限定された学習時間の中で深層学習を活用する一般物体認識手法の検討

○大 井　翔（大阪工業大学）　兒 島　宏 樹（大阪工業大学）
梁　泳 成（大阪工業大学）　佐 野　睦 夫（大阪工業大学）

1. はじめに

近年，パソコンのスペックの向上や GPU の低価格化により，深層学習を用いたオブジェクト認識の研究が盛んに行われている．深層学習によるオブジェクト認識の結果は高い精度を誇り，今後もより高精度になっていくことが期待されている．

しかし，一般的な問題として学習済みのデータに対しては高い性能を誇る深層学習であるが，学習されていな Unknown なオブジェクトに対しては，検出・認識することは難しい．例えば，ロボカップ@ホームリーグにおいても学習オブジェクトとは異なるオブジェクトが競技前に公表されている．また，本年度の ARC2017（Amazon Robotics Challenge2017）においても，競技当日の開始 30 分前に Unknown オブジェクトが指定されている． Unknown オブジェクトに対して学習する時間があれば深層学習によって解決できるが，学習する時間がない場合にどのように対応するのかというのが現状の課題の一つであると考える．

そこで，本研究では事前に既知のオブジェクトに対しては十分に学習する時間がある状態で，新たに未知のオブジェクトを識別する問題に対して，十分に学習する時間がないことを前提とする．つまり，(I) 学習されていない類似オブジェクトに対して認識する，(II) 学習されていない Unknown オブジェクトに対して認識することを目的とする．

具体的には，オブジェクトの検出・認識精度の高い YOLO v2（You Once Look One）[1]と従来のオブジェクト認識の手法であった色特徴量や Bag-of-features に着目し，Unknown オブジェクトに対して認識を行う．

2. 関連研究

ARC2017 における藤吉ら[2]のアプローチは，オブジェクトを認識するために SSD（Single Shor MultiBox Detector）[3]をベースとして，オブジェクトか非オブジェクト化を判別するための Objectness 分類器を導入し，Unknown オブジェクトを検出する方式である．その後，Unknown オブジェクトの認識に関しては，学習時間の関係上，Unknown オブジェクトの重さと色特徴量のデータベースを作成することで，認識している．奈良先端大学と Panasonic の合同チームでは，YOLO v2 で検出・認識を行い，Unknown オブジェクトに対して，色特徴量と bounding box volume を用いて対応している[4]．鳥取大学と TOSHIBA の合同チームでも YOLO v2 をはじめに用いており，Unknown オブジェクトに対しては，AKWZE を用いた検出や距離学習手法による姿勢推定などを用いて行っている[5-9].

本研究でも，関連研究に述べたように，既知オブジェクトは Deep Leaning ベース，未知オブジェクトは学習する時間が限らたれているとし，従来の一般物体認識を用いた方式で検討を行う．

3. 物体検出・認識

本研究で扱うビジョン戦略を**図 1** に示す．本研究では，通常のカメラのみを用いる．

3.1. 深層学習による物体検出・認識

深層学習による物体検出・認識として，本研究では YOLO を用いる．YOLO は，リアルタイムな物体検出・認識をする手法であり，類似している手法として R-CNN（Regions width CNN features）[10]や Fast R-CNN[11]，Faster R-CNN[12]がある．YOLO はこれらの手法とは異なり，あらかじめ画像全体をグリッド分割しておき各領域に物体のクラスと bounding box を求める方式である．YOLO は Faster R-CNN より高速で実行されるが，CNN のアーキテクチャがシンプルになったため，識別精度は少し劣る．ただし，分割されたグリッドサイズは固定かつ，グリッド内で識別できるクラスは 1 つであり，検出できる物体の数は 2 つという制約を設けているため，グリッド内に大量のオブジェクトが映ってしまうような場合は誤検出が起こることがある．

3.2. Bag-of-features

一般物体認識において，Bag-of-features モデル[12]がある．Bag-of-features は，画像を局所特徴量の集合とみなし，局所特徴量のヒストグラムをその画像の特徴量としてカテゴリを識別する手法である[13].

