# 視覚情報と意味情報からのノイズ除去によるWeb画像データベースの自動構築

# Automatic Construction of Web Image Database based on Hybrid Noise Removal from Visual and Semantic Information

岩堀 祐之 [1]　大谷 世紀 [1]　山西 良典 [2]　溝口 友喜 [1]
Yuji Iwahori [1]　Seiki Otani [1]　Ryosuke Yamanishi [2]　Yuki Mizoguchi [1]
[1] 中部大学大学院工学研究科情報工学専攻　[2] 立命館大学情報理工学部メディア情報学科
[1]Dept. of Computer Science, Chubu Univ.　[2]Dept. of Media Information, Ritsumeikan Univ.

## 1　はじめに

一般物体認識とは，計算機が実世界における制約のない画像を一般的な名称で認識することであり，画像認識の研究分野において代表的な課題の 1 つである．一般物体認識問題については，これまで主に特徴量の抽出や学習器の構築方法について多くの研究が行われてきた [Opelt et al., 2006, LeCun et al., 2004, Bergevin and Levine, 1993]．しかしながら，一般物体認識問題においては，特徴量の抽出や学習器の構築方法の他に実世界に存在する様々な画像を学習するため大規模な画像データベースが必要となる．多くの一般物体認識の研究では，実験用に研究者が人手で作成した画像データベースが用いられてきた [Griffin et al., 2007, Everingham et al., 2010]．また，人手による大規模な画像をもつ画像データベースの構築には，多大な時間的・人的コストが必要となる．近年では，Web 上の画像を検索することで，画像データセットを自動的，あるいは，半自動的に構築する Web 画像マイニングが報告されている [Song et al., 2004, Chua et al., 2009, Fergus et al., 2004]．

Web 画像マイニングにおいて，Web から得られる画像にはノイズ画像が含まれるため，画像特徴量を用いたノイズ画像除去手法が提案されてきた．文書解析や意味情報を用いた画像収集やノイズ画像除去を行っている研究 [柳井啓司, 2007, 秋間雄太 et al., 2010] も報告されているものの，視覚的特徴量を用いたものに比べ意味的特徴量を用いた手法は少なく，それらの特徴を組み合わせた研究も多くはない．特に，2 種類の異なる特徴量のどちらを重視するべきか，またどのように組み合わせてノイズ判定を行うかについての比較・議論は少ない．

本研究では，視覚情報と意味情報それぞれのノイズ画像に対する有効性，および，それらの情報のノイズ除去における組み合わせ方法を比較・検討する．そして，2 種類の異なる特徴量をハイブリッドに用いることで高い精度でノイズ画像を除去し，有用性の高い画像データベースを Web から自動構築する手法を提案する．

## 2　関連研究

Web 画像マイニングによって自動的に収集された画像には必ずノイズ（ノイズラベル，ノイズ画像）が含まれており，これをそのまま一般物体認識に応用すると学習効率が悪い．そこで，従来の研究では任意に決定した 10 種類ほどのキーワードについて画像検索を行い，出力された画像の画像特徴量を用いて精度を高めることが目的とされた．Fergus らの研究 [Fergus et al., 2004] では，画像特徴量によるノイズ除去によって平均で 58.9%の精度で画像を収集することに成功している．しかしながら，検索キーワードが人手で任意に決定されており，検索に用いたキーワードをそのまま画像ラベルとして用いるなど，ラベルがもつ意味や妥当性については触

れられていない．

近年では，Wordnet [Miller, 1995] などの知識体系を利用する方法によって，ラベルの内容に一般性を持たせた画像データセット構築が行われるようになった．Deng らが提案した *Imagenet* [Deng et al., 2009] では，Flickr からの Wordnet を用いた階層構造を持つ画像データセットの構築が行われている．しかし，*Imagenet* ではノイズ画像の除去に人手を用いており，データセットの構築や拡張に必要なコストが問題となる．これらの研究では，検索キーワードとして知識体系中の概念を用いているだけであり，画像ラベルについても検索キーワードがそのまま使用されている．Web から収集された画像に含まれるノイズ（ノイズラベル，ノイズ画像）の除去方法として，知識体系（オントロジー）がもつ意味情報と，画像情報の双方を組み合わせた研究は少ない．

