大規模なデータセットの構築のための画像のフィルタリング手法

柴山 祐輝[1,a)]　森本 直之[1,b)]　山田 良博[1,c)]　岩村 雅一[1,d)]　黄瀬 浩一[1,e)]

概要：一般物体認識において，高い認識精度を実現することは重要である．高い認識精度を実現するには，識別器に学習させるデータセットに正しくラベル付けされている必要がある．また，あるカテゴリに属する物体には，外見のパターンがいくつか存在することから，多様なデータを含んだ大規模なデータセットが必要となる．このようなデータセットを作る上で，大量にデータを集めることは比較的容易であるが，正しくラベル付けするには膨大なコストがかかり，容易でない．そこで本稿では，ラベル付けする手間を省き，多様なデータを含む大規模なデータセットを構築するための画像のフィルタリング手法を提案する．実験結果では，集める対象となる画像を約 52%含む全体で 21,786 枚のデータセットをフィルタリングすることで，その画像を約 71%含む全体で 11,158 枚のデータセットにできた．従って，提案手法では，データセット全体の画像枚数を半分にし，集める対象となる画像の割合を約 19%上げることができた．

1. はじめに

近年，携帯電話やスマートフォンのようなデバイス機器が多くの人々に普及している．このようなデバイス機器の多くにはカメラが付随しており，そのカメラから撮影された画像を入力キーとして，様々なサービスを提供できれば便利であると考えられる．このようなサービスを提供する場合，撮影された対象物体がどのカテゴリに属するのかを認識する一般物体認識技術が必要となる．このようなサービスの実現には，一般物体認識で高い認識精度を達成することが重要である．高い認識精度を実現するには，識別器に学習させるデータセットに正しくラベル付けされている必要がある．また，あるカテゴリに属する物体には，外見のパターンがいくつか存在することから，多様なデータを含んだ大規模なデータセットが必要となる．

正しくラベル付けされ，多様なデータを含む大規模なデータセットを構築するには，まずデータを収集し，次にラベル付けする必要がある．データ収集の方法としては，Web でデータが収集できる場合，Web から特定のキーワードで自動収集を行うと，大量の画像を集めるには時間効率が良い．しかし，Web 上には集める対象となる画像だけでなくそれ以外の画像も存在するため，画像の自動収集では多くのノイズを含むデータセットが構築される．従って，ラベル付けには人の介在が必要不可欠なため，膨大なコストがかかる．一方で，画像を自動でラベル付けすると，ラベルを付ける手間が省けコストを抑えることができる．

集めた画像を自動でラベル付けするには，集める対象となる画像とそれ以外の画像を正確に識別する必要があるため，高精度な識別器を用意する必要がある．高精度な識別器を作るには，正確にラベル付けされたデータセットを用いて識別器を学習する必要があり，正確にラベル付けされたデータセットを作るには高精度な識別器を用いて，元あるデータを正確に識別する必要がある．従って，データセットと識別器は，一方ができればもう一方もできるという鶏と卵の関係である．ここで，集める対象となる画像かそれ以外の画像かを正確に識別する高精度な識別器が存在しないと仮定すると，識別器を用いることなく，正確にラベル付けされたデータセットを構築する必要がある．

そこで本稿では，高精度な識別器を用いることなく，正しくラベル付けされ，多様なデータを含む大規模なデータセットを構築することを試みる．そして，そのようなデータセットを構築するために，一部のラベル付き画像を用いた画像のフィルタリング手法を提案する．提案手法によって，集めた画像がどの程度フィルタリング出来るのか，その性能を評価した．本実験では，集める対象の画像を扉画像とした．扉画像を選んだ理由としては，扉には図 1 のよう

[1] 大阪府立大学大学院工学研究科 〒 599–8531 堺市中区学園町 1–1
Graduate School of Engineering, Osaka Prefecture University 1–1, Gakuencho, Naka, Sakai, Osaka 599–8531, Japan
[a)] shibayama@m.cs.osakafu-u.ac.jp
[b)] morimoto@m.cs.osakafu-u.ac.jp
[c)] yamada@m.cs.osakafu-u.ac.jp
[d)] masa@cs.osakafu-u.ac.jp
[e)] kise@cs.osakafu-u.ac.jp