本研究では，3 種類を用いて，それぞれ Visual Word を作成し，それぞれの組み合わせについて検討した．

1) 色特徴量（HSV 色空間）
2) SIFT（Scale-Invariant Feature Transform）
3) 色特徴量＋SIFT

3.3. Unknown ラベルに対する処理

Unknown ラベルが付与されているオブジェクトとして，学習データと類似しているオブジェクト群と学習データには存在しないオブジェクト群の 2 群がある．Unknown ラベル内の類似オブジェクトに関しては，Unknown ラベル内の類似オブジェクトを入

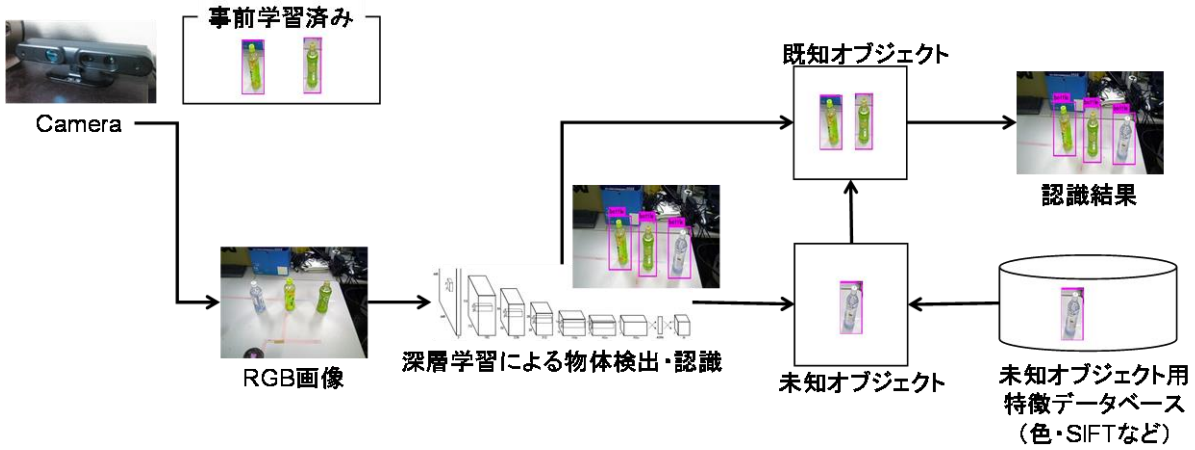


図 1　ビジョン戦略
Fig.1 Vision approach of our research.

力した際に，検出されたラベルの推定確率 $P(D)$と一般物体認識から得られた推定確率 $P(G)$を統合することで，オブジェクト $P(O)$の認識を行う．

$$P(O) = P(D)\,P(G) \qquad (1)$$

最後に，Unknown ラベル内の Unknown については，ある程度 YOLO で検出されることが予想される．そこで，検出されたエリアに対して，一般物体認識の手法を用いて，その結果をオブジェクトの結果とする．ただし，このとき，Unknown の各特徴量のデータベースは短い時間であるが，作成していることとする．　**図 2** に Unknown ラベルのオブジェクトに対する処理を示す．識別器として，本研究では SVM（Support Vector Machine）を用いた．

4. 実験

本研究で対象とするオブジェクトの画像を**図 3**，詳細を**表 1** に示す．オブジェクトのカテゴリ数は 5 とし，それぞれのカテゴリには 2 種類以上のオブジェクトがある．Unknown ラベルを付与しているオブジェクトに関しては，事前に学習できるオブジェクトと類似している画像（味や色が異なるなど）と既知オブジェクトには存在しないオブジェクトとしている．オブジェクトの選定に関しては，色特徴量だけでは識別が難しいように設定している．

実験の詳細については，以下の 3 種類の実験を行うことする．

実験 1）学習画像（既知画像）に対する精度
実験 2）Unknown ラベル内における類似オブジェクトの精度
実験 3）Unknown ラベル内における Unknown に対する精度

また，本研究では YOLO の学習のエポック数を 10,000 回行った．　学習 PC の GPU は GeForce GTX1080Ti を用いた．

5. 結果・考察

本章では，実験の結果と考察について述べる．

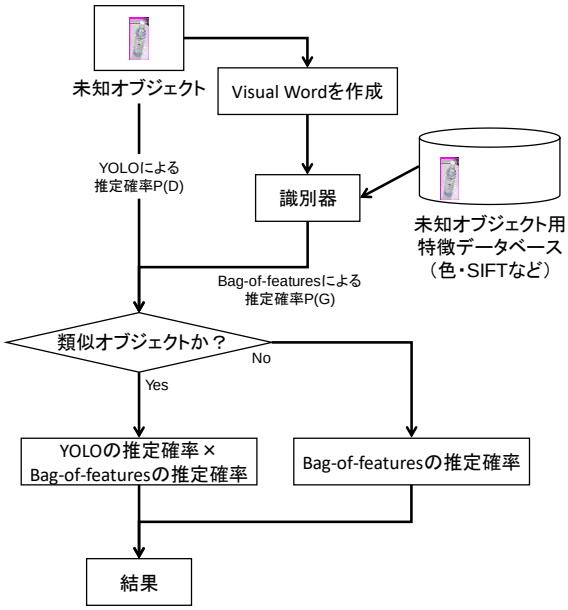


図 2　Unknown ラベルに対する処理
Fig.2 Processing of unknown label.

