4Q-07

GANを用いた動画像上のヒトとモノの除去に関する研究

梅原喜政† 中原匡哉‡ 窪田諭‡† 田中成典‡‡ 大平将也‡‡
関西大学先端科学技術推進機構† 大阪電気通信大学総合情報学部‡
関西大学環境都市工学部‡† 関西大学総合情報学部‡‡

１．はじめに

国土交通省は，i-Construction[1]を契機に建設現場の生産性向上を目指し，点群データの活用を推進している．従来の点群データの計測手段としては，地上設置型レーザスキャナや MMS（Mobile Mapping System）があるが，計測機器が高額であり導入コストが高く，活用できる現場が限定的である．そのため，安価で導入し易いビデオカメラによる動画像から点群データを生成できる Visual SLAM が注目されている．しかし，動画像内にヒトやモノの動体が映り込んだ場合，異なるフレーム間で特徴点の位置が一致せず，正確な点群データを得ることができない．そこで，本研究では，動画像内の動体を除去した後，GAN（Generative Adversarial Network）の技術である STTN（Spatial Temporal Transformer Network）[2]を用いて除去領域の背景補完を行った動画像（以下，疑似動画像）を生成することで，動体が存在する環境下においても地物の点群データを正確に生成する手法を提案する．

２．研究の概要

本システム（図 1）は，動体領域マスク機能，マスク領域補完機能と点群データ生成機能により構成される．入力データは動体が映り込んだ動画像，出力データは疑似動画像を介した地物の点群データとする．

２．１　動体領域マスク機能

本機能では，画像領域分割手法である YOLACT[3]により，入力された動画像から動体領域を画素単位で検出する．そして，検出された領域に対しマスクを付与した画像を生成する．

Research for Removing Person and Objects from Video using GAN
† Yoshimasa Umehara
Organization for Research and Development of Innovative Science and Technology, Kansai University
‡ Masaya Nakahara
Faculty of Information Science and Arts, Osaka Electro-Communication University
‡†Satoshi Kubota
Faculty of Environmental and Urban Engineering, Kansai University
‡‡Shigenori Tanaka and Masaya Ohira
Faculty of Informatics, Kansai University

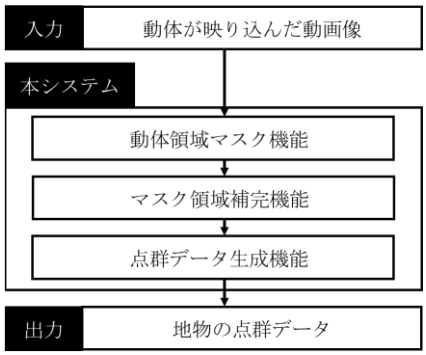

図 1　提案手法の流れ

図 2　マスク領域補完機能

２．２　マスク領域補完機能

本機能では，STTN を用いて動画像内のマスクされた領域を補完する（図 2）．STTN は，動画像内の欠損領域を周囲のピクセルや別フレームからの情報をもとに補完する技術である．まず，動体が映り込んだ動画像と動体領域マスク機能にて生成したマスク画像を STTN へ入力する．そして，マスクされた領域は欠損領域として扱われ，周囲や別フレームの情報をもとに背景を補完する．これにより，映り込んでいた動体を除去した疑似動画像を生成する．

２．３　点群データ生成機能

本機能では，マスク領域補完機能にて生成した疑似動画像を入力とし，COLMAP[4]を用いて動体を除去した地物の点群データを生成する．COLMAP は画像から特徴点を検出することで点群データを生成する技術である．

３．実証実験

３．１　実験内容

本実験では，動体の映り込んだ動画像から疑似動画像を生成し，その点群データを比較することで有用性を評価する．14m×6m の評価範囲が常にフレーム内に映り込むように撮影者が外周を一周して撮影する．この時，動体は人と車両とし，評価範囲内を（A1）人が直線状を往復，（A2）人が無作為に移動，（A3）人が立ち止まる，（B1）車両が直線状を往復，（B2）停車させた計 5 パターンを撮影する．評価指標は平坦性と一致度とする．平坦性は，評価範囲内の地面の点群データから近似平面を生成し，それに基づいた RMS 値から評価する．一致度は，風景のみの撮影で得た点群データと重畳した上で，0.050m でグリッド分割し，両点群が存在するグリッドとそうでないグリッドから適合率と再現率，F 値を算出し一致度を評価する．