(a) 引き戸　(b) 回転扉　(c) 自動扉

図 1: 様々な種類の扉

に引き戸や回転扉，自動扉など様々なパターンが存在するからである．実験では，画像投稿サイト Flickr から “door building”というキーワードで扉画像を集め，データセット中における扉画像の割合を上げることに成功した．この結果から，提案手法はデータセット中の扉画像の割合を上げることができ，データセットをフィルタリングする上で有効であることが分かった．

2. 関連研究

本節では，データセットの自動構築に関しての関連研究について述べる．

Schroff らはある物体の画像を自動で大量に集める手法を提案している [1]．この手法は，集めたい画像のテキスト情報を使って，画像のランク付けを自動で行い，画像を大量に集めていく手法である．まず，集めたい物体の名前を Web で検索し，その画像とその画像を含む Web ページを取得する．次に集めた物の中から，自動で不適切な画像を取り除き，残った適切な画像をランク付けする．このランク付けは，画像に付随しているテキスト情報や，画像のタイトルや，ファイルの名前といった情報を基に学習されるベイズ事後推定によって行われる．ランク付けされた画像の中で 1 位になった画像は，ランク付けの性能を上げるために，ベイズ事後推定に用いられる学習データとして使われる．実験の結果，Web から画像を集める従来手法より性能が良いことを示した．

データセットを構築するための手法に，Never Ending Image Learner (NEIL) と呼ばれるアルゴリズムを用いたものがある [2]．NEIL は半教師あり学習を用いたもので，Google 画像検索によって集められた画像に物体，背景，属性の 3 つのカテゴリでラベルを付け，そのラベル付けされた画像を基に検出器と識別器を学習させていく．次に，学習された検出器と識別器を用いて，ラベル付けされていない画像に対して，3 つのカテゴリと，その 3 つのカテゴリのいずれかを使って構成される関係性でラベル付けする．この関係性というのは，タイヤ (物体) は円形 (属性) であるといったものである．そしてラベル付けされた画像を学習用のデータセットに加えて，再度検出器と識別器を学習していき，40 万ものクラスに分けられたデータセットを構築している．

塚田らは，事例ベース文字検出手法と Semi-supervised learning に基づく自動ラベル付け手法の 2 つの手法を提案し，情景文字画像データベースを構築している [3] [4]．事例ベース文字検出手法で出力された文字画像を基に，自動ラベル付け手法ではラベル無しデータへのラベル付けと識別器の学習を同時に行い，多様なデータを認識できる識別器の作成を行っている．2 つの提案手法を統合したシステムに対する実験を行ったところ，文字領域の検出からデータベースへの登録までの一連の作業を自動化し，人の介在なしにデータベースを拡大させることができた．

上記の 3 つの手法はいずれも，あらかじめデータセットの一部に手動でラベルを付け，半教師あり学習を用いてデータセットを構築している．半教師あり学習では，初めに識別器に学習させるデータセットにおける，ラベル付けされる画像の枚数とラベル付けの正確さによって学習の性能が決まる．従って，性能の高い識別器を作るには，正確にラベル付けされた画像をより多く含むデータセットを手動で作る必要がある．しかし，ラベル付けは基本的に人の介在が必要なため，ラベル付けする画像が多くなるほど膨大なコストがかかる．そこで本研究では，ラベル付けする画像の枚数をできるだけ減らし，なおかつ，より正確にラベル付けされたデータセットを構築することを試みる．

3. 提案手法

3.1 提案手法における仮定

本手法の仮定は 2 つある．

(1) 同じカテゴリに属する画像は，類似した見た目を持ち，特徴抽出器で特徴量を抽出すると、特徴量間の距離は外見の類似性を反映する．

(2) 同じカテゴリに属する画像の中でも，いくつかのグループに画像は分けられ，そのグループごとに特徴量が密集している．

この 2 つの仮定が成り立つとすると，図 2 のように集める対象となる画像とそれ以外の画像は一見まばらに分布しているように見えるが，実は集める対象となる画像はいくつかのグループに分かれ，そのグループごとに特徴量が密集すると考えられる．この仮定の下，多くのノイズを含むデータセットから，集める対象となる画像のみを残すフィルタリング手法を提案する．以下では，提案手法の具体的な処理の流れについて説明する．

