# 論文要旨


## Abstract

2D/3D object detection and pose estimation is one of the essential techniques of computer vision and is critical for various real applications such as factory automation (FA) and autonomous driving. The object detection and pose estimation is classified into broad two categories, one is general object (class) detection and pose estimation and another is specific object (instance) detection and pose estimation. The target of this thesis is specific object detection and pose estimation which is mainly used in FA applications such as visual inspections and robotic manipulations.

Three kinds of algorithms for specific object detection and pose estimation are required to cover various applications. The first is 2D object detection and pose estimation of planar objects on conveyors and tabletops. The pose of the object is constrained by planes and the algorithm estimates 4 parameters (X/Y translations, in-plane rotations and scales) from a monocular image. The second is 3D object detection and pose estimation from a monocular image. This estimates rough 3D position and pose (6 parameters – X/Y/Z translations and rotations) mainly for visualization in AR/MR applications where a small and fast monocular camera is preferred. The third is 3D object detection and pose estimation from 3D point clouds captured by a 3D (range) sensor. This estimates precise 3D object position and pose (6 parameters) mainly for robotic manipulations.

The algorithms for specific object (instance) detection and pose estimation are further categorized into three kinds of approaches, a global descriptor (template matching) based approach, a local descriptor based approach and a learning based approach. The global descriptor based approach can handle any kinds of objects and is robust against background clutters, but it is fragile to occlusions and transformations. The local descriptor based approach is robust against occlusions and transformations, but can only be applied to the objects where rich features (textures and shapes) are extracted. The learning based approach is superior to the former two approaches in performance, but this requires large training dataset for each object and scene. The target of this thesis is a research on the practical algorithms which can be applied to real FA applications. In these applications, many of the target objects are rigid, occluded objects are not inspected or grasped, and it is almost impossible for customers to collect large dataset for each object and scene. For these reasons, the global descriptor (template matching) based approach is employed in this thesis.

Proposed Algorithm 1: It has been shown that the template matching based on discretized gradient orientations could handle texture-less objects. Though the matching conditions based both on gradient positions and orientations are strict and robust against background clutters, the similarity scores decrease largely even when the appearance of target object is slightly changed. To tackle this problem, we propose COF (Cumulative Orientation Feature) which is robust to appearance changes induced by object pose changes and at the same time is enough discriminative to detect target objects against cluttered backgrounds. At first, many images are generated based on 2D geometric transformations of a model image using randomized parameters for X/Y translations, rotation angles and scales. Then orientation histograms are calculated at each pixel and pixel-wise dominant orientations are extracted as features. Our proposed method was evaluated on publicly available dataset and achieved higher detection rate and faster speed compared to state of the art.

Proposed Algorithm 2: The 3D object detection and pose estimation based on the template based approach tends to be slower when the number of templates amounts to tens of thousands for handling a wider range of 3D object pose.

To alleviate this problem, we propose a novel image feature and a tree-structured model. Our proposed perspectively COF (PCOF) is developed from COF and extracted from randomly generated 2D projection images from a 3D CAD, and the template based on PCOF explicitly handle a certain range of 3D object pose. The hierarchical pose trees (HPT) is built by clustering 3D object pose and reducing the resolutions of templates, and HPT accelerates 6D pose estimation based on a coarse-to-fine strategy with an image pyramid. In the experimental evaluation on our texture-less object dataset, the combination of PCOF and HPT showed higher accuracy and faster speed in comparison with state-of-the-art techniques.

Proposed Algorithm 3: We propose PCOF-MOD (multimodal PCOF), balanced pose tree (BPT) and optimum memory rearrangement for a coarse-to-fine search in order to make the template based 3D object detection and pose estimation from a RGB-D image faster. Firstly, PCOF-MOD is developed from PCOF by adding the discretized orientations of surface normals. As with PCOF, the model templates of PCOF-MOD explicitly handle a certain range of 3D object pose and the fewer number of templates can cover wider range of 3D object pose. Secondly, a large number of templates are organized into a coarse-to-fine 3D pose tree (BPT) in order to accelerate 6D pose estimation. Predefined polyhedra for viewpoint sampling are prepared for each level of an image pyramid and 3D object pose trees are built so that the number of child nodes of every parent node are almost equal in each pyramid level. Lastly, two kinds of binary features at the lower pyramid levels are rearranged so that nearby features are linearly aligned on a memory and these vectorized features are processed at one time using SIMD instructions. In the experimental evaluation of 6D object pose estimation on publicly available tabletop and our own bin picking dataset, our template based method showed higher accuracy and faster speed in comparison with the existing techniques including recent CNN based methods.

