3次元物体座標を用いた6次元物体姿勢推定の学習

Eric Brachmann[1] , Alexander Krull[1] , Frank Michel[1] , Stefan Gumhold[1] , Jamie Shotton[2] , and Carsten Rother[1]

[1]ドレスデン工科大学（ドイツ・ドレスデン市
[2]マイクロソフトリサーチ、ケンブリッジ、イギリス

概要本研究では、1枚のRGB-D画像から特定のオブジェクトの6D Poseを推定する問題に取り組む。我々は、テクスチャのあるなしを問わず、一般的なオブジェクトを扱うことができる柔軟なアプローチを提示する。主な新コンセプトは、高密度のクラスラベルと対になった高密度の3次元物体座標ラベリングという形で学習された中間表現である。我々は、テクスチャのないオブジェクトを含む一般的なデータセットにおいて、テンプレートベースの手法が適切であり、かつ最新技術であることを示すことができる。また、テンプレートベースの手法と比較して、様々な照明条件に対するロバスト性という点でも、我々のアプローチの利点を示すことができる。この目的のために、我々は3つの異なる照明条件下でそれぞれ撮影された20のオブジェクトの1万枚の画像からなる新しいグランドトゥルースデータセットを提供する。本アプローチは、オブジェクトの数に応じてうまくスケールし、高速に実行できることを実証する。

1 はじめに

オブジェクトのインスタンス検出と姿勢推定のタスクは、コンピュータビジョンにおいてよく研究されている問題である。この研究では、入力が1つのRGB-D画像である特定のシナリオを考えます。深度チャンネルを追加することで、シーンに存在する剛体オブジェクトインスタンスの完全な6次元姿勢（3次元回転と3次元移動）を抽出することが可能となる。最終的な目標は、高速でスケーラブル、ロバストで高精度なシステムを設計することであり、乱雑な環境や変化する照明条件などの厳しい実世界設定に存在する一般物体（テクスチャあり・なし両方）に対してうまく機能することである。

長年にわたり、剛体の検出と2D/6D姿勢推定の分野では、十分なテクスチャを持つ物体に限定されてきた。11, 15]の先駆的な研究に基づいて、多数のオブジェクトインスタンスに対応する実用的で堅牢なソリューションが設計されています[18, 20]。テクスチャオブジェクトの場合、ほとんどのシステムで成功の鍵は、SIFT特徴などの手作業、またはデータから学習された局所特徴の疎な表現を使用することである。これらのシステムは、通常、2段階のパイプライン

を実行します：a) 仮の疎な特徴のマッチング、b) マッチした特徴の幾何学的検証。

近年、テクスチャのない、あるいはテクスチャの乏しい剛体 のオブジェクトインスタンス検出のタスクが検討され始めている [7, 8, 21]・この特別な課題に対して、テンプレートベースの技術が優れていることが示されている。

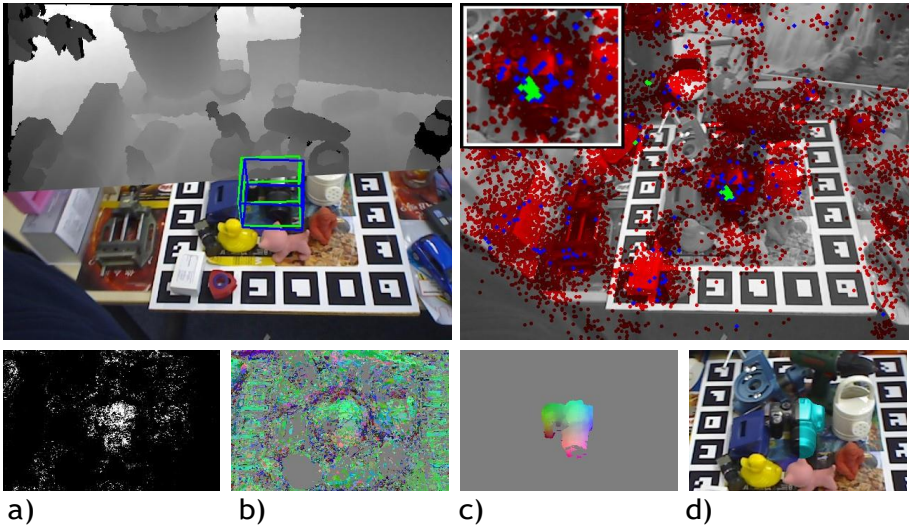

図1.本システムの概要。左上RGB-Dテスト画像（上半分が深度画像、下半分がRGB画像）。推定された6次元姿勢の物体（カメラ）を青のバウンディングボックスで、それぞれのグランドトゥルースを緑のバウンディングボックスで示す。右上。最適なポーズを探索するアルゴリズムの視覚化（入口は中心部のズーム）。このアルゴリズムは、6次元の連続した大きなポーズ空間に対して、RANSACのような方法でエネルギーを最適化します。*赤*のポーズは非常に高速なジオメトリチェックで無視され、*青*のポーズは中間高速サンプリングで我々のエネルギー関数を使用して評価され、*緑*のポーズは最もコストのかかるエネルギー精密化ステップにかけられます。下段、左から右へ：（a）クエリーオブジェクトの確率マップ、（b）RGBキューブにマッピングされた一本の木から予測される3Dオブジェクト座標、（c）対応するグランドトゥルース3Dオブジェクト座標、（d）テスト画像に*青で表示された3Dモデル*のオーバーレイ（推定ポーズに従ってレンダリングされたもの）。

これらの研究の主な焦点は、特定のエンコーディング[8]や、さらにカスケード型フレームワーク[21]を用いることで、テンプレートベースの技術を非常に高速化できることを示すことであった。テンプレートベース技術の典型的な問題点である、散乱物やオクルージョン、光条件の変化に対してロバストでないことは、慎重に手作業で作成されたテンプレートと追加の識別学習によって部分的に克服されている。しかしながら、テンプレートベースには2つの基本的な欠点があると我々は考えている。まず、テンプレートベースの手法は、テンプレートと対象画像のマッチング、つまり、ある特定のポーズの対象物を1つの「グローバル」な特徴でエンコードします。これとは対照的に、テクスチャオブジェクトのための疎な特徴に基づく表現は「局所的」であり、したがってこのシステムはオクルージョンに対してよりロバストである。第二に、テンプレートベースの技術は、必要なテンプレートの数が増えているため、オブジェクトクラスだけでなく、多関節または変形可能なオブジェクトインスタ