3.2 処理の流れ

図 3 に処理の流れを示す．初めに，画像を集め，次にその画像の特徴量を抽出する．次に，集める対象となる画像の一部に手動でラベルをつける．最後に，ラベル付けされた画像を検索質問画像 (クエリ) として，k 近傍探索をする．以下では，画像の収集，特徴量の抽出，k 近傍探索について詳しく説明する．

3.2.1 画像の収集

今回の実験では，集める対象となる画像を扉画像とす

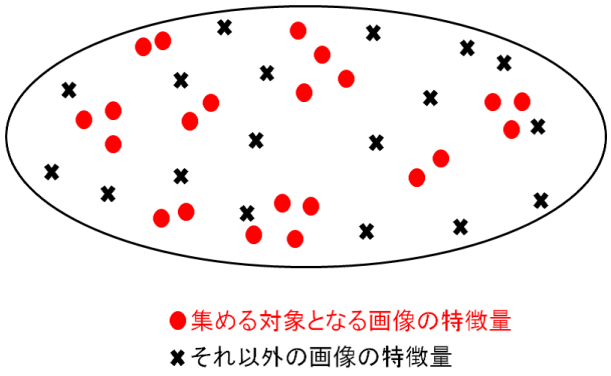


図 2: 2 次元上での集める対象となる画像とそれ以外の画像の特徴量の分布図

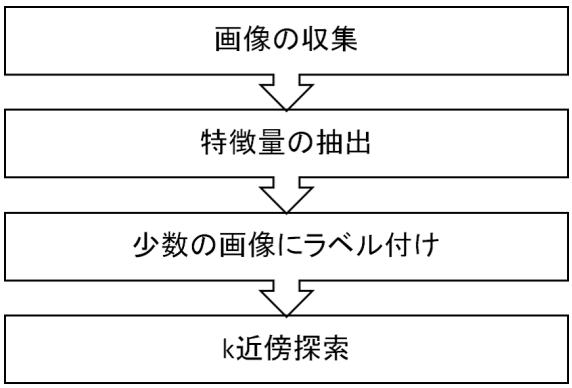


図 3: 処理の流れ

る．画像の収集は大量の画像を得るため，画像投稿サイト Flickr から “door building” というキーワードで自動で集める．図 4(a) と図 4(b) は，実際に集めた扉画像と扉以外の画像の例である．画像を集めるには，Google 画像検索や Yahoo!画像検索なども用いることができるが，API 上限を考慮して Flickr を選んだ．

3.2.2 特徴量の抽出

特徴量を抽出するために用いた特徴抽出器は Convolutional Neural Network (CNN) [5] の 1 つである Deep Residual Learning (ResNet) [6] である．ResNet は，画像から特徴的な情報を取り出し，それをもとに様々なカテゴリの画像を分類できる手法である．数ある CNN の中でも ResNet は，Imagenet Large Scale Visual Recognition Challenge 2015 (ILSVRC2015) において認識精度がトップであり，最も識別精度の高い識別器のうちの 1 つである．従って，ResNet は特徴抽出器としても優れていると考えられる．今回用いる ResNet は畳み込み層 151 層，全結合層 1 層からなる 152 層のもので，1000 クラスのオブジェクトを分類するようにあらかじめ学習されたものである．提案手法では，画像の特徴量を抽出し，扉か扉でないかを識別する識別器を用いず，あらかじめ学習された特徴抽出器として ResNet を用いる．図 5 は，ResNet の全体図である．提案手法で用いる特徴量は，全結合層の 1 つ手前の，151 層目の畳み込み層から出力される 2048 次元のものである．

3.2.3 k 近傍探索

まず，ラベル付けされた一部の画像をクエリとし，k 近傍探索をする．今回用いる k 近傍探索には近似を導入せず，クエリ画像の特徴量とデータベース中の画像の特徴量との距離計算を全ての画像について行う．次に，クエリと特徴量の距離が近い上位 k 個のデータベース特徴量を取り出す．今回は，距離尺度としてコサイン類似度を用いる．コサイン類似度は，n 次元の任意のベクトル $\vec{a} = (a_1, a_2, \cdots, a_n)$ と $\vec{b} = (b_1, b_2, \cdots, b_n)$ に対して，$\frac{\sum_{k=1}^{n} a_k b_k}{\sqrt{\sum_{k=1}^{n} (a_k)^2}\sqrt{\sum_{k=1}^{n} (b_k)^2}}$ と表せる．

