J-023

大規模言語モデルによるドローンの広域景観撮影システムの提案
Proposal for a Drone Wide-Area Landscape Imaging System Based on a Large Language Model

池山 安杜里[1)]　山内 翔[1)]　鈴木 恵二[1)]
Atori Ikeyama　Sho Yamauchi　Keiji Suzuki

1　はじめに

近年，ドローンによるドローンならではの映像撮影が積極的に行われている．特にドローンの視点を自在に動かせる利点を活用するため，山，海岸，湖沼など自然地形やダムや遺跡など大型建造物を対象とした映像が多く撮影されている．しかしこのような広大な範囲の撮影では，地形を詳細に把握することが難しく，人手による撮影が主流となっている．一方，人手で広大な地形における撮影を，トラブルなく計画通り実施することには限界があり，ドローンの特性を活かした自動飛行による撮影が期待される．

自動飛行による撮影は，測量，農業，物流さらにはアクション映画の撮影などで盛んに利用されている．それらの多くは，ウェイポイントと呼ばれる飛行経路の通過点を事前に設定する方法と，事前に登録した対象を追いかけ続ける方法に大別される．しかしどちらの方法であっても，自然地形などの広大な地形での撮影においては，カメラの向きや，ウェイポイント，トラッキング対象を詳細に設定する必要があり，エンジニアリングに関する知識や技術，経験を必要とする．さらに，魅了的な撮影を行うには，撮影面に関する技術も不可欠となる．

したがって，ドローンの特性を活かした映像撮影における課題として，以下の 2 点が挙げられる．

- 自動飛行を利用する場合であっても，ドローンの操作に関するエンジニアリング面の知識や技術が必要である．
- 魅力的な映像を撮影するための，映像撮影面の知識や技術が必要である．

これらの課題を解決しドローンによる映像撮影をより容易にするために，本研究ではドローンの自動飛行による広域景観撮影システム（図 1）を提案する．システムは自然言語で指示を受け取り，撮影を実現する飛行経路を生成する．これにより，ドローンの操作や映像撮影に関する知識を持たない人であっても，ドローンの特性を活かした映像撮影が可能になると考えられる．

上記のシステムを実現するため，我々は大規模言語モデル（以下 LLM）を中心的な技術として活用する．近年 LLM は，自然言語処理だけでなくマルチモーダルへの拡張が積極的に行われており，ロボットの分野でも利用されている．例えば PaLM-E[1] では，センサデータを専用のエンコーダが変換し，自然言語の命令につなげ，それを LLM である PaLM 入力することで，命令や文脈を理解したロボットの動作を生成することに成功している．そこで我々は，自然言語と 3 次元マップを入力として LLM へ渡し，飛行経路を出力させるアプローチをとる．

学習用データを確保することは LLM を活用するうえ

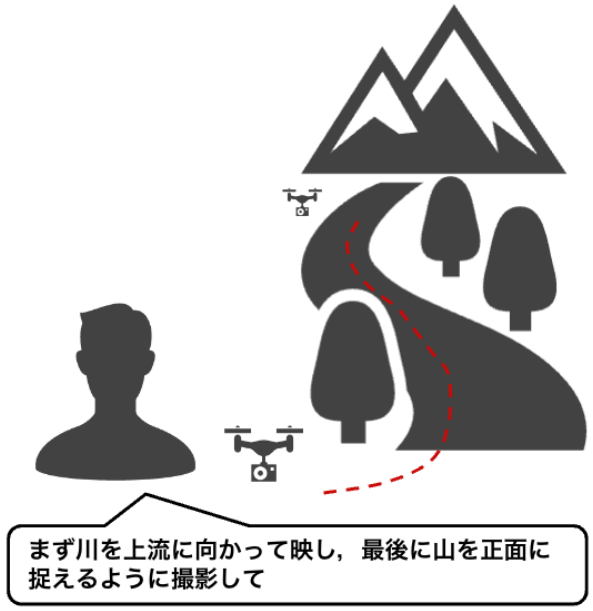

図 1　ドローンの自動飛行による広域景観撮影システムのイメージ

で重要である．しかし，自然言語＋ 3 次元マップと飛行経路がペアとなるようなデータセットは存在しない．さらに飛行経路は，言語的な指示に対応した撮影が可能であることが求められる．そこで，一般に公開されているドローンによる広大な地形の映像から，Structure from Motion（以下 SfM）により飛行経路と 3 次元マップを計算し，人手で自然言語の指示を追加する方法を考案した．そのため本稿では，学習用データの生成に関する検討も行う．

