1G1-3

確率トピックモデルによる Web 画像の分類

Web Image Mining with Probabilistic Topic Models

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

In this paper, we propose a new method to select relevant images to the given keywords from images gathered from the Web. Our method is based on generative probabilistic latent topic models such as Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA).

Firstly, we gather images related to the given keywords from the Web with Web search engines. Secondly, we choose pseudo-training images from them by simple heuristic HTML analysis, and train a probabilistic latent topic model with them. Finally, we select relevant images from all the gathered images with the learned model. The experimental results shows that the results by the proposed method is almost equivalent to the results by existing methods, although our method does not need to prepare negative training samples in advance unlike existing methods.

1. はじめに

近年のデジタルカメラの普及や WWW(World Wide Web) 上の画像の爆発的な増大によって，デジタル画像の意味内容を計算機に自動的に認識させる「一般物体認識 (Generic Object Recognition)」に対する要求が高まっており，多くの研究が行われるようになってきている [柳井 07a]．一般物体認識においては，事前に内容が既知である大量の画像が学習データとして必要であるが，そうしたデータセットを人手で構築するのは容易ではなく，現状では認識対象がある一定数に限定された学習データセットで研究が行われることが一般的である．

それに対して，我々は Web から画像を自動収集し，一般物体認識のための学習画像データとして利用することを提案している [Yanai 03, 柳井 04]．Web 上の画像 (Web 画像) は，様々な人が様々な目的で撮影した画像であり，類似している画像が多く含まれる商用の画像データベースとは異なり，実世界の一般的な画像の多様性をそのまま反映していると考えられる．また，Web 画像はそれを含んでいる Web ページの HTML 文書を解析することによって，画像に関連するキーワードを抽出することが可能であるという特徴を持つために，目的の画像を自動収集することが可能である．我々は，この一般物体認識のための Web からの画像収集を「Web 画像マイニング」と呼んでいる．他にも，Web 上の画像を学習データとして用いる一般物体認識の研究 [Fergus 05, Schroff 07] は存在しており，近年研究が盛んに行われるようになってきている．

Web 画像マイニングのための方法としては，我々は画像検索手法を用いた方法 [Yanai 03, 柳井 04], 領域分割と確率モデルを用いた方法 [Yanai 05, 柳井 07b] を提案している．また，近年，一般物体認識において極めて有効であることが示された画像表現手法である bag-of-visual-words (BoVW) 法 [Csurka 04] を画像特徴表現法に利用し，SVM を分類器とする方法も提案し，従来の手法の精度を大幅に改善できることを示した [Yanai 07]．本研究では，それとは異なる方法として，確率トピックモデルを用いた方法を提案する．具体的には，画像を bag-of-visual-words によって表現し，それを元々文書分類のために提案された確率トピック分類の手法である pLSA (probabilistic Latent Samentice Analysis) [Hofmann 01] および LDA (Latent Dirichlet Allocation) [Blei 03] を応用して，分類を行う．

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

2. 処理の概要

本研究では，画像認識の手法を用いて，与えられた単語に対応する画像を自動的に World Wide Web から収集することを目的とする．

処理は，画像収集部と画像選択部から成る．前半はテキスト処理，後半は画像処理となっている．本研究では，画像選択部に確率トピックモデルに基づく方法を導入する．

まず，画像収集部では，ユーザから与えられたキーワードに基づいて Google, Yahoo などの商用検索エンジンを利用して Web から大量の Web ページを収集し，独自の解析手法でページを解析し，キーワードと関連する可能性が高い画像ファイルのみを Web から収集する．方法は [Yanai 03, 柳井 04] と同じである．ここで，収集した画像を HTML 解析によって，A ランク画像と B ランク画像の 2 種類に分ける．経験的には，A ランク画像の適合率は 7〜8 割程度，B ランク画像の適合率は 5 割程度である．つまり，A ランク画像の 2〜3 割，B ランク画像の半分程度はキーワードに無関係なノイズ画像が含まれている．A ランク画像は，次の画像選択部において疑似学習データとして利用される．

次に，画像選択部で，キーワードに無関係なノイズ画像を画像認識手法によって取り除き，残った画像をキーワードに対応する画像セットとして出力する．ここでは，出来るだけ多くノイズ画像を除去し適合率を向上させながら，出来るだけ正解画像を除去しないようにして再現率を高く保つことが求められる．以下では，この画像選択部の手法についてその詳細を説明