4. 実験

本節では，提案手法における 2 つの仮定である，

(1) 同じカテゴリに属する画像は，類似した見た目を持ち，特徴抽出器で特徴量を抽出すると、特徴量間の距離は外見の類似性を反映する．

(2) 同じカテゴリに属する画像の中でも，いくつかのグループに画像は分けられ，そのグループごとに特徴量が密集している．

が成り立つかどうか検証する．また，提案手法によって集めた画像がどの程度フィルタリングできるのかを調べ，多様なデータを含む大規模なデータセットを作る上で，提案手法が有効であるか検証する．本実験では，集める対象の画像を扉画像とした．

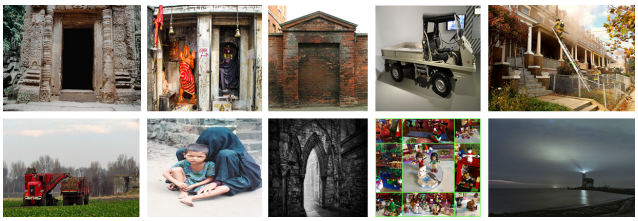
4.1 実験条件

まず，Flickr から “door building” というキーワードで扉画像の自動収集を行い，21,786 枚の画像を得た．その中には扉でないものも含まれる．次に，実験の評価に用いるためこの画像を扉があるかないかの 2 クラスに手動でラベル付けした．この時，扉画像は 11,324 枚であり，全体の約 52%であった．そして，集めた画像の中から扉画像をランダムに 100 枚選び，その 1 枚 1 枚に対して距離の近い上位 k 枚の画像を取り出した．最後に k の値を変化させながら Recall と Precision の値を得た．Recall は 11,324 枚の扉画像から何枚の扉画像が取り出せたかを示す割合であり，Precision は，取り出した上位 k 枚の画像の中に何枚の扉画像が含まれているかを示す割合である．この時取り出された画像中に重複があった場合は，重複があっても 1 枚の画像と見なした．

4.2 結果と考察

k=1, 5, 10, 15, 20, 30, 40, 50, 100, 150, 200, 300, 400, 500, 1000, 1500, 2000 と値を変化させていって Precision と Recall の値を得た．実験結果を図 6 に示す．図 6 から，k=1 の時 Precision=0.84 となった．この結果から，仮定

