Vol.2020-CVIM-221 No.44
2020/3/17

テクスチャ情報を用いない多視点インスタンス対応付け

土井 拓磨[1]　大倉 史生[1]　長原 稔樹[1]　松下 康之[1]　八木 康史[1]

概要：本研究は，物体ごとにセグメンテーションされた三次元復元などを応用とした，多視点画像におけるインスタンスセグメンテーションの実現を目的とする．特に，本論文ではテクスチャ情報や既知の三次元形状を用いない，多視点画像間でのインスタンス対応の推定手法を提案する．植物の葉のように，類似の模様や形状が繰り返し現れる場面では，画像間での特徴点対応付けが困難であるため，多視点画像間でインスタンス領域とその対応を求めることは難しかった．そこで本研究では，密な特徴点対応に依存しない，エピポーラ幾何に基づくインスタンス領域の対応付け手法を構築する．提案手法は，三次元復元のみならず，インスタンスセグメンテーションにおける領域提案の改善にも効果的であり，より正確なインスタンス検出を実現できる．実験を通じ，多視点ステレオに基づく手法と比較して，多視点画像間でのインスタンス対応付け精度が大きく向上していることを示した．さらに，提案手法の応用により，従来のインスタンスセグメンテーション手法（Mask R-CNN）と比較して，インスタンスの検出精度が改善したことを確認した．また三次元復元においては，多視点画像間でのインスタンス対応が無い場合と比較して，復元精度が大きく向上することを確認した．

1. はじめに

実世界の建造物や構造物の三次元データを取得する三次元復元技術は，コンピュータビジョンにおける重要な分野の 1 つである．二次元画像から三次元復元を行う手法として，運動からの形状復元（Structure from motion, SfM），多視点ステレオ（Multi-view stereo, MVS）などが挙げられる．SfM や MVS は伝統的な手法であるが，現在も広く利用されており，これらを拡張した手法も多く研究されている．これらの手法は，エッジやテクスチャを元に多視点画像間で対応付けを行うことで，カメラパラメータや物体の三次元点群を推定する．そのため，植物のように類似した物体（葉や枝）が多く含まれる場面においては似通った特徴が多く現れ，特徴点（あるいは画素）の対応を得ることが難しい．

一方，三次元復元と別の文脈で，インスタンスセグメンテーションが広く研究されている [1]．インスタンスセグメンテーションは様々な応用が研究されており，三次元情報を扱うものとしては，三次元点群やボリュームを入力としてインスタンスセグメンテーションを行う三次元インスタンスセグメンテーション [2,3] などが挙げられる．しかし，従来多視点の設定で広く行われてきた三次元復元研究の知見を活かすような，多視点へのインスタンスセグメンテーションの拡張は行われていない．

そこで本研究では，図 1 に示すような，インスタンスセグメンテーションを多視点設定へと拡張した多視点インスタンスセグメンテーション（Multi-view instance segmentation, MVIS）を提案する．しかし，特に類似のインスタンスや遮蔽が多く含まれる場面の場合，単一視点インスタンスセグメンテーションの多視点拡張は容易ではない．例えば植物の場合，鉢などの背景から疎な特徴点対応を検出し，カメラパラメータを推定することは可能であるが，図 2 に示すように，密な特徴点対応をとることをは難しい．

そこで本研究では，植物のように類似のインスタンスや遮蔽が多く在る場面を対象として，インスタンス領域を視点間で幾何学的に対応付ける手法を提案する．本手法の特筆点は，エピポーラ拘束を用いてインスタンス間の多視点対応を得る点である．各視点の各インスタンスをノードとし，ノード間の重みをエピポーラ線群とインスタンス領域の交差度に基づいて決定した対応付けグラフを構築することで，多視点インスタンス対応付け問題をグラフクラスタリング問題へと帰着させる．

2. 関連研究

以下，本研究と関連した，テクスチャが欠損する場面での多視点対応付けおよび，インスタンスセグメンテーションの複数画像や三次元への拡張について従来研究を概説する．

[1] 大阪大学
Osaka University

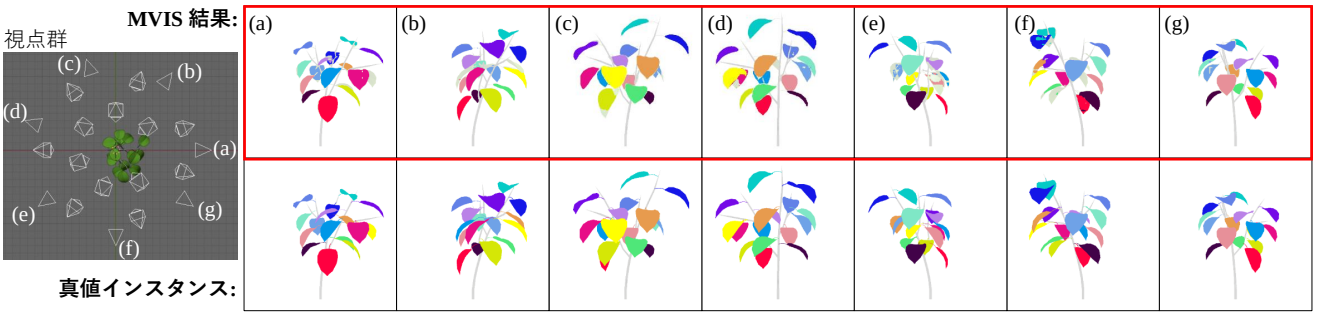

図 1　MVIS の結果例．視点間で対応するインスタンスは同じ色で塗られている．

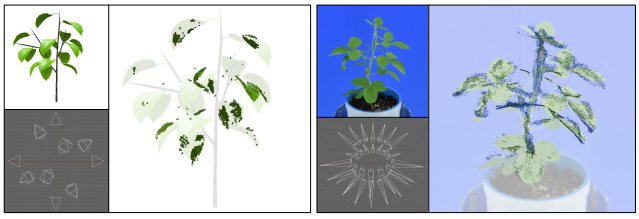

図 2　植物画像に対する MVS 結果．点群が MVS 結果を指す．

2.1　多視点対応付けにおけるテクスチャの欠損

多視点対応付けは，SfM [4] や MVS [5] などの三次元復元における基本的な問題である．多くの場面において，視点間の対応付けは特徴点 [6] やパッチベースの画素対応付け [7] が行われている．十分な対応が得られない場合、対応付けの頑健性を高めるために，画像中のある一定の面積を持った領域の特徴量に基づくワイドベースラインステレオ [8] やスーパーピクセルステレオ [9], セグメントベースのステレオ [10] といった手法が提案されている．

