Ego-Motion Estimation for
Hybrid CMOS and Event Sensor

A Project Report

submitted by

KETUL SANJAYKUMAR SHAH

in partial fulfilment of requirements
for the award of the dual degree of

BACHELOR OF TECHNOLOGY AND MASTER OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS

MAY 2018

THESIS CERTIFICATE

This is to certify that the thesis titled **Ego-Motion Estimation for Hybrid CMOS and Event Sensor**, submitted by **Ketul Shah, EE13B133**, to the Indian Institute of Technology Madras, for the award of the dual degree of **Bachelor of Technology and Master of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Kaushik Mitra
Research Guide
Assistant Professor
Dept. of Electrical Engineering
IIT Madras, 600036

Place: Chennai
Date: 9th May 2018

ACKNOWLEDGEMENTS

I would like to begin by thanking my guide, Prof. Kaushik Mitra for all his help, support and patience throughout the course of my project work. His approach and dedication to research has been incomparable and has always inspired me to keep pushing myself. All the students in the Computational Imaging Lab have been immensely cooperative, friendly and accommodating and have greatly inspired my own work. I would especially like to thank my team mates Prasan and Dhruv for helping me out through the project.

ABSTRACT

KEYWORDS: Event Sensors, Image Reconstruction, 6-DoF pose estimation, Hybrid Sensors, Deep Image Prior

Event cameras are bio-inspired vision sensors that output pixel-level brightness changes instead of standard intensity frames. They offer significant advantages over standard cameras, namely a very high dynamic range, no motion blur, and a latency in the order of microseconds. Due to the nature of event sensors, a lot of spatial intensity information is lost. Previous attempts at recovering the intensity images use only event data. We make use of an additional low frame rate conventional camera and pose the image reconstruction problem as a novel view synthesis problem. Thus, we need to estimate ego-motion of the event sensor in order to reconstruct intensity frames, at a higher rate than the CMOS sensor. This task has been solved by using an unsupervised, learning free method. We have additionally used consistency loss to further improve pose estimation. Thus, we combine the strength of both by warping the low-frame rate video to intermediate locations where we only have event sensor data, resulting in high frame rate video. We show photorealistic reconstructions using this method on a real hybrid sensor.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Problem Outline and Motivation

As a result of the nature of the event sensor, unlike the conventional cameras, the spatial intensity information is lost. The algorithms developed for frame-based computer vision like object recognition, segmentation, etc. cannot be directly applied to event sensors. There have been works that attempt to bridge this gap by developing algorithms that convert event data back to intensity frames. Due to the high temporal resolution of event sensors, videos at a higher frame rate can be potentially obtained by such methods.

All of the previous methods use only the event stream in order to reconstruct intensity images. Due to this reason, the resulting reconstructed images using such methods do not look photorealistic, and regions where there are no events can go missing in the reconstructions. In order to avoid these issues, and to obtain more photorealistic reconstruction of images, we propose to use a low frame rate conventional camera along with the event sensor. Such a hybrid camera is commercially available, called Dynamic and Active Pixel Vision Sensor (DAVIS). The high texture intensity images from the low frame rate sensor can now be used to generate intensity images, by making use of the event stream. We thus make use of the strengths of both: conventional camera which has high texture intensity images and a event sensor which is sparse but has high temporal resolution. This intensity image reconstruction pipeline can be divided into four main blocks:

1. A learning free network for scene depth estimation using low frame rate intensity images
2. An auto-encoder for mapping event data to pseudo-intensity frames
3. A learning free pose network for estimating sensor pose
4. A warping module to warp the intensity frames using estimated depth and pose

In this thesis, Parts 3 and 4 of the above mentioned pipeline are explained and discussed in detail. That is, given the pseudo-intensity frames from event data and depth for intensity images, pose of the intermediate intensity frame with respect to either of the extreme frames have been obtained. Subsequently, in the warping module, both these extreme frames have been forward warped to the intermediate temporal location and blended using alpha blending.

