全方位画像を用いたブドウの房数推定 ―歪みに応じた密度マップの生成とアラインメント―

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

概要

本研究では，全方位画像に対する物体計数手法を提案する．我々の知る限り，本研究がこの問題に対する初めての試みである．提案手法は，全方位画像をステレオ投影で 2 次元画像に変換した後，新たに提案する 2 つの手法を適用する．提案手法の評価には全方位画像を用いた物体計数データセットが必要である．そのため，ブドウの房の計数データセットを手動のラベル付けにより作成した．このデータセットで提案手法を評価したところ，ステレオ投影で 2 次元画像を作成しただけの場合に比べて，2 つの提案手法を適用した場合に平均絶対誤差（MAE）が 0.59，平均二乗誤差（MSE）が 0.54 改善した．これはそれぞれ 14.7%と 10.5%の精度改善に相当する．

1. はじめに

ブドウ栽培に特有で難しい作業に，房を間引く「摘房」がある．ブドウを始めとする果樹は実を多くつけがちだが，光合成で得られる糖の生産能力に限りがあるため，糖度を出荷できる品質に保つには実を間引いて減らす必要がある．摘房には単位面積当たりの房数で表される摘房基準があり，基準を満たすように房を切り落とす．摘房作業に慣れている篤農家は，圃場を見ると房の大体の密度が分かるため，房を数えることなく作業できる．しかし，初心者は房を切るのに躊躇し，摘房基準を上回る量の房を残す傾向にあるので，房を数えながら作業する必要がある．

本研究では，摘房作業のうち，房を数える作業に注目し，物体計数（object counting）手法を用いてその自動化を目指す．本研究で対象とするブドウは図 1 のように，人の背丈ほどの高さ（1.6m）の水平面上に育成されているため，ブドウを下から撮影する必要がある．摘房基準は 2m 四方の領域に対して定められていることから，この領域を一度に撮影しようとすれば，全方位カメラのような画角の広いカメラが必要となる．

図 1　ブドウ圃場（横から撮影）

図 2　全方位カメラでブドウ棚を下から撮影し，ステレオ投影で変換した図

ところが，既存の物体計数手法は透視投影画像を対象としており，歪みがある全方位画像にそのまま適用できない．この問題を回避する方法として，全方位画像を複数枚の透視投影画像に変換して，それぞれで計数するものがある [5]．しかし，この方法では透視投影画像の一部が重複するため，特別な後処理が必要になる．別の回避策として畳み込みの工夫がある．これは全方位画像の一部を逐次的に透視投影画像に変換してから畳み込む手法である [2], [11]．しかし，我々が実験したところ，この方法は処理が複雑で低速な上に，通常の畳み込みに比べて精度が低下した．

そこで本研究では，全方位画像に対する物体計数手法を

1　大阪府立大学 大学院工学研究科
2　大阪府立環境農林水産総合研究所
a)　akai@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

提案する．我々の知る限り，全方位画像に対する物体計数の試みは本研究が初めてである．提案手法では，全方位カメラで撮影された画像をステレオ投影（stereographic projection）画像に変換することで歪みを扱いやすくした後，新たに提案する 2 つの手法を適用する．1 つ目は，ステレオ投影画像の性質を利用して，画像中の同じ歪みを持つ部分の位置を合わせることで房数推定モデルを効率的に学習する．2 つ目は，歪みが画像の位置で決まることから，期待される房の大きさを反映して密度マップ（density map）を生成する．これらを組み合わせることで，ステレオ投影で 2 次元画像を作成しただけの場合に比べて，精度の改善が見込める．提案手法の評価には全方位画像を用いた物体計数データセットが必要である．そのため，527 枚の全方位画像から成る，ブドウの房の計数データセットを手動のラベル付けにより作成して評価に用いた．

2. 物体計数

画像中の物体を数えるタスクは物体計数と呼ばれる．顕微鏡画像中の細胞を数える cell counting [4], [9] や群衆を数える crowd counting [1] が特に盛んに研究されている．収量予測のためにブドウの実を数えた研究 [8] も存在するが，本研究で必要なブドウの房を対象とするものは無い．

物体計数手法は，物体検出に基づく手法と回帰に基づく手法の 2 つのアプローチに大別できる．前者は，対象物体の隠れが少ない場合に推定精度が良く，対象物体の密度が高く，オクルージョンが多い場合に精度が悪いことが知られている [7]．一方，後者は，前者とは対照的に，対象物体が少数の場合に精度が悪く，対象物体の密度が高い場合や隠れが多い場合に比較的精度が良いことが示されている [7]．図 2 に示すように，本研究で対象とするブドウは房が多く，房が葉に隠れることが多いという特徴がある．すなわち，密度が高く隠れが多い．そのため，本研究の推定手法としては後者に基づく手法が相応しい．