ンスに対して機能させることは、未解決の課題である。

我々のアプローチは、事前にセグメント化されたRGB-D画像からの多関節人体姿勢推定の分野における最近の研究 [28]に動機づけられている。28]の基本的な考え方は、画像から直接60自由度の人間のポーズを予測するのではなく、まず中間的ないわゆるオブジェクト座標表現を回帰させることである。これは、画像中の各画素が、Vitruvian Manifoldと呼ばれる標準的なポーズの標準的な身体上の連続座標に投票することを意味する。投票はランダムフォレストによって行われ、単純で局所的な特徴テストの訓練された集合を使用する。次のステップでは、「幾何学的検証」が行われ、これらの対応関係をパラメトリックな身体モデルと比較するエネルギー関数が定義される。最後に、エネルギー最小化により、ポーズパラメータを求める。したがって、これは、従来の疎な特徴に基づく手法の2段階のパイプラインに似ていますが、現在は、高密度に学習された特徴を用いています。24]では、同様のアイデアを6Dカメラポーズ推定に適用し、回帰の森が画像と世界の対応を正確に予測し、それをカメラポーズ推定に使用することを示しました。彼らは、疎な特徴に基づくベースラインよりもかなり正確な結果を示しました。

本システムは、これらの[28, 24]の考え方をベースに、特定の物体の6次元姿勢を推定するタスクに適用したものである。図1に本システムの概要を示す。ただし、物体分割マスクが追加で必要なため、[28, 24]をそのまま適用することはできない。しかし、[28]の手法はあらかじめセグメンテーションされた人物の形状に依存することができ、[24]はセグメンテーションを必要としないことに注意してください。このため、我々は、高密度の3Dオブジェクト座標ラベリングと高密度のクラスラベリングを共同で予測する。24]とのもう一つの大きな違いは、RANSACに基づく最適化において、不必要な偽仮説の生成を避けるための巧妙なサンプリングスキームである。

要約すると、我々の研究の**主な**貢献は、局所特徴に基づく物体検出技術の利点を持ちながら、テクスチャのない物体検出のためのテンプレートベースの技術よりも精度の面でわずかに優れた結果を達成する新しいアプローチである。これにより、多くの概念的・実用的な利点が得られます。まず、テクスチャ付きオブジェクトとテクスチャなしオブジェクトのために別々のシステムを学習する必要がない。第二に、ノートパソコンやハサミなどの剛体・非剛体、蓋のある鍋・ない鍋などの異なる状態の物体に対して、同じシステムを利用することができる。第三に、局所的な特徴を用いることで、オクルージョンに対する頑健性を獲得することができる。第四に、厳密な特徴量学習の枠組みを適用することで、照明条件の変化に対するロバスト性を獲得している。図2に本システムの利点を示す。本研究の主な技術的貢献は、高密度3次元物体座標と物体クラスラベルの結合という新しい表現を用いたことである。さらに、3つの異なる照明条件下で撮影され、正確な6次元姿勢でラベル付けされた20個の物体からなる1万画像の新しいデータセットが貢献し、これは一般に公開される予定である。

2 関連作品

姿勢推定や物体検出の分野では、インスタンスやカテゴリ認識、剛体や多関節物体など、膨大な文献が存在し

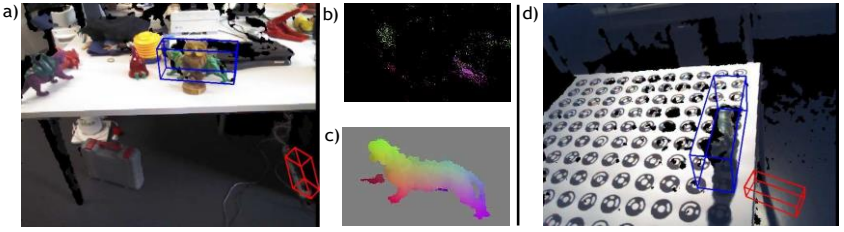

図2. テンプレートに基づく手法では失敗するような場合でも、本手法では正しい姿勢を見つけることができる。(a) 強いオクルージョンがある場合のテスト画像。本手法による推定姿勢は青色で示されている。b）1つの木から得られたaの座標予測値をRGBキューブに写像し、$pc_{,i}$を乗じたもの。 c）RGBキューブに写像されたaの真実の物体座標。(d) トレーニングセットとは異なる極端な光条件下でのテスト画像。推定された姿勢はa,b,cと同様に表示される。

粗い（量子化された）姿勢と正確な（6D）姿勢。以下の簡単なレビューでは、特に散乱したシーンにおける剛体のインスタンスの検出と、同時にその6Dポーズを推測する技術に焦点を当てます。そのうちのいくつかは、すでに上記で述べたとおりです。

テンプレートベースのアプローチ。 おそらく最も伝統的な対象物検出のアプローチは、例えば[11, 26, 7, 8]のようなテンプレートを使用することです。これは、硬いテンプレートが画像上を走査され、距離測定が計算され、最適なマッチングを見つけることを意味します。テンプレートベースのアプローチの最新技術として、[8]は3Dオブジェクトモデルの同期レンダリングを用いて、全視野半球をカバーする多数のテンプレートを生成しています。彼らは、テクスチャのないオブジェクトに対してうまく機能するエッジベースの距離メトリックを採用し、正確な6Dポーズを達成するためにICPを使用してポーズ推定を洗練します。このようなテンプレートベースのアプローチは、実際に正確かつ迅速に作業することができます。テンプレートベースのアプローチの限界については前述しました。

疎な特徴に基づくアプローチ テンプレートに代わるものとして、疎な特徴に基づくアプローチがよく知られている。これは，画像から注目点（多くの場合，スケール不変）を抽出し，それを局所記述子（多くの場合，アフィンおよび照明不変）で記述し，データベースと照合するものです・例えば，Lowe [15] は，SIFT 記述子を用いて，類似した視点からの画像をクラスタリングし，1つのモデルにしています・また，最近の高速でスケーラブルなシステムの素晴らしい例として， [16] があります・スパース技術は、膨大な語彙のマッチングにうまく対応できることが示されています[18, 20]。さらに最近の傾向として、注目点[22, 10]、記述子[30]、マッチング[14, 19, 1]を*学習するようになってきています*。しかし，このような疎なアプローチは，十分なテクスチャを持つオブジェクトを必要とするため，実際のアプリケーションでは限界がある・本手法は、テクスチャに関係なく、すべての画像画素に高密度に適用でき、利用すべき最も適切な画像特徴を学習することが可能である。テクスチャのないオブジェクトを扱うことができる輪郭や形状のマッチングに関する文献も多く存在する（例えば[29]）が、これは我々の研究とは概念的に異なることに注意されたい。

