



### Master Thesis

# Use of a DVS for High Speed Applications

Autumn Term 2018



Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

### **Declaration of originality**

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.					
I hereby confirm that I am the sole author of in my own words. Parts excepted are correct	the written work here enclosed and that I have compiled it ions of form and content by the supervisor.				
Title of work (in block letters):					
Authored by (in block letters): For papers written by groups the names of all authors a	re required.				
Name(s):	First name(s):				
sheet.  - I have documented all methods, data ar	plagiarism described in the 'Citation etiquette' information and processes truthfully.				
<ul><li>I have not manipulated any data.</li><li>I have mentioned all persons who were</li></ul>	significant facilitators of the work.				
I am aware that the work may be screened e	lectronically for plagiarism.				
Place, date	Signature(s)				

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.

### Contents

Pı	reface	iii
Al	bstract	$\mathbf{v}$
Sy	ymbols	vii
1	Introduction	1
2	6DoF Pose Tracking in Planar Scenes	3
	2.1 From Events to Frame	3
	2.1.1 Measuring the Sharpness of an Image	3
	2.1.2 Planar Homography	4
	2.2 From Frames to Map	5
	2.2.1 Map	5
	2.2.2 Tracking	5
	2.2.3 Mapping	6
	2.2.4 Relocalization	8
	2.3 Experiments	8
3	A Discussion to 6DoF Motion Estimation in General 3D Scenes	11
	3.0.1 note	11
4	Discussion	12
	4.1 Erstellen einer Tabelle	12
	4.2 Weitere nützliche Befehle	14
Bi	ibliography	15
$\mathbf{A}$	Irgendwas	17
В	Datasheets	19

### Preface

Bla bla ...

### Abstract

Hier kommt der Abstact hin ...

# **Symbols**

### Symbols

 $\phi, \theta, \psi$  roll, pitch and yaw angle

b gyroscope bias

 $\Omega_m$  3-axis gyroscope measurement

### Indices

x x axis y y axis

### Acronyms and Abbreviations

ETH Eidgenössische Technische Hochschule

EKF Extended Kalman Filter
IMU Inertial Measurement Unit
UAV Unmanned Aerial Vehicle
UKF Unscented Kalman Filter

# Introduction

Hier kommt die Einleitung DVS is

# 6DoF Pose Tracking in Planar Scenes

Throughout this work the following notation is employed: W denotes the world frame,  $C_1$  or  $C_2$  denotes a camera frame.  $T_{AB}$  is the transformation from frame A to frame B, measured in frame A.

X the position of the event with respect to world or camera frame, x the calibrated coordinates of the event.

#### 2.1 From Events to Frame

We group a set of events  $\mathcal{E} \doteq \{e_k\}_{k=1}^N$  into a temporal window, optimize the motion and scene parameters within this window, then shift the window to the next set of events and repeat this process. The temporal window size is defined by the event numbers N, which should be chosen small enough so that a constant velocity model could be applied within this window. We choose event numbers against a fixed time interval to define the window size, because this corresponds to the data-driven nature of an event-based camera: the more rapid the apparent motion of the scene is, the larger the event rate will be. If the scene stops moving, no events will be generated, the pose will also not be further updated.

An event frame is thus formed by summing up events within this window. If we simply sum along the time axis, the intensity at each pixel will be the sum of the polarities of all the events that are triggered at this pixel location within the window

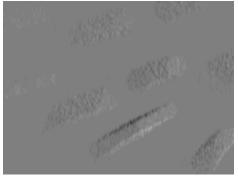
$$\mathcal{I}(\boldsymbol{x}) = \sum_{1}^{N} \pm_{k} (\boldsymbol{x} - \boldsymbol{x}_{k}), \tag{2.1}$$

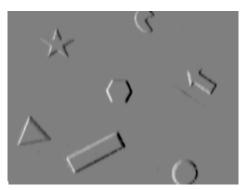
with  $\pm_k$  and  $\boldsymbol{x}_k$  denoting the polarity and pixel coordinates of the kth event, respectively. After warping the events with  $\boldsymbol{x}_k' = \boldsymbol{W}(\boldsymbol{x}_k, t; \theta)$ , we substitute  $\boldsymbol{x}_k$  in the above equation to  $\boldsymbol{x}_k'$ .