回帰に基づく物体計数手法では，Lempitsky らの提案したフレームワークが用いられている [6]．Lempitsky らの手法では，画像から物体の密度マップを推定することで，物体の位置情報を考慮しながら物体を計数する．物体の数は密度マップを積分することで得られる．本研究では，このフレームワークを用いて，回帰モデルを新たに提案した S-DCNet [10] をベースモデルとして用いる．S-DCNet はトウモロコシでの応用例があり，良好な結果を示している．このことから，同じ植物であるブドウの房の計数にも有効であることが期待できる．

3. 提案手法

提案手法は，ステレオ投影画像に対して S-DCNet [10] を適用することで，ブドウの房数を推定する．その際，ステレオ投影画像の歪みに対応するための画像のアライメント手法と，正解の密度マップ生成方法を合わせて適用する．

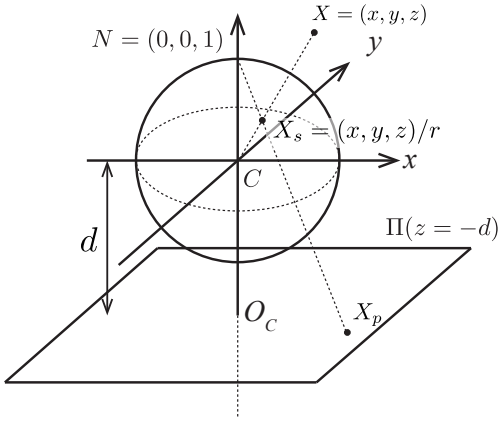

図 3 ステレオ投影のモデル

3.1 ステレオ投影

本研究で用いるステレオ投影について，3 次元空間上の点から 2 次元画像平面への投影を定式化する．図 3 のように，カメラ中心 C を原点とする 3 次元空間を仮定し，球面投影（spherical projection），ステレオ投影を通じて 3 次元空間がどのように画像平面上に投影されるかを考える．はじめに，球面投影では，3 次元空間の任意の 3 次元点 $X = (x, y, z)$ が，カメラ中心 C を中心とする単位球面上の点 X_s に投影される．$r = \sqrt{x^2 + y^2 + z^2}$ とすると，投影点の座標 X_s は次式で与えられる．

$$X_s = \left(\frac{x}{r}, \frac{y}{r}, \frac{z}{r}\right) \tag{1}$$

続いて，単位球上に投影された 3 次元点を任意の面に投影することで，全天球カメラの撮影シーンを 2 次元画像として表現する．本研究では，図 3 に示す通り，単位球の北極 $N = (0, 0, 1)$ から，単位球上の点 X_s を z 軸に垂直な投影面 $\Pi(z = -d, d \geq 1)$ に投影するものとする．このとき，単位球上の点 X_s を画像平面へ投影した点 X_p は，

$$X_p = \left(\frac{1+d}{r-z}x, \frac{1+d}{r-z}y, -d\right) \tag{2}$$

と表される．

ここで、計数対象が平面上に分布する場合を想定して，投影面と平行な面 $z = k$ が投影面 Π に投影された場合を考える．式 (2) から，r が変化すると投影の係数 $(1+d)/(r-z)$ が変化する．面 $z = k$ 上の点では，$r = \sqrt{x^2 + y^2 + k^2}$ であるため，$x^2 + y^2$ の値に依存して投影の係数が決まる．そのため，平行な面 $z = k$ が投影された画像は，画像の投影中心 O_C からの距離に応じて，歪みが変化することが分かる．

3.2 Augmented データに対するアライメント

物体計数では通常，アノテーションの困難さからデータ

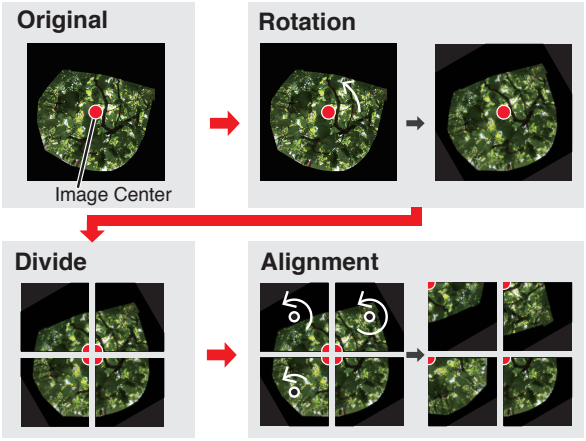

図 4　歪みのアライメント

セットの規模が小さい．そのため，画像のランダムクロッピングにより画像枚数を増やすのが一般的である．透視投影画像では，画像上で歪みが一定なため，どの位置でクロップしても画像の歪みは変わらない．しかし，ステレオ投影画像は投影中心からの距離によって歪みが変化するため，従来と同じクロッピング方法を用いることができない．

