RGB-D 画像からの物体検出における対応点集合類似度の学習

○金崎朝子 (東京大学)　Emanuele Rodolà(ミュンヘン工科大学)　原田達也 (東京大学)

1. 背景

画像からの物体検出はロボットのマニピュレーションや行動計画等において重要なタスクである．物体を検出する手法として，SIFT [6] 等の特徴点検出と記述子を用いて現在の画像フレームと物体の参照画像間の対応点集合を求める手法がよく知られている．対応点は，各特徴点の周辺の局所領域を記述する局所記述子同士を比較し，その類似度が高い点同士を結ぶことで得られる．ここで，参照画像中の特徴点に対応する特徴点が現在の画像フレーム内に存在しない場合（隠れ：オクルージョン）や，照明変化や物体の姿勢変化によって記述子自体が大きく変化した場合，よく似た別の特徴点が選択される誤対応が発生しがちである．SIFT 特徴点と記述子を用いて，参照画像中の全特徴点に対する対応点を全探索した結果を **Fig.**1 (a) に示す．非常に多くの誤対応点が得られている様子が分かる．

誤対応点を除去する方法として RANdom SAmple Consensus (RANSAC) がよく知られている．RANSAC は，対応点集合の中からある部分集合をランダムに選択して物体の姿勢変化パラメータを推定し，外れ値を除く操作を繰り返して，最適な対応点集合を選択する手法である．ただし，初期対応点集合中の誤対応点の数が比較的多い場合には推定に失敗しがちである．

もうひとつのよく知られる方法として，二組の対応点ペアの類似度の総和を最大化するような対応点の部分集合を決定するものがある．二組の対応点ペアの類似度計算には，点間距離がよく使用される．特に剛体の物体を扱う場合は，ユークリッド距離を用いる．以下，この手法をより具体的に説明する．まず，参照画像中の点 i と点 j に対して，現在の画像フレーム中の対応点をそれぞれ i'，j' とし，点 x の三次元位置座標を p_x とおく．画像中の点の三次元位置座標は特徴点のスケールから推定することも可能であるが，本研究では RGB-D 画像を使用し，深度情報から一意に求まるものとする．これら二組の対応点ペアの類似度 s_{ij} は，下式で定義できる．

$$s_{ij} \equiv \exp\left(-|\|p_i - p_j\| - \|p_{i'} - p_{j'}\||\right) \tag{1}$$

これは，参照画像中の点間距離と現在の画像フレーム中の点間距離が近いほど 1 に近く，遠いほど 0 に近い値をとる．ここで対応点の総数を M とし，$A_{ij} = s_{ij} (i \neq j)$，$A_{ii} = 0$ となる $M \times M$ の行列 $A \in \mathrm{R}^{M \times M}$ を用意する．さらに，各対応点の割り当てを表すベクトル $\boldsymbol{x} \in \{0,1\}^M$ を用意する．これらを用いて，本手法は下記の二次割当問題を解く．

$$\max \quad \boldsymbol{x}^{\mathrm{T}} A \boldsymbol{x} \tag{2}$$

$\boldsymbol{x}$ が離散値をとるとき，上式は NP 困難であることが知られている．そこで，$\boldsymbol{x}$ の各要素が $[0,1]$ の範囲の連続値をとるとし，$\|\boldsymbol{x}\|_2 = 1$ の制約の下で上式を解く手法が提案されている [2]．最終的に $\boldsymbol{x}$ の i 番めの要素がある閾値以上であれば i 番めの対応点を残し，それ以外を誤対応点として除去する．また，$\|\boldsymbol{x}\|_1 = 1$ の制約の下で上式を解き，$\boldsymbol{x}$ としてより安定した解を得る手法が提案されており [8]，本研究ではこれを用いる．このようにして得られた誤対応点除去を行った対応点集合の例を **Fig.**1 (b) と **Fig.**1 (c) に示す．ただし，**Fig.**1 (b) には正解の対応点集合を，**Fig.**1 (c) には不正解の対応点集合を載せた．