5.1.　結果

○実験 1）学習画像（既知画像）に対する精度

はじめに，実験 1 である既知画像に対する精度を**表 2** に示す．学習済みの既知オブジェクトに関しては，単体で映っている場合は Bottle を除いて正しく検出・識別されており，Bottle は写っているが検出されなかった（**図 4(a)**）．また，Cup のみ他のオブジェクトが単体で映っているときに誤検出されているケースがあった（**図 4(b)**）．また，複数のオブジェクトを並べてみた場合（**図 5**）も考慮したが，正しく検出できていない場合があり，全体の検出率が落ちる結果となった．

○実験 2）Unknown ラベル内における類似オブジェクトの精度

次に，Unknown ラベル内における類似オブジェクトの結果を**表 3** に示す．それぞれの特徴量ごとに行った結果，(3)色特徴量＋Bag-of-features を結合した方式で 69.3%という結果になった．YOLO の検出率

図 3　オブジェクト一覧
Fig.3 Object list.

表 1　オブジェクトの詳細
Table.1 Detail of objects.

既知オブジェクト		未知オブジェクト		
カテゴリ	枚数	カテゴリ	種類	枚数
Bottle	1,062	Bottle	1	24
Penholder	882	Penholder	1	24
Snack	522	Snack	1	24
Soup	1,170	Soup	1	28
Cup	504	Cup	1	22
		Unknown	4	29

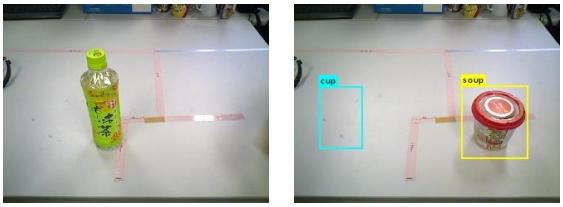
としては平均で 75.1%であった．特に，Snack に関しては YOLO で誤検出・認識したオブジェクトを正しく修正できている結果となった．

○実験 3）Unknown ラベル内における Unknown に対する精度

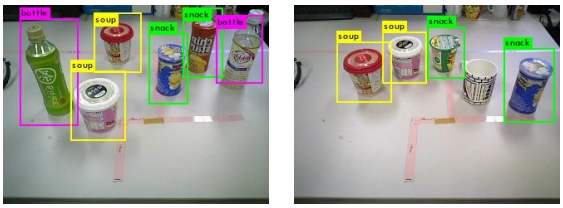
最後に，Unknown ラベル内における Unknown オブジェクトの結果を**表 4** に示す． 結果として，(1)色特徴量と(2)Bag-of-features の方式を統合している(3)の方式がお互いの良い部分をうまく統合している結果となり，平均で 81.7%の精度で分類することができている．特に，Coffee に関しては，お互いの特徴量を合わすことで精度が 100%となった．

表 2　既知オブジェクトに対する認識精度
Table.2 Accuracy of known objects.

	検出・認識精度(%)	誤検出回数
Bottle	0.815	ー
Penholder	1.000	ー
Snack	0.944	ー
Soup	0.963	ー
Cup	0.850	5
Average	0.914	ー



(a) 検出失敗　　(b) 誤検出
図 4　単体オブジェクトの例
Fig.4 Example of simple object



(a) 検出成功　　(b) 検出失敗
図 5　複数オブジェクトの例
Fig.5 Example of multiple objects.

表 3　類似オブジェクトに対する認識精度
Table.3 Accuracy of likely objects.

	(1)	(2)	(3)	YOLO の検出率
Bottle	0.739	0.870	0.870	0.870
Penholder	0.045	0.182	0.091	0.318
Snack	0.750	0.792	0.833	0.792
Soup	0.571	0.679	0.679	0.821
Cup	0.571	0.810	0.810	0.952
Average	0.535	0.666	0.693	0.751

5.2.　考察

○実験 1）学習画像（既知画像）に対する精度

表 2 より，全体的に 8 割以上な結果となった．図 5 において，Bottle が検出されない画像がいくつか存在した．これは，学習時には Bottle 単体の画像で学習したわけではなく，複数のオブジェクトが並んでいる状態＋手前の方に Bottle がある画像がなかったため，YOLO の位置の計算上，正しく検出できないモデルになっているのではないかと考える．次に，図 6 において Cup が検出された要因として，今回学習した Cup が白地を基調としたマグカップであり，テーブルの色も白地に近かったため，誤検出したと

表 4　Unknown オブジェクトに対する認識精度
Table.4 Accuracy of unknown objects.