#### 2.1.1 Measuring the Sharpness of an Image

There are several metrics one could choose from to measure the contrast of an image. A local contrast metric could be, for example, convoluting the image with a high pass filter

$$C_H = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}, \tag{2.2}$$





(a) An optimized frame using sharpening filter metric

(b) An optimized frame using variance metric

Figure 2.1: Comparison between two different metrics

and sum the pixel value of the filtered image. However, this metric only compare a pixel with its 8 neighbors, thus an image with scattered events (large noise) is also a valid configuration, as shown in fig. 2.1(a). The Michelson contrast [1], defined as  $\mathcal{C}_M = \left(\mathcal{I}_{\max} - \mathcal{I}_{\min}\right) / \left(\mathcal{I}_{\max} + \mathcal{I}_{\min}\right)$ , only considers the highest and lowest luminance in the image and is thus more suitable to quantify contrast for periodic functions.

We choose to measure the contrast by the variance of the image, defined by

$$\operatorname{Var} \doteq \frac{1}{|\Omega|} \int_{\Omega} (\mathcal{I}(\boldsymbol{x}) - \mu(\mathcal{I}(\boldsymbol{x})))^{2} d\boldsymbol{x}, \tag{2.3}$$

where  $\mu(\mathcal{I}(\boldsymbol{x}))$  is the average image intensity. Since our goal is to align events triggered from the same visual stimuli, the variance would be a very suitable metric, as a squared metric favors the configuration that projects as many events as possible to the same pixel.

#### 2.1.2 Planar Homography

The warp function  $\mathbf{x}' = \mathbf{W}(\mathbf{x}, t; \theta)$  does not only depend on the motion parameters, but also the scene parameters, which is the unknown depth. In the case of a planar scene the problems simplifies, since a plane  $\mathbf{P}$  can be parameterized by two sets of parameters:  $\mathbf{n} \in \mathbb{S}^2$  the unit surface normal of  $\mathbf{P}$  with respect to the current camera frame, and d the distance from the camera center to  $\mathbf{P}$ . The warp function then becomes

$$X' = R(t)X + T(t) \tag{2.4}$$

$$\boldsymbol{X} = \boldsymbol{R}(t)^{\top} \left( \boldsymbol{X}' - \boldsymbol{T}(t) \right) \tag{2.5}$$

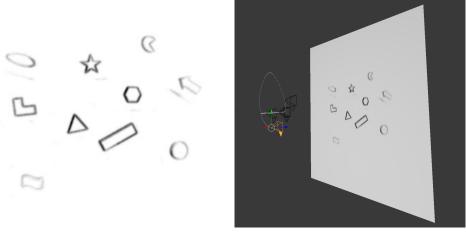
$$\boldsymbol{X} = \boldsymbol{R}(t)^{\top} \left( \boldsymbol{I} + \boldsymbol{T}(t) \boldsymbol{n}^{\top} / d \right) \boldsymbol{X}', \tag{2.6}$$

thus  $\mathbf{x}' \sim (\mathbf{R}(t)^{\top} (\mathbf{I} + \mathbf{T}(t)\mathbf{n}^{\top}/d))^{-1} \mathbf{x}$ . Here  $(\mathbf{R}(t), \mathbf{T}(t)) \in SE(3)$  denotes the relative pose between two cameras at which the current event being warped and the first event within the window happened, and t is the relative timestamp with respect to the first event. Under a constant velocity model with linear velocity  $\mathbf{v} \in \mathbb{R}^3$  and angular velocity  $\mathbf{\omega} \in \mathbb{R}^3$ , the translation is given by

$$T(t) = vt, (2.7)$$

the rotation matrix is given by the exponential map exp:  $\mathfrak{so}(3) \to SO(3)$ :

$$\mathbf{R}(t) = \exp(\boldsymbol{\omega}^{\wedge} t), \tag{2.8}$$



(a) The texture of a map

(b) A map in the global frame

Figure 2.2: Map

where  $^{\wedge}$  is the *hat* operator

$$\boldsymbol{\omega}^{\wedge} = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix} = \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix} \in \mathfrak{so}(3). \tag{2.9}$$