こうして得られた対応点集合に対して，最終的に，正しく物体を検出できているかどうかを評価する必要がある．このため，本研究では**対応点集合類似度**を設計し，正解の対応点集合に対してはこの値が高く，不正解の対応点集合に対しては低くなるような学習手法を提案する．対応点集合類似度として，最も単純には，最終的に残った対応点集合中の全対応点ペアの類似度 s_{ij} の総和を，(残った) 対応点の総数 N で正規化した下記の値が使用できる．

$$\frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \exp\left(-|\|p_i - p_j\| - \|p_{i'} - p_{j'}\||\right) \tag{3}$$

これは，任意の二組の対応点ペアの幾何的な整合性を対応点集合全体で評価した値となる．しかしながら，この方法では “偶然” 互いに幾何的な整合性のとれた対応点ペアを抑制することができない（**Fig.**1 (c) 参照）．よって，高く評価する対応点ペアと低く評価する（無視する）対応点ペアとを区別するよう物体毎に学習し，偶然的に全体の類似度が高くなってしまう事例を減らす必要がある．

本研究では，RGB-D 画像を入力とし，対応点ペア類似度行列の二次割当問題を解いて得られた対応点集合に対して，その正当性を評価する対応点集合類似度を導入し，そのパラメータを物体毎に学習する手法を提案する．対応点集合類似度は，二組の対応点ペアの類似度（式 (1)）に対し，ある一方の対応点ペアのもつ色ともう一方の対応点ペアのもつ色との組み合わせによって区別されたインデックスに対する重みを乗算し，全対応点ペアについてこの値を総和する．ここで，重みは（対象物体毎に）高く評価すべき対応点ペアに対して大きく，低く評価すべき対応点ペアに対して小さくなるように学習する．対応点集合類似度のパラメータを学習する関連研究としては，式 (1) で示した二組の対応点の二点間距離の他に，一組の対応点の記述子や法線ベクトル方向の類似度等の別のプロパティの類似度を導入し，各類似度に対する重みを最適化するものがある [1, 3, 4, 5, 7]．これらに対し，本研究の定式化は単一のプロパティの類似度（式 (1)）を用いるが，別のプロパティ（色）によって区別して重み付けする

