The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

物体メッシュモデルを用いた学習データ自動生成に基づく透明物体の深度画像予測と家事支援ロボットへの応用

Depth Image Prediction for Transparent Objects and its Application to Home Assistant Robot Based on Automatic Training Data Generation Using Mesh Model

内海 佑斗 Yuto Uchimi　和田 健太郎 Kentaro Wada　岡田 慧 Kei Okada　稲葉 雅幸 Masayuki Inaba

東京大学大学院情報理工学系研究科
The University of Tokyo, Graduate School of Information Science and Technology

Home assistant robots deal with various objects in the home. However, transparent objects such as wine glasses are difficult for robots to estimate 3D position, therefore they have trouble in manipulating such objects. This paper proposes a method of semantic segmentation and depth image prediction using CNN, which enables robots to get depth of transparent objects and reconstruct them three dimensionally. It also proposes a method of collecting training data, by which a person no longer has to annotate objects in 3D space. By using our depth prediction method, robots can handle PET bottles as well as non-transparent objects.

1. はじめに

家事支援ロボットには家庭内にある多種多様な物体を扱うことが求められるが、その中でもペットボトルやワイングラスなど透明物体の認識と操作は未だ解決されていない課題の一つである。これらは光を反射・吸収・透過するなどの性質を持つためセンシングによる三次元的な位置推定が困難であり，家事支援を遂行するための環境認識に際して大きな課題である．

近年，透明物体の三次元位置推定として，条件付き確率場や多層畳み込みネットワークを利用した機械学習による深度画像予測の手法の有効性が明らかになってきた [Eigen 14] [Liu 15] [Wang 15]．物体の特性を検知する手法 [Eren 09] [Rantoson 10] に対し，機械学習による手法の多くはシステムの入力としてカメラから取得した単一の色画像のみを用いることで，透明物体の認識のために特別な装置を要することなく深度予測を実現可能であるほか，画像中の各領域がどのような物体であるかの分類も可能である．

ところがロボットに搭載されるカメラセンサについては RGB-D の計測が可能なカメラが既に広く用いられているため，透明物体以外の深度取得にはセンサ値を採用すればよい．本研究では，多層畳み込みネットワークを用いてラベル画像予測と深度画像予測を同時に行い，センサから取得される深度画像のうち，ネットワークにより透明だと判定された領域のみを予測深度画像で置換した深度画像を出力して三次元位置推定を行うことを提案する．

しかしながら，教師あり学習を行うためには透明物体の正しい深度情報を持つ画像を収集する必要がある．NYU Depth [Silberman 12] などのデータセットは、深度情報を取得するための投射光が正常に帰ってこないという透明物体の性質のため、これらの正しい深度情報を含んでいない。この解決策として人間による三次元空間上でのアノテーションが挙げられるが，物体数が増えるに従って人間の負担も増加するため，効率的とは言えない．そこで本研究では，透明物体の三次元メッシュモデルを利用してタスク環境を仮想的に再現し，人間による三次元アノテーションの不要な効率的な学習データ生成を行うことを提案する．

連絡先: 内海佑斗，東京大学工学部情報システム工学研究室，113-8656 東京都文京区本郷 7-3-1 工学部 2 号館 7 階 73B2，03-5841-7416，uchimi@jsk.imi.i.u-tokyo.ac.jp

