

A Thesis for the Degree of Ph.D. in Engineering

**Tracking and Scene Representation
for Fast-moving Cameras
by Event Data Synthesis**

November 2024

Graduate School of Science and Technology,
Keio University

Mana Masuda

報告番号	㊦ 乙 第	号	氏 名	栢田真奈
主 論 文 題 名 : Tracking and Scene Representation for Fast-moving Cameras by Event Data Synthesis (イベントデータ合成による高速移動カメラのための運動追跡とシーン表現)				
(内容の要旨)				
機械は、運転支援システムや手術支援など、私たちの日常生活に広く浸透している。これらの技術の核となるのは、機械が周囲の三次元環境を感覚データとして解釈し、カメラ「目」を使って自分自身を位置特定する能力である。 既存のフレームベースのカメラを利用するために多くの技術が開発されているが、これらの方法は密な記録システムによる制限に直面し、より高い時間的解像度をスケールすることや計算の複雑さを減らすことが難しくなる。フレームベースの記録システムは、環境を視覚化するための最も簡単な方法であり、フレームベースの画像は解釈が容易です。しかし、これらのカメラからの情報は、動きを表現する際にしばしば冗長になる。イベントベースのカメラは、網膜を模倣した生体を模倣した視覚センサーで、実際の強度画像を定期的にキャプチャするのではなく、各ピクセルの輝度変化を非同期的に報告する。この特別な動作原理により、非常に高い時間的解像度と低いデータ消費を提供し、動きのキャプチャに最適なセンサーとなっている。 本論文の目的は、イベントベースのカメラを活用して、速い動きの視覚から機械が自分自身を位置特定できるようにすることである。この目的を達成するために、本論文ではまず、イベントベースの視覚と微分可能なシーン表現を用いた速い動きの追跡について探求する。次に、イベントベースの視覚を使用した速い動きのカメラの位置特定と追跡方法を提案する。最後に、速い動きのカメラを用いた 3D シーン再構成法を紹介し、イベントベースの視覚を取り入れる。 この目標を達成するために、本論文のいくつかの重要な貢献を示す。まず、イベントデータを用いたカメラ追跡の新しいフレームワークを提案します。このフレームワークは、事前に訓練された多層パーセプトロン (MLP) を用いてイベントを合成することにより、観測されたイベントと生成されたイベントの不一致を最小化することで、カメラの位置と速度を求める。次に、ニューラルラディアンフィールド (NeRF) シーン表現を活用してイベントを合成する技術を開発する。NeRF の時間的勾配出力を観測されたイベントと整合させることにより、この方法は、入力パラメータが実際のカメラの動きと一				

別表5
(3)

致する場合にカメラのポーズと動きを正確に求める．第三に、本論文では、速い動きのカメラからの高忠実度の 3D シーン再構成法を紹介し、数分以内に結果を得ることができます．イベントベースとフレームベースの視覚を統合することで、カメラ追跡効率を向上させ、モーションブラーを排除し、より精密な 3D 表現を実現する．最後に、速い動きのカメラからの 3D シーンを再構成する方法を提案し、カメラポーズを最適化する．提案されたカメラ追跡技術とイベント強化型 3D シーン再構成戦略を活用することで、モーションブラーを伴う入力画像でも高忠実度の再構成を確保する．

結論として、本論文は、イベントベースの視覚を使用した動きの追跡、カメラの位置特定と追跡方法、ニューラルシーン表現の分野での進展を示し、速い動きの機械が自分自身を位置特定し、3D 世界を理解するための解決策を提供する．

Thesis Abstract

No. _____

Registration Number	☒ "KOU" ☐ "OTSU" No. _____ *Office use only	Name	Mana Masuda
Thesis Title Tracking and Scene Representation for Fast-moving Cameras by Event Data Synthesis			
Thesis Summary Machines are pervasive in our daily lives, from autonomous vehicle control to surgical assistance. The core of these technologies is the ability of machines to interpret the surrounding three-dimensional environment as sensory data and localize themselves using their "eyes", the camera. While many technologies have been developed to utilize existing frame-based cameras, these methods face limitations due to the dense recording system, making it difficult to scale to finer temporal resolutions or reduce computational complexity. The frame-based recording system is the most straightforward for visualizing the environment, and frame-based images are easy to interpret. However, the information from these cameras is often redundant when representing motion. Event-based cameras are bio-inspired vision sensors that mimic the retina, asynchronously reporting per-pixel brightness changes rather than capturing actual intensity images at regular intervals. Due to this unique operating principle, they offer exceptionally fine temporal resolution and low data consumption, making them ideal sensors for capturing motion. The goal of this thesis is to digitize the 3D world and enable machines to localize themselves from fast-moving vision by utilizing event-based cameras. To achieve this, the thesis first explores fast motion tracking using event-based vision along with differentiable scene representations. Secondly, a camera localization and tracking method for fast-moving cameras using event-based vision is introduced. Finally, a 3D scene reconstruction method for fast-moving cameras is presented, incorporating event-based vision. To achieve these objectives, this thesis presents several key contributions. First, it proposes a novel framework for camera tracking using event data. This framework utilizes an implicit representation via a pre-trained multi-layer perceptron (MLP) to synthesize events, allowing for sparse recovery of camera position and velocity by minimizing discrepancies between observed and generated events. Second, it develops a technique that leverages neural radiance field (NeRF) scene representations to synthesize events. By aligning NeRF's temporal gradient output with observed events, this method accurately recovers camera pose and motion when input parameters match actual camera			

movements. Third, the thesis introduces a method for high-fidelity 3D scene reconstruction from fast-moving cameras, achieving results within minutes. Integrating event-based and frame-based vision enhances camera tracking efficiency and eliminates motion blur, resulting in more precise 3D representations. Finally, it presents a method for reconstructing 3D scenes while optimizing camera poses from fast-moving cameras. Utilizing the proposed camera tracking techniques and an event-enhanced 3D scene reconstruction strategy, this method ensures high-fidelity reconstruction even with motion-blurred input images.

In conclusion, this thesis presents advancements in motion tracking, camera localization and tracking methods, and neural scene representations using event-based vision, offering solutions that allow fast-moving machines to localize themselves and comprehend the 3D world.