Vol.2022-CVIM-229 No.5
2022/3/10

全方位カメラを使用したテクスチャレスで密集したブドウ果粒の重心位置推定

田村 泰斗[1,a)]　内海 ゆづ子[2,b)]　三輪 由佳[3,c)]　岩村 雅一[2,d)]　黄瀬 浩一[2,e)]

概要：日本のブドウ栽培に特有な摘粒作業は習熟が難しいため，情報技術による作業習得支援が望まれる．本研究では，間引くブドウ果粒を自動選定する摘粒システムの実現を目指す．その実現には，ブドウ果粒の 3 次元位置情報の推定が必要である．ブドウ房が多くぶら下がった屋外の棚状ブドウ圃場での深度センサーの使用は困難で，動画によるブドウ果粒の 3 次元位置推定が望ましい．撮影の利便性のため，画角の広い全方位カメラの使用が望まれる．また，ブドウ果粒は表面がテクスチャレスで形状の対称性が高く，かつ密集している．そのため，局所特徴量のマッチングを使用した従来の 3 次元復元と深度推定が困難である．そこで本稿では，こうした実用上の制約に対応すべく，深層学習による全方位カメラの教師なし単眼深度推定手法とカメラ位置推定によるブドウ果粒の重心位置推定の手法を提案する．実験の結果，テクスチャレスで密集した球状物体であるブドウ果粒の 3 次元位置情報が推定できることを定性的に確認した．

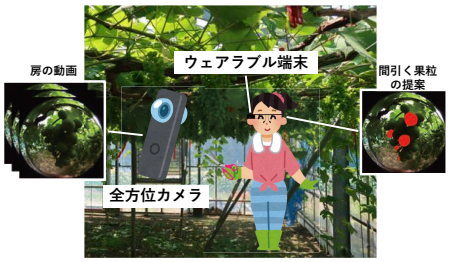
1. はじめに

日本のブドウ生産の多岐にわたる作業工程の中でも，特に生産者への負担が大きいのが摘粒である．摘粒は，ブドウが生育する過程で果粒（実）同士の間隔を広げることを目的として，ブドウの果粒が成った後の早期の段階で一定数の果粒を切り取る作業である．この作業は果粒の大きさと糖度を出荷可能な水準にし，房を逆三角形に整形するために重要である．切り取る果粒の選定には，果粒の間隔や房の各部位の果粒の数など，複数の基準がある．その上，定められた生育段階で圃場の全ての房の摘粒を完了させる必要があることから，摘粒には素早い作業も要求される．このように，摘粒は基準が多く複雑な作業であるだけなく，俊敏さが求められることから，習熟に時間がかかる．そのため，情報技術を使った摘粒作業習得支援が期待される．

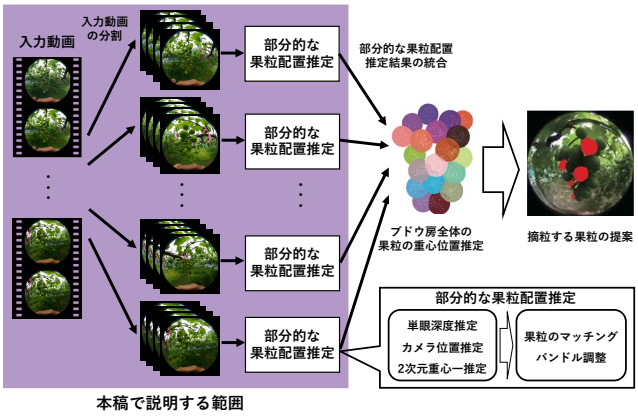
[1] 大阪府立大学工学域情報工学課程
Collage of Engineering, Osaka Prefecture University, 1–1, Gakuencho, Naka, Sakai, Osaka 599–8531, Japan
[2] 大阪府立大学大学院工学研究科
Graduate School of Engineering, Osaka Prefecture University, 1–1, Gakuencho, Naka, Sakai, Osaka 599–8531, Japan
[3] 大阪府立環境農林水産総合研究所
Research Institute of Environment, Agriculture and Fisheries, Osaka Prefecture, 442, Shakudo, Habikino, Osaka 583–0862, Japan
[a)] tamura@m.cs.osakafu-u.ac.jp
[b)] yuzuko@cs.osakafu-u.ac.jp
[c)] MiwaY@mbox.kannousuiken-osaka.or.jp
[d)] masa@cs.osakafu-u.ac.jp
[e)] kise@cs.osakafu-u.ac.jp

本研究では，切り取るべき果粒を教示することで摘粒作業習得を支援する．そして，これを実現するため，図 1(a) に示すようなシステムの構築を目標とする．このシステムでは，利用者が房を撮影して，撮影した房の望ましい摘粒結果を画像で示す．一般的な画角の狭い透視投影のカメラを使用して，人の身長ほどのブドウ棚でブドウ房をフレームアウトせずに撮影するのは困難である．そのため，本稿では撮影中に房がフレームアウトするのを防ぐために，画角の広い全方位カメラの片面のみ（視野角 180 度）を使用する．摘粒の基準を満たす果粒の選定のためには，果粒の 3 次元位置を推定する必要がある．果粒の 3 次元情報の推定には，まず深度センサーの使用が考えられる．しかし，既存手法 [1, 2] では室内環境での比較的大きな深度センサーの使用を想定しており，照明が安定しない圃場での使用は困難である．したがって，より簡単に安定してデータ取得できる単眼カメラを使用した手法がより望ましい．そして果粒の 3 次元情報取得には，複数の視点から得られた画像に基づく対象物体の 3 次元位置推定をする必要がある．

一般に単眼カメラで撮影した画像から被写体の 3 次元情報を取得するには，被写体の対応点のマッチングとカメラ位置姿勢の推定が必要である．3 次元復元の代表的な手法である Visual Simultaneous Localization and Mapping (Visual SLAM) [3] や Structure from Motion (SfM) [4] などは，対応点となりうる点を特徴点として検出し，特徴点付近のテクスチャの違いを表現する局所特徴量を頼りにフレーム間で対応点を求め，カメラ位置姿勢の推定と 3 次元



(a) システムの概要



(b) ブドウ果粒の重心位置推定手法

図 1: 動画を使用した摘粒作業習得支援システム

形状を復元する．しかし，テクスチャレスなブドウ果粒では，特徴点がそもそもあまり検出されず，フレーム間の対応点を求められない．テクスチャレスな被写体に対応した手法として，物体の輪郭やその表面から検出される曲線のマッチング [5–7]，または各フレームの物体の検出領域とエピポーラ幾何を使用したマッチング [8] が提案されている．しかし，摘粒前のブドウ房には多くの果粒が密集して互いに重なり合い，各果粒の形状の対称性や類似性が高いことから，マッチングが失敗する可能性が高い．そのため，これらの 3 次元復元手法も有効でないと考えられる．