テクスチャを用いない多視点対応付けの研究として，Dellaert ら [11] が，疎な特徴点対応付けに関して，テクスチャを用いず，特徴点の幾何学的関係のみに基づいてカメラ姿勢と疎な点対応を計算する SfM 手法を提案している．本研究は，テクスチャを用いない領域対応付けを提案する点で，新規である．

2.2　インスタンスセグメンテーションの拡張

インスタンスセグメンテーションの中でも最先端の精度を出す実装の 1 つは，Faster R-CNN [12] を用いてインスタンスの領域候補 (region of interest, RoI) を検出した後，各 RoI においてマスクを生成する Mask R-CNN [1] である．

インスタンスセグメンテーションの入力を単一視点・時間的に連続な複数画像へと拡張した手法はビデオインスタンスセグメンテーションと呼ばれる．この手法では，インスタンスセグメンテーションとともに時間方向のトラッキングを行う（ [13, 14] など）．これら一連の手法では，時系列の画像シーケンスを入力として想定しているため，フレーム間の変化が小さいことを利用して対象物のトラッキングを可能にしている．

より広いベースラインで多視点画像を取得する設定に焦点を当てると，3D インスタンスセグメンテーションが研究されている．3D セグメンテーションにおける手法では，多くの場合三次元点群 [2] や三次元ボリューム [3] などの対象物の三次元データを入力として用いる．Nassar ら [15] は，対象となる場面の三次元形状データを用いて，多視点インスタンス対応付けのためのインスタンスワーピング手法を提案している．一方，これらの手法は MVS などによる三次元復元が十分に可能であることや，インスタンス間のテクスチャ対応付けに依存するため，本研究で扱うようなシーンには不適である．

3.　提案手法

本研究では，テクスチャ情報や既知の三次元形状を用いず，幾何学的手法に基づいて多視点画像間で同一インスタンスを対応付ける手法を提案する．ただし，カメラパラメータは背景などから取得した疎な特徴点対応を用いて，SfM [4] によって求められることを仮定している．また，Mask R-CNN などによって各視点においてインスタンスセグメンテーション（視点間のインスタンス対応は未知）が利用可能であることを仮定している．

3.1　エピポーラ幾何を用いた多視点インスタンス対応付け

図 3 に示すように，本手法は他視点から投影されたエピポーラ線群（エピポーラ帯）を介してインスタンスの対応を見つける．この時，一般に複数のインスタンスが同じエピポーラ帯上に現れるため，多視点画像間のインスタンス対応付け問題をグラフクラスタリング問題に帰着させる．

3.1.1　エピポーラ帯の投影

画像中の点が対応する他の視点の点（対応点）は，エピポーラ線上に存在する．本手法では，他視点への点の投影を領域の投影へと拡張する．具体的にはまず，各インスタンス領域で密にサンプリングした点群からエピポーラ線群を求める．ここで，i 番目の視点画像を I_i，I_i 内で検出された m 番目のインスタンスを $v_{i,m} \in V$，画像座標系における $v_{i,m}$ の領域（ピクセル群）を $\mathcal{A}_{i,m} \subset \mathbb{R}^2$ と定義する．また $\mathbf{F}_{ij}$ を i 番目と j 番目の視点間の基礎行列とする．この時，I_i 中の点 $\mathbf{p} \in \mathcal{A}_{i,m}$ は，I_j においてエピポーラ線 $\mathbf{l}_j(\mathbf{p})$ を形成する．

$$\mathbf{l}_j(\mathbf{p}) = \mathbf{F}_{ij}\tilde{\mathbf{p}} \tag{1}$$

ここで，$\tilde{\mathbf{p}}$ は $\mathbf{p}$ の同次表現である．領域 $\mathcal{A}_{i,m}$ 中には，エ

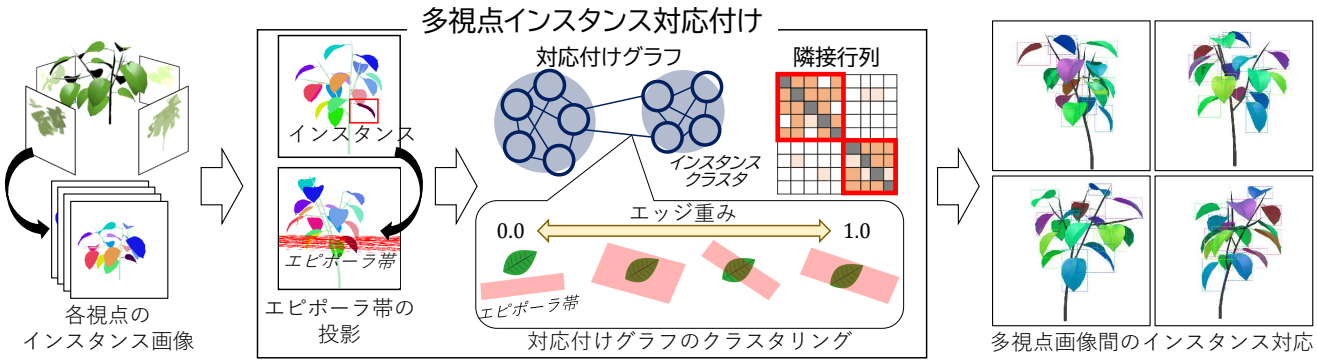

図 3 エピポーラ幾何を用いた多視点インスタンス対応付けの概要．密な特徴点対応にせず，エピポーラ拘束に基づくことで多視点画像間のインスタンス対応を得ることができる．