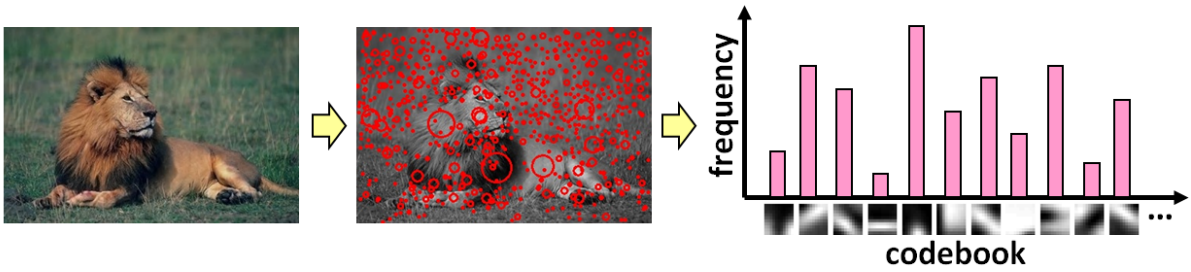

図 1: Bag-of-keypoints 表現の求め方．画像から SIFT によって特徴抽出し，コードブックに関するヒストグラムを作成する．ヒストグラムが画像の特徴量となる．

する．

3. 画像選択部

画像選択部では，(1) 入力画像の bag-of-visual-words 表現への変換，(2) 確率トピックモデルによる確率的トピッククラスタリング，(3) 疑似学習データを用いた各トピックの正解画像確率の計算，(4) 各画像の正解画像確率の推定，を順番に行い，最終的に Web から収集した画像からのノイズ画像の除去を行う．

3.1 Bag-of-Visual-Words 表現

Bag-of-visual-words [Csurka 04] は，統計的言語処理における bag-of-words のアナロジーで，bag-of-words で語順を無視して文章を単語の集合と考えるのと同様に，bag-of-visual-words では位置を無視して画像を局所特徴 (keypoints) の集合として考える．実際の処理においては，局所特徴の特徴ベクトルをベクトル量子化することによって，keypoint を word として扱えるようにする．このベクトル量子化された特徴を visual word と呼ぶこともある．つまり，bag-of-visual-words では，画像の特徴量は，画像から抽出した 100〜1000 個程度の visual word の出現頻度のヒストグラムによって表現される．なお，局所特徴の特徴表現には SIFT 法 [Lowe 04] が用いられることが一般的である．

本研究での Bag-of-visual-words による特徴表現は，(1)10 ピクセルおきの格子点による特徴点の決定, (2) 各特徴点に関する SIFT 記述子ベクトルの計算，(3) 全学習画像の全 SIFT 記述子ベクトルの k-means によるクラスタリング (実験では $k = 1000$) による visual words(代表 SIFT ベクトル) の選定，(4) 各画像に関する visual words ヒストグラムを作成，の手順によって作成することが可能である (図 1)．SIFT 記述子の計算以外は簡単な処理で，SIFT 記述子の計算も公開ソフトを利用することが可能なため，bag-of-visual-words は極めて手軽な一般物体認識のための画像表現であると言える．

3.2 確率トピック分類

Bag-of-visual-words は統計的言語処理の bag-of-words と同等の考え方であると述べたが，画像の表現として bag-of-visual-words を用いることによって，統計的言語処理の分野で提案された確率的な手法が応用可能となる．文書分類のための確率的トピック抽出の手法として提案された probabilistic Latent Semantic Analysis (pLSA)[Hofmann 01, Sivic 05], Latent Dirichlet Allocation (LDA)[Blei 03, Fei-Fei 05] などが一般物体認識に応用されている．本研究では，これを Web 画像の選択処理に利用する．

まず，pLSA もしくは LDA を用いて，あるキーワードに関する，bag-of-visual-words 表現された全ての Web 収集画像に対して，事前に指定したトピック数で確率トピッククラスタリングを実行する．収集画像枚数を N, トピック数を k，画像を $d \in (d_1, ..., d_N)$，トピックを $z \in (z_1, ..., z_k)$ とすると，各画像について各トピックへの帰属確率 $P(z|d)$ を求めることができる．