近年，深層学習を用いてカメラ位置姿勢と 3 次元情報を推定する教師なし単眼深度推定 [9–11] が提案されている．この手法では，2 つのフレーム間の対応をネットワークで学習し，2 つのフレームが入力されると，3 次元情報（深度マップ）と 2 つのフレーム間のカメラ位置姿勢（相対位置）を出力する．従来の特徴点や輪郭，または輝度値の分布を頼りにフレーム間の対応を求める手法と異なり，教師なし単眼深度推定では，学習に基づいて画像全体での対応を求める．そのため，対応点を検出できないブドウのようなテクスチャレスな画像でも 3 次元復元が可能と考えられる．多くの教師なし単眼深度推定では，2 つのフレーム間の対応を表すフレームの再投影誤差を differentiable depth-image-based-rendering (differentiable DIBR) を用いて計算する [9–11]．しかし differentiable DIBR は透視投影カメラ画像を想定しており，全方位カメラ画像に直接適用できない．

そこで本稿では，differentiable DIBR を全方位カメラに合わせて改良した DIBR for Unified Omnidirectional Camera Model (DIBR for UOCM) を提案して，テクスチャレスなブドウ果粒の 3 次元重心位置とカメラ位置姿勢を直接推定する．一般に単眼深度推定問題はスケールの曖昧さを持ち，フレームを再投影可能な深度が各ピクセルに対して何通りも存在する [10, 11]．そのため，DIBR for UOCM を使用した単眼深度推定では，スケールの曖昧さが原因での深度推定の発散が頻繁に起こる．これを防ぐために，提案手法では損失関数に Scale-Aware Constraint Loss を導入する．実験では，提案手法をブドウ果粒の重心位置推定に使用した場合の有効性を定性的に確認する．

2. 関連研究

2.1 ブドウの 3 次元形状の推定や計測

ブドウ房の 3 次元形状を推定した研究は，深度センサを用いるものと画像から 3 次元形状を推定するものの 2 つに分けられる．まず，深度センサを使用したブドウ果粒の形状取得の手法が提案されている [1, 2]．これらは室内環境を想定する．またいずれもセンサの機材が大きく，摘粒作業を想定するブドウ圃場での使用は難しい．

画像から 3 次元復元を行い，ブドウの形状を計測する手法も提案されている．[12], [13] はそれぞれ Visual SLAM や SfM を使用したブドウの計数，圃場の各領域の物体の識別手法である．これらの手法はブドウの生産量の予測を目的とした補助的な要素として 3 次元復元を使用する．そのため，ブドウ房の果粒位置を推定していない．

[14] は，Visual SLAM によるカメラ位置姿勢推定とブドウ果粒の 2 次元重心位置を使用してブドウ房の果粒の 3 次元位置を推定する．テクスチャレスなブドウでは対応をとる特徴点の検出とそのフレーム間での対応づけが難しいため，[14] は各ブドウ果粒の重心位置のマッチングを行う．しかし，[14] ではブドウ果粒の重心位置のマッチングの誤対応が起き，果粒の 3 次元位置推定の精度は低い．

2.2 テクスチャレスな物体の 3 次元復元

複数視点からの画像，もしくは動画を使用して被写体の 3 次元復元を行う代表的な手法に SfM [4] や Visual SLAM [3, 15] がある．これらの手法は，被写体に特徴点となる部分が多く含まれていることを前提としてフレーム間の対応を取り，カメラの位置姿勢や 3 次元復元する．そのためテクスチャレスな物体に対応するためには，特徴点以外の方法でフレーム間の対応を推定する必要がある．

テクスチャレスな物体に対応するため，物体の輪郭や検出される曲線を使用した SfM 手法がある [5–7]．いずれも遮蔽の無い単一物体の 3 次元復元や，検出される曲線が一意に決まることを想定している．そのため，遮蔽が多発し，類似した曲線を持つブドウの画像では曲線のマッチングが

難しく，適用が困難と考えられる．

[8] はエピポーラ幾何を基に，植物の葉のような形状が類似したテクスチャレスな物体の対応付けと 3 次元復元をする．各フレームで物体領域とカメラ位置が既知であるとして，注目フレームの各物体領域のなすエピポーラ線の集合 (エピポーラバンド) を使用する．そして，他フレームの物体領域とエピポーラバンドの重なり具合に基づいて物体領域のフレーム間でマッチングを行う．摘粒前のブドウ房は果粒の 3 次元配置が房に関してある程度の対称性をもち，密集する．そのため，同じ果粒のエピポーラバンドに対して同程度の重なりを持つ果粒が複数存在すると考えられる．そのため，この手法は本研究で想定するブドウ果粒のマッチングには向かない．

2.3 深度推定

近年，深度の正解データを必要としない深層学習ベースの単眼深度推定が大きく発展した [9–11]．これらの多くは，注目したフレームに隣接するフレームの再投影誤差を最小にするよう学習することで，入力フレームの深度推定とカメラ位置姿勢を推定する．しかし，これらの手法は一般的な透視投影カメラの使用を想定するため，全方位カメラを使用したブドウの 3 次元復元には向かない．

全方位カメラなど画角の広いカメラにおける深度推定の手法も提案されている．深層学習を使用しない手法として [16] があるが，この手法はカメラ位置が既に推定されていることを前提とした手法である．深層学習を用いた広角カメラ画像での深度推定も提案されているものの [17–19]，学習に奥行きの正解データを必要とする．また，全方位カメラを含む任意のカメラモデルでの教師なし深度推定も提案されている [20] が，カメラモデルの違い対応するため，新たなネットワークを追加する必要がある．

3. 従来手法：ピンホールカメラモデルでの教師なし単眼深度推定

3.1 概要

教師なし単眼深度推定の従来手法として，differentiable DIBR [9] を用いる手法を説明する．図 2 は differentiable DIBR を用いた教師なし単眼深度推定の全体像を示したものである．図 2(a) に示すように，この手法では動画中のある時刻 t のフレームをターゲットフレーム I_t，その直前 $t-1$ あるいは直後 $t+1$ のフレームをソースフレーム $I_{t'}$ と呼ぶ．depth network を用いれば，ターゲットフレームのみからその深度マップ D_t を推定できる．また，pose network を用いれば，ターゲットフレームとソースフレームから隣接カメラの相対位置 $\boldsymbol{T}_{t\to t'}$ が求められる．

