ディープラーニングを用いた新しい画像領域分割手法

Effective Image Segmentation Technique Based on Deep Learning

ファン クォク ヴェト PHAM Quoc Viet

近年，物流・流通分野での労働負荷増大に伴い，物流倉庫などでロボット導入による作業自動化の取り組みが進んでいる。ロボットに“何がどこにあるか”を理解させるには，搭載カメラの画像から物体の領域を抽出する画像領域分割手法が必要不可欠となる。

東芝は，ディープラーニングを用いた新しい画像領域分割手法 BiSegを開発した。BiSegは，個別の物体領域を抽出するインスタンスセグメンテーションと物体領域を種類ごとに抽出するセマンティックセグメンテーションを，計算量を削減するため一つのニューラルネットワークで同時に行う。ネットワークモデルは，主に二つのサブネットワークで構成され，それぞれでインスタンスセグメンテーションとセマンティックセグメンテーションを扱っているが，特徴量抽出部分を共有化することで計算を効率化した。更に，公開データセットを用いてほかの手法とBiSegの性能比較を行った結果，BiSegでは精度良く物体領域と種類を推定できることを確認した。

With the recent increase in the burden on workers in the logistics and physical distribution fields, the movement toward the introduction of automated processes using robots has recently accelerated in various facilities including physical distribution warehouses. To achieve precise understanding of an object's type and location, an image segmentation technique that enables the detection and segmentation of each object in an image taken by a robot-mounted camera is becoming essential.

Toshiba Corporation has developed an effective image segmentation technique called BiSeg based on deep learning. Utilizing a neural network in order to reduce the amount of calculation, BiSeg can simultaneously implement the following two image segmentation tasks: (1) instance segmentation to detect and segment each object in an image, and (2) semantic segmentation to classify each pixel in an image into the type of object. This neural network mainly incorporates two subnetworks that respectively perform these two tasks and can efficiently extract feature quantities from an image. We have conducted simulation tests using open datasets and confirmed that BiSeg makes it possible to estimate object types and shapes more accurately compared with other techniques.

1. まえがき

近年の電子商取引（EC：Electric Commerce）市場の拡大に伴って，物流・流通分野における労働負荷は増大傾向にある。少子高齢化による労働人口の減少が進む現代では，十分な人数の労働者を確保することが困難なため，AI技術などを活用して作業を効率化・省力化することが喫緊の課題となっている。例えば，物流倉庫では，荷積みや，荷降ろし，ピッキング作業などを，ロボットによって自動化する取り組みが進んでいる。ロボットによる自動化を実現するためには，ロボットに“何がどこにあるか”を理解させる技術が必要不可欠となる。

東芝は，このようなニーズに応えるため，カメラ画像から物体の領域を抽出する画像領域分割技術の研究開発に取り組んでいる。

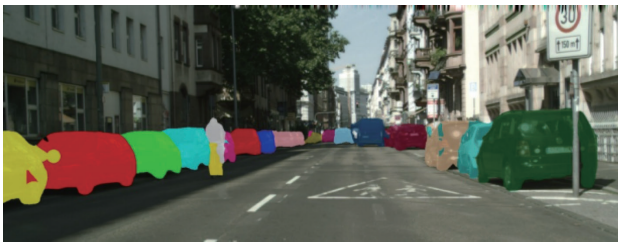
2. 画像領域分割技術

画像領域分割には，**図1**に示すように，個体ごとに領域分割するインスタンスセグメンテーションと，物体の種類ごとに領域分割するセマンティックセグメンテーションがある。図1(b)は，インスタンスセグメンテーションの例で，1人の人間と2匹の犬は個体ごとに三つの領域として抽出される。図1(c)は，セマンティックセグメンテーションの例で，人間と犬といった物体の種類ごとに二つの領域が抽出され，2匹の犬は一つの領域となる。

当社は，これら二つの領域分割を一つのネットワークで同時に行う，ディープラーニングによる新しい画像領域分割手法 BiSeg[(1)]を開発した。**図2**に示すように，インスタンスセグメンテーションは，不定形の領域を扱えないが，隣接した同種類の物体（例えば前の車と後ろの車）を区別できる。一方，セマンティックセグメンテーションは，隣接した同種類

(a) 入力画像　(b) インスタンスセグメンテーション　(c) セマンティックセグメンテーション

図1. インスタンスセグメンテーションとセマンティックセグメンテーションの例
画像領域分割には，個体ごとに領域分割するインスタンスセグメンテーションと，物体の種類ごとに領域分割するセマンティックセグメンテーションがある。
Examples of instance segmentation and semantic segmentation tasks