### 2.2 From Frames to Map

The contrast maximization procedure in the above section optimizes the relative pose between successive frames. We show in this section that the same idea can be applied to perform global pose tracking in planar scenes. We first explain how the map is defined, and how to track a known map, then we shown how this map is built by selecting a set of keyframes.

#### 2.2.1 Map

A map is a plane with three components: the normal direction  $n_w$ , the distance  $d_w$  to the origin, and the texture; the texture of a map represents all the edges on the plane. Figure 2.2 shows the an example of such map. Figure 2.2(a) also shows the set of keyframes used to construct the map. We will talk more about keyframes in section 2.2.3. The global coordinate is chosen as the camera coordinate of the first frame.

#### 2.2.2 Tracking

Suppose a map is present, then the normal direction  $n_w$  of the plane and the distance  $d_w$  to the origin are known. Also the pose of the current frame  $(\mathbf{R}_{wc}, \mathbf{T}_{wc}) \in SE(3)$  is determined by the motion estimation from the last frame (a quick note to the terminology we are using: whenever we say the *pose* of a frame, we always refer to the camera *pose* at which the first event within the frame happens). The parameters left to be estimated for each frame is  $\phi = (\omega, \mathbf{v}) \in \mathbb{R}^6$ . By substituting  $\mathbf{n}$  with  $\mathbf{n}_c = \mathbf{R}_{cw} \mathbf{n}_w$ , and d with  $d_c = d_w + \mathbf{T}_{wc} \cdot \mathbf{n}_w$  in eq. (2.6), we get the homography matrix within each frame as

$$\boldsymbol{H}_{1} = \boldsymbol{R}^{\top} \left( \boldsymbol{I} + \boldsymbol{T} \boldsymbol{n}_{c}^{\top} / d_{c} \right) \tag{2.10}$$

and  $\boldsymbol{x}_c' \sim \boldsymbol{H}_1^{-1} x_c$ .

A nonlinear optimizing problem naturally suffers from local optima. Without a good initialization, the motion computed with the method in section 2.1 could sometimes be a local optimum delivering an image that appears sharp, despite being wrongly estimated (see fig. 2.3). In order to make sure that the estimated motion from the per frame contrast maximization also conform to the global map. Thus we perform another optimization, where we project the events of the current frame to the global map. The parameter set is still  $(\omega^{\top}, v^{\top})$ , and we use the output from last procedure as an initial guess.

The procedure described in the first paragraph of this section can be understood as projecting the events on a blank canvas. Similarly, in the projecting-to-map procedure we project the events on the texture of the map, and measure the strength of the synthesized image with the same variance function as in eq. (2.3), thus finding the set of the parameters that best align the events in the current frame to their correspondences in the texture.

The projection from an event to the map is

$$x_w \sim R_n H_2^{-1} H_1^{-1} x_c,$$
 (2.11)

with  $R_n$  the transformation from the orientation of the global frame to the orientation of the map, computed by

$$\boldsymbol{K} = (\boldsymbol{n}_w \times \boldsymbol{z})^{\wedge} \tag{2.12}$$

$$\mathbf{R}_n = \mathbf{I} + \mathbf{K} + \mathbf{K}^2 / (1 + \mathbf{n}_{\boldsymbol{w}} \cdot \boldsymbol{z}), \tag{2.13}$$

where z = (0, 0, -1) denotes the plane fronto-parallel to the camera, and

$$\boldsymbol{H}_{2} = \boldsymbol{R}_{cw} \left( \boldsymbol{I} + \boldsymbol{T}_{wc} \boldsymbol{n}^{\top} / d_{w} \right) \tag{2.14}$$

is the projection from the current frame to the global frame, with  $\mathbf{R}_{wc}$ ,  $\mathbf{T}_{wc}$  being the pose of the current frame.  $\mathbf{H}_1$  is the planar homography for each frame as in eq. (2.10). But the  $\mathbf{R}(t)$  and  $\mathbf{T}(t)$  might be different since we are refining these parameters.