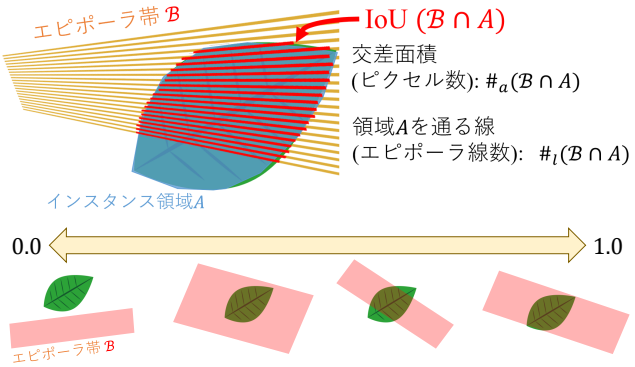

図 4 エピポーラ帯とインスタンス領域の交差度定義の概略図

ピポールを通るエピポーラ線群がある．以下では，このエピポーラ線群を**エピポーラ帯** $\mathcal{B}$ と呼ぶ．この時，画像 j 上の領域 $\mathcal{A}_{i,m}$ のエピポーラ帯 $\mathcal{B}$ は，次のように定義される．

$$\mathcal{B}(\mathcal{A}_{i,m}, I_j) = \{\mathbf{F}_{ij}\tilde{\mathbf{p}}\}, \quad \mathbf{p} \in \mathcal{A}_{i,m} \tag{2}$$

エピポーラ帯 $\mathcal{B}$ は I_j における $\mathcal{A}_{i,m}$ の存在領域を意味する．

3.1.2 対応付けグラフの構築

本提案手法は，エピポーラ拘束に基づいて無向重み付きグラフ $G = (V, E, w)$ を構築する．ただし，V は全視点のインスタンス群で構成されるノード群（各ノードは，各視点の各インスタンスと対応する），$E \subseteq V \times V$ はエッジ群，$w : E \to \mathbb{R}_+$ はエッジの重みを定義する重み関数である．ノード数 $|V|$ は，現れるインスタンスの総数に対応する．

グラフ G のエッジ重み w には，エピポーラ帯 $\mathcal{B}$ とインスタンス領域 $\mathcal{A}$ の交差度を類似度として定義し，これを用いる．ここで用いる交差度は，IoU（intersection over union）と似たアイデアであるが，IoU は領域間の一致度を評価する指標である．本研究では，これをエピポーラ帯とインスタンス領域間の交差度に拡張するため，二つの尺度を使用する．図 4 に示すように，一つはインスタンス領域を通過するエピポーラ帯のエピポーラ線の数であり，もう一つはその交差面積である．以上の尺度を用いて，グラフ G のノード $v_{i,m}$ と $v_{j,n}$ 間のエッジ重み w は次の式で得られる．

$$w(v_{i,m}, v_{j,n}) = \frac{\#_a(\mathcal{B}(\mathcal{A}_{i,m}, I_j) \cap \mathcal{A}_{j,n})}{\#_a \mathcal{A}_{j,n}} \cdot \frac{\#_l(\mathcal{B}(\mathcal{A}_{i,m}, I_j) \cap \mathcal{A}_{j,n})}{\#_l \mathcal{B}(\mathcal{A}_{i,m}, I_j)} \tag{3}$$

ここで関数 $\#_a$ は，エピポーラ帯とインスタンス領域の共通領域に属するピクセル数をカウントし，$\#_l$ はそのエリアを通過するエピポーラ線の数をカウントする．また $\#_a(\mathcal{B}(\mathcal{A}_{i,m}, I_j) \cap \mathcal{A}_{j,n})$ の計算において，実験では全てのエピポーラ線を同じ太さで処理している．他の多くの類似度指標と同様に，$w(v_{i,m}, v_{j,n})$ は範囲 $[0, 1]$ の値を取る．例えば，エピポーラ帯内の全てのエピポーラ線がインスタンス領域を通過し，かつインスタンス領域全体がエピポーラ線で埋められた場合，値は 1 となる（図 4 参照）．

その後，SymNMF [16] を用いてグラフクラスタリングを行う．各クラスタは，多視点画像間のインスタンス対応となる．つまり，あるクラスタ A のメンバは，同じインスタンスであることを意味する．ここで，本稿で述べる実験においては，クラスタ数を既知であるとして実装した．以下，クラスタを**インスタンスクラスタ**と呼び，本節で述べた多視点インスタンス対応付け手法を**エピポーラ領域対応付け**と呼ぶ．

3.2 多視点インスタンスセグメンテーションへの応用

同クラスのインスタンスが密に存在する場面においてインスタンスセグメンテーションを行う際，頻繁に生じる問題の 1 つは，比較的大きく写るインスタンスであってもしばしば検出されないことである．多くのインスタンスセグメンテーション手法 [1,13] が，物体検出部 [12] に NMS [17] を用いている．そのため，例えば葉のインスタンスセグメンテーションを行う際，同じ「葉」クラスの物体領域ではあるが，本来異なる葉を検出している領域が NMS によって抑制されることでこの問題は生じる．NMS の閾値処理によってこの問題を解決しようとした場合，本来不要な物体領域が抑制されないなどして，全体としてインスタンスセグメンテーション精度が下がる可能性があり，根本的解決とはなり得ない．本節では， 3.1 節で述べたエピポーラ領域対応付けを用いることで，従来未検出となっていたイ