(a) インスタンスセグメンテーション

(b) セマンティックセグメンテーション

図2. 領域分割の両手法の比較
インスタンスセグメンテーションは，隣接した同種類の物体を区別でき，セマンティックセグメンテーションは，空や，芝生，道路などの不定形の領域を検出できる。
Comparison of two segmentation tasks

の物体は区別できないが，空や，芝生，道路などの不定形の領域を検出できる。開発したBiSegは，一つのニューラルネットワークでインスタンスセグメンテーションとセマンティックセグメンテーションを同時に行うことで，計算量を削減できる。また，4章で説明するように，二つのセグメンテーションタスクの強い相関を利用してより有用な特徴量マップを得ることで，両方のセグメンテーションの精度を向上できる。

3. ネットワーク構造

BiSegは，インスタンスセグメンテーションとセマンティックセグメンテーションの両タスクを，一つのニューラルネットワークで実行する（**図3**）。まず，特徴抽出サブネットワークで特徴量マップを生成する。この特徴量マップを共有して，後段の三つのサブネットワークで処理を行う。すなわち，インスタンスセグメンテーションを実行する個体領域抽出サブネットワーク，セマンティックセグメンテーションを実行する物体領域抽出サブネットワーク，及び物体領域プロポーザルサブネットワーク（RPN：Region Proposal Network）[(2)]である。ここで，RPNは，候補領域（ROI：Region of Interest）を生成する。また，個体領域抽出サブネットワークは，**図4**のように，ROIごとの物体の前景尤度（ゆうど）と背景尤度を推定し，物体領域抽出サブネットワークは，**図5**のように，物体領域の存在確率を予測する。最後に，個体領域抽出サブネットワークと物体領域抽出サブネットワークの結果をROIごとに統合し，BiSeg特有のインスタンスセグメンテーションの確率（物体前景確率）を算出する。

4. 評価実験

評価実験では，PASCAL VOC 2012データセット[(3)]を用い，学習セット（5,623枚）に対する学習と，評価セット（5,732枚）に対する評価を行った。評価尺度としては，mAP（Mean Average Precision）@50[(注1)]を用いた。

まず，インスタンスセグメンテーションの性能を評価する

（注1） 正解領域と50％以上一致すれば正解と判定する基準上で，クラスごとの平均適合率を平均したもの。

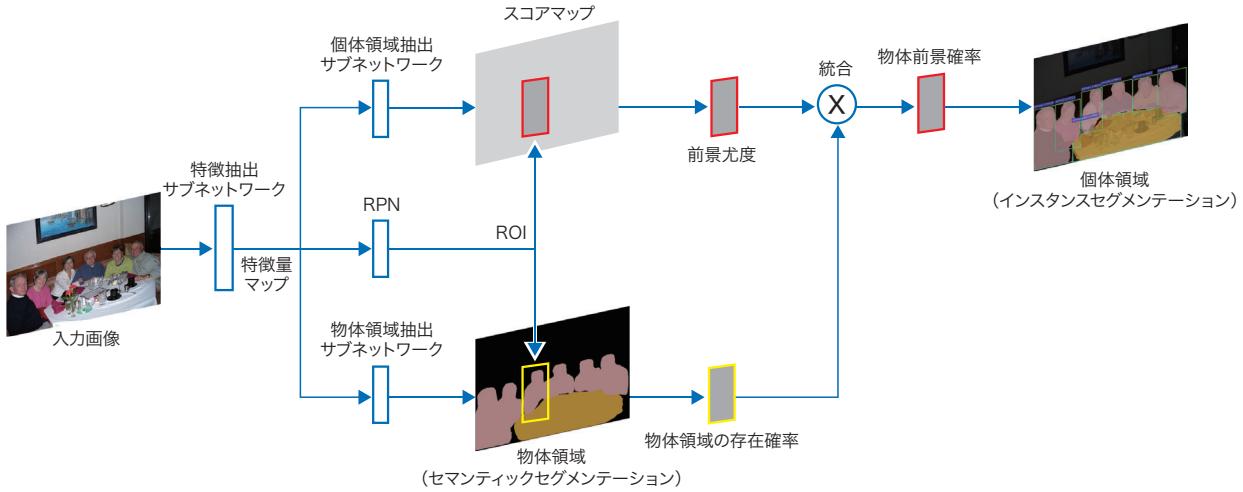


図3. BiSegのネットワーク構造

一つのニューラルネットワークで，インスタンスセグメンテーションとセマンティックセグメンテーションの両タスクを実行する。

Network architecture of BiSeg

図4. 個体領域抽出サブネットワークによるROIごとの物体の前景尤度と背景尤度の推定例

個体領域抽出サブネットワークが，ROIごとの物体の前景尤度と背景尤度を推定する。

Foreground and background likelihoods for each region of interest (ROI) predicted by instance segmentation subnetwork