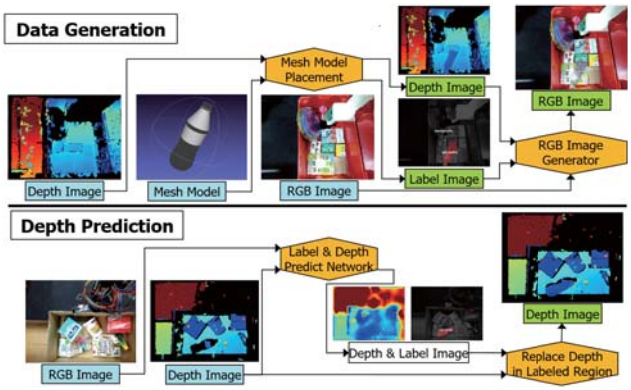

図 1: 学習データ生成 (上) と深度予測システム (下) の概観

本研究の提案手法の有用性を示すため，ロボットが家事支援において透明物体を扱うタスクに提案システムを適用する実験を行う．

2. 物体モデル配置による学習画像生成

機械学習による深度予測を行うために，透明物体の教師ラベル画像と教師深度画像を取得する必要がある．本研究では，まず実際の様々なタスク環境について，ロボットが各シーンにつき複数視点から撮影することで，透明物体の置かれていない画像データを収集する．次に，環境の点群上に透明物体の CAD モデルを仮想的に自動配置することで，教師画像を生成する．また，自動収集した教師画像に対応する色画像を生成するため，本研究では光の反射・物体に生じる陰・半透明性による曇りを再現するアルゴリズムと，GAN を組み合わせることで色画像を自動生成することを提案する．さらに，ラベル画像予測の精度向上のために，タスク環境上で実際の透明物体の色画像を撮影して，二次元アノテーションを行うことで教師信号としてラベル画像のみを持つ少数の学習データセットを用意する．

2.1 データ収集

ロボットに搭載された RGB-D カメラセンサを用いて，タスク環境に透明物体の置かれていない色画像 I_{obj-} および深度

画像 D_{obj-} を撮影する．撮影の際には色画像と深度画像の他にも，カメラ内部パラメータ K とロボットのベースリンクからカメラへの座標変換 R_{cam}^{base} も同時に保存しておく．

2.2 深度画像およびラベル画像の生成

透明物体の写っていない色画像 I_{obj-} に対し，物体メッシュモデルを置くことのできる領域 (これを可配置領域とする) を二次元アノテーションにより与える．続いて深度画像 D_{obj-} のうち可配置領域以外を N/A として，点群を生成する領域を限定する．これにより作られた深度画像と，収集済みのカメラパラメータ K および座標変換 R_{cam}^{base} を用いて，ロボットのベースリンク座標系での点群を生成し，その点群上に物体メッシュモデルを仮想的に配置する．

配置の際の物体モデルの姿勢は，モデルを直方体に近似して上を向く面を 6 面からランダムに決定し，その後鉛直軸周りにランダムな角度だけ回転させることで決定される．またモデルの位置に関しては，水平方向には可配置領域内でランダムに決定され，鉛直方向には重力に従って可配置領域内の面に接するように決定される．

メッシュモデルを仮想配置した後にレイキャスティングを行ってカメラ視点での物体モデルの深度画像とラベル画像 L_{obj_g} を取得し，そしてこの深度画像を収集済みの深度画像 D_{obj-} に重ねて，学習に用いる深度画像 D_{obj_g} を生成する．

2.3 色画像の生成

ここでは，学習時に用いるため自動生成された深度画像およびラベル画像に対応するタスク環境の色画像を自動生成する．

2.3.1 光の反射・物体の陰・半透明による曇りの再現

色画像生成の第一段階として，光の反射・物体に生じる陰・物体の半透明性による曇りを人工的に再現した画像 I_{obj_g} を生成する．この段階には透明物体の写っていない色画像 I_{obj-}，生成された深度画像 D_{obj_g} およびラベル画像 D_{obj_g} を用いる．

再現アルゴリズムは以下のように説明される．まず最初に，色画像 I_{obj-} とラベル画像 L_{obj_g} を用いて，色画像のうち透明物体の背景に相当する部分を抽出する．第二に，深度画像 D_{obj_g} に Canny アルゴリズムを適用してエッジを抽出し，各画素のエッジからの L2 距離を計算した画像を用意する．これと並行して，ラベル画像を透明か否かによって二値化したマスク画像に水平右方向の Sobel フィルタを半時計回りに 45° だけ傾けたフィルタを適用し，さらに平滑化して物体左下部のエッジ付近が抽出された画像を用意する．