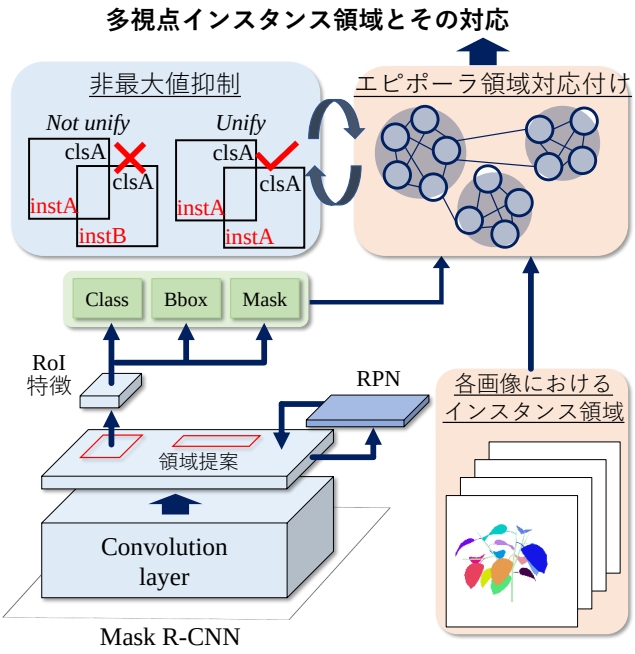

図 5 提案手法の MVIS 実装例．インスタンス検出と多視点対応付けを反復的に行うことで，両精度を改善する．

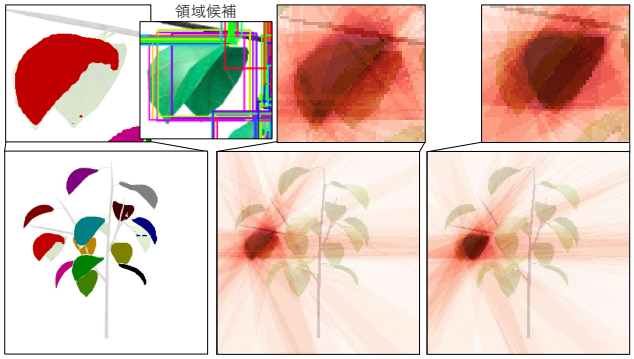

図 6 未検出インスタンスに対するエピポーラマップの例．エピポーラマップによって近傍インスタンスを区別することができる．

ンスタンスを保持する手法について述べる．

3.2.1 反復処理とエピポーラマップ

本手法は，Mask R-CNN のようなインスタンスセグメンテーションフレームワークの結果と，エピポーラ領域対応付けによって得られた多視点インスタンス対応が初期値となる．図 5 に示すように，インスタンスセグメンテーションと領域対応付け処理を 1 セットとし，これを反復的に実行することで，セグメンテーションと対応付けの両方の結果を改善することができる．具体的には，まずインスタンスセグメンテーションを行い， 3.1 節の手法を用いて多視点インスタンス対応群を得る．次に，多視点インスタンス対応群から後述するエピポーラマップを生成し，各検出領域にクラスタ ID を割り振ることでインスタンスセグメンテーション精度を改善する．その後再びエピポーラ領域対応付けを行い，同様の処理を反復的に実行する．

エピポーラマップの生成：各検出領域のクラスタ ID を推定するために，多視点インスタンス対応群（前の反復でのエピポーラー領域対応付け結果であるインスタンスクラスタ）を，3.1 節で述べたインスタンス投影と同じ方法で各視点に再投影する．ここで，インスタンスクラスタ k を $\{\mathcal{A}_k\} \subset \{\mathcal{A}\}$ とする．ただし $\mathcal{A}$ は，1 つ前の反復結果である多視点インスタンス対応群である．この時，画像 i 上に投影された $\{\mathcal{A}_k\}$ を，エピポーラーマップ $\mathcal{M}_{i,k}$ と定義する．$\mathcal{M}_{i,k}$ はクラスタ内のインスタンスのエピポーラー帯が重なり合ったものであり，各視点画像には，インスタンスクラスタの数だけエピポーラマップが生成される．$\mathcal{M}_{i,k}$ は以下の式で表せる．

$$\mathcal{M}_{i,k} = \frac{\sum_{\{\mathcal{A}_k\}} \mathcal{B}(\mathcal{A} \in \{\mathcal{A}_k\}, I_i)}{\#_c\{\mathcal{A}_k\}} \tag{4}$$

ここで，$\#_c\{\mathcal{A}_k\}$ はクラスタ $\{\mathcal{A}_k\}$ 内のインスタンスの数をカウントする関数である．

各検出領域のクラスタ ID 推定：図 6 に示すように，エピポーラマップは隣接する本来異なるインスタンスに対し，異なるクラスタ ID を割り当てることができる．実装では，各検出領域に対して，エピポーラーマップと検出領域との交差度の最高値に基づいてインスタンスにクラスタ ID を割り当てている．ここで，画像 i におけるインスタンスマスク r を $\mathcal{A}'_{i,r}$ とすると，2 つの領域 $\mathcal{M}_{i,m}$ と $\mathcal{A}'_{i,r}$ の類似度は，IoU の計算と同様の方法で得ることができる．エピポーラマップは $\{0,1\}$ の値をとらないため，2 つの非負領域における IoU の拡張である Ruzicka（weighted Jaccard）類似度 [18] を用いる．

$$s(\mathcal{M}_{i,k}, \mathcal{A}'_{i,r}) = \frac{\Sigma_{\mathbf{x}} \min(\mathcal{M}_{i,k}(\mathbf{x}), \mathcal{A}'_{i,r}(\mathbf{x}))}{\Sigma_{\mathbf{x}} \max(\mathcal{M}_{i,k}(\mathbf{x}), \mathcal{A}'_{i,r}(\mathbf{x}))}, \tag{5}$$

ここで，$\mathbf{x}$ は画像座標系におけるピクセルの座標を示す．式 (5) は，各インスタンスマスク領域 $\mathcal{A}'_{i,r}$ を $\{0,1\}$ 値を取る領域として扱うため，ピクセルがマスク内にある場合 1，外にある場合は 0 の値を取るものとして実装した．また IoU の定義は，Ruzicka 類似度において両方の領域が $\{0,1\}$ 値を取る特殊な場合である．検出領域 $\mathcal{A}'_{i,r}$ のインスタンス ID は，$s(\mathcal{M}_{i,k}, \mathcal{A}'_{i,r})$ が最大となる k が選択される．

3.3 三次元ボリューム復元への応用

提案する MVIS 手法から得られた多視点インスタンス対応群を用いることによって，インスタンス毎の多視点画像を容易に生成することができる． 図 7 に示すように，これを用いて三次元空間上にインスタンスの存在投票空間を生成することで，三次元ボリューム復元を行う．本手法は，伝統的なトモグラフィー手法である逆投影（back projection）法 [19] を踏襲したものである．ここで，ある場面における i 番目のインスタンスを $\mathcal{I}_i$ とし，三次元ボクセル空間から $\mathcal{I}_i$ の多視点画像の内，k 番目の画像への投影を関数 θ_k: $\mathbb{R}^3 \rightarrow \mathbb{R}^2$ で表す．この時，あるボクセル $\mathbf{x}_{3D} \in \mathbb{R}^3$ に