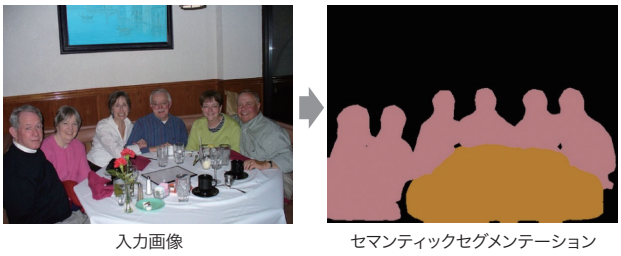


図5. 物体領域抽出サブネットワークによる物体領域の存在確率予測の例

物体領域抽出サブネットワークが，物体領域の存在確率を予測する。

Example of semantic segmentation probability predicted by semantic segmentation subnetwork

ため，BiSegを幾つかの手法と比較した結果を，**表1**に示す。開発手法のBiSegは，精度の高い従来手法FCIS[4]より1.6ポイント上回った。また，個体領域抽出サブネットワークだけを用いた場合は，mAP@50が64.2％とBiSegを3.1ポイント下回った。更に，インスタンスセグメンテーションとセマンティックセグメンテーションを独立に推定する一般のマルチタスク手法では，mAP@50が65.2％となり，BiSegを2.1ポイント下回った。BiSegでは，インスタンスセグメンテーションとセマンティックセグメンテーションの強い相関を利用することで，インスタンスセグメンテーションの精度が向上したと考えられる。

次に，セマンティックセグメンテーションの性能を評価した結果を，**表2**に示す。ここでは，Mean accuracy（クラ

表1. 手法別のインスタンスセグメンテーションの性能比較

Comparison of performance of instance segmentation using BiSeg and other techniques

画像領域分割手法	mAP@50（%）
FCIS[4]	65.7
インスタンスセグメンテーションだけ	64.2
一般のマルチタスク手法	65.2
BiSeg	67.3

表2. 手法別のセマンティックセグメンテーションの性能比較

Comparison of performance of semantic segmentation using BiSeg and conventional technique

画像領域分割手法	Mean accuracy（%）	Mean IU（%）
一般のマルチタスク手法	69.0	59.5
BiSeg	70.2	60.8

(a) 入力画像　(b) インスタンスセグメンテーション　(c) セマンティックセグメンテーション

図6. BiSegによる処理結果の例

下側の2行の画像のように，物体が隠れた場合でも，良いインスタンスセグメンテーション結果が確認できた。

Examples of results of segmentation obtained by BiSeg

スごと正解画素数の割合の平均）と，Mean IU（Intersection over Union：クラスごと正解領域とのオーバーラップ率の平均）という二つの尺度を用いた。表2から，BiSegは，インスタンスセグメンテーションだけではなくセマンティックセグメンテーションに対しても一般のマルチタスクより有効であることが確認できた。

最後に，公開データセット[3], [5]の中の幾つかに対する，BiSegによるインスタンスセグメンテーションとセマンティックセグメンテーションの処理結果を**図6**に示す。

5. あとがき

インスタンスセグメンテーションとセマンティックセグメンテーションを同時に行う，ディープラーニングによる画像領域分割手法 BiSegを開発した。両方のセグメンテーションタスクを一つのネットワークに実装することで，物体領域やカテゴリーが精度良く推定できることを確認した。

文　献

(1) Pham, Q. V. et al. "BiSeg: Simultaneous Instance Segmentation and Semantic Segmentation with Fully Convolutional Networks". The Proceedings of the British Machine Vision Conference 2017. London, UK, 2017-09, British Machine Vision Association (BMVA). 2017, Tuesday, 5 Sep. 58.

(2) Ren, S. et al. "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks". 28th Conference on Neural Information Processing Systems (NIPS 2015). Montreal, CANADA, 2015-12, NIPS. 2015.

(3) Everingham, M. el al. The Pascal Visual Object Classes (VOC) Challenge. The International Journal of Computer Vision. 2010, **88**, 2, p.303–338.

(4) Li, Y. et al. "Fully Convolutional Instance-aware Semantic Segmentation". The Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu, HI, 2016-07, IEEE. 2016, p.4438–4446.

(5) Lin, T. Y. et al. "Microsoft COCO: Common objects in context". The 13th European Conference on Computer Vision – ECCV 2014. Zurich, Switzerland, 2014-09, Springer. 2014, p.740–755.

ファン クォク ヴェト　PHAM Quoc Viet, Ph.D.
研究開発本部 研究開発センター
メディア AI ラボラトリー
博士（情報理工学）
Media AI Lab.