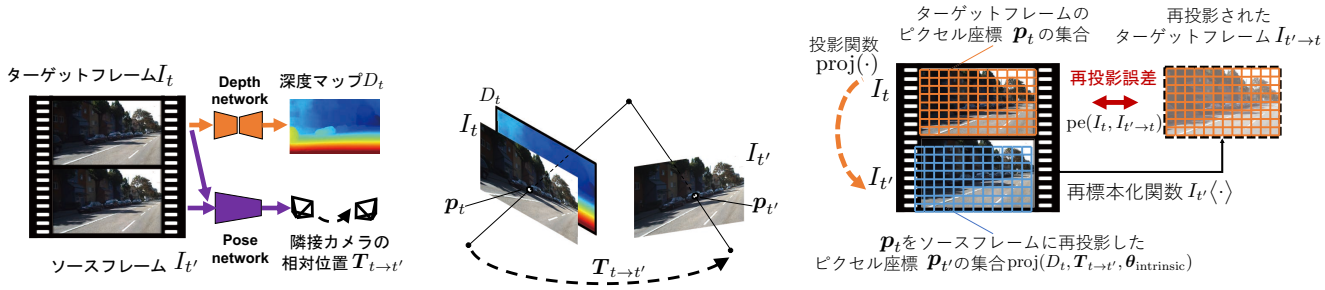
これら 2 つのネットワークの学習には，図 2(c) に示す再投影誤差 $\mathrm{pe}(I_t, I_{t'\to t})$ を用いる．これにより，教師データを用いずに学習できる．その仕組みを図 2(b)，図 2(c) と順を追って説明する．まず，図 2(b) に示すように，深度マップ D_t と 6 自由度の相対的なカメラ位置 $\boldsymbol{T}_{t\to t'}$ を用いれば，ターゲットフレーム I_t 中の画素 $\boldsymbol{p}_t$ が対応付く画素 $\boldsymbol{p}_{t'}$ をソースフレーム $I_{t'}$ 中で見つけられる．この対応付けをターゲットフレーム I_t 内の全ての画素について行うのが図 2(c) の投影関数 $\mathrm{proj}(\cdot)$ である．図 2(c) は，投影関数 $\mathrm{proj}(\cdot)$ により，ターゲットフレーム I_t とソースフレーム $I_{t'}$ の間で対応する画素を見つける．そして，それらの点を基にソースフレームの再標本化を行い，ターゲットフレームを再投影する．それらターゲットフレームの画素の画素値の差（再投影誤差）を算出して，同じになるように誤差逆伝播法で学習することを表している．ここで，投影関数 $\mathrm{proj}(\cdot)$ は深度マップを用いて対応を求めるのに対して，再投影誤差は RGB チャネルの輝度値で計算することに注意が必要である．投影関数 $\mathrm{proj}(\cdot)$ の出力は，対応する画素の座標（実際はその集合）であるが，それは整数では無く，実数である．そのため，そのままでは画素と画素の間を指すことが多い．そこで，バイリニア補間の要領で，画素と画素の間の画素値を計算する．この処理を行うのが再標本化関数 $I_{t'}\langle\cdot\rangle$ であり，その出力が「再投影されたターゲットフレーム」$I_{t'\to t}$ である．つまり，実際に再投影誤差を計算する際には，ターゲットフレーム I_t の画素値と再投影されたターゲットフレーム $I_{t'\to t}$ の画素値の差を用いる．

3.2 定式化

前節で述べたように，図 2(c) に示す Differentiable DIBR では，再投影誤差を計算する際に投影関数 $\mathrm{proj}(\cdot)$ と再標本化関数 $I_{t'}\langle\cdot\rangle$ を用いる．再投影誤差を定式化する準備として，それら 2 つの関数を最初に説明する．

投影関数は $\mathrm{proj}(D_t, \boldsymbol{T}_{t\to t'}, \boldsymbol{\theta}_{\mathrm{intrinsic}})$ の形で表される．前節では，投影関数が深度マップ D_t と 6 自由度の相対的なカメラ位置 $\boldsymbol{T}_{t\to t'}$ によって決まると述べたが，実際には既知の内部パラメータである $\boldsymbol{\theta}_{\mathrm{intrinsic}}$ の関数でもある．投影関数は，ターゲットフレーム I_t の全ての画素 $\{\boldsymbol{p}_t\}$ について，ソースフレーム $I_{t'}$ 内で対応付く画素の座標 $\{\boldsymbol{p}_{t'}\}$ を返す関数である．そのため，ターゲットフレームの画素の数だけ 2 次元座標を出力する．なお，ソースフレームの画面外に対応付いた画素は無視される．投影関数は，使用するカメラモデルによって決まる関数である．従来手法ではピンホールカメラモデルを用いており，本稿の提案手法では全方位カメラモデルを用いる．再標本化関数 $I_{t'}\langle\cdot\rangle$ は，$\mathrm{proj}(D_t, \boldsymbol{T}_{t\to t'}, \boldsymbol{\theta}_{\mathrm{intrinsic}})$ の出力に基づいてソースフレーム $I_{t'}$ を再標本化する．本稿では [9] に倣って，この再標本化に Spatial Transformer Networks [21] の微分可能なバイリニア補間を用いる．上記 2 つの関数を用いると，「再投影されたターゲットフレーム」$I_{t'\to t}$ は次式で与えられる．

$$I_{t'\to t} = I_{t'}\big\langle \mathrm{proj}(D_t, \boldsymbol{T}_{t\to t'}, \boldsymbol{\theta}_{\mathrm{intrinsic}})\big\rangle \qquad (1)$$



(a) 教師なし単眼深度推定　(b) フレームの再投影　(c) Differentiable DIBR

図 2: Differentiable DIBR を用いる教師なし単眼深度推定

再投影誤差は次式で与えられる．

$$L_{\mathrm{reprojection}} = \sum_{t'} \mathrm{pe}(I_t, I_{t'\to t}) \tag{2}$$

ここでソースフレームの時刻 t' は $t' \in \{t-1, t+1\}$ である．また $\mathrm{pe}(\cdot)$ は次式に示すように，[10] と同様に L1 ノルムと SSIM [22] をハイパーパラメータ α で重みづけした．

$$\mathrm{pe}(I_a, I_b) = \frac{\alpha}{2}(1 - \mathrm{SSIM}(I_a, I_b)) + (1-\alpha)||I_a - I_b||_1 \tag{3}$$

さらに，[10] と同様に，深度推定における以下の smoothness loss を損失関数に加える．

$$L_{\mathrm{smoothness}} = |\partial_x d_t^*| e^{-|\partial_x I_t|} + |\partial_y d_t^*| e^{-|\partial_y I_t|} \tag{4}$$

ただし d_t^* は正規化された逆深度のマップである．この正規化は [11] で提案されているように，深度推定結果の発散を防ぐために使用される．

3.3 ピンホールカメラモデルにおける投影関数

ピンホールカメラモデルを用いる場合の投影関数を説明する．前述のように，投影関数は $\{\boldsymbol{p}_t\}$ から $\{\boldsymbol{p}_{t'}\}$ への写像を与える関数である．したがって，ソースフレームの画素 $\{\boldsymbol{p}_t\}$ がターゲットフレームの画素 $\{\boldsymbol{p}_{t'}\}$ に写像される法則をピンホールカメラモデルに基づいて計算すれば良い．ピンホールカメラモデルでは，D_t は I_t の深度 z_t のマップで，$\boldsymbol{\theta}_{\mathrm{intrinsic}}$ は内部パラメータ $\boldsymbol{K}$ の要素を並べたベクトルとなる．なお，4 で説明する提案手法ではこの部分を全方位カメラモデルで置き換える．