第三の工程では物体に陰をつける．第二の工程で作成したエッジからの距離画像の画素 i の画素値 x_i に対し，式 (1) の第 2 式を用いて明度を決定する係数 v_i を計算する．ただし，第二の工程で作成したエッジ付近の抽出された画像の画素値 y_i が閾値 $thre$ 以上ならば式 (1) の第 1 式を用いて v_i を計算する．

$$v_i = \begin{cases} \dfrac{1}{(1+\exp(-\beta(x_i+\theta)))^2} & (if\ y_i > thre) \\ \dfrac{1}{1+\exp(-\beta(x_i+\theta))} & (otherwise) \end{cases} \tag{1}$$

この係数 v_i を最初の工程で抽出した色画像の R，G，B 各チャンネルに乗じることで物体に陰をつけ，その後陰をつけた画像で入力色画像 I_{obj-} のマスク領域を置き換える．

第四の工程では透明物体の曇り，すなわち半透明性を表現する．第三の工程で生成された画像のうちマスク領域について，各画素 i の R，G，B 各チャンネルの画素値 x_i に不透明度 $opacity$ を考慮した式 (2) を適用して，画素値 y_i を得る．

$$y_i = (1 - opacity) \cdot x_i + opacity \cdot 255 \tag{2}$$

第五の工程では光源が画像上方向に存在すると仮定した時の白色光源からの光の反射を表現する．垂直上方向の Sobel フィルタを深度画像 D_{obj_g} に適用し，さらに平滑化を行って光が反射して見える領域を抽出する．物体マスク領域に関して，これにより得られる画像の各画素 i の画素値 x_i にパラメータ $reflect$ を乗じた上で，第四の工程で生成した色画像の R，G，B 各チャンネルの画素値 y_i を加えることで，出力色画像 I_{obj_g} の R，G，B 各チャンネルの画素値 z_i を得る．

2.3.2 敵対的生成ネットワークを用いた人工色画像の変換

色画像生成の第二段階として，CycleGAN [Zhu 17] を用いて色画像 I_{obj_g} を変換し，変換後の色画像 I_{obj_g2} を訓練データセットの色画像として採用する．

本研究では，人工画像ドメインとして透明物体再現アルゴリズムにより生成された色画像 I_{obj_g} から透明物体に注目して切り出した ROI 画像を用いる．一方，実在画像ドメインには実在する透明物体に注目した色画像群 I_{obj+} を用意する．これらを用いて学習したネットワークによって色画像 I_{obj_g} を変換し，物体マスク領域のみを変換された画像で置き換えることで，訓練データセットに用いる色画像 I_{obj_g2} を得る．

以上により得られる色画像 I_{obj_g2}，ラベル画像 L_{obj_g}，深度画像 D_{obj_g} の例を図 2 に示す．

色画像 I_{obj_g2}　ラベル画像 L_{obj_g}　深度画像 D_{obj_g}

図 2: 自動生成されたデータセットの例

3. ラベル画像予測と深度画像予測

生成された学習データセットを用いて，透明物体のラベル画像予測および深度画像予測を行う多層畳みこみネットワーク (CNN) に学習させる．本研究の提案ネットワークは，図 3 に示すようにセンサから取得できる色画像と深度画像の 2 つを入力とし，透明物体のラベル画像と深度画像を出力する．

3.1 ラベル画像予測

本項および次項では，本研究で提案するネットワーク構造について説明する．

ネットワークモデルのうちラベル画像予測を行う部分については，生成された色画像 I_{obj_g2} を入力として学習させ，FCN-8s at-once [Shelhamer 17] を用いてフォワーディングを行う．なお各層の重みやバイアスについては，事前学習された VGG16 モデル [Simonyan 14] を用いて初期化することとする．