そこで本研究では，ランダムクロッピングに代わって，図 4 に示すように，投影中心 O_C を中心として 0° から 90° の範囲で画像を回転させた上で，4 分割してから投影中心が左上になるように画像を回転して位置を合わせる．このように投影中心の位置を合わせることで，画像の歪みを揃えることができ，歪みがあっても学習を妨げることなく，画像枚数を増やすことができる．

3.3　画像歪みに対応した密度マップの生成

2 で述べた通り，回帰ベースの物体計数では，学習データの物体アノテーション結果に基づいて生成された密度マップを回帰で推定する．密度マップは等方性 2 次元 Gaussian の重ね合わせで表現される．密度マップは回帰の正解データとして用いられるため，密度マップが画像上の物体の密度を忠実に反映することが精度の向上につながる．

ステレオ投影には，単位球（図 3 の球）上の円は，投影面（図 3 の II）上でも円であるという性質がある [3]．そのため，透視投影画像上での等方性 2 次元 Gaussian カーネルは，ステレオ投影画像上でも円型の形状を保つ．ただし，円の中心位置がずれる事から，ステレオ投影画像上では等方性 2 次元 Gaussian カーネルにはならないが，同じカーネルをそのまま使うことが良い近似となると期待できる．

また，3.1 で述べた通り，投影面と平行な画像が投影されると，中心からの距離に応じた歪みが発生する．このことから，提案手法では，中心の距離に反比例した分散を持つ等方性 2 次元 Gaussian カーネルを用いて密度マップを生成する．これを画像歪みに対応した密度マップと呼ぶ．

表 1　データセットの詳細

ビニールハウス	房数平均	房数分散	画像枚数
A	46.3	14.2	268
B	34.6	12.0	259

4.　評価用データセット

提案手法の評価のためにデータセットを作成した．データセットは 2,688×2,688 pixels の 527 枚の全方位画像から構成され，房にバウンディングボックスによるアノテーションが付いている． データの房数の平均・分散は表 1 の通りである．

4.1　撮影

撮影は大阪府立環境農林水産総合研究所の果樹圃場にて行った．ブドウの品種はデラウェアである．2019 年 5 月 13, 20 日に撮影を行い，機材には RICOH THETA S を用いた．天候は両日とも晴れだった．

撮影方法について述べる．ブドウ棚を 2 m × 2 m の格子状の領域に区切り，それぞれの領域において数回ずつ，ブドウ棚との距離はおよそ 0.5 m〜1.5 m の間で高さを変えて，目測で測った中心位置から撮影した．撮影は 2 つのビニールハウスについて行い，それぞれ 268 枚，259 枚の合計 527 枚の画像が集まった．

4.2　撮影後の処理

まず，撮影したデータを Exif タグのジャイロセンサの傾きの情報に基づき天頂補正を行い，カメラの傾きを補正した．ブドウ棚はカメラの上部にあるが，このまま平面にステレオ投影を行うと，投影中心から離れた位置に射影され、大きな歪みの影響を受ける．そのため，鉛直下向きが z 軸の正の方向になるよう回転させた後，ステレオ投影を行った．写像の際の画素値の補完には bicubic 法を用いた．

続いて，房のアノテーションと計数対象領域を切り出した．まず，計数対象範囲を知るため，2 m × 2 m の領域の 4 頂点をつないだ矩形をステレオ投影画像に描画した．そして，この領域内のブドウ房に対して手動でバウンディングボックスのアノテーションを行った．最後に，対象領域と領域内の房のバウンディングボックスの凸包を房数推定に用いる部分としてマスク処理を行った．

5.　実験

5.1　実験条件

ベースモデル S-DCNet [10] は，著者らが公開している実装を用いた．学習に関する全てのパラメータは [10] に従った．全画像 527 枚のうち，ビニールハウスごとに学習用の画像として 268 枚，テスト用として 258 枚用いた．Data augmentation は，3.2 で述べた方法で行った．

表 2　実験結果

密度マップ生成法	アライメント	MAE	MSE
Geometry-adaptive kernel		3.99	5.12
（従来手法）	✓	3.72	4.93
歪みに対応した Gaussian		3.46	4.58
（提案手法）	✓	3.40	4.58

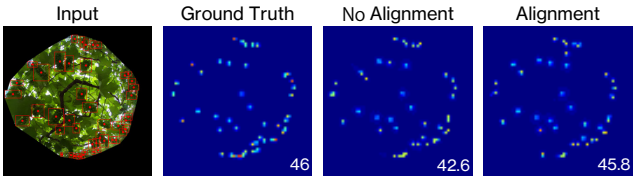

図 5　密度マップの推定結果．入力画像には房のアノテーションを描画．

具体的には，回転を 2 回行い，画像の反転も行って，画像を 4 分割したものを学習データとして用いた．Augmentation 後の学習データ数は回転を行わない場合も含めて $268 \times (1+2) \times 2 \times 4 = 6,432$ 枚となった．