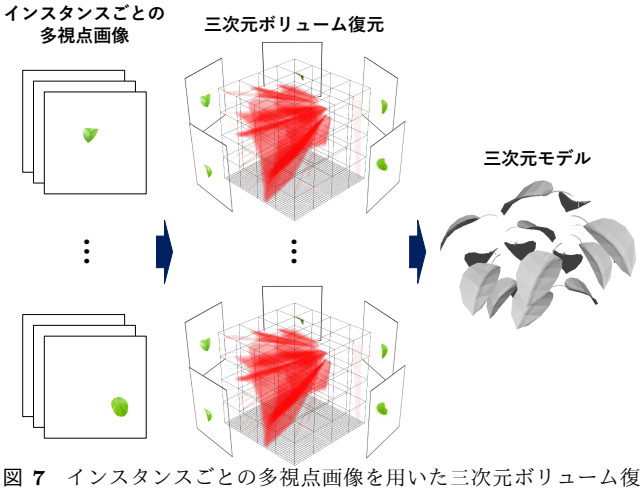

図 7　インスタンスごとの多視点画像を用いた三次元ボリューム復元の概略図

おける $\mathcal{I}_i$ の投票率 $\mathcal{I}_{3D_i}$ は次のように表せる．

$$\mathcal{I}_{3D_i}(\mathbf{x}_{3D}) = \sum_k \mathcal{L}_{2D_k}(\theta_i(\mathbf{x}_{3D})) \qquad (6)$$

ただし，$\mathcal{L}_{2D_k}$ はインスタンス $\mathcal{I}_i$ の多視点画像内 k 番目のものであり，領域内・外の画素値がそれぞれ 1，0 である画像を指す．式 (6) の計算後，投票率 0.5 以上の三次元ボクセル $\mathbf{x}_{3D}$ を，構築されたボクセル形状として出力する．

4. 実験

提案手法および三次元ボリューム復元手法の精度を評価するために，それぞれ実験を行った．

4.1 データセット

本節では，実験に用いたデータセットの詳細を示す．図 8 は各データセットのサンプル画像を示している．

CG 植物：植物モデリングに関する研究 [20] で使用されたデータセットに，一部修正を加えた CG 植物モデルのデータセットを用いた．本データセットの植物モデルは，自己組織木モデル [21] を使用して枝の分岐パターンと葉の位置を変更することで作成されている．Mask R-CNN の学習のために，14 種の CG 植物モデルそれぞれにおいて 32,760 視点からレンダリングを行った（画像の合計数：458,640）．評価用として，葉の数が異なる 4 つの植物モデル（4，8，16，32）を使用した．評価用の植物画像は，各植物モデルにつき 20 視点からレンダリングを行ったが，実験に用いた視点数を適宜間引いて使用した．実験には真値のカメラパラメータを用いた．

実植物 (大豆)：本データセットは，植物科学に関する研究用に作成された自動撮影システム [22] を用いて撮影された大豆植物の多視点画像である．Mask R-CNN の学習においては，COCO データセットで学習済みのものから同システムで撮影された 8 種の大豆植物の画像を使用して転移学習を行った．評価には 28 視点を用い，評価用植物の葉の数は 28 枚だった．以下，実画像を用いたデータセットにおいて，視点間のインスタンス対応，各視点のインスタンスマスクは手動で作成し，カメラパラメータは SfM [4] を用いて推定した．

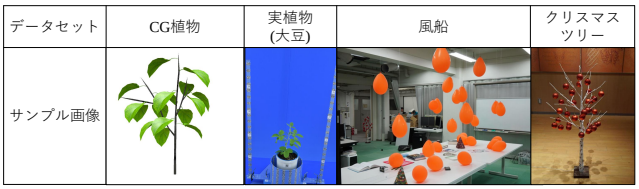

図 8　各データセットのサンプル画像

図 9　多視点対応付けに用いたカメラ位置 (表 1)

風船：植物以外の実環境を用いて，提案手法とその MVIS への応用手法の定量的評価，三次元復元の定性的評価を行った．Mask R-CNN の学習には，COCO データセットで学習済みのものから風船データセット[*1]の 61 枚の風船画像を使用して転移学習を行った．評価には，18 視点から 27 個の風船があるシーンを撮影し，その画像を用いた．

クリスマスツリー：クリスマスツリーに球形の 33 個の装飾を施して実験を行った．多視点インスタンス対応付けには 21 視点を用いた．装飾は球形であるため，風船データセットで学習済みの Mask R-CNN から，25 枚のクリスマスツリー画像を用いて転移学習を行った．

4.2 多視点インスタンス対応付け結果

上記データセットを用いて，提案手法であるエピポーラ領域対応付けと，その応用である MVIS 手法の多視点インスタンス対応付け精度を評価した．提案手法は，各視点のインスタンス画像を入力として想定している．そこで，純粋に提案手法の精度を測るために，真値のインスタンスマスクを入力した場合の精度を用いて評価し，通常のインスタンスセグメンテーション結果を入力とした MVIS と対比して後述している．また，実環境である大豆，風船，クリスマスツリーの真値は手動で作成した．

4.2.1 ベースライン：MVS ベース対応付け

MVIS は新しい問題であるため，多視点インスタンス対応付けのベースライン手法は確立されていない．そこで特徴点対応に基づいた手法として広く使われている MVS を採用し，提案手法と MVS に基づくインスタンス対応付け手法（以下，**MVS ベースの対応付け手法**）と比較した．