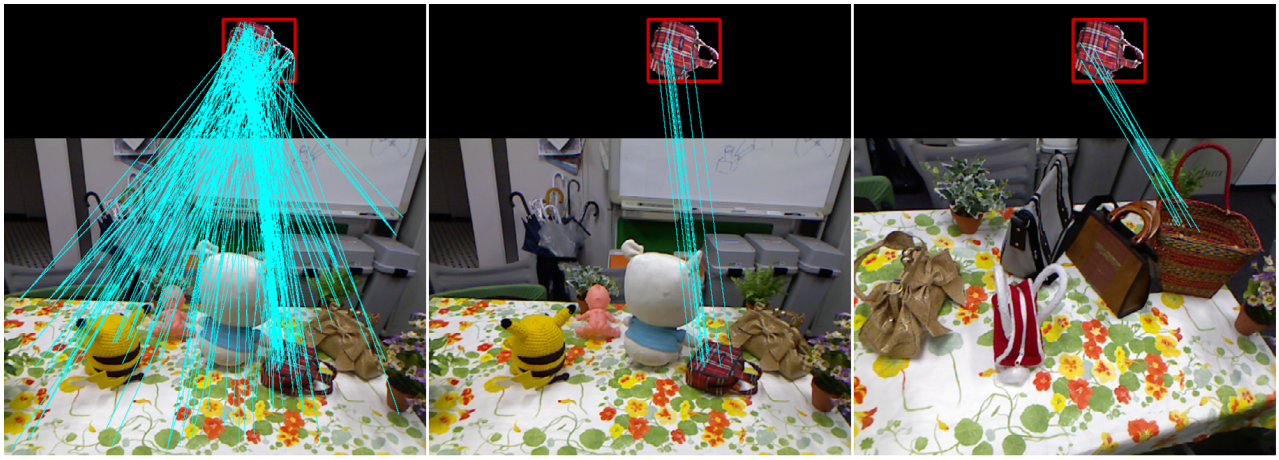

図 1 SIFT 特徴点と記述子で得られた物体参照画像との対応点集合の例．(a) 全探索により得られた初期対応点集合．(b) 誤対応点除去を行った正解の対応点集合．(c) 誤対応点除去を行った不正解の対応点集合．

というアプローチをとる．そして本研究は，学習データにおける性能を評価する損失関数を設計し，これを最小化する問題における閉じた解を求めた．

2. 手法

提案手法の概念図を **Fig.2** に示す．提案手法は，対応点集合類似度が，正解の対応点集合は高く，不正解の対応点集合は低くなるよう，二組の対応点ペアに対する重みを学習する．ここで二組の対応点は色の組み合わせによって区別する．具体的には，まず Hue 値を k 個のビンに分割し，参照画像中の各点の Hue 値がどのビンに所属するかを示すベクトルを $\boldsymbol{h} \in \mathrm{R}^k$ とする．所属するビンの値は 1，それ以外のビンの値は 0 とする．そして，参照画像中の点 i のベクトル $\boldsymbol{h}_i$ と点 j のベクトル $\boldsymbol{h}_j$ について行列 $H \equiv \boldsymbol{h}_i \boldsymbol{h}_j^{\mathrm{T}}$ を求め，さらに $H'_{mn} = H_{mn} + H_{nm} (m \neq n)$，$H'_{nn} = H_{nn}$ となる行列 H' を計算する．そして H' の上三角成分（対角成分を含む）を並べたベクトルを $\boldsymbol{q}_{ij} \in \mathrm{R}^{k(k+1)/2}$ とする．本研究の目的は，このベクトル $\boldsymbol{q}_{ij}$ に対する重み $\boldsymbol{w} \in \mathrm{R}^{k(k+1)/2}$ を物体毎に最適化することである．

提案する対応点集合類似度 $g(\boldsymbol{w})$ を下式で定義する．

$$\frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(1 - \exp\left(\frac{-\alpha \cdot \boldsymbol{w} \cdot \boldsymbol{q}_{ij}}{|\|p_i - p_j\| - \|p_{i'} - p_{j'}\|| + \epsilon} \right) \right) \tag{4}$$

ここで，α は指数関数の出力が 0 か 1 に張り付かないよう調整する定数であり，ϵ は指数関数の中身の分母が 0 になることを防ぐための小さな値の定数である．本研究では $\alpha = 10^{-3}$，$\epsilon = 10^{-20}$ とした．

提案手法は学習データセットに対して，正解の対応点集合の $g(\boldsymbol{w})$ が大きく，不正解の対応点集合の $g(\boldsymbol{w})$ が小さくなるような $\boldsymbol{w}$ を決定する．このために，SVM 等の識別器学習と同様に学習サンプルを適切に分離する超平面を考え，そのマージン内に学習サンプルが入ってしまう損失の量であるヒンジロスを最小化させることを考える．このためにはスコア関数の値域が $-\infty$ から $+\infty$ までの値を取る必要がある．そこで，本研究では対応点集合の正解らしさを評価するスコア関数を下式で定義する．

$$\begin{aligned} f(\boldsymbol{w}, b) &\equiv \mathrm{logit}\left(g(\boldsymbol{w})\right) + b \\ &= \log(g(\boldsymbol{w})) - \log(1 - g(\boldsymbol{w})) + b \end{aligned} \tag{5}$$