CHAPTER 2

BACKGROUND

2.1 Introduction to Event Sensors

Event sensors are neuromorphic sensors which are a paradigm shift away from the traditional cameras. These sensors capture only the intensity changes in the scene, much like a human retina. Hence, they have advantages like ultra-low response latency, low data rate, high dynamic range and low power consumption over conventional cameras. Standard CMOS cameras send full frames at fixed frame rates. On the other hand, event-based sensors such as the DVS, have independent pixels that fire events at local relative brightness changes in continuous time. Specifically, if $I(x, y)$ is the brightness or intensity at point $u(x, y)^T$ in the image plane, the DVS generates an event at that location if the change in logarithmic brightness is greater than a threshold (typically 10-15$\%$ relative brightness change). These events are timestamped and transmitted asynchronously at the time they occur using sophisticated digital circuitry. Each event is a tuple $e_k =< x_k, y_k, t_k, p_k >$, where x_k, y_k are the pixel coordinates of the event, and $p_k \in -1, +1$ is the polarity of the event, which is the sign of the brightness change. This representation is sometimes also referred to as Address Events Representation (AER).

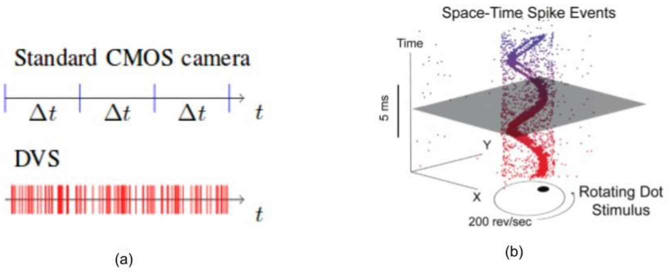

Figure 2.1: (a) Comparison of standard CMOS camera with event sensor output, (b) visualization of output of event sensor

CHAPTER 3

PRIOR WORK

3.1 Traditional Method

Traditionally, 6-DoF pose estimation between two views from images is generally carried out by first finding point or feature correspondences ($x_1 \leftrightarrow x_2$) in the images, and subsequently estimating the essential matrix from the obtained point correspondences. Essential matrix relates the point correspondences by enforcing the epipolar constraint (3.1). Once the essential matrix has been estimated, relative camera rotation (R) and translation (T) can be extracted from it using (3.2).

$$x_2^T E x_1 = 0 \tag{3.1}$$

$$E = \hat{T} R \tag{3.2}$$

3.2 Event Based algorithms for pose estimation

Due to high temporal resolution and low bandwidth requirements of event sensors, they have been used for localization and ego-motion estimation, for various applications ranging from drones to robots. Along similar lines, there has been an interest in Visual Odometry / SLAM algorithms research using event sensors. Kim *et al.* (2016) estimates 6-DoF camera motion, log intensity gradient and inverse depth using three decoupled probabilistic filters. Weikersdorfer *et al.* (2013) use only the event data for visual SLAM. EVO (Event-based Visual Odometry) by Rebecq et al. combines an event-based tracking approach based on image-to-model alignment with a recent event-based 3D reconstruction algorithm to achieve 6-DoF tracking in real time. Recently, a dataset by Mueggler *et al.* (2017) was proposed to benchmark event based pose estimation, visual odometry and SLAM algorithms. The dataset contains multiple video sequences captured with DAVIS and sub-millimeter accurate ground truth camera motion acquired using a motion-capture system.

3.3 Deep Learning algorithms for pose estimation

There has been significant interest in recent times to tackle simultaneous pose and depth estimation using supervised and unsupervised deep neural networks. Ummenhofer *et al.* (2017) formulate structure from motion as a supervised learning problem. They employ multiple encoder-decoder networks and additionally estimated surface normals and optical flow, along with structure and motion. Zhou *et al.* (2017) use a PoseNet and DispNet to estimate pose and depth in an unsupervised manner. Other recent works by Mahjourian *et al.* (2018) and Kendall *et al.* (2015) tackle similar problems of pose estimation using deep networks.