3 次元点 $\boldsymbol{x}_t = (x_t, y_t, z_t)^\top$ を正規化画像平面にピンホールカメラモデルで投影した点 $\boldsymbol{p}_t$ とその同時座標 $\tilde{\boldsymbol{p}}_t$ は

$$\tilde{\boldsymbol{p}}_t = \begin{pmatrix} \boldsymbol{p}_t \\ 1 \end{pmatrix} = \frac{1}{z_t}\boldsymbol{K}\boldsymbol{x}_t \tag{5}$$

で与えられる．ターゲットフレーム I_t とソースフレーム $I_{t'}$ の相対位置 $\boldsymbol{T}_{t\to t'}$ が回転行列 $\boldsymbol{R}$ と並進ベクトル $\boldsymbol{t}$ で次式のように表せるとする．

$$\boldsymbol{T}_{t\to t'} = \begin{pmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0}^\top & 1 \end{pmatrix} \tag{6}$$

この時，$\boldsymbol{x}_t$ と $\boldsymbol{x}_{t'}$ の間には次の関係が成り立つ．

$$\boldsymbol{x}_{t'} = \boldsymbol{R}\boldsymbol{x}_t + \boldsymbol{t} \tag{7}$$

式 (5) はソースフレーム $I_{t'}$ でも成り立つので，式 (5) と式 (7) より，$\{\tilde{\boldsymbol{p}}_t\}$ から $\{\tilde{\boldsymbol{p}}_{t'}\}$ への写像を与える次式を得る．

$$\tilde{\boldsymbol{p}}_{t'} = \frac{1}{z_{t'}}\boldsymbol{K}(z_t \boldsymbol{R}\boldsymbol{K}^{-1}\tilde{\boldsymbol{p}}_t + \boldsymbol{t}) \tag{8}$$

4. 全方位カメラでの教師なし単眼深度推定

本節では，前節で述べた differentiable DIBR に 2 つの改良を加えることで，全方位カメラでの教師なし単眼深度推定である DIBR for UOCM を提案する．1 つ目の改良は differentiable DIBR への Unified Omnidirectional Camera Model (UOCM) の導入であり，その結果として全方位カメラモデルにおける投影関数を導く．2 つ目は，DIBR for UOCM を使用した depth network の訓練の安定化のために，Scale-Aware Constraint Loss を提案する．

4.1 Unified Omnidirectional Camera Model

Unified Omnidirectioanl Camera Model (UOCM) は全方位カメラなどの画角の広いカメラのモデルである．図 3 はその概略図であり，ターゲットフレーム I_t は，3 次元空間の点を C_t を中心とする単位球に投影された後，カメラ中心 C_t^{c} の正規化平面にピンホールモデルで投影される．

UOCM の投影と逆投影を定式化する前に，UOCM の投影中心である C_t^{c} と C_t を原点とする 2 つの座標系の座標変換を考える．ターゲットフレーム t において，C_t^{c} はカメラ中心，C_t が球中心である．C_t の座標系は，C_t^{c} の座標系を z 軸の正の方向に ξ だけ平行移動させたものである．ここで ξ は全方位カメラの内部パラメータの 1 つである．3 次元空間のある点 P を C_t の座標系で表すと $\boldsymbol{x}_t = (x_t, y_t, z_t)^\top$ となる．また，$\boldsymbol{t}_\xi = (0, 0, \xi)^\top$ と定義して，P を C_t^{c} の座標系で表すと，$\boldsymbol{x}_t^{\mathrm{c}} = \boldsymbol{x}_t + \boldsymbol{t}_\xi = (x_t, y_t, z_t + \xi)^\top$ となる．

UOCM の投影と逆投影を定式化する．まず，3 次元空間

の点 P を，C_t を中心とする単位球に投影した点を P' とすれば，その座標 $\hat{\boldsymbol{x}}_t$ が次式で与えられる．

$$\hat{\boldsymbol{x}}_t = \frac{\boldsymbol{x}_t}{||\boldsymbol{x}_t||} = \begin{pmatrix} x_t/||\boldsymbol{x}_t|| \\ y_t/||\boldsymbol{x}_t|| \\ z_t/||\boldsymbol{x}_t|| \end{pmatrix} \tag{9}$$

この点を C_t^{c} の座標系で表すと，次式が得られる．

$$\hat{\boldsymbol{x}}_t^{\mathrm{c}} = \frac{\boldsymbol{x}_t}{||\boldsymbol{x}_t||} + \boldsymbol{t}_\xi = \begin{pmatrix} x_t/||\boldsymbol{x}_t|| \\ y_t/||\boldsymbol{x}_t|| \\ z_t/||\boldsymbol{x}_t|| + \xi \end{pmatrix} \tag{10}$$

次に，式 (5) のピンホールモデルを用いて，C_t^{c} の正規化画像平面に $\hat{\boldsymbol{x}}_t^{\mathrm{c}}$ を投影する．式 (5) の $\boldsymbol{x}_t$ に式 (10) の $\hat{\boldsymbol{x}}_t^{\mathrm{c}}$ を代入し，式 (5) の z_t に $\hat{\boldsymbol{x}}_t^{\mathrm{c}}$ の z 座標である $z_t/||\boldsymbol{x}_t|| + \xi$ を代入すれば，次式が得られる．

$$\tilde{\boldsymbol{p}}_t = \frac{1}{z_t/||\boldsymbol{x}_t|| + \xi} \boldsymbol{K} \hat{\boldsymbol{x}}_t^{\mathrm{c}} = \frac{1}{z_t + \xi||\boldsymbol{x}_t||} \boldsymbol{K} \left(\boldsymbol{x}_t + \boldsymbol{t}_\xi ||\boldsymbol{x}_t||\right) \tag{11}$$

また，C_t を中心とした座標への逆投影は同次座標を用いて次のように閉じた形で計算できる [3]．

$$\begin{aligned} \boldsymbol{x}_t =& ||\boldsymbol{x}_t|| \frac{\xi + \sqrt{1 + (1 - \xi^2)(||\boldsymbol{K}^{-1}\tilde{\boldsymbol{p}}_t||^2 - 1)}}{||\boldsymbol{K}^{-1}\tilde{\boldsymbol{p}}_t||^2} \boldsymbol{K}^{-1}\tilde{\boldsymbol{p}}_t \\ & - ||\boldsymbol{x}_t|| \boldsymbol{t}_\xi \end{aligned} \tag{12}$$