ただし b はオフセットであり，$\boldsymbol{w}$ と共に学習する変数である．ここで正解対応点集合のラベルを $y = 1$，不正解対応点集合のラベルを $y = -1$ とすると，ヒンジロス関数は下記のとおりとなる．

$$l(\boldsymbol{w}, b; (f, y)) = \begin{cases} 0 & yf(\boldsymbol{w}, b) \geq 1 \\ 1 - yf(\boldsymbol{w}, b) & otherwise \end{cases} \tag{6}$$

提案手法は $\boldsymbol{w}$ の初期値を零ベクトルに近いベクトル $\boldsymbol{w}_0 = (\epsilon', \ldots, \epsilon'), \epsilon' \sim 0$，$b$ の初期値を $b_0 = 0$ とし，対応点集合を一つ観測する度にこれらを更新する逐次学習を行う．t 回めの更新後のパラメータを $\boldsymbol{w}_t$，b_t とすると，$\boldsymbol{w}_{t+1}$，b_{t+1} は下式で計算される．

$$\{\boldsymbol{w}_{t+1}, b_{t+1}\} = \arg\min_{\boldsymbol{w}, b} \frac{1}{2} \left(\|\boldsymbol{w} - \boldsymbol{w}_t\|^2 + \|b - b_t\|^2 \right) \tag{7}$$

$$\text{s.t. } l(\boldsymbol{w}, b; (f_t, y_t)) = 0 \tag{8}$$

本式は閉形式である．解法を下記に示す．まず，$y_t f_t(\boldsymbol{w}_t, b_t) \geq 1$ であれば，$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t, b_{t+1} = b_t$ となる．このため，以下，$y_t f_t(\boldsymbol{w}_t, b_t) < 1$ の場合のみを考える．ラグランジュの未定乗数法を用いて

$$\begin{aligned} L(\boldsymbol{w}_t, b_t, \lambda) = &\frac{1}{2} \|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 + \frac{1}{2} \|b_{t+1} - b_t\|^2 \\ &+ \lambda (1 - y_t f_t(\boldsymbol{w}_t, b_t)) \end{aligned} \tag{9}$$

$\frac{\partial L}{\partial \boldsymbol{w}_t} = 0$, $\frac{\partial L}{\partial b_t} = 0$ より，下式が得られる．

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \lambda y_t \frac{\partial f_t(\boldsymbol{w}_t, b_t)}{\partial \boldsymbol{w}_t} \tag{10}$$

$$b_{t+1} = b_t - \lambda y_t \tag{11}$$

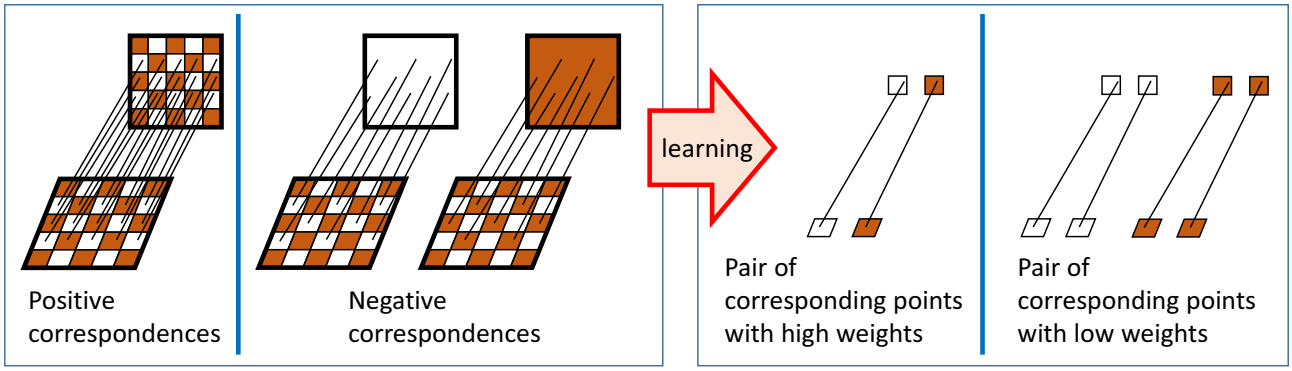

図 2 提案手法の概念図．学習データとして正解の対応点集合と不正解の対応点集合を用意する．提案手法は，対応点集合類似度が，正解の対応点集合は高く，不正解の対応点集合は低くなるよう，二組の対応点ペアに対する重みを学習する．二組の対応点は色の組み合わせによって区別する．この例では，白-茶，白-白，茶-茶の 3 種類となる．

