2N-07

複数視点からのCNN出力結果の統合による実時間セグメンテーション及び物体認識

中島 由勝† 斎藤英雄†
慶應義塾大学†

1. はじめに

一般物体認識はコンピュータビジョン，ロボット工学において主要な研究分野の一つであり，自動運転，インタラクション等への応用がなされている．従来手法の多くは，Hinterstoisser の手法[1]に代表されるように特徴点を用いたフレーム毎の認識手法であり，各物体を多方向から認識しないため，一般に遮蔽物や対象物体の見えに頑健でない．Pillai らは SLAM に基づき各物体及びそれらの物体に対するカメラ位置姿勢を追跡することで複数視点からの物体認識を行ったが[2]，その手法は毎フレームの認識結果を単純に加算し最終的な認識結果とするため，カメラが対象物体の見えが悪い位置姿勢で停滞した場合，その認識結果の精度は低下する．そこで本研究ではシーン中の各物体周りの視点を均等に分割し，分割された各視点からの認識結果が同じ重みになるよう公平に統合することで，実時間で動作する高精度な一般物体認識手法を提案する．

2. 提案手法

シーン中の各物体を複数視点からカメラの動きに依存せず高精度に評価するためには各物体及びカメラ位置姿勢を毎フレーム追跡する必要及び，偏りのない複数視点からの認識結果を統合する必要がある．本章ではその詳細を述べる．

2.1 各物体及びカメラ位置姿勢の追跡

本手法における各物体及びカメラ位置姿勢の追跡には，Tateno らの RGB-D SLAM[3]を基にしたシステムを用いる．Tateno らの RGB-D SLAM は RGB-D 画像を入力として受け取り，頂点・法線の情報を用いて毎フレームセグメンテーションを行い，SLAM により生成される 3 次元マップを，計算量$O(n^2)$を保ちつつ更新する．本手法ではそのセグメンテーション結果を基に入力画像中の各物体周りを矩形状に切り抜き，構築した RGB 及び法線の 2 入力を受け取る CNN モデルに，次節で述べる条件を満たした場合に入力し，その