3.3 正解画像確率の計算

次に，各画像の正解画像確率 $P(pos|d)$ を計算する．本研究では，各トピックに属する画像のうち，画像収集部で A ランク画像として収集された画像の枚数と，B ランク画像として収集された画像の枚数の比を利用して，まず各トピックの正解画像確率 $P(pos|z)$ を計算し，最終的に $P(pos|d)$ を求める．具体的には以下の式によって求める．

$$p_0 = \frac{1}{N_A} \sum_{d \in A} P(d|z) \quad (1)$$

$$p_1 = \frac{1}{N_B} \sum_{d \in B} P(d|z) \quad (2)$$

$$P(pos|z) = p_0/(p_0 + p_1) \quad (3)$$

$$P(neg|z) = p_1/(p_0 + p_1) \quad (4)$$

ただし，

$$P(d|z) = \frac{P(z|d)P(d)}{\sum_{d \in D} P(z|d)P(d)} \quad (5)$$

とし，N_A, N_B をそれぞれ A ランク画像の枚数，B ランク画像の枚数とする．

最終的には，$P(pos|z)$ を確率トピック分類の手法で求めた $P(z|d)$ を用いて全トピックについて以下のように周辺化することによって，$P(pos|d)$ を求め，その値が一定の閾値以上の画像を正解画像として出力する．

$$P(pos|d) = \sum_{z \in Z} P(pos|z)P(z|d) \quad (6)$$

4. 実験

表1に示す10種類のキーワード（夕暮れ (sunset)，山 (mountain)，滝 (waterfall)，海岸 (beach)，花 (flower)，ライオン (lion)，リンゴ (apple)，赤ちゃん (baby)，ノート PC(notebook PC)，ラーメン (Chinese noodle)）について実験を行った．表1には，参考までに，Google Image Search による結果の上位100枚の適合率，画像収集部で収集した A ランク画像および A ランクと B ランクを合わせた画像の枚数と適合率，従来手法による方法の結果を示した．従来手法の結果については，“GMM” が領域確率モデルを用いた手法 [Yanai 05, 柳井 07b]，“SVM” が bag-of-visual-words を SVM を用いた手法 [Yanai 07] の結果の適合率をそれぞれ表す．なお，評価は [Schroff 07] で用いられている方法と同様の，再現率 15%時の適合率で行うこととする．再現率は (正しく分類された画像の枚数)/(収集画像中の正解画像の枚数)，適合率は (正しく分類された画像の枚数)/(選択された画像の全枚数) である．

表 2に確率トピックモデルとして pLSA を用いた場合の結果，表 3に確率トピックモデルとして LDA を用いた場合の結果を示す．それぞれ，トピック数 k を 10, 20, 30, 50, 100, 150, 200 と 7 通りに変化させて実験し，その結果を全て示している．表の一番右の列の BEST は，7 通りのうち最も良い結果の値とその時のトピック数を表す．

LDA の結果は，pLSA の結果に比べて，lion の結果が良くなっているが，一方 baby の結果が悪くなっており，BEST の平均値としては 1.1 ポイント上昇している．

各キーワードの BEST のトピック数は，10 から 200 までであり，一定の傾向を見られない．キーワード毎に適切なトピック数は異なっている．Waterfall, baby, notebook PC の 3 種類を除いて，GMM よりも pLSA, LDA の値の方がいい結果が得られている．Waterfall, baby は，A ランク画像の適合率と B ランク画像の適合率の差が小さかったために，各トピックの正解確率 $P(pos|z)$ がうまく推定出来なかったことによると考えられる．また，notebook PC は，そもそも A ランク画像も B ランク画像も適合率が低すぎるために，同様に $P(pos|z)$ がうまく推定出来なかったと考えられる．一方，SVM の結果には，pLSA, LDA 共に及ばなかった．SVM のソフトマージンによる不完全学習データからの学習の性能が優れているいるためと考えられる．

図 2に sunset, flower, lion, Chinese noodle の出力結果の上位 24 枚の画像を参考までにそれぞれ示す．なお，結果の一部は，`http://mm.cs.uec.ac.jp/yanai/jsai08/` で見ることができる．