After projecting, we obtain an image composed of the map texture and the events of the current frame. We maximize the contrast of this image using the cost function defined in eq. (2.3). The optimized velocity is used for propagating the pose to the next incoming frame via

$$T_{wc_2} = R_{wc_1} v \Delta t + T_{wc_1} \tag{2.15}$$

$$\mathbf{R}_{wc_2} = \mathbf{R}_{wc_1} \exp(\boldsymbol{\omega}^{\wedge} \Delta t), \tag{2.16}$$

where  $c_1$  and  $c_2$  denotes the current frame and the next frame, respectively, and  $\Delta t$  is the temporal size of the current frame.

#### 2.2.3 Mapping

After having collected the first N events, we start the mapping process. For the first frame we estimate the full parameter set  $\phi = (\boldsymbol{\omega}^\top, \boldsymbol{v}^\top/d_w, \varphi, \psi)^\top \in \mathbb{R}^8$ , where  $(\varphi, \psi)$  parametrize the unit vector of the plane normal, and  $\boldsymbol{v}^\top/d_w$  account for the scale ambiguity problem introduced in linear velocity estimation from monocular camera. We can for example determine the scale of the scene by setting  $d_w$  to 1m, then scales of the consecutive frames are also determined by  $d_c = d_w + T_{wc} \cdot \boldsymbol{n}_w$ . From the optimized first frame we initialize the map, by projecting the frame to the planar scene via a rotation  $\boldsymbol{R}_n$  as in eq. (2.13) (you might have implemented this

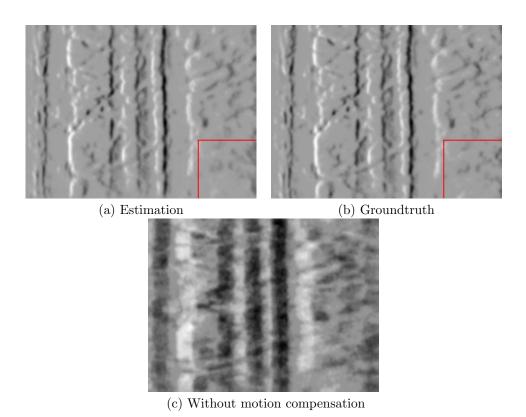


Figure 2.3: An example of local optima. This is the dataset  $slider\_hdr\_close$  with a window size of 50000 events. (a) shows the optimized image with  $linear\ velocity\ (0.231,0.109,0.256),\ angular\ velocity\ (0.405,-0.130,-0.278)$  and  $plane\ normal\ (-0.579,0.282,-0.765)$ . (b) shows the result using groundtruth parameters with  $linear\ velocity\ (0.163,0,0),\ angular\ velocity\ (0,0,0)$  and  $plane\ normal\ (0,0,-1)$ . Both images appear mostly identical, though at the lower right corner, for example, one can still recognize the difference. Also both images look much sharper than the image without motion compensation in (c). It is worth mentioning that the contrast of the estimation is actually slightly larger than that of the groundtruth

part wrong). Then we can track the next frames with this map using the method described in section 2.2.2.

As the camera moves, there might be new information available in the scene, so that the map needs to be expanded. After optimization for each single frame, we also measure the *per-pixel-overlap* between the frame and the map, that is, how many percent pixels in the frame can be explained by the map. The overlap might be small, when the camera is just exploring a new area so that only part of the frame and the map is overlapping; another possibility is that the map is not accurate enough to explain the current frame, which often happens when only one frame is used until now to construct the map. When the overlap reaches a certain threshold (0.8 for example), we try to insert a new *keyframe*.

