摘要

面向混合脉冲相机的高时空分辨率成像技术，旨在结合脉冲相机和传统数字相机在时空分辨率上的优势，实现对高速运动场景的连续精细记录，为后续视觉任务提供准确的视觉信息，是高速视觉感知领域的全新技术路线，推动了机器视觉从“离散帧分析”向“连续过程认知”的范式升级。“空-时双域优势互补”特性使面向混合脉冲相机的成像技术不仅能在体育科学分析和车辆安全性测试等领域中为分析环节提供全时序稠密动态序列，还能在无人机和自动驾驶领域中通过构建动态障碍物轨迹预测模型来增强无人机和车辆的避障能力，在高速场景应用中具有重要的研究意义和应用价值。

脉冲相机输出的脉冲流和传统数字相机输出的图像是两种不同模态的数据，充分利用各自在时空分辨率上的优势并将二者统一在同一模型中实现高时空分辨率成像存在以下几个挑战。首先，脉冲流和图像间的模态异质性以及分辨率差异给跨模态数据处理及有效融合带来了难题。其次，现有的光流估计方法在估计不同时刻间的运动时通常只能构建一阶模型，不能准确估计场景中出现的复杂非线性运动，导致最终的成像质量下降。最后，当重构时刻发生变化，现有方法通常基于输入图像间的运动关系线性地调整重构时刻与输入数据对应时刻间的运动信息，难以灵活可控地高质量重建带有复杂运动的影像。为此本文围绕混合数据高时空分辨率成像问题，从不同模态数据的融合成像、可泛化的脉冲流运动解析和重构时刻可控的成像三个方面出发设计解决方法。本文主要贡献包括：

针对混合数据间的模态异质性及分辨率差异，本文提出了一个能处理混合数据并有效利用各自优势的统一高时空分辨率成像模型。鉴于脉冲流数据包含场景中的连续光学信息，本文在所提的成像模型中设计了一条脉冲支路，用于从时域稠密的脉冲流中提取场景的瞬时光强特征。考虑到当前脉冲流的空间分辨率通常低于高清图像，在脉冲支路中加入了瞬时光强特征空间分辨率对齐模块，使来自脉冲流的瞬时光强特征与重构影像处于同一高空间分辨率下。为降低上采样过程中引入的误差对最终成像质量的影响，在脉冲支路中只保留高空间分辨率下光强特征中的场景结构、物体轮廓等低频信息。为利用输入图像在空间分辨率上的优势，所提模型还包括一条图像帧支路，用于提取图像中的高频纹理细节和色彩信息。此外，为有效融合来自两种模态数据的特征，本文还提出一种双向自适应隐式对齐的融合策略。实验证明所提模型充分利用了脉冲流和图像各自的高时间分辨率和高空间分辨率的优势，重构出高质量影像，为后续混合数据的高时空分辨率成像研究提供了基础。

针对复杂非线性运动建模精度有限的难点，本文提出了一个基于高效表征的无监

督光流估计模型。高时间分辨率的脉冲流能提供稠密的场景变化信息，本文基于脉冲流数据对场景中的复杂非线性运动建模。因为运动信息隐式包含于光强变化中，所提模型包含了一个对光强信息的高效表征模块。该模块结合脉冲流的时域相关性采用时域多扩张卷积的表征方法，利用一维扩张卷积代替常用的二维卷积在时域上以较小参数量获取较大感受野，建立更长范围内的时域相关性分析模型。该模型还包含一种全新的层注意力机制，增强时域上的上下文分析能力，动态选择适合参与运动估计的表征，以此提高运动建模的精度。此外，本文还提出首个基于脉冲流数据的无监督光流估计损失函数，突破了现有有监督方法对有标签数据的依赖，提高模型在新场景中的泛化能力。实验证明所提模型能有效提高运动建模的精度。

不同时刻上的重构影像和输入数据间存在不同的时间偏移，给混合数据融合下的连续场景刻画带来了挑战。针对这一挑战，本文基于上述混合数据成像模型和对脉冲流的运动分析模型，提出重构时刻可控的混合相机高时空分辨率成像模型。该模型包含一个可引入外部重构时刻参数的脉冲时域控制编码器，灵活地提取与重构影像相关的脉冲流和脉冲流特征，提高了特征提取的可控性。为高效地获取目标时刻上的高频细节和色彩，本文提出基于相似性的图像特征映射模块，利用来自脉冲流的运动信息将输入图像的特征映射至重构时刻并根据内容相似度融合。本文在双模态特征融合阶段设计了一个跨模态的互注意力融合机制。广泛实验证明所提模型可在任意指定时刻重构纹理细节清晰的影像，提高了高时空分辨成像的重构时刻可控性和成像质量。

基于上述三项工作，本文对混合脉冲相机的高时空分辨率成像问题展开探究，形成有效的高时空分辨率成像方法。上述方法的研究有效提高了面向高速复杂场景的高时空分辨率成像质量，对高速态势感知领域的发展提供有效支持。

关键词：混合脉冲相机，高时空分辨率成像，运动分析，重构时刻可控

High Spatio-Temporal Resolution Imaginig for Spike-RGB Camera

Lujie Xia (Computer Applied Technology)
Directed by: Prof. Ruiqin Xiong

ABSTRACT

High Spatio-Temporal Resolution Imaging (HSTRI) for spike-rgb camera aims to combine the advantages of spike cameras and rgb cameras in spatial and temporal resolution to record high-speed dynamic scenes with high spatio-temporal resolution, which can provide accurate visual information for subsequent tasks. This novel approach in high-speed visual perception advances machine vision from "discrete frame analysis" to "continous process cognition". The complementary "spatial-temporal dual-domain" characteristics of spike-rgb camera imaging not only provide full temporal dynamic sequences for analysis in fields such as sports science and vehicle safety testing but also enhance obstacle trajectory prediction models in drones and autonomous driving, demonstrating significant research value and application potential.