# 3 提案手法

提案手法では，大きく 1）画像検索，2）概念関係によるノイズラベル除去，3）視覚情報と意味情報を用いたハイブリッドノイズ除去，の 3 つの処理によって画像データセットを自動生成している．

## 3.1 Web からの画像収集

初めに Flickr においてキーワード検索を行うことで，画像の収集を行う．収集時にはオントロジー内に存在するクラスのクラスラベルを検索クエリとして用いることで，検索キーワードを自動的に決定する．このとき検索に用いられたクラスを検索によって得られた画像に対する検索クラスと呼ぶ．

上記の方法で収集した画像に付与された画像タグには，無意味な文字列などのノイズタグが存在する．そこで，画像に付与されていた各タグとオントロジー概念を照合し，検索クラスと IS-A 関係にないタグをノイズタグとして除去する．これにより，画像に付与されたタグをラベルとして採用するときのラベリングの精度を向上させる．

## 3.2 視覚情報に基づくノイズ画像除去

得られた画像のうち，マイノリティな画像特徴を有している画像は自動的に決定される閾値に従ってノイズ画像として除去する．画像特徴量は Dense SIFT 特徴量を Fisher Vector として表現した特徴ベクトルする．本稿では，Dense SIFT 特徴量の抽出する際には，抽出間隔を 8px，抽出窓の大きさを 16px とした．また画像から Fisher Vector を算出する際の visual word 数 K は 512 とした．

まず，各画像特徴量 $\boldsymbol{G(i)}$ と全画像の重心ベクトル $M^q$ 間のユークリッド距離 $FD^q(i)$ を得る．クエリ $l$ によって収集された画像を $FD^l(i)$ に基づいて，以下のように分類する．

$$\boldsymbol{VI}^l = \begin{cases} \text{適合画像} & (FD^q(i) <= VT^q) \\ \text{ノイズ画像} & (FD^q(i) > VT^q) \end{cases} \quad (1)$$

ここでの $VT^q$ は閾値を示し，各 $FD^q(i)$ の平均値とする．

## 3.3 意味情報に基づくノイズ画像除去

意味特徴量として，Word2Vec[1]によって算出したタグの単語ベクトルを用いる．単語ベクトルは，2016 年 2 月時点での Latest Wiki Dump データを学習した 1000 次元のベクトルを用いる．本稿では，Word2Vec 実行時のパラメータとして，ウィンドウサイズは最大 5 単語，ネガティブサンプルの個数は 5 とて設定し，モデル構築と正規化手法は，Skip-gram と階層的ソフトマックスをそれぞれ用いた．

まず，ある画像 $i$ に付与されたタグ $c$ の単語ベクトルを得る．そして，ある画像 $i$ に付与された各タグ $c$ の重心ベクトルを，画像の意味特徴量 $SF^q(i)$ とする．次に，各画像の意味特徴量 $SF^q(i)$ と，その画像を得るために用いたラベル $q$ の単語ベクトルとの，ユークリッド距離 $SD^q(i)$ を算出する．クエリ $q$ によって収集された画像を $SD^q(i)$ に基づき以下のように分類する．

$$\boldsymbol{SI}^l = \begin{cases} \text{適合画像} & (SD^q(i) <= ST^q) \\ \text{ノイズ画像} & (SD^q(i) > ST^q) \end{cases} \quad (2)$$

ここでの $ST^q$ は閾値を示し，各 $SD^q(i)$ の平均値とする．

[1] https://code.google.com/p/word2vec/

### 3.4 ノイズ画像除去の組み合わせ

まず，画像特徴量と意味特徴量をハイブリッドに用いてノイズ除去を行う際には，各特徴量によるノイズ除去の組み合わせ方法を考えなければならない．ノイズ除去の組み合わせ方としては，パラレル型とシリアル型の 2 種類が考えられる．パラレル型では，ノイズの除去時に各特徴量空間のノイズ判定結果の積集合と和集合のどちらに着目すべきかを検討する必要がある．一方でシリアル型では，各特徴量によるノイズ除去の適用順序を考えなければならない．