密なアプローチテンプレートやスパースアプローチに代わるものとして、デンスアプローチがある。これらのアプローチでは，各ピクセルが望まれる出力について何らかの予測を行う・一般化ハフ投票法では，すべてのピクセルが量子化された予測空間（例えば，2Dオブジェクトの中心やスケール）において投票を行い，最も多くの票を得たセルが勝者とされる・27, 5]では，ハフ投票が物体検出に用いられ，粗い物体の姿勢を予測することができることが示された・我々の研究では，Gallら[5]のアイデアを採用し，Hough投票と物体分割の両方に対して目的語を共同で学習させる・しかし，[5]とは異なり，我々は出力（我々の場合は3次元物体座標と物体クラスラベル）に対する単純な結合分布が，[5]で提案された変形よりも良い性能を示すことを見いだした・Drostら[4]もまた、投票アプローチをとり、探索空間を縮小するために指向性のある点ペア特徴を使用します。完全な6Dポーズを得るために、*グローバルに*量子化された6Dポーズに対して、すべてのピクセルが直接投票する[5]の変種を想像することができます。しかし、探索空間の次元が高いため（したがって、高度な量子化が必要）、姿勢の推定がうまくいかない可能性があります。そこで、本アプローチでは、各画素が3次元モデルとの*局所的な*対応関係のみを3次元的に連続的に予測する。これにより、探索空間が大幅に縮小され、識別的な予測を学習するために、物体表面の各点をあらゆる角度から見る必要がなくなるため、学習セットを大幅に削減することができる。我々は、この3次元物体対応関係が、その後のモデルフィッティング段階を効率的に駆動し、高精度な6次元物体姿勢を達成する方法を示す。

最後に、我々の3Dオブジェクト座標表現と同様のアイデアを使用するオブジェクトクラス検出のためのアプローチがあります。最初のシステムの1つは、3D LayoutCRF [9]であり、決定森を使用して、3D剛体オブジェクトをカバーする、密な部分ラベルを予測するタスクを考慮します。その後、Vitruvian Manifoldが登場した。
[28]は人間の姿勢推定のために導入され、最近ではRGB-D画像におけるカメラの再局在化のためにシーン座標回帰森が導入された[24]。両作品については前述した通りである・

3 方法

まず、3Dオブジェクトの座標とオブジェクトのインスタンス確率の両方を予測する決定森を説明する。次に、Forestの出力に基づくエネルギー関数について説明する。最後に、RANSACに基づく最適化スキームについて述べる。

3.1 ランダムフォレスト

RGB-D 画像のピクセルを分類するために、1つの決定森を使用します。決定森は、決定木 T_j の集合 T です。画像のピクセルは、それぞれの木 T_j によって分類され、木の葉 l_j の1つに入ります。この森は、ピクセル i がどのオブジェクト c_i に属するかもしれないという情報を得ることができる方法で訓練されています。
と、そのオブジェクト上の位置がどうなっているのか。ここでは、ピクセル

の位置を $\mathbf{yi}$ で表し、それをピクセル・オブジェクト*座標*と呼びます。各リーフ l_i には、可能性のあるオブジェクト所属に対する分布 $p(c|l_i)$と、可能性のあるオブジェクト所属 c に対するオブジェクト座標 $\mathbf{y}_c(l_i)$のセットが格納されます。$\mathbf{y}_c(l_i)$は、*座標予測*と呼ばれます。以下では、座標予測についてだけ説明する。
また，この問題に特化した興味深い設計上の決定を行ったので，詳細な説明は補足資料を参照してください・

森の設計と訓練 我々は、標準的なランダム化された訓練手順を用いて決定木を 構築した[2]。連続分布 $p(\mathbf{y}_i|c_i)$ を $5 \times 5 \times 5 = 125$ 個の離散ビンに量子化した。我々は背景クラスに対して追加のビンを使用する。この量子化により、標準的な情報量である
これは、回帰目的よりも、多モデル分布 $p(\mathbf{y}_i|c_i)$に対処する能力が高い。我々の離散分布の両方に対応するノード分割目的としては $p(c_i|l_i)$ と $p(\mathbf{y}_i|c_i, l_i)$ は、共同分布に対する情報利得を使用します。
これは、125|C| + 1のラベルを持つ可能性があり、オブジェクトインスタンスと背景のために、多くのビンは空であり、速度のためにヒストグラムを疎に格納することができますが。
私たちは、2つの別々の情報利得基準を混ぜるという[5]の提案は、私たちのデータでは劣っていることを発見しました。

重要な問題は、木の分割で評価する特徴量の選択である。我々は、法線、色など多数の特徴を調べた。我々は、[24]の非常に単純で計算が速い特徴がよく機能し、特徴の種類を追加しても精度が上がらない（しかし遅くなる）ことを見出した。直感的な説明としては、学習された木における単純な特徴の組み合わせが、学習データと分割の目的によって定義されたタスクに特化した複雑な特徴を作り出すことができる、というものである。の特徴量は
[24]は，画素 i の近傍の画素からの深度または色差を考慮し，文脈の局所的なパターンを捕らえる・この特徴量は，深度に対してほぼ不変となるように深度適応される[23]・各オブジェクトは学習のために分割される。特徴検査がオブジェクトマスクの外にまで及ぶ場合、特徴応答を計算するために、ある種の背景をモデル化する必要がある。我々の実験では、一様なノイズや、オブジェクトが乗っている模擬平面を使用する。これはうまく機能し、新しい未知の画像にもうまく汎化できることがわかりました。オブジェクトを平面の上に置くことで、フォレストは文脈情報を学習することができる。

学習には、セグメント化されたオブジェクト画像とRGB-D背景画像のセットからランダムにサンプリングされたピクセルを使用する。量子化されたオブジェクト座標に基づく木構造を学習した後、全てのオブジェクトからの学習ピクセルを木を通して押し出し、各葉において各オブジェクト$\mathbf{c}$の連続位置$\mathbf{y}$を全て記録する。次に、ガウシアンカーネルと帯域幅を用いた平均シフトを実行する
2.5cmである。トップモードを予測値 $\mathbf{y}_c(l_i)$ として、リーフに格納する。さらに、各葉において、オブジェクトの所属の分布 $p(c|l_i)$を近似するために、各オブジェクト c から来るピクセルのパーセンテージを保存します。また、l_i に到着した背景集合からのピクセルの割合を格納し、$p(bg|l_i)$ と呼ぶ。
Forestの利用 学習が完了したら、RGB-D画像の全画素をForestの各木に通すことで、各画素iにそれぞれ

分布 $p(c|l_i^j)$ と、各木 j と各オブジェクト c に対して 1 つの予測 $\mathbf{y}_c(l_i^j)$ がある。

ここで、l_i^j は、tree j における画素 i の葉の結果である。
画素iは、ベクトル$\mathbf{l}_i = (l_i^0, \ldots, l_i^j, \ldots, l_i^{|T|})$にまとめられる。の葉の結果は