CHAPTER 4

6-DoF POSE ESTIMATION FROM EVENTS

In this chapter, we describe the approach we take to estimate the pose of intermediate location. We use pseudo-intensity images in order to obtain 6-DoF pose of the sensor in intermediate location, by using deep image prior as in Ulyanov *et al.* (2017). We assume that depth map is available at all low frame rate intensity images. We describe the problem formulation and the approach to solve it in detail in the following section.

As the event data contains noise which is dependent on various factors such as threshold, scene illumination, etc and cannot be modelled, it is not suitable for matching points or features. Hence, we map the event frames to pseudo-intensity frames (Fig 4.1). In order to learn this mapping, we propose an auto-encoder based deep learning model, which has been explained in more detail in section 3.2.1 in "Photorealistic Image reconstruction from Hybrid Intensity and Event based Sensor".

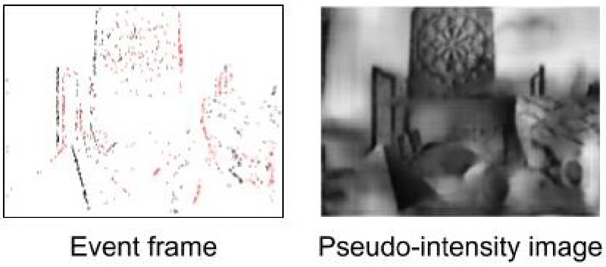

Figure 4.1: Event frame and corresponding pseudo-intensity frame

We propose to estimate pose by direct matching of these pseudo-intensity frames, as detecting and matching feature points in two views for the noisy event data is very challenging. We aim at reconstructing the intermediate intensity frames using this estimated pose obtained from event data, and using the texture rich intensity frames from the low frame rate conventional camera.

4.1 Overall Problem Formulation

Let two consecutive intensity images from low frame rate (25 fps) conventional camera be I_k and I_{k+1}. The event data in the time between two frames (i.e. 40 ms), is divided into N time blocks (each of time duration 40/N ms). Events in each of these time blocks are accumulated to form an event frame. This way, the event stream between the two frames I_k and I_{k+1} can be divided into N event frames denoted as e_k^j where $j = 1, 2, ..., N$. These event frames mapped to pseudo-intensity frames E_k^j where $j = 1, 2, ..., N$, using a learned auto-encoder based deep learning model as mentioned above. In the following pose estimation algorithm, it is assumed that the pseudo-intensity images E_k^j where $j = 1, 2, ..., N$ corresponding to the event frames e_k^j where $j = 1, 2, ..., N$ and E_k^0 and E_{k+1}^0 temporally corresponding to I_k and I_{k+1} are available. Furthermore, depth for intensity frames I_k and I_{k+1} are also assumed to be available. The dense depth maps used in this work have been obtained as described in detail in (paper). The overall problem has been depicted in Fig 4.2.

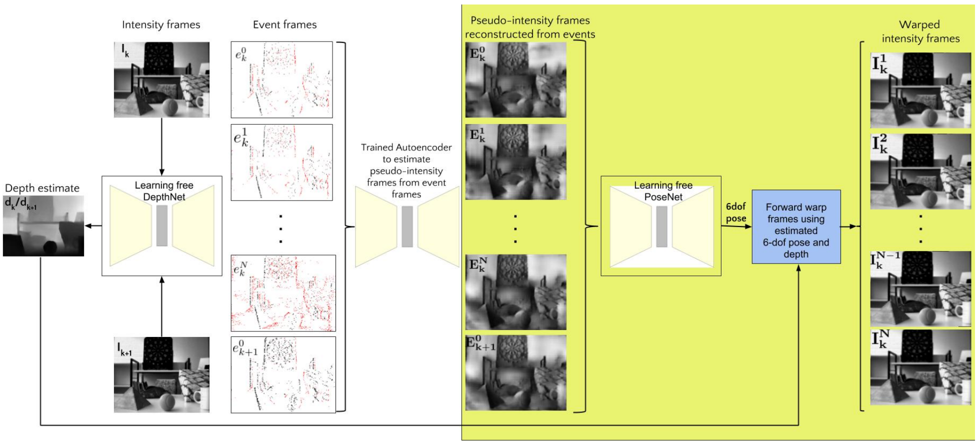

Figure 4.2: Overall problem diagram. The highlighted part in yellow is the subject of this thesis