最終層の出力 y を用いて計算を行うセグメンテーションに対する損失関数 $Loss_{seg}$ は，式 (3) に示すようにラベル画像 L_{obj_g} を教師信号とする Softmax Cross Entropy を用いる．$I(L_{obj\_g}^{i,\ j})$ は指示関数で，教師ラベル画像の i 番目の画素にラベル $j \in \{0, 1, \cdots, cls-1\}$ が付されていれば 1，それ以外なら 0 を返す．また $y^{i,\ j}$ は最終層の出力 y の i 番目の画素の第

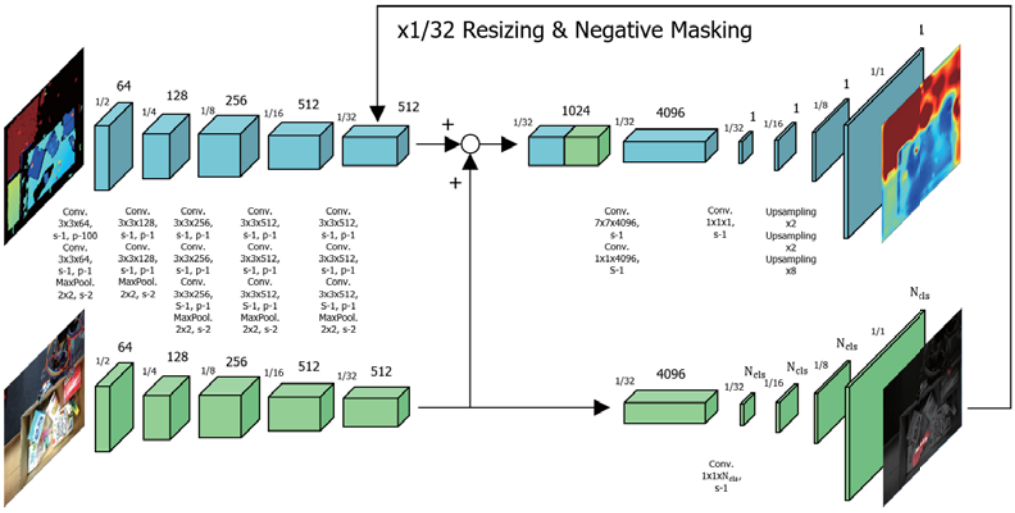

図 3: 提案ネットワークモデル．左上：入力深度画像，左下：入力色画像，右上：出力深度画像，右下：出力ラベル画像

j チャンネルの値を表す．

$$Loss_{seg} = -\frac{1}{n_{pix}} \sum_{i=0}^{n_{pix}-1} \sum_{j=0}^{cls-1} I(L_{obj\_g}^{i,\ j}) \log(Softmax(y^{i,\ j})) \tag{3}$$

3.2 深度画像予測

深度画像予測を行う部分については，生成された色画像 I_{obj_g2} と透明物体の置かれていない深度画像 D_{obj-} を入力として学習させる．入力深度画像 D_{obj-} は JET カラーマップを用いて 3 チャンネルの色画像に変換しておく．この画像に畳み込みを行って得られる 512 チャンネルの特徴量画像 $pool5_{depth}$ を得た後に，教師ラベル画像 L_{obj_g}(テスト時は予測ラベル画像 L_{pred}) をこれと同じサイズになるよう 1/32 にリサイズする．そして予測ラベルが 0(背景) の画素のみを残してそれ以外の画素値を 0 とするよう，$pool5_{depth}$ の各チャンネルごとにマスキングを行う．その後，入力色画像に畳み込みおよび最大プーリングを順次施して画像サイズ 1/32 になった段階で最大プーリングを行った後の 512 チャンネル特徴量画像とマスキングされた画像を連結させて，1024 チャンネル特徴量画像を生成し，この特徴量画像から深度画像予測を行う．