ここでは $\frac{\partial f_t(\boldsymbol{w}_t,b_t)}{\partial b_t}=1$ であることに注意されたい．上式を (9) に代入し，下式が得られる．

$$L(\lambda)=\frac{1}{2}\lambda^2\left\|\frac{\partial f_t(\boldsymbol{w}_t,b_t)}{\partial \boldsymbol{w}_t}\right\|^2+\frac{1}{2}\lambda^2+\lambda(1-y_t f_t(\boldsymbol{w}_t,b_t)) \tag{12}$$

ここで $\frac{\partial L}{\partial \boldsymbol{\lambda}}=0$ より，下式が得られる．

$$\lambda=\frac{y_t f_t(\boldsymbol{w}_t,b_t)-1}{\left\|\frac{\partial f_t(\boldsymbol{w}_t,b_t)}{\partial w_t}\right\|^2+1} \tag{13}$$

したがって，$\boldsymbol{w}_{t+1}$ と b_{t+1} の更新式は下式となる．

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_t+\frac{(y_t-f_t(\boldsymbol{w}_t,b_t))}{\left\|\frac{\partial f_t(\boldsymbol{w}_t,b_t)}{\partial \boldsymbol{w}_t}\right\|^2+1}\cdot\frac{\partial f_t(\boldsymbol{w}_t,b_t)}{\partial \boldsymbol{w}_t} \tag{14}$$

$$b_{t+1}=b_t+\frac{(y_t-f_t(\boldsymbol{w}_t,b_t))}{\left\|\frac{\partial f_t(\boldsymbol{w}_t,b_t)}{\partial \boldsymbol{w}_t}\right\|^2+1} \tag{15}$$

$f(\boldsymbol{w},b)$ の勾配は下式のとおり計算できる．

$$\begin{aligned}\frac{\partial f(\boldsymbol{w},b)}{\partial \boldsymbol{w}}&=\frac{\partial}{\partial \boldsymbol{w}}\mathrm{logit}(g(\boldsymbol{w}))\\&=\frac{\partial}{\partial \boldsymbol{w}}\left(\log(g(\boldsymbol{w}))-\log(1-g(\boldsymbol{w}))\right)\\&=\frac{1}{g(\boldsymbol{w})(1-g(\boldsymbol{w}))}\frac{\partial g(\boldsymbol{w})}{\partial \boldsymbol{w}}.\end{aligned} \tag{16}$$

紙面の都合上導出は省略するが，$|\|p_i-p_j\|-\|p_{i'}-p_{j'}\||$ を d_{ij} とおき，$\frac{\partial g(\boldsymbol{w})}{\partial \boldsymbol{w}}$ は下式のとおり計算できる．

$$\frac{\partial g(\boldsymbol{w})}{\partial \boldsymbol{w}}=\frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(\frac{\alpha\cdot\boldsymbol{q}_{ij}}{d_{ij}+\epsilon}\cdot\exp\left(\frac{-\alpha\cdot\boldsymbol{w}\cdot\boldsymbol{q}_{ij}}{d_{ij}+\epsilon}\right)\right) \tag{17}$$

以上の更新を全学習サンプルについて行う操作を 1 試行とし，T 試行行った後に学習を終了させる．(途中でヒンジロスが十分に小さくなった場合はそこで学習を打ちきっても良い．) 本研究では $T=100$ とした．