4.2 Pose estimation using deep image prior

We propose to estimate 6-DoF relative pose between pseudo-intensity frames E_k^0 (temporally corresponding to I_k) and any intermediate E_k^j ($j = 1, 2, ..., N$) by direct matching. We use a convolutional encoder-decoder architecture (PoseNet)(Fig 4.3) which has been widely used for pose estimation.

As opposed to learning the weights of the parameter from a vast amount of data, we use this network as a handcrafted prior. Treating deep networks as handcrafted prior has shown excellent results for tasks like denoising, inpainting, etc by Ulyanov *et al.* (2017) in Deep Image Prior. The weights of the network (θ) act as the parameterization of the restored image. A loss function is appropriately defined according to the task at hand, and the network parameters are randomly initialized. A uniform noise (z) is given as input to the network, and the loss function is optimized w.r.t the network parameters till convergence.

$$x = f_\theta(z) \tag{4.1}$$

$$\theta^* = \underset{\theta}{\text{argmin}} E(f_\theta(z); x_0) \tag{4.2}$$

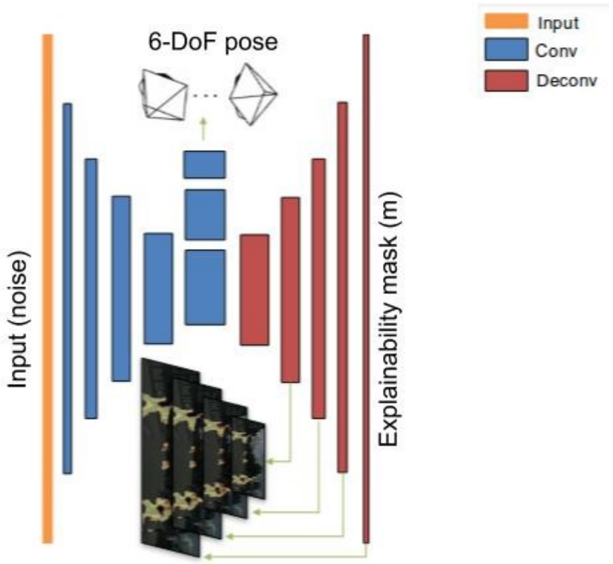

Figure 4.3: Architecture of PoseNet which is used as a handcrafted prior

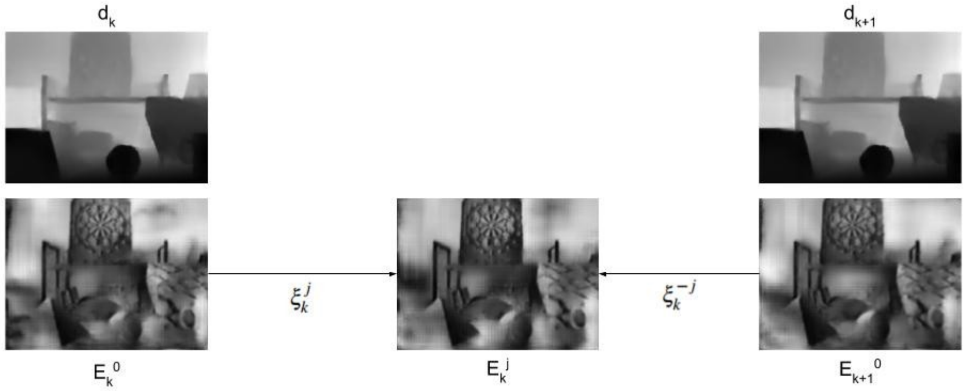

Figure 4.4: Relative pose ξ_k^j and ξ_k^{-j} have been estimated using PoseNet, via a learning free unsupervised method

