Visual Modelling and Collision Avoidance in *Dynamic Environments* from Monocular Video Sequences

Darius Burschka

Machine Vision and Perception Group (MVP)
Department of Computer Science

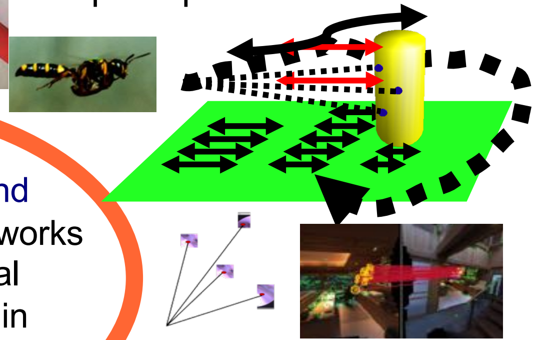
Technische Universität München

Research of the MVP Group

Perception for manipulation

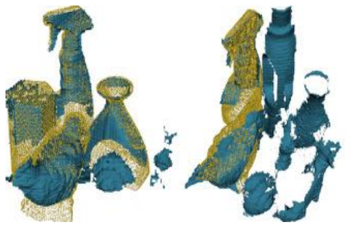
Visual navigation

Biologically motivated perception



The Machine Vision and Perception Group @TUM works on the aspects of visual perception and control in medical, mobile, and HCI applications

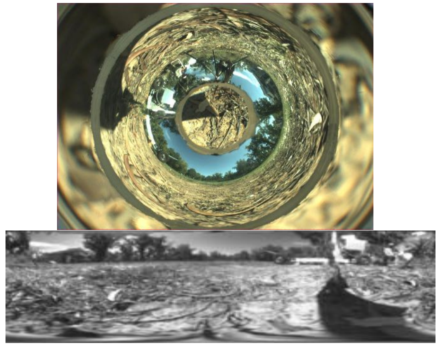
Rigid and Deformable Registration



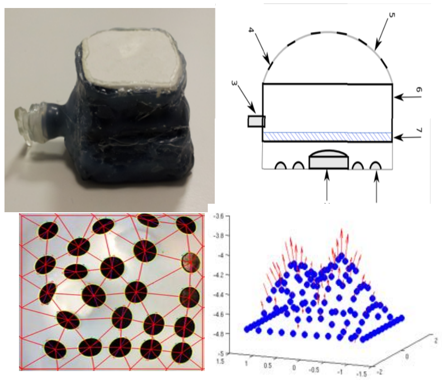
Photogrammetric monocular reconstruction

Visual Action Analysis

Research of the MVP Group



Sensor substitution

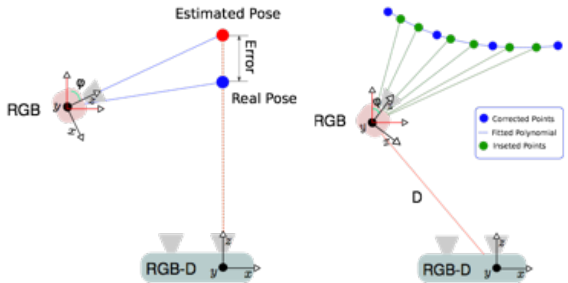


The Machine Vision and Perception Group @TUM works on the aspects of visual perception and control in medical, mobile, and HCI applications

Development of new Optical Sensors

Multimodal Sensor Fusion

Exploration of physical object properties



Applications (past German Aerospace (DLR) collaborations)