[*1] https://github.com/matterport/Mask_RCNN/tree/master/samples/balloon

# of cameras / Method / # leaves		20 視点	10 視点	5 視点	平均
MVS-based matching	4	0.960	0.979	N/A	0.565
	8	0.882	0.912	N/A	
	16	0.711	0.892	0.329	
	32	0.696	0.424	N/A	
MVIS w/ Mask R-CNN (iter #1)	4	**1.000**	**1.000**	0.750	0.865
	8	0.974	0.948	**0.973**	
	16	0.903	0.893	0.844	
	32	0.776	0.710	**0.606**	
MVIS w/ Mask R-CNN (iter #3)	4	**1.000**	**1.000**	**1.000**	**0.900**
	8	**0.987**	**0.961**	**0.973**	
	16	**0.953**	**0.916**	**0.905**	
	32	**0.827**	**0.729**	0.541	
MVIS w/ GT mask	4	1.000	1.000	1.000	0.943
	8	1.000	1.000	1.000	
	16	0.991	0.943	0.962	
	32	0.840	0.814	0.771	

表 1 CG 植物を用いた多視点インスタンス対応付け精度．N/A は MVS に失敗したことを意味する．また太字は，各列（視点）において同じ植物に対する手法間の精度を比較した際の最も高い値を示す．

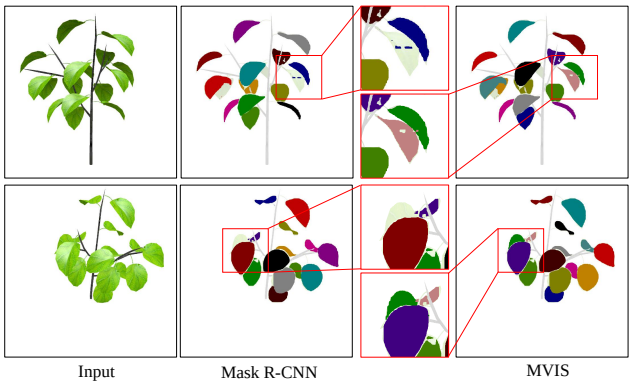

図 10 インスタンス検出の改善例．従来 NMS によって抑制されていたインスタンスを MVIS では検出できている．

MVS ベースの対応付け手法では，カメラパラメータを使用して，COLMAP [5] によって再構築された密な三次元点群を使用した．ただし，カメラパラメータには，CG 植物のデータセットは真値を，実環境データセットは SfM [4] を用いて推定したものを用いた．

ベースラインとなる MVS ベースの対応付け手法について概説する．まず提案手法と同じ入力画像から SfM/MVS を行い，三次元点群を抽出する．次に，i 番目の視点における m 番目のインスタンスである $\mathcal{A}_{i,m}$ に対して，三次元点 $\mathbf{p}_{\mathcal{A}_{i,m}}$ を選択する．ただし，遮蔽が存在を考慮し，$\mathbf{p}_{\mathcal{A}_{i,m}}$ は，$\mathcal{A}_{i,m}$ の領域内に投影された点群の中でカメラに最も近い点とする．次に，$\mathbf{p}_{\mathcal{A}_{i,m}}$ を他の視点に投影する．例えば，$\mathbf{p}_{\mathcal{A}_{i,m}}$ が j 番目の視点における n 番目のインスタンス $\mathcal{A}_{j,n}$ に投影される場合，$\mathcal{A}_{i,m}$ および $\mathcal{A}_{j,n}$ は対応するインスタンスと判定する．その後，提案手法と同様に，$|V|$ ノードを持つ対応付けグラフを構築し，SymNMF を使用してグラフクラスタリングを行った．ただし，エッジの重みは対応を意味している．上記例の場合，$\mathcal{A}_{i,m}$ と $\mathcal{A}_{j,n}$ の間のエッジは 1 で重み付けされる．

Method	大豆 28 obj. 28 視点	風船 27 obj. 18 視点	クリスマスツリー 33 obj. 21 視点
MVS-based	N/A	N/A	N/A
MVIS (iter #1)	0.464	0.691	0.358
MVIS (iter #3)	0.461	0.732	0.363

表 2 実環境における多視点インスタンス対応付け精度：MVIS は，Mask R-CNN を用いた提案手法（MVIS w/ Mask R-CNN）

4.2.2 評価尺度

対応付けの精度 s_{match} は，全インスタンスの数 $|V|$ に対する正しく分類されたインスタンスの割合を用いて評価を行った．クラスタリング問題として多視点対応付け問題を解くために，推定したインスタンスクラスタと真値クラスタを関連付ける必要がある．そこで k_{est} に属するインスタンス間で k_{gt} のモードを計算することにより，推定クラスタ k_{est} の ID に対応する真値クラスタの ID，k_{gt} を決定した．

$$s_{match} = \frac{\sum_{k_{est}} \max_{k_{gt} \in \{k_{gt}\}} |\mathrm{id}(\mathcal{A}_{k_{est}}) \cap k_{gt}|}{|V|} \quad (7)$$

ここで $\mathrm{id}(\mathcal{A}_{k_{est}})$ はインスタンス $\mathcal{A}_{k_{est}}$ の推定 ID を指し，$\{k_{gt}\}$ はインスタンスクラスタ ID の真値の集合を指す．式 (7) は，正しい対応付けの数を多視点画像間のインスタンスの総数である $|V|$ で除算している．s_{match} はクラスタリングの評価尺度である純度と同等の意味を持つ．

Mask R-CNN の MVIS 実装においては，対応付け精度の評価とは別に，生成されたインスタンスマスクと真値インスタンスマスク間の対応を決定する必要もある．そこで Mask R-CNN が生成した各インスタンスごとに，最大の IoU を持つ真値インスタンスを対応するインスタンスとして実装した．

4.2.3 結果

CG 植物を用いた対応付け精度を 表 1 に示す．提案手法の精度と視点数との関係を調べるため， 図 9 に示すようにカメラ数を変化させた結果を示している．ここで，カメラは 20 視点，10 視点，5 視点それぞれにおいて，半球状かつ水平・鉛直の両方向へ均等に分散するよう配置した．また，手法 **MVIS w/ GT mask** は真値のインスタンスマスクを入力として用いた手法であり，提案手法の上界を評価するものである．Mask R-CNN を用いた提案手法の MVIS 実装（**MVIS w/ MaskR-CNN**）は，最初（1 反復目）と最後（3 反復目）の精度を示している．