4.2 全方位カメラモデルにおける投影関数

続いて，UOCM における DIBR を行うために，全方位カメラモデルにおける投影関数を求める．UOCM を用いる場合，$\boldsymbol{\theta}_{\mathrm{intrinsic}}$ は 4.1 の内部パラメータ ξ と $\boldsymbol{K}$ の要素を並べたベクトルである．D_t は I_t の各ピクセル $\boldsymbol{p}_t$ に対応する深度 $||\boldsymbol{x}_t||$ を並べた深度マップとする．すなわち，深度は C_t から点 $\boldsymbol{x}_t$ への距離に等しい．ここでソースフレーム t' の単位球中心である $C_{t'}$ の座標系とカメラ中心である $C_{t'}^{\mathrm{c}}$ の座標系でそれぞれ点 P を表した $\boldsymbol{x}_{t'}$ と $\boldsymbol{x}_{t'}^{\mathrm{c}}$ は，$\boldsymbol{R}$ と $\boldsymbol{t}$ を用いて次式のように書ける．

$$\boldsymbol{x}_{t'} = \boldsymbol{R}\boldsymbol{x}_t + \boldsymbol{t} \tag{13}$$

$$\boldsymbol{x}_{t'}^{\mathrm{c}} = \boldsymbol{R}\boldsymbol{x}_t^{\mathrm{c}} + \boldsymbol{t} = \boldsymbol{R}(\boldsymbol{x}_t + \boldsymbol{t}_\xi) + \boldsymbol{t} \tag{14}$$

式 (11) より，ソースフレーム t' の $\tilde{\boldsymbol{p}}_t'$ は次式で表せる．

$$\tilde{\boldsymbol{p}}_{t'} = \frac{1}{z_{t'}/||\boldsymbol{x}_{t'}|| + \xi} \boldsymbol{K} \hat{\boldsymbol{x}}_{t'}^{\mathrm{c}} \tag{15}$$

ここで $z_{t'}$ は，式 (13) の z 座標である．式 (15) に式 (13) と (14) を代入し，さらに式 (12) を代入すれば，$\{\tilde{\boldsymbol{p}}_t\}$ から $\{\tilde{\boldsymbol{p}}_{t'}\}$ への写像が得られる．

4.3 Scale-Aware Constraint Loss

Depth network の学習時の発散を防ぐために，損失関数に Scale-Aware Constraint Loss を導入する．具体的には，

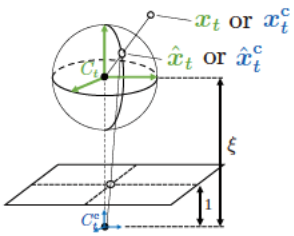


図 3: Unified Omnidirectional Camera Model（C_t の座標系は緑色，C_t^{c} は青色で記載）

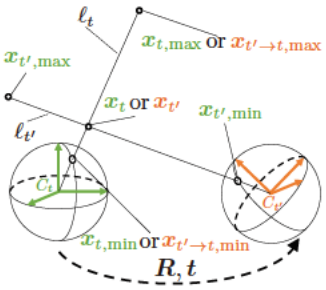


図 4: Scale-Aware Constraint（C_t の座標系は緑色，$C_{t'}$ はオレンジ色で記載）

図 4 に示すように，ターゲットフレームの単位球中心 C_t と $\boldsymbol{x}_t$ を結ぶ視線 ℓ_t を使用した制約条件を加える．C_t から見た ℓ_t 上の最近点と最遠点をそれぞれ $\boldsymbol{x}_{t,\min}$ と $\boldsymbol{x}_{t,\max}$ とする．この時，$\boldsymbol{x}_{t,\min}$ と $\boldsymbol{x}_{t,\max}$ をそれぞれソースフレームの球中心 $\mathrm{C}_{t'}$ の座標系で見た座標は次式で与えられる．

$$\boldsymbol{x}_{t'\to t,\max} = \boldsymbol{R}\boldsymbol{x}_{t,\max} + \boldsymbol{t} \tag{16}$$

$$\boldsymbol{x}_{t'\to t,\min} = \boldsymbol{R}\boldsymbol{x}_{t,\min} + \boldsymbol{t} \tag{17}$$

この時 $C_{t'}$ を中心とする座標系において，ℓ_t と $\ell_{t'}$ が交わるという条件を加える．つまり，$\boldsymbol{x}_{t'\to t,\min}$ と $\boldsymbol{x}_{t'\to t,\max}$ の線形和が $\boldsymbol{x}_{t',\min}$ と $\boldsymbol{x}_{t',\max}$ の線形和と等しくなるという制約を与える．そしてその線形和の重みを深度 $||\boldsymbol{x}_t||$ と $||\boldsymbol{x}_{t'}||$ を基に計算する．

ℓ_t は各 $\boldsymbol{p}_t$ に対して一つずつ決まるので，Scale-Aware Constraint Loss は次のように計算できる．

$$L_{\mathrm{scale}} = \sum_{\boldsymbol{p}_t \in \Omega} ||\boldsymbol{v}_{\mathrm{scale}}(\boldsymbol{p}_t)|| \tag{18}$$

ただし，$\boldsymbol{v}_{\mathrm{scale}}(\boldsymbol{p}_t)$ は次のように計算する．

$$\begin{aligned} \boldsymbol{v}_{\mathrm{scale}}(\boldsymbol{p}_t) =& \{(1 - \alpha_t)\boldsymbol{x}_{t'\to t,\min} + \alpha_t \boldsymbol{x}_{t'\to t,\max}\} \\ & - \{(1 - \alpha_{t'})\boldsymbol{x}_{t',\min} + \alpha_{t'} \boldsymbol{x}_{t',\max}\} \end{aligned} \tag{19}$$

ここで，重み α_t，$\alpha_{t'}$ はそれぞれ以下のように計算する．

$$\alpha_t = \frac{||\boldsymbol{x}_t|| - d_{\min}}{d_{\max} - d_{\min}}, \quad \alpha_{t'} = \frac{||\boldsymbol{x}_{t'}|| - d_{\min}}{d_{\max} - d_{\min}} \tag{20}$$

ここで $d_{\min}$ と $d_{\max}$ は深度マップ D_t の値の最小値と最大値である．この Scale-Aware Constraint Loss を加えた場合，pose network，depth network の訓練の最終的な損失関数 L は，損失関数を重みづけるハイパーパラメータを μ，λ，ν として次のように計算する．

$$L = \mu L_{\mathrm{reprojection}} + \lambda L_{\mathrm{smoothness}} + \nu L_{\mathrm{scale}} \tag{21}$$