5. おわりに

本研究では，Web 画像マイニングにおいて，一般物体認識において有効であることが示されている画像表現手法である bag-of-visual-words (BoVW) 法 [Csurka 04] を画像特徴表現法に利用し，確率トピックモデルを用いた方法によって確率的に画像を分類し，ノイズ画像を除去する方法を提案した．確率トピックモデルとしては，元々文書分類のために提案された pLSA (probabilistic Latent Samentice Analysis) [Hofmann 01] および LDA (Latent Dirichlet Allocation) [Blei 03] を用いた．実験結果では，SVM による方法 [Yanai 07] には及ばなかったものの，領域確率モデルによる方法 [Yanai 05, 柳井 07b] を上回る結果が得られた．

今後の課題としては，トピック数の自動決定が挙げられる．本研究では，最適なトピック数はキーワードによって異なり，一律に決めることは難しい事が明らかになった．今後はキーワード毎に自動的にトピック数を決めることが求められる．

参考文献

[Blei 03] Blei, D., Ng, A., and Jordan, M.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[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)

[Fei-Fei 05] Fei-Fei, L. and Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories, in *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 524–531 (2005)

[Fergus 05] Fergus, R., Fei-Fei, L., Perona, P., and Zisserman, A.: Learning Object Categories from Google's Image Search, in *Proc. of IEEE International Conference on Computer Vision*, pp. 1816–1823 (2005)

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

[Lowe 04] Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110 (2004)

[Schroff 07] Schroff, F., Criminisi, A., and Zisserman, A.: Harvesting Image Databases from the Web, in *Proc. of IEEE International Conference on Computer Vision* (2007)

[Sivic 05] Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., and Freeman, W. T.: Discovering Objects and their Localization in Images, in *Proc. of IEEE International Conference on Computer Vision*, pp. 370–377 (2005)

[Yanai 03] Yanai, K.: Generic Image Classification Using Visual Knowledge on the Web, in *Proc. of ACM International Conference Multimedia*, pp. 67–76 (2003)

[Yanai 05] Yanai, K. and Barnard, K.: Probabilistic Web Image Gathering, in *Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 57–64 (2005)

[Yanai 07] Yanai, K.: Image Collector III: A Web Image-Gathering System with Bag-of-Keypoints, in *Proc. of the International World Wide Web Conference* (2007)

[柳井 04] 柳井 啓司：一般画像自動分類の実現へ向けた World Wide Web からの画像知識の獲得, 人工知能学会論文誌, Vol. 19, No. 5, pp. 429–439 (2004)

[柳井 07a] 柳井 啓司：一般物体認識の現状と今後, 情報処理学会論文誌: コンピュータビジョン・イメージメディア, Vol. 48, No. SIG16 (CVIM19), pp. 1–24 (2007)

[柳井 07b] 柳井 啓司：確率的 Web 画像収集, 人工知能学会論文誌, Vol. 21, No. 1, pp. 10–18 (2007)

表 1: 従来手法の結果の適合率 (再現率 15%時)，および Google Image Search の上位 100 枚の結果の適合率.

concepts	Google	A-rank	A-rank + B-rank	GMM	SVM
sunset	85	790 (67)	1500 (55.3)	100.0	98.0
mountain	57	1950 (88)	5837 (79.2)	96.5	100.0
waterfall	78	2065 (71)	4649 (70.3)	82.0	90.7
beach	67	768 (69)	1923 (65.5)	75.0	99.0
flower	71	576 (72)	1994 (69.6)	78.5	91.9
lion	52	511 (87)	2059 (66.0)	74.6	85.7
apple	49	1141 (78)	3278 (64.3)	81.0	90.7
baby	39	1833 (56)	3571 (54.5)	70.7	65.7
notebook PC	70	781 (57)	2537 (43.6)	70.5	52.3
Chinese noodle	68	901 (78)	2596 (66.6)	70.9	95.3
TOTAL/AVG.	63.6	11316 (72)	29944 (62.2)	80.0	86.9

表 2: pLSA を用いた場合の結果の適合率 (再現率 15%時). k はトピック数を示す.