When the system decides that a new keyframe is needed, we collect all the keyframes until now, plus the current frame, which is a keyframe candidate, and optimize the poses and velocities of these frames, as well as the map together. Suppose there are k frames (including the keyframe candidate), the to be optimized parameter set is

$$\phi = (\mathbf{R}_{(1\sim k)}, \mathbf{T}_{(1\sim k)}, \boldsymbol{\omega}_{(0\sim k)}, \boldsymbol{v}_{(0\sim k)}/d_w, \varphi, \psi) \in \mathbb{R}^{12k-4}.$$

Here  $1 \sim k$  means from frame 1 to frame k, and we skip the pose optimization of frame 0, which is the first frame, since it defines the global coordinate. We project all the events of these k frames to the plane parametrize by  $(\varphi, \psi)$  with eq. (2.11), and again optimize the contrast of the synthesized image the obtain the optimal parameter set.

Note that in the mapping process we drop the polarity of the event, since different keyframes might include events from the same visual stimuli but triggered when moving in different directions, when summed together the polarities might cancel each other out. For the same reason, we also don't use the polarity when matching a frame to the map.

After this step, we also measure the *per-pixel-overlap* between the keyframe candidate and the image synthesized by the other keyframes, with the newly optimized parameter set. If this is a good match (define a good match), we continue the tracking. Otherwise, we consider the pose estimation of the current frame to be already off. In this case we preserve the former map, and use this map to start the *relocalization* procedure.

#### 2.2.4 Relocalization

### 2.3 Experiments

Despite of the rapid movements, in the above datasets the camera actually only moves in a relative small region in front of the textured plane, as depicted in fig. 2.4. However, this method also applies when camera travels a longer distance with respect to the scene depth, as shown in fig. 2.5. Also, although this method is designed for planar scenes, in a complex scene with rotation only movement, where no scene parameters are needed, this method can still be applied to build a global map, delivering a result similar to *image stitching*. Figure 2.6 is such an example.

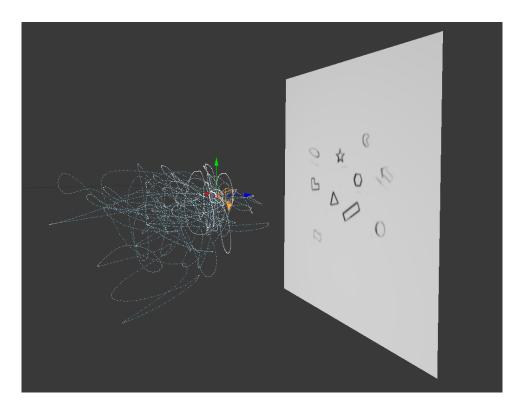


Figure 2.4: The motion path of the dataset  $shapes\_6dof$ . Despite a trajectory length of about 50m, in the whole 60s the camera actually stays in a relative small region compared to the scene depth, and there is no observable increase of drift using the method in this work. ref to error estimation



Figure 2.5: The dataset  $slider\_hdr\_far$ , with a scene depth of 0.584m, and a camera movement in the positive x direction only. Note that this is a much wider map than what we have seen before. The figure shows the result at 4.8sec within the whole range of 6.3sec; afterwards the algorithm is lost in the forest, where almost all the textures are vertical, causing severe local optima problem. If we constrain the motion estimation to translation only, it delivers a much better result.

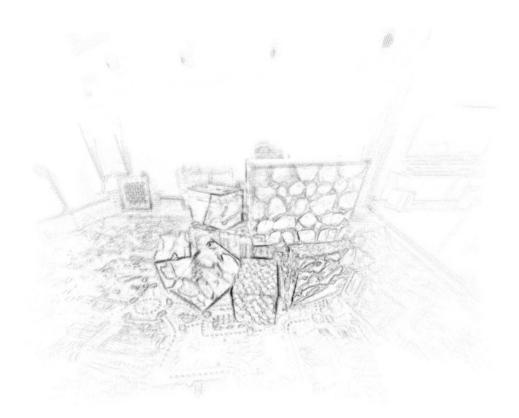


Figure 2.6: The dataset  $boxes\_rotation$ . The motion in this dataset is rotation-dominated

# A Discussion to 6DoF Motion Estimation in General 3D Scenes

#### 3.0.1 note

tried initialize with multiple frames, didn't work very well. similarly sliding window didn't work; too many events didn't work

### Discussion

One possible reason why the algorithm sometimes get lost is that there are too few textures available, as shown in fig. 4.1(a). At this time we usually need to rely on the relocalization or bundle adjustment pipeline. The current implementation needs a good initial guess, so it does not always work reliably.