は、$L = (\mathbf{l}_1, \ldots, \mathbf{l}_n)$ に要約されます。画素が分類された後
の各画素iについて計算する。
は、葉 l_i^j に格納された $p(c|l_i^j)$ を組み合わせて、数 p をイメージする。$_{c,i}$

$p_{c,i}$ は、ある画素 i が以下の画素に属する近似確率 $p(c_i|l_i)$ と見なすことができる。
をオブジェクトcのすべてのリーフノードで終了したと仮定する $\mathbf{l}_i = (l_i^0, \ldots, l_i^j, \ldots, l_i^{|T|})$.我々は

は、このように数$p_{c,i}$、オブジェクトの確率と呼ぶ。オブジェクトという確率

$$p_{c,i} = \frac{\prod_{j=1}^{|T|} p(c|l_i^j)}{\sum_{c'} \prod_{j=1}^{|T|} p(c'|l_i^j)}. \quad (1)$$

式1の詳細な演繹は、補足資料にあります。

3.2 エネルギー機能

我々の目標は、物体cの6自由度姿勢H_cを推定することである。姿勢H_cは、ある点を物体空間からカメラ空間に写す剛体変換（3次元回転と3次元並進）と定義される。我々は、ポーズ推定をエネルギー最適化問題として定式化する。エネルギーを計算するために、H_c を用いてレンダリングした合成画像と、観測された深度値 $D = (d_1, \ldots, d_n)$ およびフォレストの結果 $L = (\mathbf{l}_1, \ldots, \mathbf{l}_n)$ を比較する。我々のエネルギー関数は、3つの成分に基づいている。

$E(H_c) = \lambda_c^{\text{depth}} E^{\text{depth}}(H_c) + \lambda_c^{\text{coord}} E^{\text{coord}}(H_c) + \lambda_c^{\text{obj}} E^{\text{obj}}(H_c)$ となる。 (2)

$E^{\text{depth}}(H_c)$ という成分は、観測された深度画像と理想的なレンダリング深度画像との間の偏差を罰するものであるが、$E^{\text{coord}}(H_c)$ と $E^{\text{obj}}(H_c)$ という成分は、観測された深度画像と理想的なレンダリング深度画像との間の偏差を罰するものである。
の予測値との乖離が大きい。図3は、各コンポーネントの利点を視覚化したものである。パラメータ λ_c^{depth}, λ_c^{coord}, λ_c^{obj} は、異なる観測の信頼性を反映している。ここで、各コンポーネントの詳細を説明する。
深度成分は次のように定義されます。

$$E_c^{\text{depth}}(H_c) = \frac{\sum_{i \in M_c(H_c)} f(d_i, d_i^*(H_c))}{|M_c(H_c)|}, \quad (3)$$

ここで、$M_c(H_c)$ は物体 c に属する画素の集合である。これは、物体を画像にレンダリングして、姿勢 H_c から導き出したものである。深度観測値d_i を持たない画素は除外される。項$d_i^*(H_c)$ は、ポーズH_c でレンダリングされたオブジェクトcについて記録された3Dモデルのピクセルiにおける深度です。 3Dモデルの不正確さを扱う

ために、ロバスト誤差関数：$f(d_i, d^*(H)) =$ を使用します。
$\min(||x(d_i) - \mathbf{x}(d^*(H))||, \tau_d) / \tau_d$, ここで $\mathbf{x}(d_i)$ は，3次元座標を表す。深度dから得られるカメラシステム$_i$。 定義の分母は
は，深度成分を正規化し，カメラに対するオブジェクトの距離に依存しないようにします。

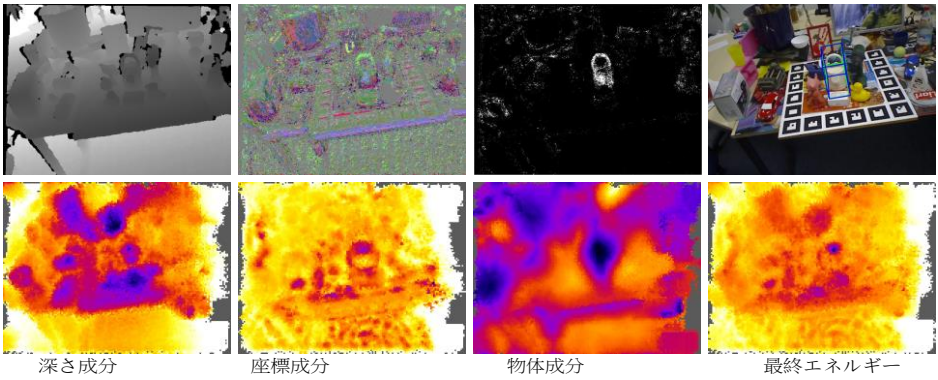

図3.異なるエネルギー成分の利点。異なるエネルギー成分は強いローカルミニマムを示すが、それらの組み合わせは通常、正しいポーズで最も強いミニマムを示す。エネルギーは異なるポーズで計算され、最小射影を用いて画像空間に投影された。白は高エネルギー、紺は低エネルギーを表しています。各成分は、関連するデータの下に表示されます。左から右へ：テスト画像の奥行き成分と $E_c^{\text{depth}}(H_c)$、予測される物体座標を $E_c^{\text{coord}}(H_c)$、オブジェクトの確率を $E_c^{\text{object}}(H_c)$、RGBの c の成分が最終的なエネルギー $E_c(H_c)$ と共に表示される。*推定姿勢（青）*と*真実姿勢（緑）*はバウンディングボックスとして表示される。

Object Component は、M^D の理想的なセグメンテーションの中にある、Forest によればオブジェクト c に属する可能性が低い画素を罰するものである。これは次のように定義される。

$$E_c^{\text{obj}}(H_c) = \frac{\sum_{i \in M^D(H_c)} -\log p(c \mid \mathbf{f}_i)}{|M^D(H_c)|} \tag{4}$$

座標成分は、フォレストによって予測されるオブジェクト座標 $\mathbf{y}_{i,c}(\mathbf{l}_i^j)$ と、レンダリング画像から得られる理想的なオブジェクト座標 $\mathbf{y}_{i,c}(H_c)$ との間の偏差を罰する。このコンポーネントは次のように定義されます。

$$E_c^{\text{coord}}(H_c) = \frac{\sum_{i \in M^L(H_c)} \sum_{j=1}^{m} g\left(\mathbf{y}_{i,c}(\mathbf{l}_i^j), \mathbf{y}_{i,c}(H_c)\right)}{|M^L(H_c)| \cdot m}. \tag{5}$$