concepts	k=10	k=20	k=30	k=50	k=100	k=150	k=200	BEST
sunset	99.0	98.0	98.0	98.0	98.0	99.0	98.0	**99.0 (k=10)**
mountain	97.2	82.7	81.8	81.3	83.2	85.8	84.2	**97.2 (k=10)**
waterfall	63.8	68.5	71.6	68.5	70.9	66.1	68.2	**71.6 (k=30)**
beach	89.9	88.3	91.6	92.5	88.3	92.5	91.6	**92.5 (k=50)**
flower	83.9	83.9	81.8	81.3	79.9	80.4	82.3	**83.9 (k=10)**
lion	75.9	71.7	64.7	66.7	71.0	69.5	75.9	**75.9 (k=10)**
apple	87.0	89.3	90.5	89.3	85.9	87.0	87.0	**90.5 (k=30)**
baby	44.1	44.8	42.7	54.2	52.9	50.8	48.1	**54.2 (k=50)**
notebook-PC	27.7	25.9	29.7	35.7	44.9	42.3	41.0	**44.9 (k=100)**
Chinese-noodle	90.9	68.2	71.4	68.2	95.2	95.2	95.2	**95.2 (k=100)**
AVG.	75.9	72.1	72.4	73.6	77.0	76.9	77.2	**80.5**

表 3: LDA を用いた場合の結果の適合率 (再現率 15%時).

concepts	k=10	k=20	k=30	k=50	k=100	k=150	k=200	BEST
sunset	97.0	96.0	97.0	96.0	96.0	96.0	97.0	**97.0 (k=10)**
mountain	96.5	97.2	96.5	97.2	97.2	97.2	97.2	**97.2 (k=20)**
waterfall	64.0	64.3	65.2	64.6	65.5	65.2	75.7	**75.7 (k=200)**
beach	91.6	92.5	91.6	92.5	94.2	93.3	91.6	**94.2 (k=100)**
flower	88.2	86.0	81.8	80.4	82.3	79.9	79.0	**88.2 (k=10)**
lion	75.9	81.5	83.5	84.6	88.0	85.7	88.0	**88.0 (k=100)**
apple	89.3	88.2	88.2	85.9	85.9	89.3	71.3	**89.3 (k=10)**
baby	47.8	44.8	43.5	40.5	40.5	38.1	38.3	**47.8 (k=10)**
notebook-PC	26.0	34.5	31.6	36.7	43.7	42.6	41.5	**43.7 (k=100)**
Chinese-noodle	89.5	92.3	93.8	92.3	92.3	95.2	95.2	**95.2 (k=150)**
AVG.	76.5	77.9	77.3	77.1	78.5	78.2	77.0	**81.6**

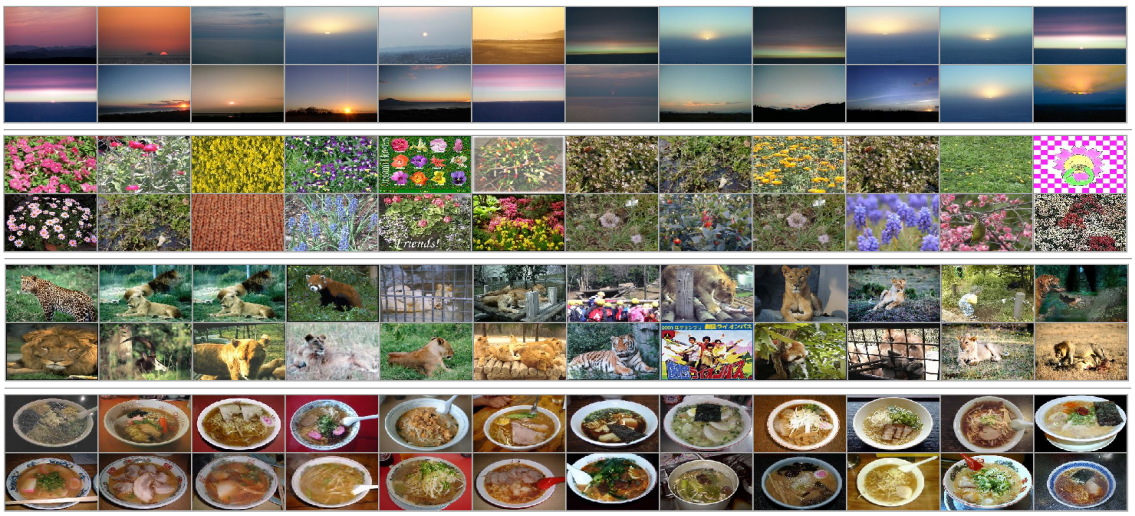

図 2: Sunset, flower, lion, Chinese noodle の出力結果の上位 24 枚の画像.