5. ブドウ房の果粒の 3 次元配置の推定

本節では，単眼深度推定結果を使用した，動画によるブドウ房の果粒の重心位置の推定方法を説明する．そして，

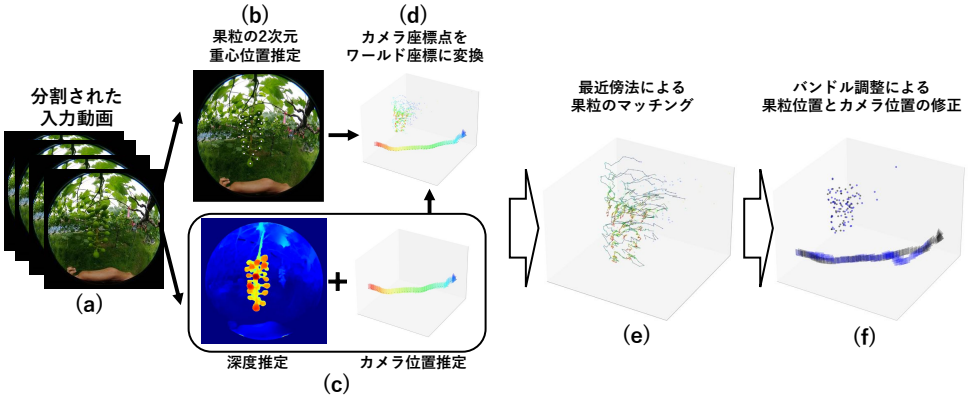


図 5: ブドウ房の部分的な果粒配置推定の流れ

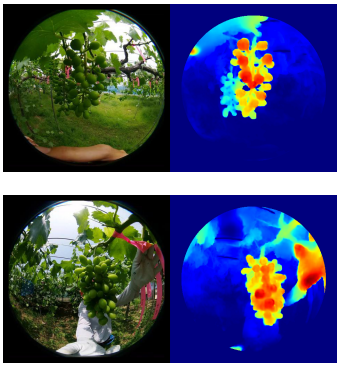


図 6: 単眼深度推定の結果

その手法を本稿での摘粒支援での使用を想定する全方位カメラに適用するために，UOCM でのバンドル調整とフレーム間での特徴点のマッチングについて説明する．

5.1 ブドウ果粒の重心位置の推定方法

本研究では，最終的な目標として，ブドウ房全体の果粒の 3 次元位置を推定する．その前段階として，一部分のブドウの果粒の配置を推定する手法を説明する．

ブドウ房の部分的な果粒配置推定の流れを図 5 に示す．まず初めに，ブドウ房の果粒の部分的な重心位置を推定するために，図 5(a) に示した分割された入力動画の各フレームの果粒に対して，図 5(b) のように果粒の 2 次元の重心位置を推定する．重心位置の推定には，果粒のインスタンスセグメンテーション [23] 等を使用する．それと並行して，4 で述べた DIBR for UOCM により，図 5(c) のように各フレームの単眼深度と隣接フレームの相対カメラ位置を推定する．これらの推定結果から，3 次元空間上での果粒の重心位置を求め，図 5(d) のように果粒の重心位置を世界座標に変換する．そして，図 5(e) のように最近傍法を使用してフレーム間での果粒のマッチングを行い，バンドル調整により図 5(f) のようにブドウ房の部分的な 3 次元重心位置を推定する．

5.2 UOCM でのバンドル調整

バンドル調整では，ワールド座標上の特徴点を各フレームの画像上に再投影した誤差を用いてカメラ位置と特徴点の位置を最適化する．本研究は，特徴点をブドウ果粒の重心とし，最適化の収束性の観点から [16] と同様に C_t を中心とする単位球上への射影点の誤差を再投影誤差とする．

本稿での UOCM でのバンドル調整の定式化を行う．任意に選んだカメラ位置の単位球中心を C_1 として，C_1 を中心とする座標をワールド座標とする．各カメラ位置に対応する球中心が C_1 から順に $\mathrm{C}_i,\ i = (1,,\ldots,N_I)$ であるとする．N_I はバンドル調整を行うカメラ位置の総数である．図 5 の (c) にそれぞれの球中心に対応するカメラの位置の例を示す．C_i のカメラの外部パラメータ $\boldsymbol{T}_i$ は回転行列と並進ベクトルをそれぞれ $\boldsymbol{R}_i, \boldsymbol{t}_i$ とすると，式 (6) と同様に表現可能である．ただし，$\boldsymbol{T}_1$ は単位行列である．$\boldsymbol{w}_j$，$\boldsymbol{x}_{i,j}$ をワールド座標系，C_i を中心とする座標系での各特徴点の座標とし，$\langle\cdot\rangle$ は任意の $\boldsymbol{x} \in \mathbb{R}^3$ に対して $\langle \boldsymbol{x} \rangle = \boldsymbol{x}/||\boldsymbol{x}||$，$||\cdot||_{\mathrm{H}}$ は Huber 損失とする．この時，UOCM でのバンドル調整は以下の最適化問題として定式化される．

$$\underset{\boldsymbol{w}_j, \boldsymbol{T}_i}{\operatorname{argmin}} \sum_{i,j} ||\langle \boldsymbol{x}_{i,j} \rangle - \langle \boldsymbol{R}_i \cdot \boldsymbol{w}_j + \boldsymbol{t}_i \rangle||_{\mathrm{H}} \qquad (22)$$

式 (22) のバンドル調整はスケールの曖昧さの影響が大きく，カメラと特徴点のスケールが C_1 の位置によって大きく変化する．この問題に対処するため，本稿では特徴点とカメラ位置それぞれを固定した状態で順に式 (22) のバンドル調整を行う．また，$\boldsymbol{T}_i$ の初期値として 3 の単眼深度推定の結果推定された $\boldsymbol{T}_{t \to t'}$ を掛け合わせた値を使用する．

5.3 ブドウ果粒のマッチング

従来のバンドル調整では，局所特徴量などを用いて，各フレーム間で特徴点の対応をとる．しかし，本研究で特徴点として用いるブドウの果粒には適用が出来ないため，各フレーム上の果粒重心位置，単眼深度推定結果，カメラ位置推定結果を使用して果粒をフレーム間でマッチングする．

深度推定結果を使用して，図 5(d) のようにカメラ位置 i のフレーム上の k_i 番目の特徴点を，C_i を中心とする座標系に逆投影した点を $\boldsymbol{x}_i^{(k_i)},\ k_i = (1,\ldots,K_i)$ とする．ただし，K_i はカメラ位置 i のフレーム上で検出された特徴点の個数である．図 5(d) では，逆投影された各特徴点 $\boldsymbol{x}_i^{(k_i)}$ とカメラ位置 C_i の対応関係を点とカメラの色で表している．$\boldsymbol{x}_i^{(k_i)}$ を，C_1 を中心とするワールド座標に変換した点 $\boldsymbol{w}_i^{(k_i)}$ は同次座標で以下のように計算できる．

$$\begin{pmatrix} \boldsymbol{w}_i^{(k_i)} \\ 1 \end{pmatrix} = \boldsymbol{T}_i^{-1} \cdot \tilde{\boldsymbol{x}}_i^{(k_i)} = \boldsymbol{T}_i^{-1} \cdot \begin{pmatrix} \boldsymbol{x}_i^{(k_i)} \\ 1 \end{pmatrix} \qquad (23)$$