We need to find the pose of the intermediate pseudo-intensity frame E_k^j w.r.t to E_k^0, given the depth map d_k for E_k^0. We use PoseNet as our network architecture, which corresponds to f in the above equation. The weights of the network correspond to parameterization θ, and the same uniform noise (z) is given as input to the network at each iteration of optimization. The network gives an estimate of relative pose ξ_k^j of E_k^j w.r.t E_k^0, along with an explainability mask m_k^j. Using this estimated pose and given depth d_k, we warp the intermediate pseudo-intensity frame E_k^j to E_k^0 to obtain $\hat{E}_k^0$. The photometric error $\mathcal{L}_p(E_k^0, \hat{E}_k^0, m_k^j) = \|(\hat{E}_k^0 - E_k^0) \odot m_k^j\|_1$ is minimized to obtain the estimate of the relative pose ξ_k^j. In order to avoid holes in the reconstructed images due to disocclusions and for increased consistency, we jointly estimate relative pose ξ_k^j of E_k^j and E_k^0 along with relative pose ξ_k^{-j} of E_k^j w.r.t E_{k+1}^0. For robustness, we compose the two relative pose ξ_k^j and ξ_k^{-j} to obtain the pose between E_k^0 and E_k^j, using which we warp I_k to I_{k+1} to obtain $\hat{I}_k$. The photometric loss $\mathcal{L}_{cons} = \|(\hat{I}_{k+1} - I_{k+1})\|_1$ is included in the total loss function. Relative pose estimates ξ_k^j and ξ_k^{-j} are obtained as

$$\xi_k^j, \xi_k^{-j}, m_k^j, m_k^{-j} = \underset{\xi_k^j, \xi_k^{-j}, m_k^j, m_k^{-j}}{\arg\min} \; \mathcal{L}_1 + \mathcal{L}_2 + \lambda_{cons}\mathcal{L}_{cons} \tag{4.3}$$

$$\mathcal{L}_1 = \mathcal{L}_p(E_k^0, \hat{E}_k^0, m_k^j) + \lambda_{reg}^p \mathcal{L}_{reg}(m_k^j); \;\; \mathcal{L}_2 = \mathcal{L}_p(E_{k+1}^0, \hat{E}_{k+1}^0, m_k^{-j}) + \lambda_{reg}^p \mathcal{L}_{reg}(m_2) \tag{4.4}$$

Here, $\lambda_{reg}^p(m)$ is binary cross entropy loss with a constant label 1 at each pixel location. This is to ensure that the explainability mask doesn't produce a zero mask to minimize the loss.

CHAPTER 5

WARPING AND BLENDING

5.1 Forward Warping of Intensity Images to Intermediate Location

We now have the relative pose ξ_k^j and ξ_k^{-j} of the intermediate location w.r.t the frames I_k and I_{k+1}. We also have depth maps d_k and d_{k+1} at these frames. Using these depth maps and 6-DoF pose, we can perform forward warping of the frames I_k and I_{k+1} to obtain intensity frames at intermediate location, I_k^{j1} and I_k^{j2} respectively. The knowledge of intrinsic matrix of the camera is essential for forward warping, which was provided in the dataset. The holes due to forward warping are filled by splatting the intensity values. In order to perform splatting, we use implementation given by scipy.interpolate.griddata.