MVS ベースの対応付け手法の平均精度が 56.5% だった一方で，Mask R-CNN 拡張の MVIS 手法は平均精度 90.0%

# objects		シルエット	平均	提案手法	平均
20 視点	4	1.692	1.593	**0.086**	**0.132**
	8	1.248		**0.086**	
	16	2.303		**0.125**	
	32	1.127		**0.230**	
10 視点	4	1.858	1.707	**0.082**	**0.138**
	8	1.396		**0.087**	
	16	2.269		**0.129**	
	32	1.306		**0.252**	
5 視点	4	2.004	1.874	**0.118**	**0.203**
	8	1.540		**0.116**	
	16	2.361		**0.190**	
	32	1.593		**0.389**	
平均		1.978		**0.210**	

表 3 CG 環境における三次元復元誤差（相対値）．各行においてより良い精度を示す値を太字でハイライトしている．

となり，真値マスクを用いた場合に近い精度に達した．これは，本研究が対象とする場面における提案手法の有用性を示している．さらに，ほとんどの場合において反復処理の結果対応付け精度が向上していることから，インスタンス検出と多視点対応付けの反復処理は対応付け精度を向上させ得ることがわかった．しかし，32 枚の葉がついた植物を 5 視点から撮影した場合は反復処理によって対応付け精度が減少していることから，反復処理を効果的に機能させるためには，インスタンス数に対して一定の視点数を確保する必要があると考えられる．また，図 10 に示すように，本 MVIS 実装を用いることで，従来 NMS によって抑制されていたインスタンスを検出することができた．

また実環境での実験においても， 表 2 に示すように CG 環境での実験と同様の傾向が見られた．風船やクリスマスツリーに比べて大豆データに対する精度が大きく低い理由として，画像サイズに対して対象物が小さかったことが，インスタンス検出精度に影響したためであると考えられる．

4.3 三次元復元結果

本提案手法によって得た多視点画像間のインスタンス対応から，各インスタンスごとの多視点画像を作成し， 3.3 節にて述べた三次元復元手法を用いて実験を行った．

4.3.1 ベースライン：シルエットベース三次元復元

三次元復元精度に対する多視点画像間でのインスタンス対応を得る効果を検証するために，視点間でインスタンス対応をとらず，各視点で検出したインスタンス群すべてを 1 つのシルエットとして用いてボクセル復元を行う手法をベースラインとして採用し，比較を行った．

4.3.2 評価尺度

復元したモデルおよび真値モデルの双方向ユークリッド距離 [23] の平均を評価尺度として用いる．復元したモデルおよび真値モデルそれぞれの頂点を $\mathbf{g} \in \mathcal{G}$ および $\mathbf{t} \in \mathcal{T}$ で表すと，評価尺度は以下のように表せる．

$$d(\mathcal{G},\mathcal{T}) = \frac{1}{2}\left(\frac{\sum_{\mathcal{G}} ||\mathbf{g} - N_{\mathcal{T}}(\mathbf{g})||}{|\mathcal{G}|} + \frac{\sum_{\mathcal{T}} ||\mathbf{t} - N_{\mathcal{G}}(\mathbf{g})||}{|\mathcal{T}|}\right) \tag{8}$$

ここで，$N_{\mathcal{G}}(\mathbf{x})$ および $N_{\mathcal{T}}(\mathbf{x})$ は，点 $\mathbf{x}$ から点群 $\mathcal{G}$ および $\mathcal{T}$ の最近傍点を取得する関数であり，$|\mathcal{G}|$ および $|\mathcal{T}|$ はそれぞれ点群 g および T の頂点数を表す．ただし，本研究で復元される三次元モデルはスケールが可変であるため，評価尺度の数値はスケールによって変動する．

4.3.3 結果

表 3 は， 表 1 の各結果を入力として三次元復元を行った際の精度を示している．いずれの設定においても提案手法はベースライン手法の誤差を大きく上回っている．提案手法の誤差は，20 視点を入力した場合，葉の長さの 6.8% 程度であった．ベースライン手法を用いた場合，本来葉が存在しない領域を除くことができないため，復元誤差が増大したと考えられる．本結果は, 提案手法が多視点画像間でインスタンスを対応付けることによって，インスタンスが存在し得ない領域を効果的に排除できることを示している．

4.4 失敗例

実験結果は，提案する MVIS 手法が多視点画像間でインスタンスの対応を効果的に推定できることを示している．しかし一方で，いくつかの失敗例も見られた．インスタンス検出における典型的な失敗例として，ほぼ完全に遮蔽されたインスタンスのように，Mask R-CNN のフレームワークよって低い物体クラススコアが付いたインスタンスの未検出が挙げられる．このような場合，NMS で抑制されなかったとしても，クラススコアが低いことから検出されない．インスタンスクラスのスコアは，元の Mask R-CNN フレームワーク内のインスタンスのクラス分類を行う部分で付与されており，本実装でもそのまま用いている．

5. まとめ

本稿では，類似のインスタンスや遮蔽が多く存在する場面を対象として，インスタンス領域を視点間で幾何学的に対応付ける手法を提案した．提案手法はテクスチャ情報に依存しておらず，エピポーラ拘束に基づいて領域対応付けを行うことができる．また本論文では，インスタンスセグメンテーションを多視点設定へと拡張した多視点インスタンスセグメンテーション（MVIS）および，インスタンスごとの三次元復元の応用を紹介した…実験では，16 枚の葉を持つ植物に対して，真値のマスクを入力として用いた場合，提案手法の精度は 20 視点で 99.1%，5 視点で 96.2% に達した．この結果から，各視点において正確なインスタンス領域が得られた場合，提案手法は十分に対応付けを行えることを示した．さらに三次元ボリューム復元に関する実験では，シルエットベースの三次元復元に対して大きな