ここで、$M^L(H)$ は、物体 c に属するピクセルの集合で、深度観測 d がないピクセル i と、$p_{c,i} < \tau_{pc}$ を除く。後者は、$p_{c,i}$ が小さいピクセルでは、信頼できる座標予測 $\mathbf{y}_{i,c}(\mathbf{l}_i^j)$ ができないことがわかったので必要である。$\mathbf{y}_{i,c}(H_c)$ という用語は、姿勢 H_c でレンダリングされたオブジェクト c の3Dモデルのピクセル i における オブジェクト空間での座標を表します。

ロバスト誤差関数 $g(\mathbf{y}_{i,c}(\mathbf{l}_i^j), \mathbf{y}_{i,c}(H_c)) = \min\left(\left\|\mathbf{y}_{i,c}(\mathbf{l}_i^j) - \mathbf{y}_{i,c}(H_c)\right\|^2, \tau_y\right) / \tau_y$

最終的なエネルギー関数 エネルギー項はすべて正規化されているため、考慮する画素数が非常に多くなると安定性が問題になることがあります。

を小さくする。この問題に対処するために、我々は以下の安定な定式化を用いる。

$$E_c(H_c) = \begin{cases} \text{fE}\hat{}c(H_c), & \text{if } |M^L(H_c)| > 100 \\ \infty, & \text{さもなければ} \end{cases} \quad (6)$$

3.3 最適化

式6のタスクの解を求めるために、我々はRANSACベースのアルゴリズムを使用する。このアルゴリズムは、観測された深度値とフォレストからの座標予測に基づき、ポーズ仮説をサンプリングする。その後、これらの仮説は評価され、改良される。図 1 はこの過程を視覚化したものである。ここでは、その手順を詳細に説明する。

ポーズ仮説のサンプリングは、まず画像から1つの画素i_1 を、事前に計算したpに比例した重みで描画することで行われる$_{c,i}$ 各画素の$i.i_1$ を中心とする正方形の窓から、さらに2つの画素 i_2 と i_3 を同じ方法で描画する。窓の幅は、物体の直径と画素の観測深度値 d_{i_1} から計算される $w = f\delta_c / d_i$ ここで、f = 575.816 画素は焦点距離である。サンプリングは$p_{c,i}$ の積分画像を用いて効率的に行われる。 各ピクセルに対して、ツリーインデックスj_1 、j_2 、j_3 をランダムに選択する。最後に、Kabsch アルゴリズムを使用して、姿勢仮説 H_c を以下から計算します。

3D-3D-対応表 $(\mathbf{x}(i_1), \mathbf{y}_c(l_{j^1}))$、$(\mathbf{x}(i_2), \mathbf{y}_c(l_{j^2}))$、$(\mathbf{x}(i_3), \mathbf{y}_c(l_{j^3}))$ があります。

3つの予測位置$\mathbf{y}_c(l_{j^-})$のそれぞれをカメラ空間にマッピングする。ing H_c 、変換誤差 $e_{i,j}(H_c) = ||x(i_\bullet) - H_c \mathbf{y}_c(l_{j^-})||$ を計算します。

であり、その相手とのユークリッド距離を単純化したものである。我々は、ポーズ仮説H_c は、3つの距離のいずれもオブジェクトの直径 δ_c の5%より大きくない場合のみである。このプロセスは、一定数の210個の仮説が受け入れられるまで繰り返される。受け入れられたすべての仮説は、式(6)に従って評価される。

洗練は、上位25の受け入れられた仮説に対して実行される。ポーズ H_c を洗練させるために、エネルギー計算で行ったように、オブジェクト c に属すると思われるピクセル $M^D(H_c)$ のセットに対して反復処理を行います。各ピクセル $i \in M^D(H)$に対して、以下の計算を行う。

誤差$e_{i,j}(H_c)$ すべての木jについて、最小の誤差$e_{i,j}(H_c) \leq$を持つ木をjとする。

$e_{i,j}(H_c) \forall j \in \{1, .., |T|\}$画素$i$について、$e_{i,j}(H_c) < 20$mmであるすべての画素iを対象とする。

をインライアと見なす。すべてのインライヤー pix- について、対応関係 $(\mathbf{x}(i_1), \mathbf{y}_c(l_{ij}))$ を格納する。

を計算し、それを使ってKabschアルゴリズムでポーズを再推定します。このプロセスは、式6に従った姿勢のエネルギーが減少しなくなるか、インライアピクセルの数が3以下になるか、または反復の合計が100回に達するまで繰り返されます。

最終推定値 精密化後に最もエネルギーが低い姿勢仮説を最終推定値として選択する。図1〜図3の推定値、および実験セクションの定量的結果は、上記の正確なアルゴリズムを用いて得られたものである。しかし、我々のタスクはエネルギー最適化問題として定式化されているため、推定値の精度をより向上させるために、一般的な最適化アルゴリズムを使用することが可能である。

4 実験風景

過去にいくつかの物体インスタンス検出データセットが 公開されているが[21, 3]、その多くは2次元の姿勢のみを扱っ ている。Laiら[13]は、300個のオブジェクトからなる大規模なRGB-Dデータセットを公開し、近似的な回転角の形でグランドトゥルースのポーズを提供しています。しかし、残念ながら、このような注釈は、我々が解決しようとする正確な姿勢推定タスクには粗いものである。我々は、最近導入されたHinterstoisserら[8]のデータセットと我々自身のデータセットで我々のアプローチを評価した。Hinterstoisserのデータセットでは、合成トレーニングと実テストデータが提供されている。我々のデータセットでは、現実的なノイズパターンと困難な照明条件による実トレーニングと実テストデータを提供する。両データセットにおいて、我々は[8]のテンプレートに基づく方法と比較した。また、我々の手法のスケーラビリティを検証し、実行時間についてコメントする。補足資料として、オクルージョンデータセット、検出タスク、及び、我々の個々のエネルギー項の貢献度に関する追加の実験結果を示す。我々は以下のパラメータで決定森を学習する。各ノードにおいて、500個の色彩特徴と深度特徴をサンプリングする。各反復において、学習画像あたり1000個のランダムなピクセルを選択し、現在の葉に集め、到着したピクセルが50個以下の場合は分割を停止する。木の深さは制限されません。パラメータの完全なセットは付録で見ることができます。

