K., Kawakubo, H., and Qiu, B.: A Visual Analysis of the Relationship between Word Concepts and Geographical Locations, in *Proc. of ACM International Conference on Image and Video Retrieval* (2009)