深度画像予測の損失関数は，式 (4) に示す画像全領域の損失関数 $Loss_{depth}^{all}$ と，式 (5) に示す教師マスク領域の損失関数 $Loss_{depth}^{mask}$ からなる．D_{pred}^{i} および $D_{obj\_g}^{i}$ はそれぞれ予測深度画像，教師深度画像の i 番目の画素を表す．また M^i は教師ラベルが 1 以上 (透明物体) ならば画素値 1，それ以外 (背景) ならば 0 となる教師マスク画像の i 番目の画素を表す．

$$Loss_{depth}^{all} = \frac{1}{n_{pix}} \sum_{i=0}^{n_{pix}-1} (D_{obj\_g}^{i} - D_{pred}^{i})^2 \tag{4}$$

$$Loss_{depth}^{mask} = \frac{1}{sum(M^i)} \sum_{i=0}^{n_{pix}-1} (D_{obj\_g}^{i} - D_{pred}^{i})^2 M^i \tag{5}$$

3.3 ネットワークの重み更新

ネットワーク全体の損失関数として，式 (6) に示すように損失関数 $Loss_{total}$ を設計する．

$$Loss_{total} = Loss_{seg} + \lambda_1 (Loss_{depth}^{all} + \lambda_2 Loss_{depth}^{mask}) \tag{6}$$

本研究ではパラメータ $\lambda_1 = \lambda_2 = 10$ とする．この損失関数を用いて誤差逆伝播を行い，ネットワークの重みを更新する．ただし，本研究ではラベル予測精度の向上のため，自動生成した主データセットと実環境を二次元アノテーションして用意した少数の副データセットを交互に学習する．副データセットは教師深度画像を持たないので，これの学習時は $\lambda_1 = \lambda_2 = 0$ として深度予測ネットワークの誤差逆伝播を行わないことにする．

深度画像予測ネットワークのマスキング処理に対しては誤差逆伝播を行わないこととするが，特徴量の連結処理については誤差逆伝播を行う．従って，ラベル予測ネットワークでは教師ラベル画像と教師深度画像を用いた特徴量抽出が可能となる．

4. 検証実験

検証実験では，透明物体としてペットボトルを用いる．生成したデータセットおよびネットワークモデルの有用性を検証する実験を行い，さらに透明物体のピッキング実験を行う．

4.1 生成色画像データの検証実験

自動生成した色画像での学習が実環境画像を用いたテスト時にも有効であることを検証する．本検証実験では，主データセット (Main)1610 組，副データセット (Sub)120 組を用意する．これらのデータを既存の FCN-8s at-once [Shelhamer 17] ネットワークモデルに学習させ，実環境画像を三次元アノテーションして用意した 130 組のテストデータに対してラベル画像予測を行った結果を表 1 に示す．なお，ここでは有色と無色を分けて結果を記載している．130 組のテストデータの内訳は，(A) 有色透明物体 22 組，(B) 無色透明物体 108 組となっている．

表 1: FCN-8s at-once による主データセットと副データセットのラベル画像予測結果．(A) 有色透明物体，(B) 無色透明物体．

Dataset	Mean IU		
	(A)	(B)	(A) + (B)
Main	0.582	0.511	0.523
Sub	0.796	0.713	0.727
Main + Sub	0.793	0.673	0.693

表 1 から，有色透明・無色透明のいずれの場合においても，生成した主データセットと少数の副データセットを交互に学習することで，実環境画像である副データセットのみの学習時に近い Mean IU を得られたと言える．

4.2 ラベルと深度を予測するネットワークの検証実験

本研究の提案ネットワークとの比較対象として，色画像からラベル画像を予測する CNN および深度画像とラベル画像から深度画像を予測する CNN の二つを用意する．これら二つのネットワークが予測したラベル画像および深度画像の精度と，本研究の提案ネットワークが予測したラベルおよび深度画像の精度を比較することで評価を行う．比較対象のうち深度予測ネットワークの学習時には提案手法同様にマスキング処理を行うが，特徴量画像の連結処理については行わないこととする．