$\boldsymbol{w}_i^{(k_i)}$ を図 5(e) のように最近傍法でクラスタリングし，そ

のクラスタの平均座標を $\boldsymbol{w}_j$ とする．この時，$\boldsymbol{w}_j$ のクラスタの要素の点は全て対応しているとみなす．つまり，$\boldsymbol{x}_i^{(k_i)} = \boldsymbol{x}_{i,j}$ となる．対応点の探索では，1 つのフレーム画像において，ある 1 点に対応する特徴点は 1 つしか存在しない制約がある．この制約を満たすようにクラスタリングするため，カメラ位置 i のフレーム上の特徴点の最近傍点をカメラ位置 $i+1$ のフレーム上の特徴点の中のみから探索する．そして，カメラ位置 $i+1$ の特徴点の最近傍点をカメラ位置 i の特徴点のみから探索する．双方の最近傍点が一致した場合のみ，同じクラスタとみなす．この処理は図 5(e) のように，カメラの色の変化の順に各点を結合していく操作に相当する．そして各クラスタの要素数が閾値以上だったものをバンドル調整に用いる特徴点と判定し，そのクラスタの $\boldsymbol{w}_i^{(k_i)}$ の座標平均を $\boldsymbol{w}_j$ として追加する．

6. 実験

6.1 実験条件

教師なし単眼深度推定の実装は，[10] の実装を基にした．また，ブドウの動画の撮影はキャリブレーション済みの Richo THETA S の片側のレンズで撮影された画像のみを使用した．そして，各ブドウ房の周囲を一周分，ブドウを動かさないように撮影した．実験で使用した動画のサイズは 480×480 px，フレームレートは 30 fps であった．単眼深度推定の学習には訓練データセットに合計 128,068 枚，検証用データセットに合計 31,554 枚のフレームを使用した．[10] の実装のハイパーパラメータは，式 (21) の ν は 10^{-5}，$d_{\min}$ と $d_{\max}$ をそれぞれ 1 と 100 とし，それ以外はデフォルトの設定を用いた．ネットワークの訓練には NVIDIA TITAN Xp を 1 台使用し，10 エポック行った．

検証用データセット中の約 200 フレームからなる動画 1 つに対して，各フレームでのブドウの 2 次元の重心位置を手動でプロットしてブドウ房を部分的に 3 次元復元した．入力画像を 4 分割して，カメラ位置の個数を $N_I = 50$ とした．果粒のマッチングを行う際に 5.2 の $\boldsymbol{w}_i^{(k_i)}$ のクラスタの閾値を 5 とした．また，式 (22) の $\boldsymbol{w}_j$ と $\boldsymbol{T}_i$ を同時に最適化する場合と順番に 1 回ずつ最適化する場合の両方を行った．バンドル調整の最適化には SciPy を用いた．

6.2 深度推定の結果

単眼深度推定の結果の例として，検証データでの逆深度のヒートマップを図 6 に示す．被写体がカメラから近いほど色が赤く，像が映される範囲以外の逆深度は 0 としている．図 6 に示したもの以外でも，ほとんどの検証データでブドウ果粒の前後関係に応じた深度が推定できるのを定性的に確認した．また，Scale-Aware Constraint Loss を損失関数に加えずに訓練を行った結果，多くの場合，1 エポックの訓練後に depth network の深度推定の結果が $d_{\max}$ の一様分布となった．しかし，Scale-Aware Constraint Loss を損失関数に加えることでこの現象を回避できたため，Scale-Aware Constraint Loss の有効性を定性的に確認した．

6.3 バンドル調整の結果

バンドル調整に使用した動画の 0 から 49 フレームに対してバンドル調整をした例の結果を図 7 と 8 に示す．図 7 は，式 (22) の残差 $\langle \boldsymbol{x}_{i,j} \rangle - \langle \boldsymbol{R}_i \cdot \boldsymbol{w}_j + \boldsymbol{t}_i \rangle$ の各要素のグラフである．また，図 8 は，バンドル調整前後での果粒位置とカメラ位置の変化を表す．図 7(b) より，カメラ位置と特徴点を同時に最適化した場合に，各フレームに特徴点を再投影した残差の減少が確認できた．しかし，図 8(b) のように，この場合バンドル調整後にカメラ位置とブドウ房のスケールが大きく変化した．一方でカメラ位置と特徴点位置を交互に 1 回ずつ最適化した場合，図 7 と 8(c) より残差の減少は比較的小さかったが，特徴点とカメラ位置のスケールはほとんど変化しなくなった．動画の他の分割部分のバンドル調整においても同様の傾向が確認できた．これにより，ブドウ房の果粒の配置が推定できることが，ブドウの動画データを使用して定性的に確認できた．

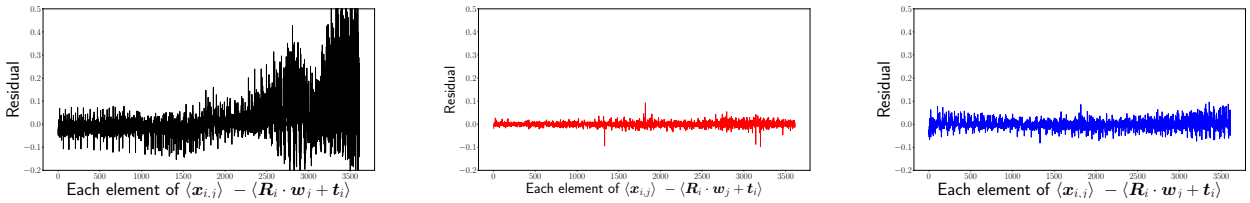
7. まとめ・今後の予定

本稿では，全方位カメラで撮影されたテクスチャレスで密集したブドウ果粒の重心位置を推定した．この実現のために，教師なし単眼深度推定手法である differentiable differentiable DIBR を全方位カメラに拡張した DIBR for UOCM を提案した．DIBR for UOCM には depth network の訓練が発散する問題があったため，2 つの時刻における見えの制約である Scale-Aware Constraint Loss を提案し，問題を解決した．ブドウ果粒の動画データセットを使用した実験では，カメラ位置とブドウ果粒位置を同時に最適化することで，マッチング点の少なさ，スケールの曖昧さによるバンドル調整のスケールの不安定化を緩和した．本稿では提案手法の定性的評価に留まったため，今後は提案手法の定量的評価や Scale-Aware Constraint Loss の深度推定への影響の検証が必要と考えている．そして，バンドル調整の結果を統合して動画 1 つの房全体の果粒の配置の推定を行う予定である．