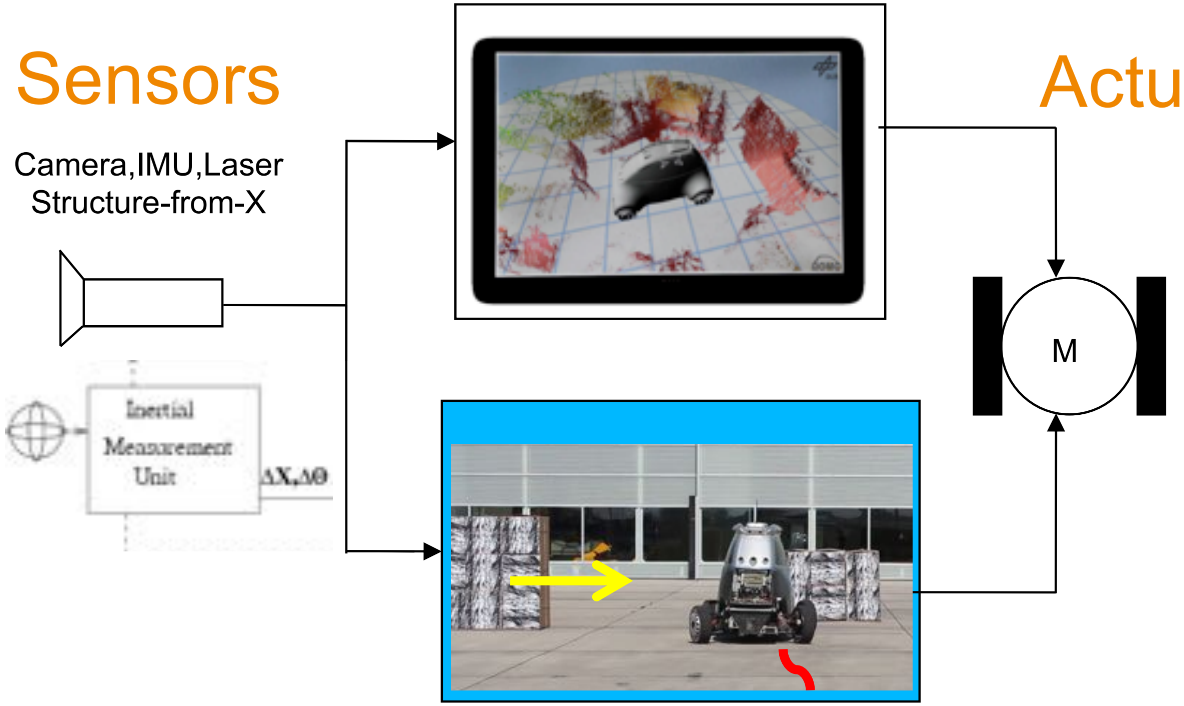
Coupling Alternatives for Perception Modules

Map-based action planning (not real-time)
(Metric Representation)