#### 4.1 Erstellen einer Tabelle

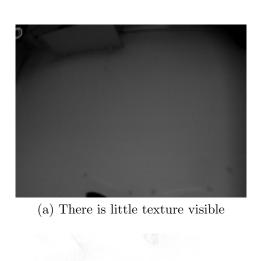
Ein Beispiel einer Tabelle:

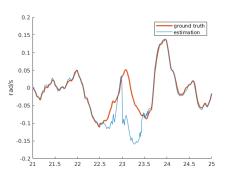
Table 4.1: Daten der Fahrzyklen ECE, EUDC, NEFZ.

Kennzahl	Einheit	ECE	EUDC	NEFZ
Dauer	S	780	400	1180
Distanz	$\mathrm{km}$	4.052	6.955	11.007
Durchschnittsgeschwindigkeit	$\mathrm{km/h}$	18.7	62.6	33.6
Leerlaufanteil	%	36	10	27

Die Tabelle wurde erzeugt mit:

```
\begin{table}[h]
\begin{center}
  \caption{Daten der Fahrzyklen ECE, EUDC, NEFZ.}\vspace{1ex}
  \label{tab:tabnefz}
  \begin{tabular}{11|ccc}
    \hline
    Kennzahl & Einheit & ECE & EUDC & NEFZ \\ hline
    Dauer & s & 780 & 400 & 1180 \\
    Distanz & km & 4.052 & 6.955 & 11.007 \\
    Durchschnittsgeschwindigkeit & km/h & 18.7 & 62.6 & 33.6 \\
    Leerlaufanteil & \% & 36 & 10 & 27 \\
    \hline
  \end{tabular}
  \end{center}
\end{table}
```





(b) Relocalized after get lost (roll component)



#### (c) Polluted map after tracking is lost

Figure 4.1: When tracking the dataset *shapes\_translation*, we always get lost at around 23sec, where only a small part of the oval is visible (a). When later there are more textures available, the camera has already moved some distance so that the current estimation of the pose might already be far off. When the algorithm confirms that there is little overlap between the current frame and the map at this pose, it tries to insert a new keyframe. However, because of the lack of a good initial guess, this new keyframe won't necessarily conform to other already inserted keyframes. The current implementation doesn't reinitialize the map after the tracking is lost. The advantage is that there is still chance to come back to the original map after tracking a wrong *local map* (b); the disadvantage is that the global map will most likely be polluted by the *local map* (c)

### 4.2 Weitere nützliche Befehle

Hervorhebungen im Text sehen so aus: hervorgehoben. Erzeugt werden sie mit dem  $\epsilon$ . Befehl.

Einheiten werden mit den Befehlen \unit [1] {m} (z.B. 1 m) und \unitfrac [1] {m} {s} (z.B. 1 m/s) gesetzt.

## **Bibliography**

- [1] A. A. Michelson, Studies in optics. Courier Corporation, 1995.
- [2] M. Raibert, Legged Robots That Balance. Cambridge, MA: MIT Press, 1986.
- [3] M. Vukobratović and B. Borovac, "Zero-moment point thirty five years of its life," *International Journal of Humanoid Robotics*, vol. 1, no. 01, pp. 157–173, 2004.
- [4] G. A. Pratt and M. M. Williamson, "Series elastic actuators," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1995, pp. 3137–3181.

Bibliography 16

# Appendix A

# Irgendwas

Bla bla ...

# Appendix B

# Datasheets