Hinterstoisserらのデータセット Hinterstoisserら[8]は、学習用に13個のテクスチャレスオブジェクトのカラー3Dモデル[3] と、散らかった机上の各オブジェクトの1000以上のテスト画像と、グランドトゥルースポーズを提供します。テスト画像は、異なるスケールと$\pm 45^\circ$ in- の範囲で、上面視半球をカバーしています。
面回転を行う。の姿勢推定精度を評価することである。
オブジェクトが存在します。どのオブジェクトが存在するかは既知である。我々は、オブジェクトのポーズが正しく推定されたテスト画像の割合として精度を測定することにより、正確に[8]のテストプロトコルに従います。厳しいポーズ許容度は補足資料で定義されています。8]では、著者らは平均96.6%の正しく推定されたポーズという強力なベースラインを達成しました。我々は彼らの手法を再実装し、これらの数字を再現することができました。彼らのパイプラインは、効率的なテンプレートマッチングスキーマから始まり、2つの外れ値除去ステップと反復的な最接近点調整と続く。2つの外れ値除去ステップは、報告された結果を達成するために非常に重要である。本質的には、現在の推定値とテスト画像との間の、色と深さの差に関する2つの閾値からなる。残念ながら、正しい値はオブジェクトによって大きく異なるため、各オブジェクトに対して手作業で設定する必要があります[4] 。また，性能と速度を上げるためにHinterstoisserテンプレートを識別的に最適化する[21]と比較します。彼らはまた、同じ2つの異常値除去チェックに依存していますが、オブジェクトに依存する閾値を識別的に学習しています。

本手法の学習データを作成するために，13 個の物体モデルすべてを [8] と同じ視点サンプリングでレンダリングしたが，スケールバリエーションをスキップしたのは，以下の理由による。

[3]2つのオブジェクトは、適切な3Dモデルがないため、省略することになりました。
[4]著者との私信で確認した。これらの値は論文には記載されていない。

奥行き不変の特徴量の我々の特徴は学習中にオブジェクトのセグメンテーションの外にまで及ぶ可能性があるため、感覚的な特徴応答を計算するために背景モデルが必要である。色彩特徴には、背景画像群からランダムに抽出された色を用いる。背景画像は、私たちが撮影した約1500枚のRGB-D画像から構成されています。深度特徴には、背景モデルとして無限の合成地面を用いる。テストシーンでは、全てのオブジェクトはテーブルの上に立っているが、密集した乱雑さの中に埋め込まれている。したがって、我々は合成平面を許容可能な事前分布と見なす。さらに、一様な深度ノイズと一様なRGBノイズの背景モデルに対する結果も示す。決定森は13個のオブジェクト全てと背景クラスに対して同時に学習される。背景クラスについては、オフィスの背景セットからRGB-Dパッチをサンプリングした。純粋な合成訓練画像と実際のテスト画像との間の外観のばらつきを考慮し、色特徴の応答にガウスノイズを追加する[25]。エネルギーを最適化した後、[8, 21]とは対照的に、外れ値除去のステップを導入しない。

表1. Hinterstoisserらのデータセットにおける、合成訓練データ、実訓練データ、異なる背景モデル（平面、ノイズ）を用いた結果。我々のアプローチが一貫して[8, 21]より優れていることが分かる。

	シンセサイザートレーニング				リアルトレーニング	
	LINEMOD[8]の場合	DTT-3D[21]の場合	私たちの(平面)	私たちの(ノイズ)	私たちの(平面)	私たちの(ノイズ)
Avg.	96.6%	97.2%	98.3%	92.6%	98.1%	97.4%
メド。	97.1%	97.5%	98.9%	92.1%	99.6%	98.8%
マックスです。	99.9%	99.8%	100.0%	99.7%	100.0%	100%
Min.	91.8%	94.2%	95.8%	84.4%	91.1%	89.2%

その結果を表1にまとめる。同期平面背景モデルで平均98.3%のスコアを出すことができた。したがって、[8]と[21]の両方のシステムを改善することができた。定性的な結果については、図4を参照してください。また、背景モデルとして一様なノイズを用いた場合でも、平均92.6%の正解率を得ることができ、優れた結果を得ることができました。

本アプローチが合成学習データに限定されないことを確認するため、各オブジェクトの実画像で学習を行う実験を行った。データセットには1つのオブジェクトにつき1シーンしか含まれていないため、各シーケンスを訓練とテストに分割する必要がありました。学習画像は、少なくとも15°
の角度距離で、上部半球をほぼ規則的にカバーすることができます。
Hinterstoisserらのセットアップと同様の球体。訓練画像の最大距離

は≈25° であり、このテストは合成セットアップよりも若干難しくなっています。他の画像は全てテスト画像です。学習画像の背景を除去するために、3Dオブジェクトのバウンディングボックスとの交差テストを行います。このテストでは
背景画素に、既に述べた2つの背景モデル変種を適用した。また、特徴量にはノイズを加えない。この実験では、単純なノイズ背景モデルでも安定した優れた精度を得ることができた（表1の右2列を比較）。

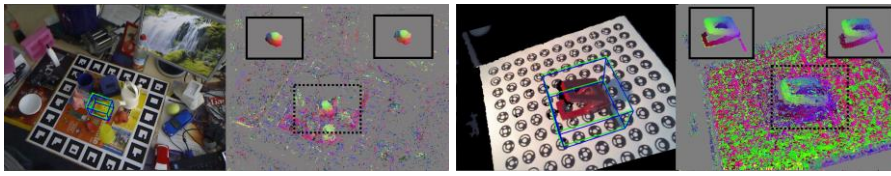

図4.我々のシステムによる姿勢推定（青のバウンディングボックス）とグランドトゥルース姿勢（緑のバウンディングボックス）の比較例。左のテスト画像はHinterstoisserらのデータセット[8]のオブジェクト、右のテスト画像は我々のデータセットのオブジェクトを表している。各テスト画像の横には、森の1本の木から予測される物体座標**yが示されている。**ここで、"最良 "とは、地上真実に対して全ての木の中で最良の予測である（説明のため）。

我々のデータセット明るい人工光（*bright* ）、暗い自然光（*dark* ）、指向性のあるスポット光（*spot* ）の3つの照明条件下で、テクスチャを持つ物体と持たない物体の計20個を記録した。各照明条件において、マーカーボード上の各物体を、その上面視半球を覆うような動きで記録した。このとき、物体までの距離は変化させたが、面内回転は一定に保った。の範囲で後から人為的に面内回転を加えた。$\pm 45^\circ$ 。我々はKinectFusion [17, 12]を使用して、各フレームの外部カメラパラメータを記録しました。これはポーズのグランドトゥルースとして機能し、決定森をトレーニングするためのピクセルごとのオブジェクト座標を生成するために使用される。の記録は
また、同じ物体で照明の異なる画像をマーカボードで登録した。トレーニングに使用した画像は、3Dオブジェクトバウンディングボックスを用いてセグメンテーションを行った。データセットの概要と記録方法の詳細は付録を参照されたい。トレーニング画像は、少なくとも15° の角度距離でサンプリングしました。トレーニング画像の最大角度距離は
$\approx$ 25° .合成平面上にオブジェクトを配置しなかったのは、オブジェクトが
すでに平面基板に記録されている対象物の外まで届く深度特徴
のマスクは、元の学習画像の深度のみを使用する。色彩特徴については、オブジェクトを含まない別のオフィス背景のセットからランダムにサンプリングしました。