Sensors

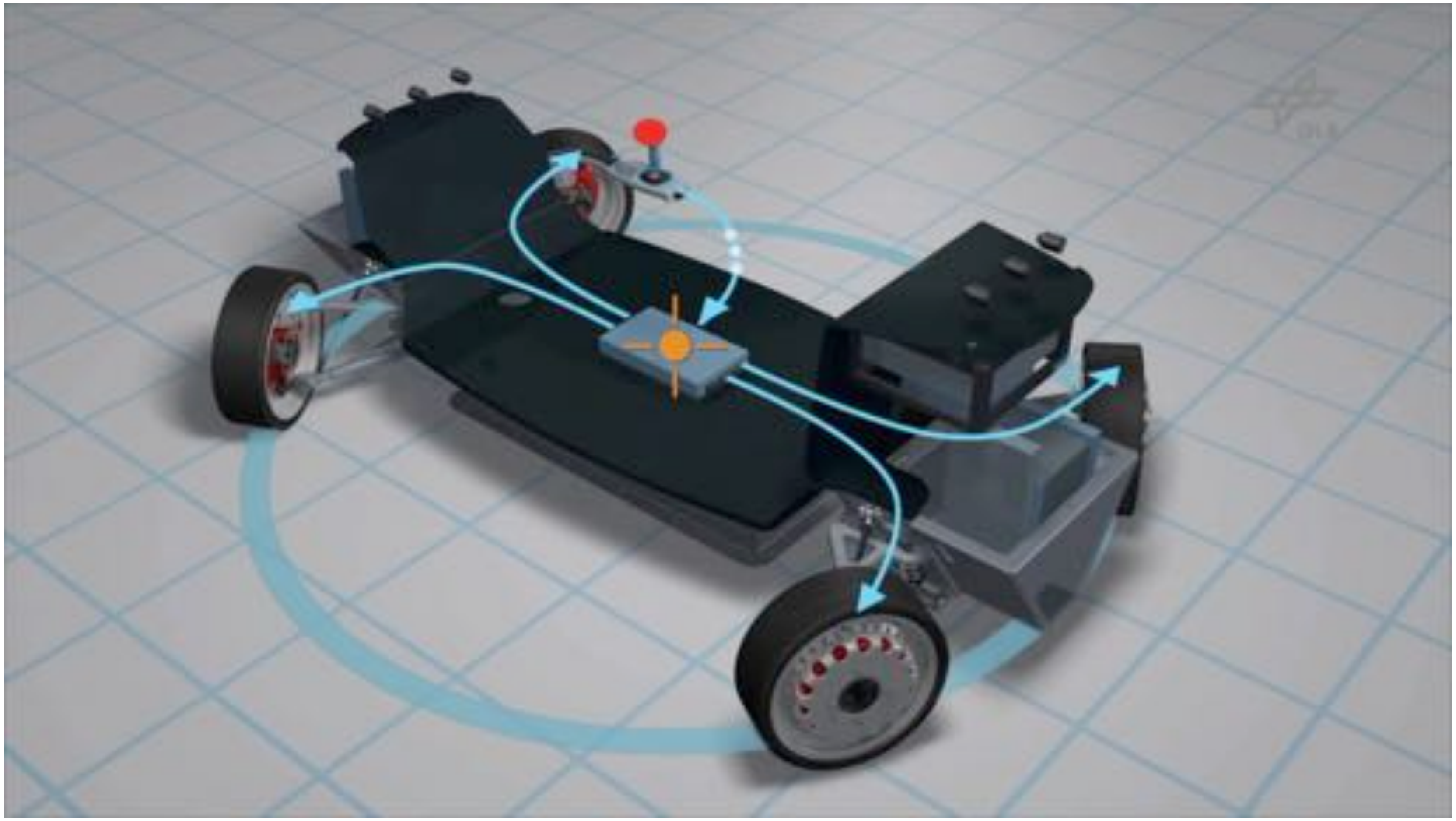
Camera,IMU,Laser
Structure-from-X

Actuators



Reactive behavior(Instincts), e.g., Obstacle avoidance,…
(real-time for control)
do we really need metric representation here?

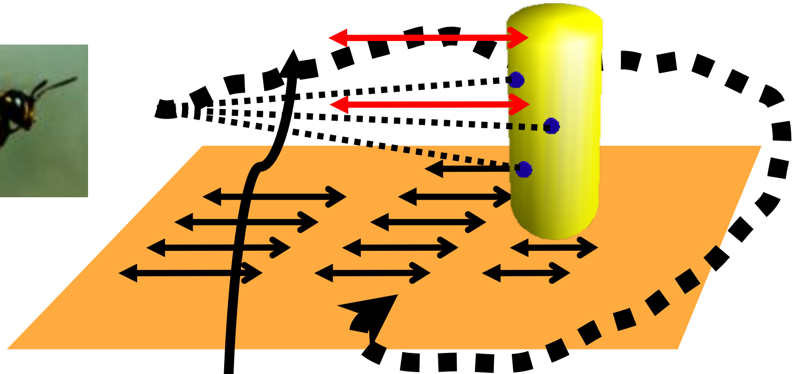
Our Experimental Platform (RoboMobil DLR)

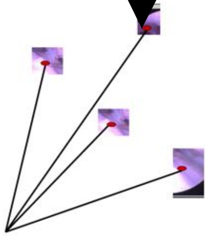


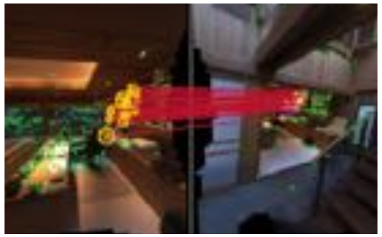
Early Monocular Navigation Approaches VGPS (IROS 2003)