3. 実験

Kinect センサで撮影した RGB-D 画像を用いて 10 個の物体の検出性能を検証する．学習データセットとして，各対象物体につき一つの RGB-D 参照画像と，少し視点をずらして撮影した 9 枚の RGB-D 画像を用意し (**Fig.**3)，正解の対応点集合を求めた．また，対象物体が写っていない RGB-D 画像を各対象物体毎に 70 枚用意し，不正解の対応点集合を求めた．対応点集合は，参照画像からランダムに 2,000 点を選択し，各点について各学習用画像から RGB 値が最近傍となる 5 個の点を求め（計 10,000 対応点），[8] の手法で絞り込みを行うことで求めた．ここで，最後に得られた割当ベクトル $\boldsymbol{x}$ の二値化を行うが，その閾値は $\boldsymbol{x}$ の最大要素に 0.5 をかけた値とした．比較手法として，参照画像と各学習画像から SIFT 特徴点を求め，Brute-Force 法でマッチングを行った後に [8] の手法で絞り込みを行うものも試す．なお，Hue 値の分割数を $k=3$ とした．($k=2,4,5$ も試したが，結果は大きく変わらなかった．)

テスト画像 120 枚に対する Precision-Recall カーブを **Fig.**4 に，Average Precision を **Fig.**5 に示す．青線は SIFT 特徴点と式 (3)（ベースライン）の対応点集合類似度を用いた場合，緑線は RGB 最近傍探索と式 (3)（ベースライン）の対応点集合類似度を用いた場合，そして赤線は RGB 最近傍探索と式 (4)（提案手法）の対応点集合類似度を用いた場合である．まず，テクスチャが多く平面的である対象物体 No. 1 と No. 5 を除いて，SIFT 特徴点を用いた場合の性能が低いことが分かる．そして，RGB 最近傍探索を用いた場合，対象物体 No. 4 を除いて，全ての物体について提案手法の性能がベースラインよりも高いことが分かる．全対象物体の Average Precision の平均値は，SIFT 特徴点を用いた場合が 0.13，RGB 最近傍探索を用いた場合のベースラインが 0.27，そして提案手法が 0.31 であった．

4. 結論

本研究では，RGB-D 画像を入力として検出対象物体の RGB-D 参照画像との対応点を求め，対応点ペア類似度行列の二次割当問題を解いて得られた対応点集合に対して，その正当性を評価する対応点集合類似度を導入し，パラメータを物体毎に学習する手法を提案し

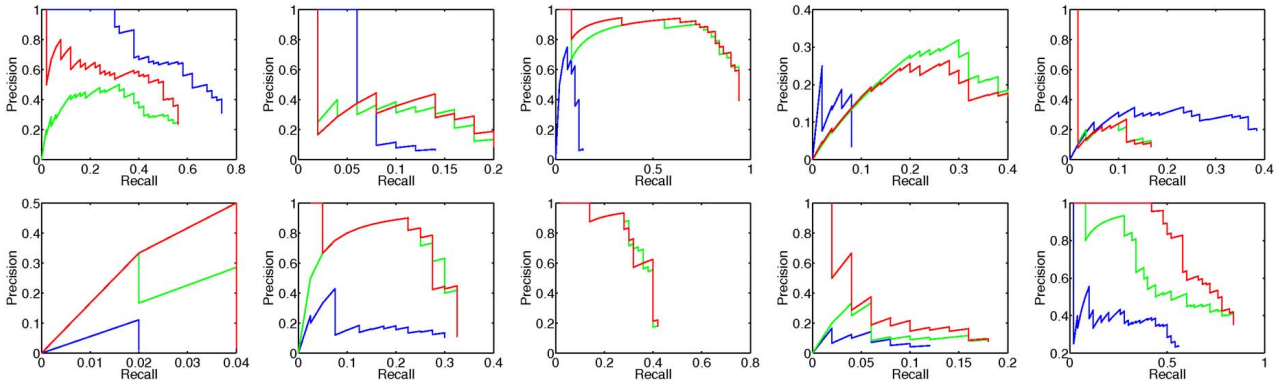

図 4 Precision-Recall カーブ．左上から順に，対象物体 No. 1，...，No. 10 である．青線は SIFT 特徴点と式 (3)（ベースライン）の対応点集合類似度を用いた場合，緑線は RGB 最近傍探索と式 (3)（ベースライン）の対応点集合類似度を用いた場合，そして赤線は RGB 最近傍探索と式 (4)（提案手法）の対応点集合類似度を用いた場合である．