各特徴量によるノイズ除去，および，特徴量の組み合わせ方をまとめると以下の 6 種類の手法が考えられる．

(1) **V-method** : 画像特徴量を用いた視覚的評価のみによって画像評価をする手法．

(2) **S-method** : 意味特徴量を用いた意味的評価のみによって画像評価をする手法．

(3) **PAND-method** : 各特徴量に基づく画像評価を独立に行い，適合と評価された画像の積集合を採択する．

(4) **POR-method** : 各特徴量に基づく画像評価を独立に行い，適合と評価された画像の和集合を採択する．

(5) **SVS-method** : “視覚?意味”の順序で画像を評価する．このとき，画像特徴に基づいて適合画像と評価された画像のみに対して，意味特徴量を算出し，意味的評価を行う．

(6) **SSV-method** : SVS-method においての評価順序を “意味?視覚”とし，シリアルにノイズ除去を行う手法．

## 4 評価実験と考察

節 3.4 で示した 6 種類のノイズ画像除去法を適用したそれぞれの画像集合に対して，適当なラベルが付加されているかを評価実験により確認した．実験で使用する Web 画像データセットは，2015 年 2 月時点において Flickr から収集した画像集合を基にして構築した．収集に用いるラベルは，*DBpedia* 内の一般的な単語から任意に取り出した 12 種類のクラスを用いた．表 1 に各ラベルについて，Web から得られた時点での画像の枚数と精度を示す．

表 1: Web から収集した時点での各ラベル毎の画像枚数と評価精度

| ラベル | 画像枚数 | 精度 (%) |
|---|---|---|
| cat | 444 | 68.0 |
| coffee | 328 | 52.1 |
| crab | 567 | 80.6 |
| dog | 272 | 97.4 |
| earth | 150 | 30.7 |
| fencing | 277 | 84.5 |
| fern | 132 | 74.7 |
| fungus | 135 | 93.2 |
| green | 167 | 65.9 |
| lion | 101 | 95.0 |
| lizard | 452 | 98.5 |
| squirrel | 140 | 94.3 |
| AVG | 263.7 | 77.9 |

表 2: 各ノイズ除去手法の主観評価による実験結果 [%]

| Method | Precision | Recall | F-measure |
|---|---|---|---|
| V-method | 79.8 | 55.6 | 64.0 |
| S-method | 85.1 | 52.7 | 63.3 |
| PAND-method | 87.4 | 28.7 | 41.7 |
| POR-method | 80.8 | **79.0** | **78.2** |
| SVS-method | **87.6** | 50.2 | 62.4 |
| SSV-method | 84.4 | 52.2 | 63.8 |

表 2 に，各ノイズ除去手法の評価実験結果を示す．ここでの評価値は，12 種のラベルでの平均値を示している．同表から双方の手法によって初期状態からの Precision の若干の増加が確認される．V-method は画像特徴量に基づく視覚的評価であり，全画像特徴量の重心ベクトルに基づいて各画像を評価しているため，マイノリティな画像が除去されている．つまり，より一般的な画像が抽出されていると考えられる．例えば，図 1 に示すように ‘cat’ の場合の ‘tiger’ や，‘crab’ の場合の ‘カニ料理’ などのマイノリティな画像がノイズとして評価されている．一方で，S-method の適合画像を考察すると，画像中のオブジェクトが大きく写った画像が多く抽出されていた．これは，オブジェクトが画像全体を占めており，周辺環境等のタグがあまり付加されないことで，概念距離として近い画像だけが抽出されていると考察する．パラレル型とシリアル型の 4 種類の組み合わせ手法において，PN-method は各組み合わせ手法の中で最も高い F-mesure の値（78.2%）を示し，低い誤検出および高い精度でのノイズ除去が確認された．V-method や S-method も高い Precision を示