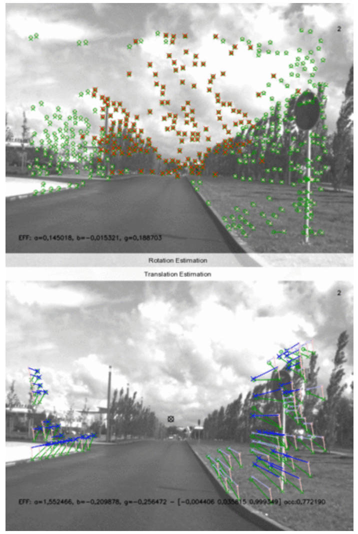
Biology helps to increase robustness

Mair, Burschka
Mobile Robots Navigation, book chapter, In-Tech, 2010

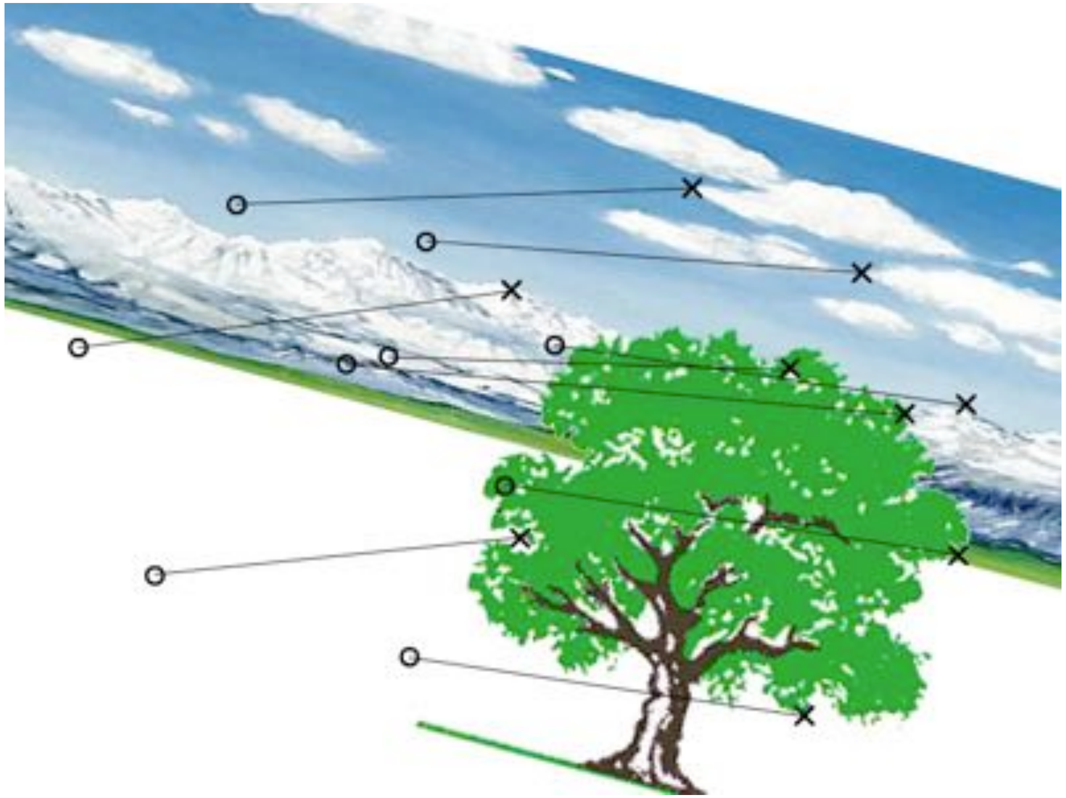








Can we navigate directly from monocular video? (Zinf system, Burschka et al. 2008)