謝辞 本研究は 2019 年度電気普及財団研究調査助成，大阪府信用農業協同組合連合会令和 3 年度産学連携研究支援事業による研究成果に基づく．

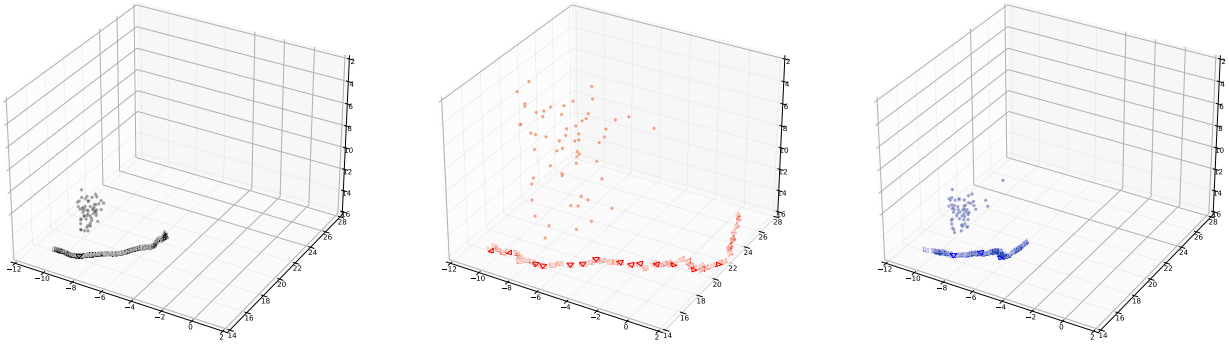
参考文献

[1] Rist, F., Herzog, K., Mack, J., Richter, R., Steinhage, V. and Töpfer, R.: High-Precision Phenotyping of Grape Bunch Architecture Using Fast 3D Sensor and Automation, *Sensors*, Vol. 18 (2018).
[2] Mack, J., Lenz, C., Teutrine, J. and Steinhage, V.: High-Precision 3D Detection and Reconstruction of Grapes from Laser Range Data for Efficient Phenotyping Based



(a) 再投影残差の初期値　(b) カメラ位置と果粒位置を同時最適化　(c) カメラ位置，果粒位置の順に最適化

図 7: バンドル調整による再投影誤差残差の変化



(a) カメラと果粒位置の初期位置　(b) カメラと果粒位置を同時最適化　(c) カメラと果粒位置の順に最適化

図 8: バンドル調整によるカメラと果粒位置の変化

on Supervised Learning, *Computers and Electronics in Agriculture*, Vol. 135 (2017).

[3] Caruso, D., Engel, J. and Cremers, D.: Large-Scale Direct SLAM for Omnidirectional Cameras, *Proc. IROS* (2015).

[4] Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S. M. and Szeliski, R.: Building Rome in a Day, *Comm. of the ACM*, Vol. 54, No. 10 (2011).

[5] Furukawa, Y., Sethi, A., Ponce, J. and Kriegman, D. J.: Structure and Motion from Images of Smooth Textureless Objects, *Proc. ECCV* (2004).

[6] Wong, K.-Y. and Cipolla, R.: Structure and Motion from Silhouettes, *Proc. ICCV* (2001).

[7] Nurutdinova, I. and Fitzgibbon, A. W.: Towards Pointless Structure from Motion: 3D Reconstruction and Camera Parameters from General 3D Curves (2015).

[8] Doi, T., Okura, F., Nagahara, T., Matsushita, Y. and Yagi, Y.: Descriptor-Free Multi-View Region Matching for Instance-Wise 3D Reconstruction, *Proc. ACCV* (2020).

[9] Zhou, T., Brown, M., Snavely, N. and Lowe, D. G.: Unsupervised Learning of Depth and Ego-Motion from Video, *Proc. CVPR* (2017).

[10] Godard, C., Mac Aodha, O., Firman, M. and Brostow, G.: Digging Into Self-Supervised Monocular Depth Estimation, *Proc. ICCV* (2019).

[11] Wang, C., Miguel Buenaposada, J., Zhu, R. and Lucey, S.: Learning Depth from Monocular Videos Using Direct Methods, *Proc. CVPR* (2018).

[12] Nellithimaru, A. K. and Kantor, G. A.: ROLS : Robust Object-Level SLAM for Grape Counting, *Proc. ECCV Workshops* (2019).

[13] Dey, D., Mummert, L. and Sukthankar, R.: Classification of Plant Structures from Uncalibrated Image Sequences, *Proc. WACV* (2012).

[14] 内海ゆづ子, 三木啓輔, 尾形亮輔, 大林拓実, 三輪由佳, 岩村雅一, 黄瀬浩一: 画像を用いた果房の３次元構造推定に基づくブドウの摘粒支援, 農業情報学会年次大会 (2018).

[15] Mur-Artal, Raúl, M. J. M. M. and Tardós, J. D.: ORB-SLAM: a Versatile and Accurate Monocular SLAM System, *IEEE Trans. Robotics*, Vol. 31, No. 5 (2015).

[16] Im, S., Ha, H., Rameau, F., Jeon, H.-G., Choe, G. and Kweon, I. S.: All-Around Depth from Small Motion with a Spherical Panoramic Camera, *Proc. ECCV* (2016).

[17] Zioulis, N., Karakottas, A., Zarpalas, D. and Daras, P.: OmniDepth: Dense Depth Estimation for Indoors Spherical Panoramas., *Proc. ECCV* (2018).

[18] xiang Chen, H., Li, K., Fu, Z., Liu, M., Chen, Z. and Guo, Y.: Distortion-Aware Monocular Depth Estimation for Omnidirectional Images, *IEEE Signal Processing Letters*, Vol. 28 (2021).

[19] Zioulis, N., Karakottas, A., Zarpalas, D., Alvarez, F. and Daras, P.: Spherical View Synthesis for Self-Supervised 360^o Depth Estimation, *Proc. 3DV* (2019).

[20] Vasiljevic, I., Guizilini, V. C., Ambrus, R., Pillai, S., Burgard, W., Shakhnarovich, G. and Gaidon, A.: Neural Ray Surfaces for Self-Supervised Learning of Depth and Ego-motion, *Proc. 3DV* (2020).

[21] Jaderberg, M., Simonyan, K., Zisserman, A. and Kavukcuoglu, K.: Spatial Transformer Networks, *Advances in Neural Information Processing Systems* (2015).

[22] Wang, Z., Bovik, A. C., Sheikh, H. R. and Simoncelli, E. P.: Image Quality Assessment: From Error Visibility to Structural Similarity, *IEEE Trans. Image Processing*, Vol. 13, No. 4 (2004).

[23] He, K., Gkioxari, G., Dollar, P. and Girshick, R.: Mask R-CNN, *Proc. ICCV* (2017).