３．２　結果と考察

元動画像と疑似動画像から生成した点群データの可視化結果の一部を図 3（A1）と図 4（B1）に示す．A1（図 3）の疑似動画像を見ると，人の除去領域に軽度のぼやけを確認できたが，背景の色や形状は認識できる程度まで補完された．そのため，元動画像による点群データは人が通過したことで点の散らばりが目立つが，疑似動画像による点群データではほとんど見られなかった．B1（図 4）も同様に，元動画像で視認できた車両の点群は疑似動画像による点群データでは見られなかった．しかし，黒い靄が空中に生成された．これは，除去対象ではない車両直下の影を補完対象と誤認識し，除去領域内に影が生成されたためと考えられる．

地面の平坦性として算出した RMS 値を表 1 に示す．いずれのパターンにおいても，車両の影の有無に依らず，疑似動画像の方が元動画像よりも RMS 値を低減できることがわかった．

最後に，一致度の評価結果を表 2 に示す．どのパターンにおいても適合率と再現率，F 値のすべての値で改善が見られた．ただし，適合率が再現率よりも低い傾向となった．これは補完後の領域のぼやけが新たな特徴点として検出され，空中に軽度な点の散らばりが生成されたためと考えられる．

４．おわりに

本研究では，動体を除去した疑似動画像を生成することで，より正確な地物の点群データの生成を目指した．そして，実証実験により，提案手法は動体の影響を低減できることがわかった．今後は，影の影響も考慮した実験を行う．

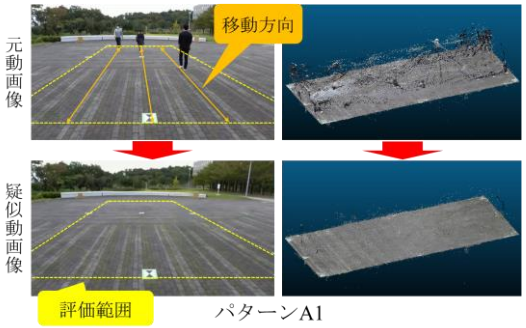

図 3　A1 の動画像と点群データ

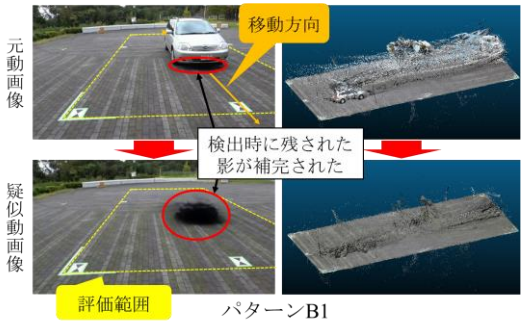

図 4　B1 の動画像と点群データ

表 1　RMS 値による平坦性の評価結果

パターン	元動画像	疑似動画像
A1	0.263m	0.063m
A2	0.275m	0.087m
A3	0.487m	0.092m
B1	0.518m	0.121m
B2	0.452m	0.117m

表 2　一致度の評価結果

	適合率		再現率		F 値	
	a	b	a	b	a	b
A1	0.530	0.566	0.807	0.850	0.639	0.679
A2	0.504	0.544	0.792	0.832	0.616	0.658
A3	0.503	0.558	0.775	0.808	0.610	0.660
B1	0.365	0.545	0.701	0.833	0.480	0.659
B2	0.429	0.550	0.702	0.787	0.533	0.647

a：元動画像，b：疑似動画像

参考文献

[1] 国土交通省：i-Construction 推進コンソーシアム，国土交通省（オンライン），入手先〈https://www.mlit.go.jp/tec/i-construction/3d_wg/index.html〉（参照 2022-1-7）．

[2] Zeng, Y., Fu, J. and Chao, H.: Learning Joint Spatial-Temporal Transformations for Video Inpainting, *The Proceedings of the European Conference on Computer Vision*, pp.528-543 (2020).

[3] Bolya, D., Zhou, C., Xiao, F. and Lee, Y.: YOLACT: Real-time Instance Segmentation, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.9156-9165 (2019).

[4] Schönberger, J. and Frahm, J.: Structure-from-Motion Revisited, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4104-4113 (2016).