**日本語要旨**

画像や３次元点群から物体の２次元あるいは３次元の位置と姿勢を認識する技術は、工場自動化や自動運転など様々なアプリケーションで必要とされる画像認識の基本技術の一つである。物体位置姿勢認識技術は大きく二つに分類することができ、一つは顔や人体など物体クラスを対象とする一般物体位置姿勢認識、もう一つは特定の物体（インスタンス）を対象とする特定物体位置姿勢認識である。本論文では工場での外観検査やロボットによる把持・組立に用いられることが多い特定物体位置姿勢認識を対象とする。

特定物体位置姿勢認識はアプリケーションによって三種類のアルゴリズムが必要であると考えられる。一つ目はコンベアや机の上に置かれた平面的な物体の位置姿勢を認識するアルゴリズムである。この場合の物体姿勢変化は平面上に限定されるため、単眼カメラのみを用いて並進（XY 成分）、回転、スケールの４つのパラメータを推定する。二つ目は AR/MR など３次元表示を目的とした物体の概略３次元位置姿勢を認識するアルゴリズムである。この場合は処理速度や可搬性の観点から単眼カメラを用い、並進（XYZ 成分）、回転（XYZ 成分）の６つのパラメータを推定する。三つ目はロボットによる把持・組立等を目的としたより高精細な３次元位置姿勢を認識するアルゴリズムである。高精度認識のため距離センサにより計測した３次元点群を入力として用い、３次元位置姿勢の６つのパラメータを推定する。本論文ではこれら三つのアルゴリズムに関し、テクスチャ無しや単純形状を含むあらゆる物体に適用可能で高速かつ外乱に対してロバストな手法を提案する。

特定物体位置姿勢認識アルゴリズムは、大きく三つの手法に分類することができる。一つ目はテンプレートマッチングに基づく手法、二つ目は局所特徴量に基づく手法、三つ目は機械学習に基づく手法である。テンプレートマッチングに基づく手法は、あらゆる物体に適用可能で外乱に対してロバストであるが

変形や隠れに弱い。局所特徴量に基づく手法は、隠れや物体の変形に対してロバストであるがテクスチャや形状などの特徴量が多く抽出できる物体にしか適用できない。機械学習に基づく手法は性能面では前者二つの手法と比較して優位に立っているものの、対象となる物体や背景について多くの学習データを収集する必要がある。本論文では工場自動化やロボットビジョンといった実アプリケーションに適した特定物体位置姿勢認識アルゴリズムの研究を目的としている。こういったアプリケーションにおいては対象とする物体は幅広いがその多くは剛体である、隠れている物体は検査や把持の対象とならない、物体や環境ごとに大量の学習データを収集することは現実的でないといった理由から、本論文ではテンプレートマッチングに基づく手法を採用した。

提案手法１：テクスチャの少ない物体にも適用可能な２次元物体位置姿勢認識手法として、輝度勾配方向特徴量を用いたテンプレートマッチングが提案されてきた。しかし勾配方向を照合条件として用いることで複雑背景下においても頑健な照合が可能である一方、対象物体自身の見えがわずかに変化した場合には照合スコアが大きく低下してしまうという課題があった。そこで本論文では、物体の姿勢変動による見えの変化を考慮した累積勾配方向特徴量（COF: Cumulative Orientation Feature）を提案する。提案手法ではまず、一定範囲内でランダムに発生させた平行移動、回転角度、スケールパラメータを用い、１枚のモデル画像に対して幾何学的変換を適用して多数の画像を生成する。次に各画像において算出した量子化勾配方向特徴量を用いて画素毎に勾配方向ヒストグラムを作成し、頻度の大きい勾配方向のみを用いて特徴量を抽出した。実際の画像に対して照合処理行い、提案手法が対象物体と背景を識別する性能を維持したまま物体自身の見えの変動を許容できることを確認した。またテクスチャレス物体の公開画像データセットを用いた２次元物体位置姿勢認識の実験を行い、提案手法が認識正確性及び処理速度において既存手法を上回ることを示した。