我々のアプローチが様々な照明条件に対してどの程度汎化できるかを評価するために、*明るい*学習セットと*暗い*学習セットで決定森を学習させた。また、ロバスト性のために、色の特徴量の反応にガウスノイズを加えた。最初のテストでは、トレーニングに使用しなかった*明るい*セットの画像でテストを行った。このとき、forestは新しい照明条件に対して汎化する必要はなく、見たことのない景色に対してのみ汎化することができ、その精度は非常に高い（平均95%、表2参照）。前回と同様に、テスト画像ごとに常に存在する1つのオブジェクトのポーズを正しく推定した割合で性能を測定しました。2回目のテストでは、難しい新しい照明条件への汎化能力を実証するために、完全な*スポット*セットでテストを行った。その結果、平均88.2%の正解率を得ることができた。

Linemod[8]のテンプレートベースのアプローチが照明の変化に対してあまり一般化しないことを示すために、我々の再実装を使用して、上記のトレーニングセットに基づいて1つのオブジェクトのテンプレートを抽出しました。ここで、学習セットには1つのスケールで表示された各ビューが含まれていることに注意してください。これは、テストデータがトレーニングデータでカバーされていないスケールバリエーションを持つ場合、Linemodにとって問題となる可能性があります。そこで、各トレーニング画像を10cm刻みで大小2つの距離からレンダリングします。これにより、[8]と同様に各トレーニング画像のスケールは6種類となります。8]と同様に、外れ値除去のパラメータは手作業で調整しました。しかし、新しい照明条件下では、これらのテストを完全に無効にしないと検出されないことがわかった。検証では、*明るい*学習セットからテンプレートを抽出し、*明るい*テストセットでテストを行いました。8]の手順に従い、80.1%の画像で正しいポーズを推定することができました。Hinterstoisserデータセット[8]の性能との差は、オブジェクトにテクスチャがあることと、画像にノイズがあることに起因しています。同じテンプレートを使って*スポット*セットでテストした場合、性能は57.1%に低下します。Linemodは*明るい*ところと*暗い*ところの両方を学習しているため、以下のテスト方法を適用し、公平に比較します。*暗い*学習セットからテンプレートを抽出し、*スポット*のテストセットに適用したところ、55.3%のパフォーマンスが得られました。最終的なスコアは、Linemodが解決した画像のうち、*暗い*テンプレートと*明るい*テンプレートのいずれかが正しいポーズを導いた場合、その画像を解決したと見なします。その結果、70.2%の精度が得られました。つまり、Linemodが有利なテスト条件下でも、性能は10%低下します。同じオブジェクトで、*明るい*テストセット（トレーニング照明に含まれる）で96.9%、*スポット*テストセット（トレーニングに含まれない）で91.8%の精度を報告します。

表2.異なる照明条件でのテスト時のデータセットでの精度。トレーニングセットには*明るい照明*が登場し、*スポット照明は登場しない*。20個のオブジェクトの平均と中央値を報告する。また、1つのオブジェクトについてLinemod[8]との比較も行った。詳細は本文で述べる

	すべて		おもちゃ(バトルキャット)			
テスト条件	Avg.	メド。	私たちの	[8](*暗*)	[8](*明るい*)	[8](合算)
あたまのかいてんがはやい	95.6%	97.7%	96.9%	-	80.1%	-
スポット	88.2%	93.0%	91.8%	55.3%	57.1%	70.2%

スケーラビリティ我々は、スケーラビリティに関して、2つの異なる方法、すなわち、オブジェクト数のスケーラビリティとポーズの空間のスケーラビリティに関して、本手法の可能性を示す。前者はシステムが識別できるオブジェクトの数に関係し、後者はシステムが認識できるポーズの範囲に関係する。まず、5つのオブジェクトについて学習したフォレストと、*暗い*照明条件と*明るい照明条件*、最

小の角度距離でサンプリングした学習画像群から始める。45°。我々は
は、各トレーニング画像に±45°の面内回転を加える。テスト時には
の画像は、最も近い学習画像から最大10°離れたスポットセットである。この結果、テストの難易度はこれまでの実験と同じになる。性能は1つの物体（猫のぬいぐるみ）に対して測定されている。我々はこの設定を2つの方法で修正した。1つ目は、我々のデータセットを組み合わせてオブジェクト数を30に増やし

を、Hinterstoisserデータセットの実画像と比較しました。Hinterstoisserデータセットは、我々のオブジェクトとHinterstoisserの追加オブジェクトに対して、ほぼ同じ量のトレーニング画像を持つようにサンプリングしました。次に、面内回転させた学習画像の枚数を4倍に増やし、*±180*°まで増やしました。その結果は以下の通りです。
を図5に示します。

	デフォルト	6×オブジェクト	4×ポーズ
森の時間	102ms	138ms	113ms
最適時間	443ms	424ms	517ms
精度	95%	94%	95%

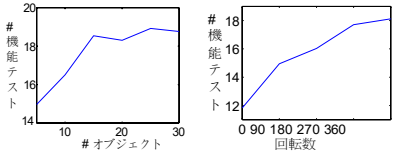

図5. 上図：オブジェクト数とポーズ範囲が増加した場合の本システムの実行時間。精度は安定している。下図。決定フォレストのサブリニアな成長を示す図。

オブジェクトの数とポーズの範囲が増えると、木の評価時間は若干増加するが、6×、4×よりかなり小さい。また、エネルギー最適化の実行時間は、(1)オブジェクトの数、(2)ポーズ数の変動により若干の影響を受ける。
森林予測で、システムの精度は安定したままである。図5の表の下に、オブジェクトの数とポーズの範囲の増加に伴い、ピクセルあたりの平均特徴検査数がサブリニアに増加する様子をプロットしています。我々の提案するパイプラインはオブジェクトの数に対して線形であるが、フォレストでは我々の識別的に訓練された手法の最初の重要なステップがオブジェクトの数に対してサブ線形に振る舞うことを実証している。
実行時間我々の姿勢推定アプローチの完全な実行時間は、フォレスト予測時間とエネルギー最適化時間の合計である。フォレスト予測はフレームごとに1回生成され、その結果はそのフレーム内のすべてのオブジェクトに対して再利用される。ランダムフォレストのCPU実装は、[8]のデータセットでフレームあたり平均160msを要する。これらの予測に基づき、オブジェクトごとにエネルギー最適化が行われる。我々はGPUでエネルギー評価を行い、[8]のデータセットにおいて、上記で提案したパラメータ設定でオブジェクトあたり平均398msを報告した。しかし、我々は、削減されたパラメータのセットが、精度を維持したまま、大きなスピードアップをもたらすことを発見した。我々は、仮説の数を210から42に減らし、洗練のステップ数を100から20に減らし、最良の3つの仮説のみを洗練した。これにより、[8]のデータセットにおいて、平均96.4%の精度を達成すると同時に、エネルギー最適化の平均時間を61msに短縮することに成功した。