表 2 に，比較対象の二段階ネットワークと提案ネットワークによるラベル予測の Mean IU と深度画像予測の精度を評価した実験の結果を示す．深度画像予測精度は，ネットワークの出力深度画像のうち教師マスク領域について，教師深度画像との誤差の絶対値が x 未満である画素数の割合を表している．

表 2: 二段階予測ネットワークと提案ネットワークのラベル画像および深度画像予測結果

Network	Mean IU	Depth accuracy ($\lvert error \rvert < x$)		
		x = 1cm	x = 2cm	x = 5cm
Staged	0.693	0.264	0.487	0.850
Ours	0.670	0.242	0.454	0.850

表 2 から二段階予測ネットワークを用いても本研究の提案手法を用いても精度に大きな差はないと言える．しかし，二段階予測のうち深度予測ネットワークの学習はラベル予測の学習が終わってからでないと進行できないため，提案ネットワークは二段階予測に比べ学習時間を低減できるという長所がある．

4.3 透明物体の把持への適用実験

最後に，ロボットによる物体把持行動および把持収納行動に対し本研究で提案した深度画像予測システムを適用する実験を行う．透明物体の代表としてペットボトルを家庭用冷蔵庫から取り出すタスクを実現し，5 回中 4 回成功した．また，ペットボトルを含む 5 つの家庭内物体を連続して籠から荷箱へ移し替えるタスクを実現した．図 4 および図 5 にそれぞれペットボトルを取り出している様子，移し替えている様子を示す．

図 4: 冷蔵庫からペットボトルを取り出す様子

図 5: 籠からペットボトルを取り出し，荷箱へ収納する様子

5. 結論

本研究では透明物体のピッキング動作を行うための物体三次元位置推定として，CNN を用いてラベル画像と深度画像を同時に予測し，センサの深度画像のうち透明領域のみを予測された深度画像で置換して出力することを提案した．また学習データを効率的に収集するため，透明物体の三次元メッシュモデルを用いてタスク環境を再現し，さらに光の反射・陰・物体の曇りの再現アルゴリズムと GAN を組み合わせることで，人間による三次元アノテーションの不要なデータセット生成手法を提案した．検証として，透明物体であるペットボトルを家庭用冷蔵庫から取り出す実験を行い，5 回中 4 回成功した．また，ペットボトルを含む 5 つの家庭内物体を連続して籠から荷箱へ移し替える実験を行い，成功した．

透明なショーケースの中にある物体のマニピュレーションには，ショーケースと内部の物体の両方の位置を認識することが必要となる．今後は本研究を応用して，これらの三次元位置推定を実現することを目指す．

参考文献

[Eigen 14] Eigen, D., et al.: Depth Map Prediction from a Single Image using a Multi-Scale Deep Network, in *Advances in Neural Information Processing Systems 27*, pp. 2366–2374, Curran Associates, Inc. (2014)

[Eren 09] Eren, G., et al.: Scanning from heating: 3D shape estimation of transparent objects from local surface heating, *Opt. Express*, Vol. 17, No. 14, pp. 11457–11468 (2009)

[Liu 15] Liu, F., et al.: Deep Convolutional Neural Fields for Depth Estimation From a Single Image, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015)

[Rantoson 10] Rantoson, R., et al.: 3D reconstruction of transparent objects exploiting surface fluorescence caused by UV irradiation, in *2010 IEEE International Conference on Image Processing*, pp. 2965–2968 (2010)

[Shelhamer 17] Shelhamer, E., et al.: Fully Convolutional Networks for Semantic Segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 4, pp. 640–651 (2017)

[Silberman 12] Silberman, N., et al.: Indoor Segmentation and Support Inference from RGBD Images, in *ECCV* (2012)

[Simonyan 14] Simonyan, K., et al.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *CoRR*, Vol. abs/1409.1556, (2014)

[Wang 15] Wang, P., et al.: Towards Unified Depth and Semantic Prediction From a Single Image, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015)

[Zhu 17] Zhu, J., et al.: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, *CoRR*, Vol. abs/1703.10593, (2017)