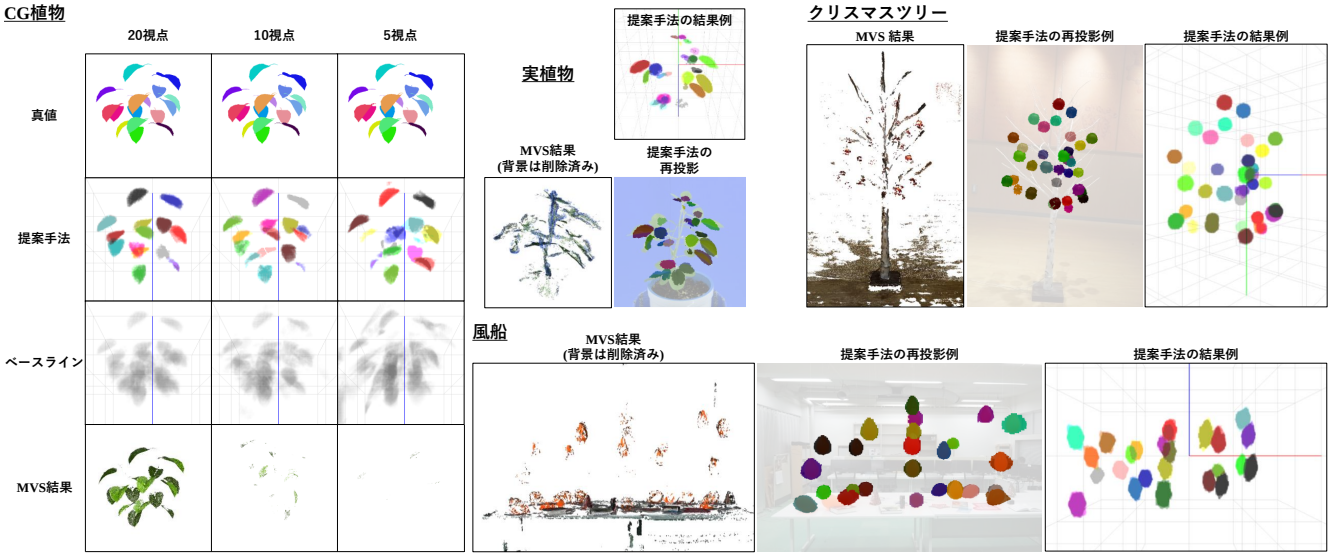

図 11 三次元復元結果の例：CG 植物のベースラインはシルエットベースの手法，実画像は MVS の結果を参考として示す．

改善が見られた．本研究はコンピュータビジョン研究としての側面以外にも，潜在的な応用先として植物の成長分析などの植物科学・農学応用が考えられる．

謝辞 本研究の一部は，JST さきがけ JPMJPR17O3 の支援を受けたものである．

参考文献

[1] He, K., Gkioxari, G., Dollár, P. and Girshick, R.: Mask R-CNN, *ICCV* (2017).
[2] Wang, W., Yu, R., Huang, Q. and Neumann, U.: SGPN: Similarity group proposal network for 3D point cloud instance segmentation, *CVPR* (2018).
[3] Lahoud, J., Ghanem, B., Pollefeys, M. and Oswald, M. R.: 3D instance segmentation via multi-task metric learning, *ICCV* (2019).
[4] Schönberger, J. L. and Frahm, J.-M.: Structure-from-motion revisited, *CVPR* (2016).
[5] Schönberger, J. L., Zheng, E., Pollefeys, M. and Frahm, J.-M.: Pixelwise view selection for unstructured multi-view stereo, *ECCV* (2016).
[6] Lowe, D. G.: Distinctive image features from scale-invariant keypoints, *Int. J. on Computer Vision*, Vol. 60, No. 2, pp. 91–110 (2004).
[7] Bleyer, M., Rhemann, C. and Rother, C.: PatchMatch Stereo - Stereo Matching with Slanted Support Windows., *BMVC* (2011).
[8] Matas, J., Chum, O., Urban, M. and Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions, *BMVC* (2002).
[9] Li, L., Zhang, S., Yu, X. and Zhang, L.: PMSC: PatchMatch-based superpixel cut for accurate stereo matching, *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 28, No. 3, pp. 679–692 (2016).
[10] Klaus, A., Sormann, M. and Karner, K.: Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure, *ICPR* (2006).
[11] Dellaert, F., Seitz, S. M., Thorpe, C. E. and Thrun, S.: Structure from motion without correspondence, *CVPR* (2000).
[12] Ren, S., He, K., Girshick, R. and Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks, *NeurIPS* (2015).
[13] Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B. B. G., Geiger, A. and Leibe, B.: MOTS: Multi-object tracking and segmentation, *CVPR* (2019).
[14] Narita, G., Seno, T., Ishikawa, T. and Kaji, Y.: PanopticFusion: Online volumetric semantic mapping at the level of stuff and things, *IROS* (2019).
[15] Nassar, A. S., Lefèvre, S. and Wegner, J. D.: Simultaneous multi-view instance detection with learned geometric soft-constraints, *ICCV* (2019).
[16] Kuang, D., Ding, C. and Park, H.: Symmetric nonnegative matrix factorization for graph clustering, *SIAM Int. Conf. on Data Mining (SDM)* (2012).
[17] Girshick, R., Iandola, F., Darrell, T. and Malik, J.: Deformable part models are convolutional neural networks, *CVPR*, pp. 437–446 (2015).
[18] Deza, M. M. and Deza, E.: *Encyclopedia of Distances*, Springer (2009).
[19] Brooks, R. A. and Chiro, G. D.: Theory of image reconstruction in computed tomography., *Radiology*, Vol. 117, pp. 561–72 (1975).
[20] Isokane, T., Okura, F., Ide, A., Matsushita, Y. and Yagi, Y.: Probabilistic plant modeling via multi-view image-to-image translation, *CVPR* (2018).
[21] Palubicki, W., Horel, K., Longay, S., Runions, A., Lane, B., Měch, R. and Prusinkiewicz, P.: Self-organizing tree models for image synthesis, *ACM Trans. on Graphics*, Vol. 28, No. 3, Article No. 58 (2009).
[22] Tanabata, T., Hayashi, A., Kochi, N. and Isobe, S.: Development of a semi-automatic 3D modeling system for phenotyping morphological traits in plants, *Conf. IEEE Industrial Electronics Society (IECON)* (2018).
[23] Zhu, J., Du, S., Yuan, Z., Liu, Y. and Ma, L.: Robust affine iterative closest point algorithm with bidirectional distance, *IET Computer Vision*, Vol. 6, pp. 252–261 (2012).