提案手法２：単眼カメラ画像から３次元物体位置姿勢を高速に認識する手法においては、認識対象となる３次元姿勢範囲が広い場合に照合に用いるテンプレートの数が膨大になり処理速度が低下するという課題があった。この課題に対して本論文では、透視投影に基づく累積勾配方向特徴量（PCOF: Perspectively COF）と階層的姿勢探索木（HPT: Hierarchical Pose Tree）の二つの手法を提案する。PCOF は COF を拡張した特徴量であり、対象物体の３次元 CAD を様々な視点から見た２次元投影画像を生成して特徴抽出を行う。このことにより、３次元姿勢変化による対象物体の見えの変化に対する許容性と複雑背景に対する頑健性の両立を実現した。HPT は様々な視点において作成された大量のテンプレートに対し、類似度に基づいたクラスタリングとテンプレートの低解像度化を繰り返すことで作成する。HPT を用いて画像ピラミッド上を探索することにより、数万個の３次元姿勢候補の中から最も類似度の高いテンプレートを高速に絞り込むことが可能になる。９種類の金属部品を様々な方向から撮影したデータセットを用いて評価実験を行い、PCOF と HPT を組み合わせた提案手法が３次元物体位置姿勢認識の高速性・正確性両面において既存手法を上回ることを確認した。

提案手法３：距離画像や RGB-D 画像から３次元物体位置姿勢認識を行う場合においても、単眼カメラからの認識と同様に照合に用いるテンプレートの数が多く処理速度が遅くなるという課題があった。この課題に対して本論文では、透視投影に基づく RGB-D 累積勾配方向特徴量（PCOF-MOD: Multimodal PCOF）、平衡姿勢探索木（BPT: Balanced Pose Tree）、特徴量再配置による粗密探索高速化の三つの要素技術からなる高速・ロバストな３次元物体位置姿勢認識手法を提案する。一つ目の PCOF-MOD は PCOF にデプス画像特徴量を加えた特徴量であり、一定範囲内においてランダムに設定した３次元視点位置から対象物体の３次元 CAD を見た場合のデプス画像を多数生成し、それらからデプス勾配方向と表面法線方向について画素ごとに方向ヒストグラムを作成し特徴抽出を行う。これにより、PCOF-MOD は視点を設定

した範囲内の 3 次元姿勢変化による見えの変化のみを照合時に許容可能な特徴量となる。二つ目の要素技術である BPT は、画像ピラミッドの階層ごとに頂点数の異なる多面体の頂点を視点位置として使用することで、画像内 2 次元位置の粗密探索と 3 次元姿勢の粗密探索を同時に実施可能とした探索木である。全ての探索木の深さは等しく、親ノードに連結する子ノードの数もほぼ均一であるため探索効率が高いという特徴を備えている。三つ目の特徴量再配置は、画像ピラミッドの最上位階層以外では一つ上の階層で検出された正解候補周辺の画素においてのみ特徴量照合を行うという粗密探索の特性を活用している。即ち、照合対象となる周辺画素の特徴量がメモリ上で連続するように再配置した特徴量マップを作成し、連続データに適用可能な CPU 命令（SIMD 命令）を用いて一括照合を行うことで粗密探索の高速化を実現する。これら三つの要素技術を組み合わせた 3 次元物体位置姿勢認識手法について公開 RGB-D データセットと我々が構築したバラ積み部品データセットにおいて性能評価を行い、提案手法が 3 次元物体位置姿勢認識の高速性・正確性両面において近年の CNN ベース手法を含む既存手法を上回ることを示した。