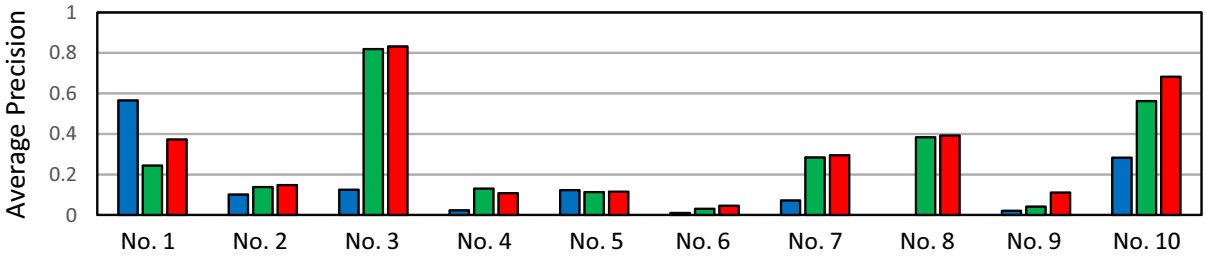

図 5 Average Precision．青線は SIFT 特徴点と式 (3)（ベースライン）の対応点集合類似度を用いた場合，緑線は RGB 最近傍探索と式 (3)（ベースライン）の対応点集合類似度を用いた場合，そして赤線は RGB 最近傍探索と式 (4)（提案手法）の対応点集合類似度を用いた場合である．

図 3 対象物体の参照画像と 9 枚の学習用 RGB-D 画像.

た．提案手法は，対応点集合類似度が，正解の対応点集合は高く，不正解の対応点集合は低くなるよう，二組の対応点ペアに対する重みを学習する．実験では，全ての対応点ペアを均一に評価する場合（ベースライン）に対して，提案手法を用いた場合に約 15%の性能向上が確認できた．今後は，より大規模なデータセットを用いた評価を行っていく．

参考文献

[1] Tibério S. Caetano, Julian J. McAuley, Li Cheng, Quoc V. Le, and Alex J. Smola. Learning graph matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 31(6):1048–1058, 2009.

[2] Marius Leordeanu and Martial Hebert. A spectral technique for correspondence problems using pairwise constraints. In *Proc. IEEE ICCV*, 2005.

[3] Marius Leordeanu and Martial Hebert. Smoothing-based optimization. In *Proc. IEEE CVPR*, 2008.

[4] Marius Leordeanu, Rahul Sukthankar, and Martial Hebert. Unsupervised learning for graph matching. *Int. J. of Computer Vision*, 96(1):28–45, 2012.

[5] Marius Leordeanu, Andrei Zanfir, and Cristian Sminchisescu. Semi-supervised learning and optimization for hypergraph matching. In *Proc. IEEE ICCV*, 2011.

[6] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. IEEE ICCV*, 1999.

[7] Deepti Pachauri, Maxwell Collins, Vikas Singh, and Risi Kondor. Incorporating domain knowledge in matching problems via harmonic analysis. In *Proc. ICML*, 2012.

[8] Emanuele Rodolà, Andrea Albarelli, Filippo Bergamasco, and Andrea Torsello. A scale independent selection process for 3d object recognition in cluttered scenes. *Int. J. of Computer Vision - Special Issue on 3D Imaging, Processing and Modeling Techniques*, 19, 2012.