2　提案システム

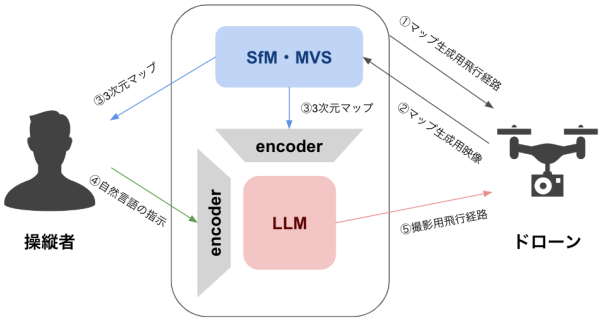

図 2　ドローンの自動飛行による広域景観撮影システムの概要

本研究で提案するシステムの概要を図 2 に示す．

システムはまず，撮影対象の周辺地形を 3 次元マップ化する（①，②）．マップの生成は，マップ化したい

1)　Future University Hakodate

地形をドローンに撮影させ，その映像を SfM および Multi-View Stereo（以下 MVS）を利用することにより行う．生成されたマップは，操縦者とマップ用のエンコーダを通して LLM へ渡される（③）．この際 LLM へは，ボクセル形式へ変換した後，固定サイズのパッチへ分割したものが渡される．3 次元マップの扱いは，Perceiver-Actor[2] を参考にしている．操縦者は，生成された 3 次元マップや目視をもとに，撮影したい対象やイメージに関する自然言語の指示を出す（④）．指示は自然言語用のエンコーダを通して LLM へ渡される．LLM は 3 次元マップと自然言語の指示を受け取り，飛行経路を生成する（⑤）．飛行経路には，ウェイポイントとカメラのパラメータが含まれる．ウェイポイントとは，その点を追いかけることで目的の飛行経路となるような 3 次元の点群である．またカメラのパラメータは，カメラの向きと拡大・縮小を決定するパラメータである．LLM は事前に 3 次元マップ＋自然言語と飛行経路のペアを学習することによって，指示に対応した飛行経路の生成を可能にする．

3　LLM 学習データの生成

LLM の学習データに必要となる，3 次元マップの生成方法について検討する．

3.1　3 次元マップの生成

3 次元マップの生成は以下の手順で行った．

1. ドローンにより撮影された映像のダウンロード
2. 映像から ffmpeg[3] による画像の切り出し
3. COLMAP[4, 5] による SfM および MVS の実施
4. Open3D[6] を用いたボクセル化

映像（図 3）には Youtube にて公開されている「Matterhorn with an FPV Drone | Cinematic Long Range - Swiss Alps」[7] を用いた．映像は，字幕などドローンの撮影した映像以外の写り込みが少ないこと，撮影対象が大きいことを基準に選んだ．SfM・MVS は画像からしか実施できないため，ffmpeg を利用し映像から 10fps で画像を切り出した．ffmpeg は動画や音声の変換を行えるオープンソースソフトウェアである．SfM を実施する際のマッチング方法としてシーケンシャルマッチングを用いた．シーケンシャルマッチングは動画など連続する画像の多いデータに向いているマッチング方法である．

実行は，プロセッサ：Intel(R) Core(TM) i7-9750H，RAM：32.0GB，GPU：NVIDIA GeForce RTX 2060 を搭載したマシンで行った．実行時間は，画像の切り出しに 72 秒，SfM および MVS に 13015 秒（約 3.6 時間）であった．図 4 は生成結果である．

3.2　生成結果の検討

生成された 3 次元マップは，色および形状に関して，元の映像データの特徴を大まかに再現できている．ただし，影のかかっている地形については 3 次元化できていない部分が多い．また雲などの地形とは関係のない情報も取り込んでしまっている．これらが LLM を学習する際にどのように影響するか調査することは今後の課題である．

本実験において映像の選定は人手で行った．しかし全データを人手で選定することは現実的ではない．そのため，自動化または半自動化する仕組みを検討する必要がある．

図 3　YouTube の映像 [7] の切り抜き（21 秒付近）．

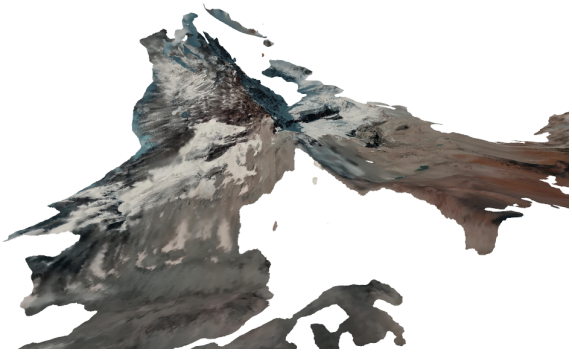

図 4　COLMAP により生成した 3 次元マップ．

4　おわりに

本研究では，大規模言語モデルによるドローンの広域景観撮影システムを提案した．また，学習データの生成方法を検討した．今後は大規模に LLM 学習用のデータを生成し，LLM の学習を行う必要がある．その上で，定量的および定性的にシステムの有効性を検証していきたい．

参考文献

[1] Driess, Danny, et al. ”Palm-e: An embodied multimodal language model.” arXiv preprint arXiv:2303.03378 (2023).
[2] Shridhar, Mohit, Lucas Manuelli, and Dieter Fox. ”Perceiver-actor: A multi-task transformer for robotic manipulation.” Conference on Robot Learning. PMLR, 2023.
[3] ”FFmpeg.” FFmpeg, https://ffmpeg.org/. (最終アクセス：2023 年 6 月 13 日)
[4] Schonberger, Johannes L., and Jan-Michael Frahm. ”Structure-from-motion revisited.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[5] Schönberger, Johannes L., et al. ”Pixelwise view selection for unstructured multi-view stereo.” Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. Springer International Publishing, 2016.
[6] Zhou, Qian-Yi, Jaesik Park, and Vladlen Koltun. ”Open3D: A modern library for 3D data processing.” arXiv preprint arXiv:1801.09847 (2018).
[7] ”YouTube.” YouTube, https://www.youtube.com/watch?v=5xqgvRIUffI. (最終アクセス：2023 年 6 月 13 日)