	(1)	(2)	(3)	テスト画像数
Can	1.000	0.000	0.600	5
Coffee	0.875	0.625	1.000	8
Honey	0.667	0.167	0.667	6
Pack	0.500	1.000	1.000	10
Average	0.760	0.448	0.817	―

考えられる．最後に，図 7 においてコップが検出されていない例があるが，これは学習数の差が影響しているのではないかと考える．学習数の差が 2 倍程度の差が出ているため，正しく検出できなかったのではないかと考える．

○実験２）Unknown ラベル内における類似オブジェクトの考察

表 3 より，YOLO での検出率が 31.8%と低い結果になっているが，ほとんどが Cup と間違って検出されていた．見え方によっては取ってのついていない Cup に類似しているため，誤検出したと考えられる．そのため，一般物体認識の手法においても Penholder は改善することが難しかったと考えられる．

○実験３）Unknown ラベル内における Unknown に対する考察

表 4 より，Bag-of-features における Can と Honey の精度が悪く，Pack へ識別されていた．これは，Pack が他の key-point と類似していたためであると考える．また，Can のみ統合モデルの精度が落ちているが，これは Bag-of-features の精度が悪かったため，その影響を受けてしまい，精度が落ちたと考えられる．今回は YOLO が未知オブジェクトを検出できていたが実際には学習していないオブジェクトを検出することは難しいので，Point Cloud などを用いる必要があると考える．

6. 結論

今回，2 次元データのみを用いた，未知オブジェクトに対してあまり学習時間のない状況を想定した未知オブジェクト認識に関する検討を行った．実際のロボットのピッキングを行う際には画像認識だけでロボットがピッキングすることは難しく，3 次元情報を付加し，オブジェクトの触覚や重さなどの情報を加えることで実践的なロボットの把持ができるのではないかと考える．

謝辞　本研究の一部は，JSPS KAKENHI Grant Number JP15K00368 の支援を受けた．

参　考　文　献

[1] Redmon, J.，Divvala, S.，Girshick, R.，and Farhadi, A.：“You only look once: Unified, real-time object detection.”，In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition（CVPR），pp.779-788.

[2] H. Fujiyoshi, T. Yamashita, Y. Yamauchi, R. Murata, T. Hasegawa, M. Kaneko, Y. Murai, M. Hashimoto, S. Akizuki, M. Nagase, Y. Sakuramoto, S. Takei, S. Itoh, Y. Domae, R. Kawanishi, K. Shiratsuchi, R. Haraguchi, and M.Fujita, “Combined Point Cloud and Appearance-Based Object Detection for Grasping Rigid and Non-Rigid Objects, ”International Workshop on Recovering 6D Object Pose at ICCV, 2015.

[3] W. Liu, D. Anguelov, and D. Erhan, “SSD: Single Shot MultiBox Detector,” European Conference on Computer Vision, pp.21–37, 2016.

[4] 藤吉弘亘，松元叡一，岡田慧，“[特別講演] Amazon Picking Challenge 2016 の参加レポート,” パターン認識・メディア理解研究会，pp.123‐129，2017.

[5] W. Liu, D. Anguelov, and D. Erhan, “SSD: Single Shot MultiBox Detector,” European Conference on Computer Vision, pp.21–37, 2016.

[6] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit, “Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes,” International Conference on Computer Vision, pp.858–865, 2011.

[7] J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” Conference on Computer Vision and Pattern Recognition, pp.7263–7271, 2017.

[8] F.A. Pablo, B. Adrien, and J.D. Andrew, “KAZE Features,” European Conference on Computer Vision, pp.214-227, 2012.

[9] P. Wohlhart and V. Lepetit, “Learning descriptors for object recognition and 3D pose estimation,” Conference on Computer Vision and Pattern Recognition, pp.3109–3118, 2015.

[10] Girshick, Ross, et al.:“Rich feature hierarchies for accurate object detection and semantic segmentation.”, Proceedings of the IEEE conference on computer vision and pattern recognition，2014.

[11] Girshick，Ross：“Fast r-cnn.”Proceedings of the IEEE International Conference on Computer Vision，2015.

[12] Ren, Shaoqing, et al. :“Faster R-CNN: Towards real-time object detection with region proposal networks.”, Advances in neural information processing systems，2015.

[13] Li, F.-F. and Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories, CVPR ’05: roceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05) – Volume 2, Washington, DC, USA, IEEE Computer Society, pp.524–531 (2005).

[14] 永橋知行，伊原有仁，藤吉弘亘：“画像分類における Bag-of-features による識別に有効な特徴量の傾向”，コンピュータビジョンとイメージメディア(CVIM) Vol.2009-CVIM-169，2009.