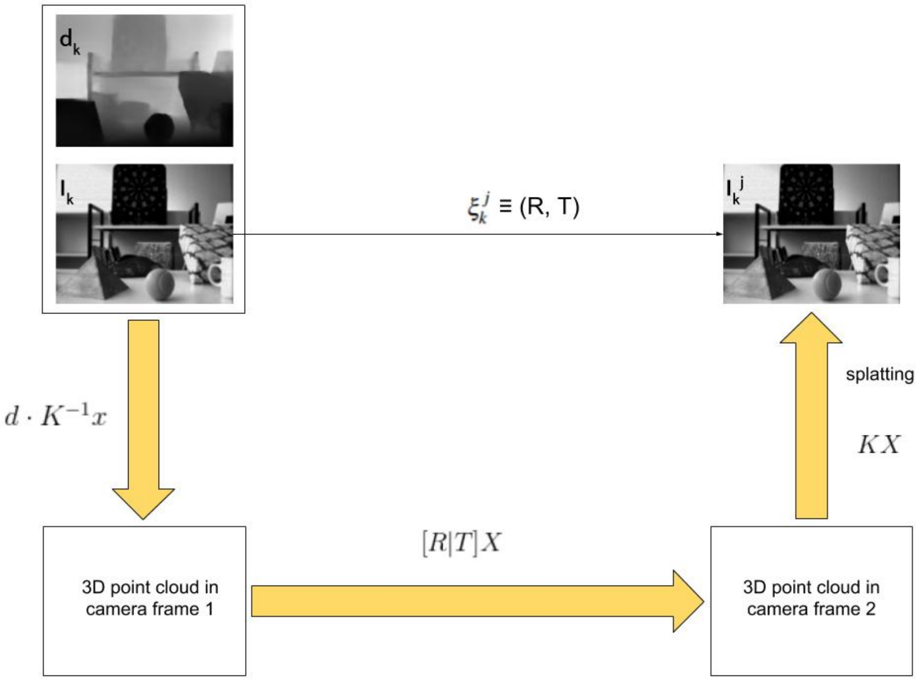

Figure 5.1: The figure shows forward warping in three steps. x denotes a 2-D image point in homogeneous coordinates, whereas X denotes a 3-D point in homogeneous coordinates. K=camera intrinsic matrix, R,T=relative camera pose

5.2 α Blending to get final intermediate intensity frame

Finally, we need to merge the intensity frames at intermediate location, I_k^{j1} and I_k^{j2} obtained by forward warping I_k and I_{k+1} respectively, in order to obtain the final intensity frame at intermediate location I_k^j. We use alpha blending in order to merge these images.

CHAPTER 6

RESULTS

We have used the recently proposed dataset by Mueggler *et al.* (2017) which consists of several video sequences captured using DAVIS. We divide the event stream between two frames I_k and I_{k+1} into 10 event frames, with the last event frame corresponding to the second intensity frame I_{k+1}. We have used the PyTorch implementation of PoseNet provided in Zhou *et al.* (2017). We found that we get optimal results when we use $\lambda^p_{reg} = 0.05$ and $\lambda_{cons} = 0.007$.