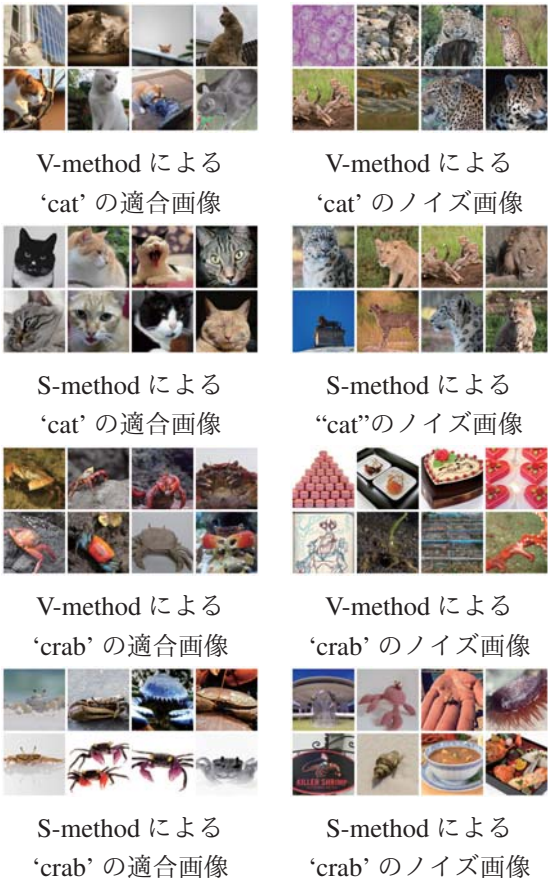


図 1: V-method と S-method による画像評価の例

したが，どちらも Recall は低く，多くの画像をノイズとして誤検出してしまっている．しかしながら，PN-method は結果的に視覚情報と意味情報のそれぞれの評価の和集合をとることになるため，ノイズ画像を除去しつつも網羅性が高い画像集合が抽出されたと考えられる．

## 5 おわりに

本稿では，全自動での Web からの画像データセット生成手法を提案した．提案手法では，視覚情報と意味情報にハイブリッドに用いることで，高精度なノイズ画像除去を実現した．そして，提案手法によってノイズ除去された画像データセットが，一般物体認識のための学習データセットとして実用可能な精度をもっていることを示した．今後は，一般物体認識において学習データとしての有効性をさらに検討していく．

謝辞　本研究の一部は中部大学情報科学研究所研究費，特別研究費ならびに科研費助成金基盤研究 (C) (#2633 0210) 及び，立命館大学研究高度化制度の支援による．ここに感謝申し上げる．

## 参考文献

[Bergevin and Levine, 1993] Bergevin, R. and Levine, M. D. (1993). Generic object recognition: Building and matching coarse descriptions from line drawings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(1):19–36.

[Chua et al., 2009] Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., and Zheng, Y. (2009). Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, page 48. ACM.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009*, pages 248–255. IEEE.

[Everingham et al., 2010] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338.

[Fergus et al., 2004] Fergus, R., Perona, P., and Zisserman, A. (2004). A visual category filter for google images. In *Computer Vision-ECCV 2004*, pages 242–256. Springer.

[Griffin et al., 2007] Griffin, G., Holub, A., and Perona, P. (2007). Caltech-256 object category dataset.

[LeCun et al., 2004] LeCun, Y., Huang, F. J., and Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004*, volume 2, pages II–97. IEEE.

[Miller, 1995] Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

[Opelt et al., 2006] Opelt, A., Pinz, A., Fussenegger, M., and Auer, P. (2006). Generic object recognition with boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):416–431.

[Song et al., 2004] Song, X., Lin, C.-Y., and Sun, M.-T. (2004). Autonomous visual model building based on image crawling through internet search engines. In *Proc. of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 315–322.

[秋間雄太 et al., 2010] 秋間雄太, 柳井啓司, et al. (2010). Folksonomy による階層構造画像データベースの構築. 研究報告コンピュータビジョンとイメージメディア *(CVIM)*, 2010(24):1–8.

[柳井啓司, 2007] 柳井啓司 (2007). 確率的 web 画像収集. 人工知能学会論文誌, 22(1):10–18.