Spike streams and images are multimodal data with distinct characteristics. Achieving HSTRI by fully utilizing their respective advantage in spatial and temporal resolution and unifying them within a single processing model faces the following challenges. Firstly, the modal heterogeneity between spike streams and images, along with their resolution differences, poses significant challenges to cross-modal data processing and effective fusion. Secondly, existing optical flow estimation models usually rely on first-order approximations when analyzing motions. These models fail to accurately estimate complex non-linear motions in dynamic scenes, leading to degraded imaging quality. Finally, when the reconstruction timestamp changes, existing methods typically adjust motion information between the target image and input data linearly, based on the motions derived from input images. So These methods are difficult to controllably and flexiblly reconstruct high-quality images with complex non-linear motions. This thesis focuses on HSTRI and proposes methods from three aspects: a dual-modal hybrid imaging model integrating spike streams and images, generalizable motions analysis from spike streams and reconstruction time-controllable robust imaging. The main contirbutions of this thesis include:

To address the modal heterogeneity and resolution discrepancies among hybrid data, this

thesis proposes a unified HSTRI model capable of processing multimodal data while effectively leveraging their respective advantages. Given that spike streams contain continous optical information of the scene, the proposed model contain a spike branch to extract transient light intensity features from temporally dense spike streams. Considering that the spatial resolution of spike streams is typically lower than that of images, a spatial resolution alignment module is introduced in the spike branch. This module aligns the transient light intensity features extracted from spike streams with the high spatial resolution of the reconstructed images. To reduce the impact of errors introduced during upsampling on final imaging quality, the spike branch retains only low-frequency information, such as scene structures and object contours, from the high spatial resolution light intensity features. To leverage the spatial resolution advantages of input images, the proposed model also incorporates an image branch dedicated to extracting high-frequency texture details and color information. Furthemore, to effectively fuse features from the two modalities, this model introduces a bidirectional adaptive implicit alignment fusion strategy. Expriment results demonstrate that the proposed model fully leverages the respective resolution advantages of spike streams and images to reconstruct high-quality images. This work establishes a foundational data flow for subsequent research on spike-rgb hybrid data in HSTRI.

In terms of addressing the limited accuracy in modeling complex nonlinear motion, this thesis proposes an unsupervised optical flow estimation model based on efficient features representation. Leveraging the high temporal resolution of spike streams, which provide dense scene variation information, the model analyzes complex nonlinear motions in dynamic scenes. Since motion information is implicity contained in varying light intensity, the proposed model includes an efficient light intensity representation module for robust feature extraction. This module integrates the temporal correlations of spike streams by employing a temporal multi-dilated convolution representation. It replaces 2D convolutions with 1D dilated convolutions along the temporal dimension, achieving a larger receptive field with fewer parameters. This design enables the construction of a long-span temporal correlation analysis model. The proposed model also incorporates a novel layer attention mechanism to enhance temporal contextual analysis, dynamically selecting representations relevant to motion estimation. This significantly improves the precision of motion modeling. Furthermore, this thesis proposes the first unsupervised optical flow estimation loss function based on spike streams, eliminating the reliance of existing supervised methods on labeled data and thereby enhancing the model's generalization capablility in new scenarios. Experiment results demonstrate that the proposed

model significantly improves the accuracy of motion estimation.

In terms of the temporal offsets between reconstructed images at different timestamps and input data pose a challenge for feature extraction and mapping during reconstruction, this thesis proposes a hybrid spike-rgb camera HSTRI model with controllable reconstruction times based on the aforemationed hybrid data processing model and spike motion analysis model. The model contains a spike temporal control encoder that incorporates an external reconsturction time parameter, enabling flexible extraction of spike streams and spike features relevant to the target reconstrcuted images. This design significantly enhances the controllability of features extraction. To efficiently obtain high-frequency details and color information at the target reconstruction time, this model includes a similarity-based image feature mapping module. This module leverages motion estimated from spike streams to map the features of two input images to the reconstruction time and fuses them based on content similarity. Additionally, a cross-modal mutual attention fusion mechanism is designed for the dual-modal feature fusion stage, ensuring robust integration of heterogeneous data. Extensive experiments demonstrate that the proposed model can reconstruct high-quality images at arbitrarily times. This significantly imporves both the controllability of reconstrcution times and the robustness of HSTRI.

Based on the three aforementioned works, this thesis investigates the problem of high spatio-temporal resolution imaging for spike-rgb camera, forming effective high spatio-temporal resolution imaging methods. The research on these methods significantly ehhances the quality of high spatio-temporal resolution imaging in high-speed complex scenes and provides effective support for advancing the field of high-speed situational awareness.

KEY WORDS: Spike-RGB Camera, High Spatio-Temporal Resolution Imaging, Motion Analysis, Controllable Reconstruction Time