Acknoledgements:

この研究は，欧州社会基金とザクセン州のプロジェクト VICCI (#100098171) によって部分的に支援されている・Linemodの再実装を

行ったHol- ger Heidrichに感謝する。また、Stephan Ihrke、Daniel Schemala、Patrick Sprungには、さまざまなデータセットの準備に協力してもらった。

参考文献

1. Bo, L., Ren, X., Fox, D.: RGB-D ベースの物体認識における教師なし特徴学習.In:ISER.(2012)
2. クリミニシ、A.、ショットン、J.：コンピュータビジョンと医療画像解析のための決定フォレスト.シュプリンガー (2013)
3. Damen, D., Bunnun, P., Calway, A., Mayol-Cuevas, W.: Real-time learning and detection of 3D texture-less object:このような場合、「曖昧さ」を解消することが重要である。In:BMVC.(2012)
4. Drost, B., Ulrich, M., Navab, N., Ilic, S.: Model globally, match locally:効率的でロバストな3次元物体認識.In:CVPR.(2010)
5. IEEE Trans.33(11) (2011)
6. Girshick, R., Shotton, J., Kohli, P., Criminisi, A., Fitzgibbon, A.: Efficient regres- sion of general-activity human poses from depth images.In:ICCV.(2011)
7. Hinterstoisser, S., Cagniart, C., Ilic, S., Sturm, P., Navab, N., Fua, P., Lepetit, V.: Gradient Response Map for real-time detection of texture-less object（テクスチャーレスオブジェクトのリアルタイム検出のための勾配応答マップ）.において。IEEE Trans.(2012)
8. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scene（モデルに基づく、乱雑なシーンにおけるテクスチャのない3Dオブジェクトの学習、検出、姿勢推定）.In:を用いた。(2012)
9. Hoiem, D., Rother, C., Winn, J.: 3D LayoutCRF for multi-view object class recognition and segmentation（多視点オブジェクトクラス認識とセグメンテーションのための3D LayoutCRF）.In:CVPR.(2007)
10. ホルツァー、S.、ショットン、J.、コーリ、P.。深度データにおける繰り返し可能な関心点の効率的な検出のための学習.In:ECCV.(2012)
11. Huttenlocher, D., Klanderman, G., Rucklidge, W.: Comparing images using the hausdorff distance.IEEE Trans.on PAMI.(1993)
12. このような場合、「萌え萌え」なのは、「萌え萌え」なのですが、「萌え萌え」なのは、「萌え萌え」なのですが、「萌え萌え」なのは、「萌え萌え」なのですが、「萌え萌え」なのは、「萌え萌え萌え」なのです。In:UIST.(2011)
13. Lai, K., Bo, L., Ren, X., Fox, D.。大規模な階層的多視点RGB-Dオブジェクトデータセット.で。ICRA.IEEE (2011)
14. Lepetit, V., Fua, P.。ランダムツリーを用いたキーポイント認識.IEEE Trans.on PAMI.28(9) (2006)
15. Lowe, D.G.: 3次元物体認識のための局所特徴ビュークラスタリング.In:CVPR.(2001)
16. マルティネス、M、コレット、A、スリニバサ、S.S.・Moped:また、このような技術的な問題点を解決するため に、「Scientific Process System」（以下、「SCS」という。で。ICRA.(2010)
17. Newcombe, R., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A.: KinectFusion.リアルタイム高密度表面マッピングとトラッキング。リアルタイムの高密度なサーフェスマッピングとトラッキング。In:ISMAR.(2011)
18. Nist´er, D., Stew´enius, H.: 語彙木によるスケーラブルな認識.In:CVPR.(2006)
19. Ozuysal, M., Calonder, M., Lepetit, V., Fua, P.（オズィサル、M.、カロンダー、ルプティ、V.、フア、P.）。ランダムなシダを用いた高速なキーポイント認識.IEEE Trans.on PAMI.(2010)
20. このような場合、「曖昧模糊」と呼ばれる。In:CVPR.(2007)
21. Rios-Cabrera, R., Tuytelaars, T.: 3D オブジェクト検出のための識別的に学習されたテンプレート。実時間スケーラブルアプローチ.において。

ICCV.(2013)
22. Rosten, E., Porter, R., Drummond, T.: FASTER and better: コーナー検出のための機械学習アプローチ。IEEE Trans.on PAMI.32 (2010)

23. また、このような場合、「俯瞰的な視点」を持つことが重要である。In:CVPR.(2011)
24. このような場合、「曖昧さ」を解消するために、「曖昧さ」を解消するために、「曖昧さ」を解消するために、「曖昧さ」を解消するために、「曖昧さ」を解消するために、「曖昧さ」を解消するために、「曖昧さ」を解消する必要があります。In:CVPR.(2013)
25. また、"萌え "と "癒し "をキーワードに、"癒し "と "癒し "をキーワードに、"癒し "と "癒し "をキーワードに、"癒し "と "癒し "をキーワードに、"癒し "と "癒し "をキーワードに、"癒し "をキーワードに、"癒し "をキーワードに、"癒し "をキーワードにした、新たな "癒し "の提案をしています。
26. Steger, C:オクルージョン、クラッタ、イルミネーションに影響されない物体認識のための類似性尺度.で。DAGM-S.(2001)
27. また、このような場合、「曖昧さ」を解消するために、「曖昧さ」を解消した上で、「曖昧さ」を解消するために、「曖昧さ」を解消した上で、「曖昧さ」を解消した上で、「曖昧さ」を解消した上で、「曖昧さ」を解消した上で、「曖昧さ」を解消した上で、「曖昧さ」を解消した上で、「曖昧さ」を解消する。In:ECCV.(2010)
28. このような場合、「曖昧模糊」と呼ばれる。において。CVPR.(2012)
29. V.フェラーリ、F.J.、シュミッド、C:物体検出のための画像から形状モデルへ.In:IJCV.(2009)
30. Winder, S., Hua, G., Brown, M.: Picking the best DAISY（最高のDAISYを選ぶ）.In:CVPR.(2009)