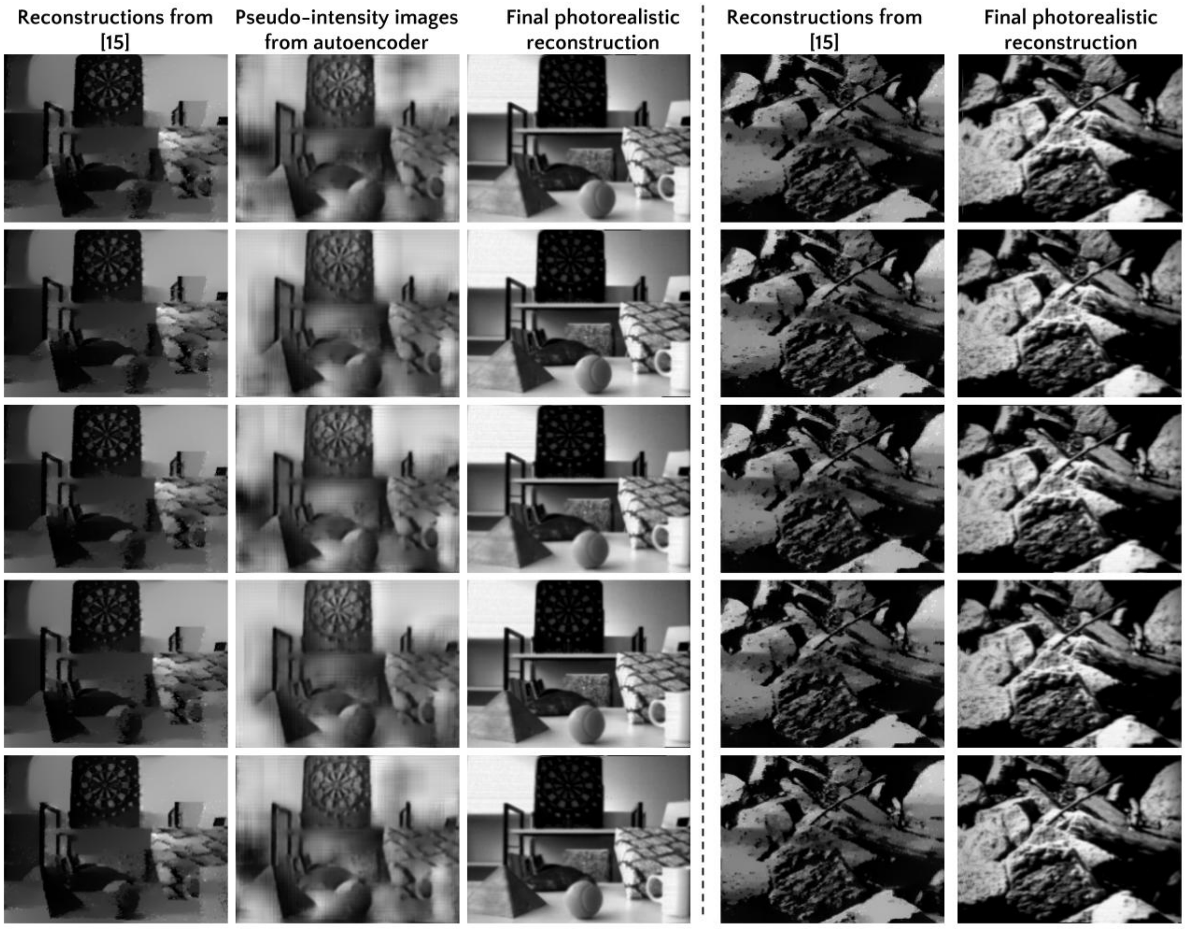
Reconstructions from [15]
Pseudo-intensity images from autoencoder
Final photorealistic reconstruction
Reconstructions from [15]
Final photorealistic reconstruction

Figure 6.1: Comparison of our results with Reinbacher *et al.* (2016)

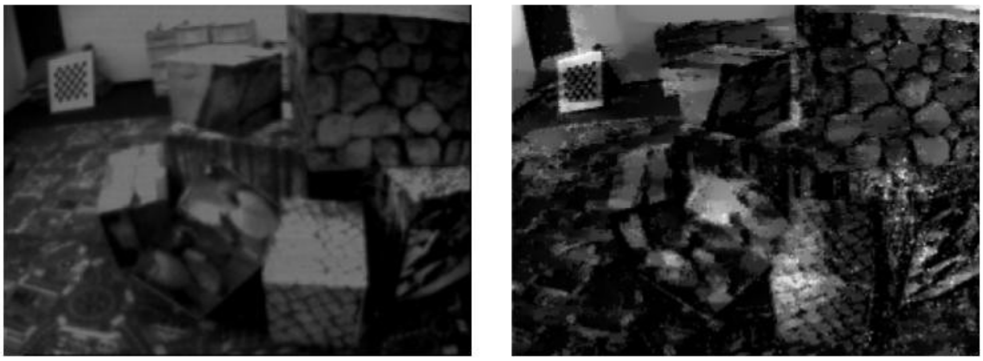

Figure 6.2: Results on DAVIS Dataset compared to [4]