(a) 扉画像



(b) 扉以外の画像

図 4: Flickr から集めた画像

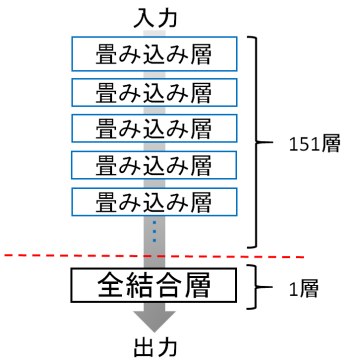


図 5: ResNet の全体図

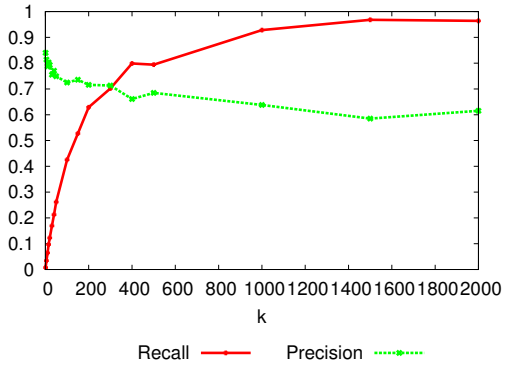


図 6: k を変化させたときの Precision と Recall の値

(a) 扉画像

(b) 図 7(a) に一番特徴量が似ている画像

図 7: 一番特徴量の似ている画像のうち扉画像でない場合の例

(1)(2) は成り立つことが多いが，必ず成り立つわけではないことが分かった．図 8(a) と図 8(b) は，一番特徴量の似ている画像のうち扉画像でない場合の例である．この結果から，画像の色や壁の模様が似ていると扉がない場合でも特徴量が近いと認識されることがわかった．実験によって算出した Precision と Recall の値を用いると，Precision-Recall Curve は図 8 のようになった．Recall はフィルタリング後のデータセットの大きさを表し，Precision は扉画像の割合を表す．図 8 から，Recall が低い時は Precision が高くなるが，Recall が高い時は Precision が低くなる．従って，フィルタリング後のデータセットの規模が小さければ小さいほど，その中に含まれる扉画像の割合が高くなることがわかった．Recall=0.1，0.2，0.3，0.4，0.5，0.6，0.7，0.8，0.9，1 の時，提案手法を用いると，扉画像と扉以外の画像を何枚ずつ含んだデータセットを構築できるかを調べ，図 9 に示した．図 9 から，k の値が小さいときは、大きいときに比べると、扉の画像が比較的多くとれ，k が大きくなればなるほど、得られる扉画像に対して、より多くの扉以外の画像が得られることがわかった．また，Recall がどの値の時をとっても，元のデータセットよりも扉画像の割合が高い．特に，Recall=0.7 の時，扉画像枚数が 7,927 枚で割合が約 71%の全体で 11,158 枚のデータセットを構築できている．元のデータセットは全体で 21,786 枚で，その中に含まれる扉画像の割合は約 52%であったので，データセット全体の画像枚数が半分になり，約 19%扉画像の割合が上昇している．この結果から，提案手法はデータセット中の扉画像の割合を上げることができ，データセットをフィルタリングする上で有効であることが分かった．従って，提案手法を用いることによって，多様なデータを含む大規模なデータセットを構築できると考えられる．

5. おわりに

大規模なデータセットを構築するためのフィルタリング手法を提案した．実験の結果，データセットをフィルタリングし，多様なデータを含む大規模なデータセットを作る上では，提案手法が有効であることが分かった．今後の課題は，識別器の性能向上には、データ数の増加と正解画像の割合の増加の両方が寄与するが、提案手法を使った時に，どのパラメータを使うのが識別性能の向上に寄与するかを知ることである．また，提案手法を改良し精度を上げ，より多くの画像に正解ラベルが付いたデータセットを構築す

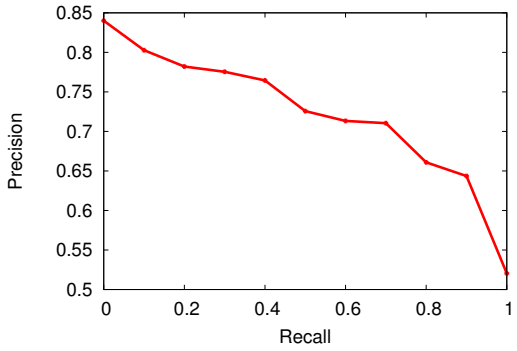


図 8: Precision-Recall Curve

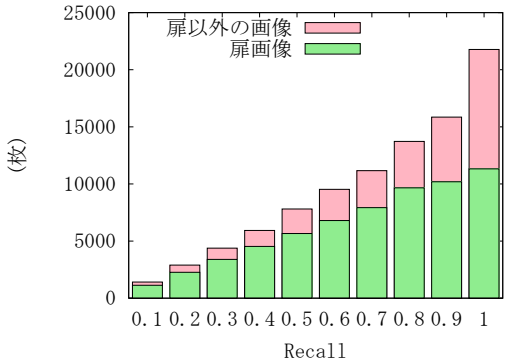


図 9: k を変化させたときにできるデータセットの中身

ることも課題の 1 つである．

謝辞 本研究は，JSPS 基盤研究 (A)25240028 ならびに基盤研究 (B)17H01803 の助成を受けたものである．

参考文献

[1] Schroff, F., Criminisi, A. and Zisserman, A.: Harvesting image databases from the web, *IEEE transactions on pattern analysis and machine intelligence*, Vol. 33, No. 4, pp. 754–766 (2011).

[2] Chen, X., Shrivastava, A. and Gupta, A.: Neil: Extracting visual knowledge from web data, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1409–1416 (2013).

[3] Iwamura, M., Tsukada, M. and Kise, K.: Automatic labeling for scene text database, *Proceedings of the IEEE International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1365–1369 (2013).

[4] Tsukada, M., Iwamura, M. and Kise, K.: Expanding recognizable distorted characters using self-corrective recognition, *Proceedings of the IAPR International Workshop on Document Analysis Systems (DAS)*, pp. 327–332 (2012).

[5] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, pp. 1097–1105 (2012).

[6] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016).