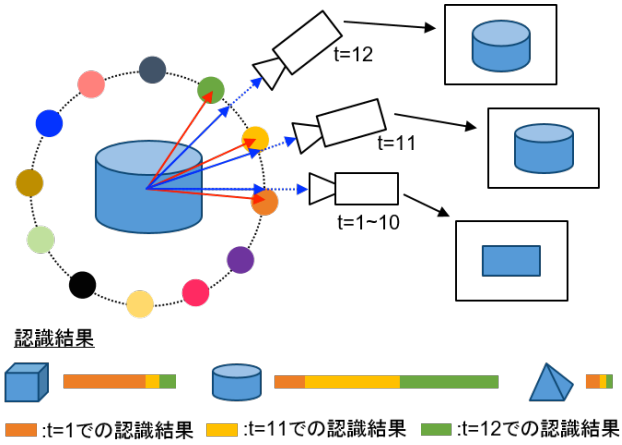


図 1 CNN による認識結果の統合の概念図

フレームでの認識結果を得る．

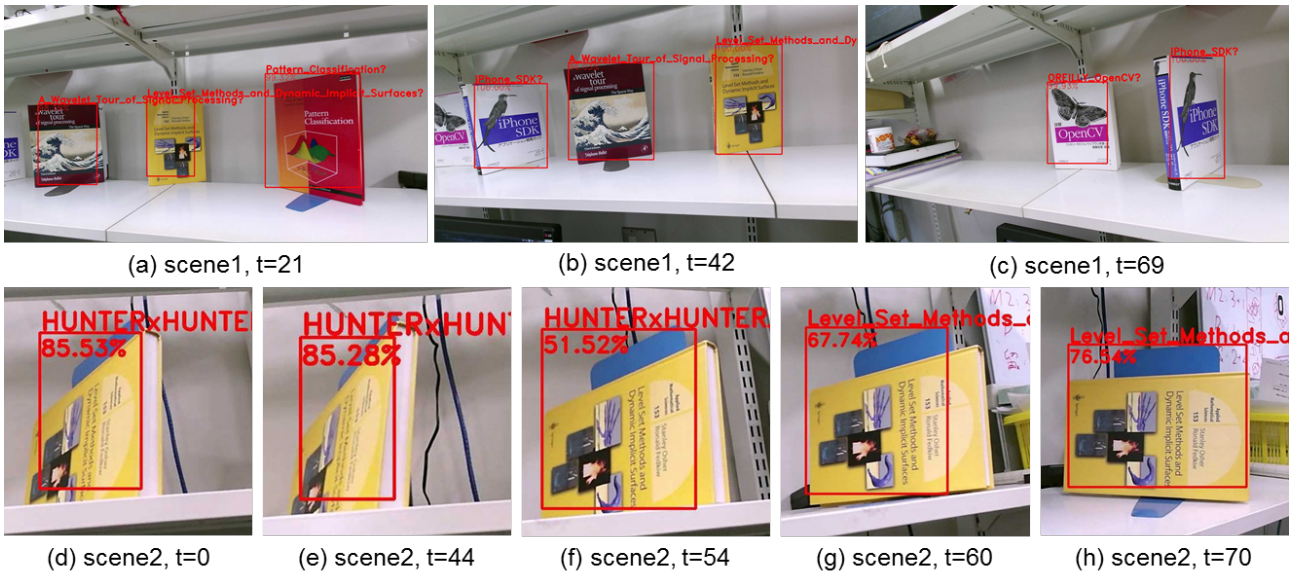
2.2 認識結果の統合

本節では，各フレームにおける CNN による認識結果を，SLAM により生成されるセグメンテーションが施された 3 次元マップ中の各物体の認識結果に統合する手法を述べる(図 1 参照)．本手法では Tateno らの SLAM を改良し，3 次元マップ中のセグメンテーションされた各物体の重心を毎フレーム計算量$O(n^2)$を保ちつつ更新する．それらの各重心を中心とし，Saff らの式[4]に基づき球上に等間隔に N 個の球を配置する．本手法では各物体周りに配置された各球を視点クラスと呼ぶ．ただしここで，図 1 では簡単のため 2 次元の円上に視点クラスを配置している．次に各フレームにおいて，各物体の重心から各球へのベクトル(図 1 赤ベクトル)と，重心からカメラへのベクトル(図 1 青ベクトル)を比較し，カメラ位置に対する近傍の球が変化したときのみ 2.1 で述べた CNN モデルへ入力する．最後に式(1)によりそのフレームまでの各物体の認識結果を統合し，物体 y がカテゴリ k である確率を得，確率が最大であるカテゴリ k をその物体の認識結果とする．

$$y_k = \frac{\exp\left(\sum_{i \in V} u_{(k,\ i)}\right)}{\sum_{j=1}^{K} \exp\left(\sum_{i \in V} u_{(j,\ i)}\right)} \quad (1)$$

ここで，Kはカテゴリ数であり，Vは認識を行った視点クラス，$u(s,\ t)$は視点クラスtにおけるカテゴリsの出力である．

Simultaneous Object Segmentation and Recognition by Merging CNN Outputs from Multiple Viewpoints
Yoshikatsu Nakajima† and Hideo Saito†
†Keio University



(a) scene1, t=21　(b) scene1, t=42　(c) scene1, t=69

(d) scene2, t=0　(e) scene2, t=44　(f) scene2, t=54　(g) scene2, t=60　(h) scene2, t=70

図 2 本手法による一般物体認識結果

3. 評価実験

本章では提案手法の有効性を評価するために行った実験について記す．本実験では認識対象を3次元形状が既知である本とし，33カテゴリの本の表紙画像に対し平面射影変換を用い様々な角度から見た画像群，法線マップ群をそれぞれ500組生成した．またこの際，照明及び背景の変化に頑健になるようノイズをランダムに発生させ， それらのデータセットを用いCNNの深層学習を行った．次に環境下に学習を行った本を適当に配置し，Kinect v2により取得したRGB−D画像を入力とし実験を行った．また，視点クラスの数Nは2000とした．図2上段に実験結果のフレームtにおける結果を示す．この結果を見ると，本手法による複数視点からの認識，また，RGB画像及び法線マップの2入力を持つCNNモデルにより，高精度に物体認識が行えたことがわかる．図2下段では対象物体に対する角度が浅く，更にその一部が遮蔽される位置にカメラが停滞した場合での実験結果である．この結果を見ると，本手法による視点クラスに基づく多視点からの均等な認識により，誤差の生じた前半部での認識結果が蓄積されることなく最終的に高精度な認識結果を得ることが出来たことがわかる．

また図3に物体認識に要した処理時間を示す．処理時間が上下した理由は2.2節で述べた，CNNに入力する条件を満たす物体の数がフレーム間で変動したためである．この図を見ると本手法によりフレームが進んでも計算量が増加することなく認識を行えたことがわかる．認識に要した処理時間の平均値は 69.4msec であり，十分に実時間で動作することを確認した．

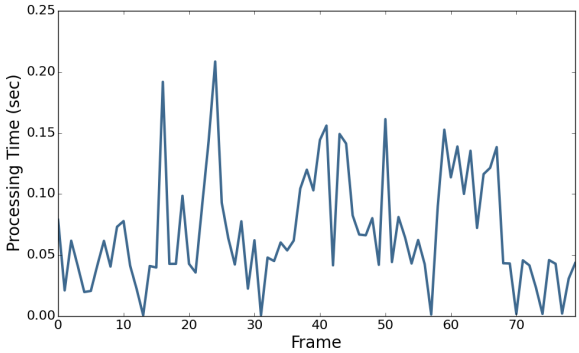


図 3 物体認識に要した処理時間

4. 結論

本稿ではシーン中の各物体周りの視点を均等に分割し，分割された各視点からの認識結果を公平に統合することによる物体認識手法の提案，評価を行った．実験を通しその有効性及び実時間で動作することを確認した．今後の課題としてより大規模なシーンでの実験等が挙げられる．

参考文献

[1] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lepetit, “Gradient Response Maps for Real-Time Detection of Textureless Objects,” TPAMI, vol. 34, no. 5, pp. 876–888, 2012.

[2] S. Pillai and J. Leonard. “Monocular slam supported object recognition” arXiv preprint arXiv:1506.01732, 2015.

[3] K. Tateno, F. Tombari, and N. Navab, “Real-time and scalable incremental segmentation on dense slam,” in Int. Conf. on Intelligent Robots and Systems (IROS), pp. 4465-4472, 2015.

[4] E. Saff and A. Kuijlaars, “Distributing many points on a sphere,” Math Intelligencer, vol.10, pp.5-11, 1997.