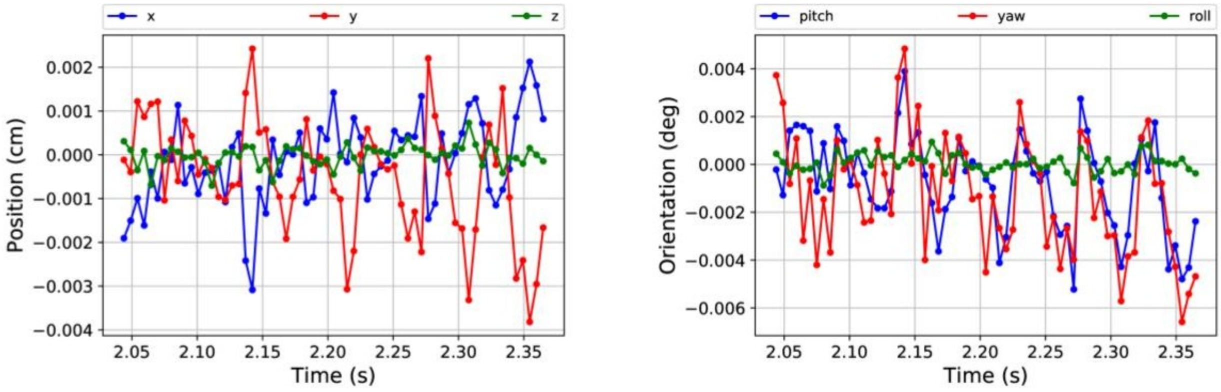

Figure 6.3: Position and Orientation error for the sequence "slider far"

CONCLUSION

We combine the strength of a texture-rich low frame rate conventional camera with a high temporal rate events camera to obtain photorealistic images at high temporal resolution. We achieve this by warping the low frame rate intensity frames from the conventional image sensor to intermediate locations. We compute dense depth maps from the low frame rate images and sensor ego-motion from the events data by direct matching of pseudo-intensity frames reconstructed from event frames. A three way consistency loss has been used to improve robustness of estimated pose. We warp images from both ends and blend them using alpha blending in order to avoid holes due to disocclusions. We show photorealistic results on DAVIS dataset.

REFERENCES

1. **Kendall, A.**, **M. Grimes**, and **R. Cipolla** (2015). Convolutional networks for real-time 6-dof camera relocalization. *CoRR*, **abs/1505.07427**. URL `http://arxiv.org/abs/1505.07427`.

2. **Kim, H.**, **S. Leutenegger**, and **A. J. Davison**, Real-time 3d reconstruction and 6-dof tracking with an event camera. *In European Conference on Computer Vision*. Springer, 2016.

3. **Mueggler, E.**, **H. Rebecq**, **G. Gallego**, **T. Delbruck**, and **D. Scaramuzza** (2017). The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research.*

4. **Reinbacher, C.**, **G. Graber**, and **T. Pock**, Real-Time Intensity-Image Reconstruction for Event Cameras Using Manifold Regularisation. *In 2016 British Machine Vision Conference (BMVC)*. 2016.

5. **Ulyanov, D.**, **A. Vedaldi**, and **V. Lempitsky** (2017). Deep image prior. *arXiv:1711.10925.*

6. **Ummenhofer, B.**, **H. Zhou**, **J. Uhrig**, **N. Mayer**, **E. Ilg**, **A. Dosovitskiy**, and **T. Brox**, Demon: Depth and motion network for learning monocular stereo. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

7. **Weikersdorfer, D.**, **R. Hoffmann**, and **J. Conradt**, Simultaneous localization and mapping for event-based vision systems. *In International Conference on Computer Vision Systems*. Springer, 2013.

8. **Zhou, T.**, **M. Brown**, **N. Snavely**, and **D. G. Lowe**, Unsupervised learning of depth and ego-motion from video. *In CVPR*. 2017.