実験では，提案するアライメントを行った場合と行わない場合を比較し，さらに，密度マップ作成では，従来手法である S-DCNet で用いた Geometry-adaptive kernel [12] と，提案手法の歪みに対応した Gaussian のそれぞれ 2 通りを比較した．Geometry-adaptive kernel の σ は [12] に基づいて実験を行った．

5.2　実験結果

実験の結果を表 2 に示す．ここで MAE は平均絶対誤差，MSE は平均二乗誤差を示す．アライメントの有無で精度を比較すると，密度マップの生成方法が，従来手法，提案手法のどちらの場合でも精度が向上している．このことから，歪みのアラインメントは密度マップの生成方法に依らず，精度を向上させる汎用的な手法であることが明らかになった．また，密度マップの生成方法による比較でも，提案手法が従来手法より精度が向上していることが分かる．双方の提案手法を組み合わせることで，従来手法に比べて，MAE が 0.59，MSE が 0.54 向上した．これはそれぞれ 14.7%と 10.5%の精度改善に相当する．

図 5 は，提案手法の密度マップ生成方法で学習し，密度マップを推定した結果を表したものである．図 5 から，アライメントを用いることで，アライメントを用いない場合よりも，6 時方向や 10 時方向にある画像端に近い房が正確に推定できていることが分かる．画像の外側になるほど歪みの影響で解像度が低くなることを考えると，アライメントによって画像端の大きな歪みに対し頑強になり，精度向上に寄与したと考えられる．

6.　まとめ

本研究では，全方位画像を用いて摘房時におけるブドウの房数を推定する手法を提案した．提案手法を評価するため，摘房前後のブドウ棚を撮影したデータセットを作成した．ステレオ投影画像に変換し，画像の投影中心位置のアライメントと，画像の歪みに合わせた密度マップの生成を行うことで，全方位画像を用いる上で避けることができない歪みの影響を低下させた．実験の結果，ステレオ投影で 2 次元画像を作成しただけの場合と比較して最大で MAE が 0.59, MSE が 0.54 改善した．これはそれぞれ 14.7%と 10.5%の精度改善に相当する．

謝辞

本研究は I-O DATA 財団 2018 年度 研究開発助成，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 2 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] Cenggoro, T. W.: Deep Learning for Crowd Counting: A Survey, *EMACS Journal*, Vol. 1, No. 1, pp. 17–28 (2019).

[2] Coors, B., Condurache, A. P. and Geiger, A.: SphereNet: Learning Spherical Representations for Detection and Classification in Omnidirectional Images, *Proc. ECCV* (2018).

[3] Goldberg, D. M. and Gott, J. R.: Flexion and Skewness in Map Projections of the Earth, *Cartographica: The Intl. J. for Geographic Information and Geovisualization*, Vol. 42, No. 4, pp. 297–318 (2007).

[4] He, S., Minn, K. T., Solnica-Krezel, L., Li, H. and Anastasio, M.: Automatic Microscopic Cell Counting by Use of Unsupervised Adversarial Domain Adaptation and Supervised Density Regression, *Medical Imaging 2019: Digital Pathology*, Vol. 10956, pp. 1–8 (2019).

[5] Iwamura, M., Hirabayashi, N., Cheng, Z., Minatani, K. and Kise, K.: VisPhoto: Photography for People with Visual Impairment as Post-Production of Omni-Directional Camera Image, *Proc. CHI Extended Abstracts* (2020).

[6] Lempitsky, V. and Zisserman, A.: Learning To Count Objects in Images, *Proc. NIPS* (2010).

[7] Liu, J., Gao, C., Meng, D. and Hauptmann, A. G.: DecideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation, *Proc. CVPR* (2018).

[8] Nuske, S., Achar, S., Bates, T., Narasimhan, S. and Singh, S.: Yield Estimation in Vineyards by Visual Grape Detection, *Proc. IROS* (2011).

[9] Xie, W., Noble, J. and Zisserman, A.: Microscopy Cell Counting and Detection with Fully Convolutional Regression Networks, *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pp. 1–10 (2016).

[10] Xiong, H., Lu, H., Liu, C., Liu, L., Cao, Z. and Shen, C.: From Open Set to Closed Set: Counting Objects by Spatial Divide-and-Conquer, *Proc. ICCV* (2019).

[11] Yang, W., Qian, Y., Kämäräinen, J.-K., Cricri, F. and Fan, L.: Object Detection in Equirectangular Panorama, *Proc. ICPR* (2018).

[12] Zhang, Y., Zhou, D., Chen, S., Gao, S. and Ma, Y.: Single-Image Crowd Counting via Multi-Column Convolutional Neural Network, *Proc. CVPR*, pp. 589–597 (2016).