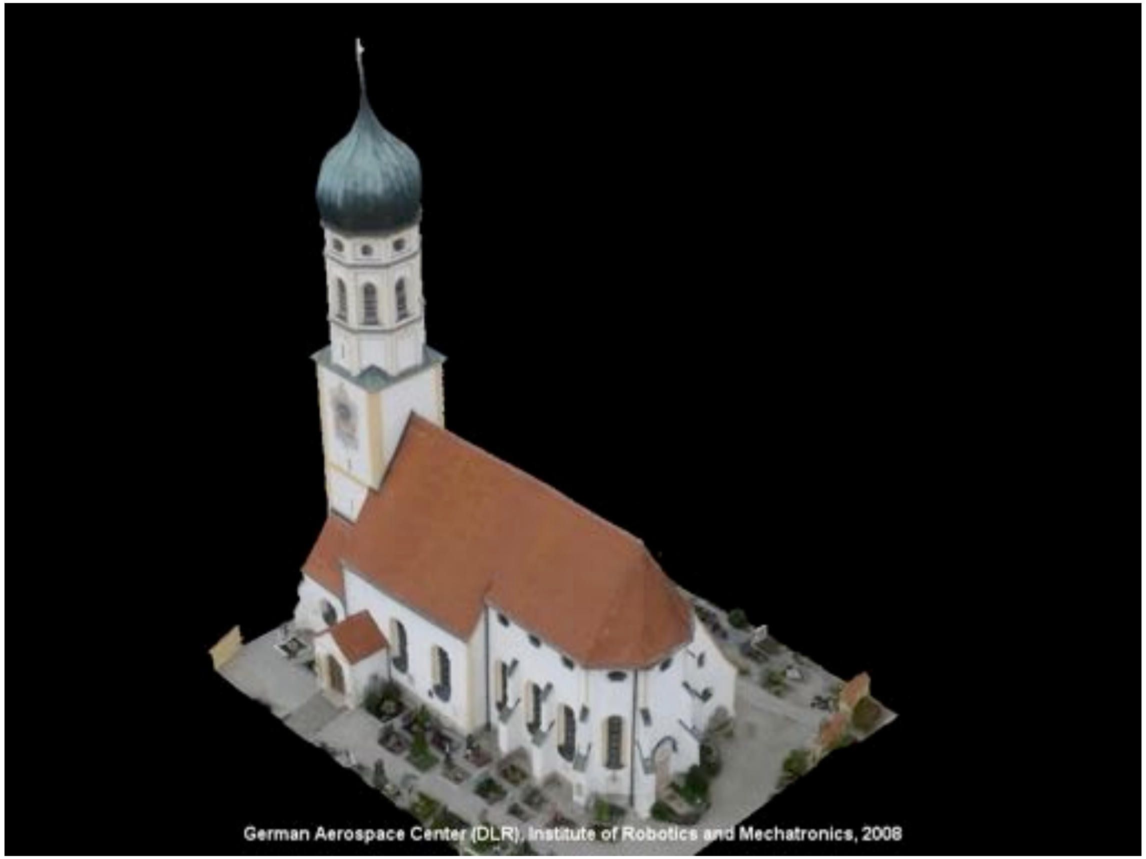
Visual Static Modelling with a Drone (2007)

Mount for an 8MP
Digital Camera

Real-Time Navigation Data from an Image Sequence

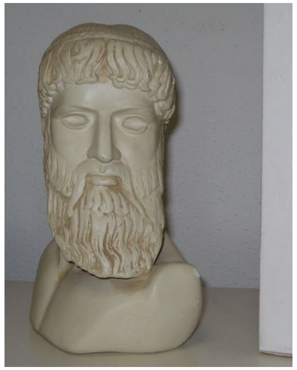
Estimation of the 6 Degrees of Freedom

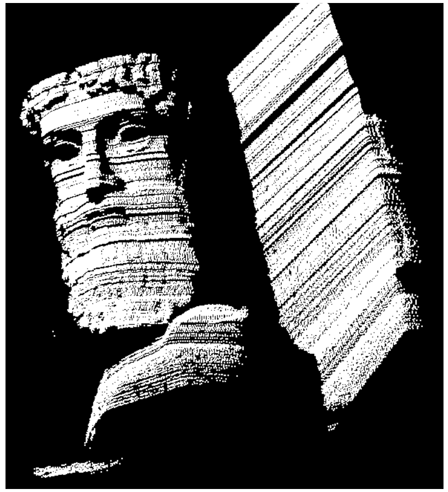
We used to reconstruct static scenes from monocular in 2007… (with DLR)

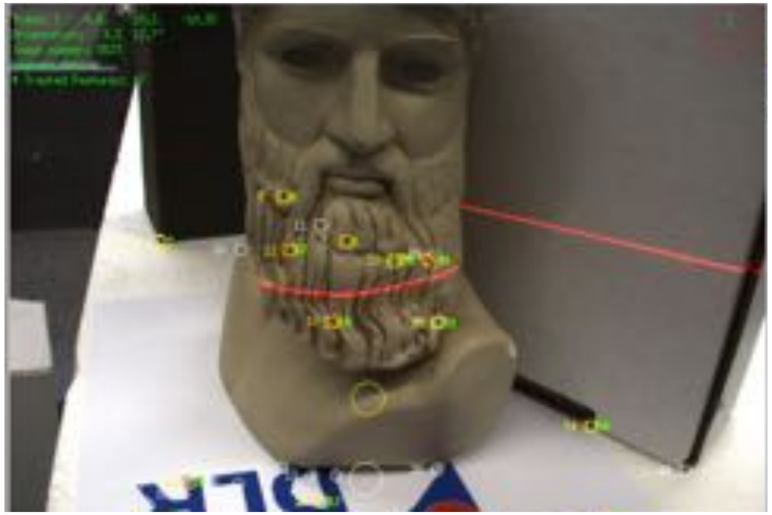


Accuracy:1.5cm

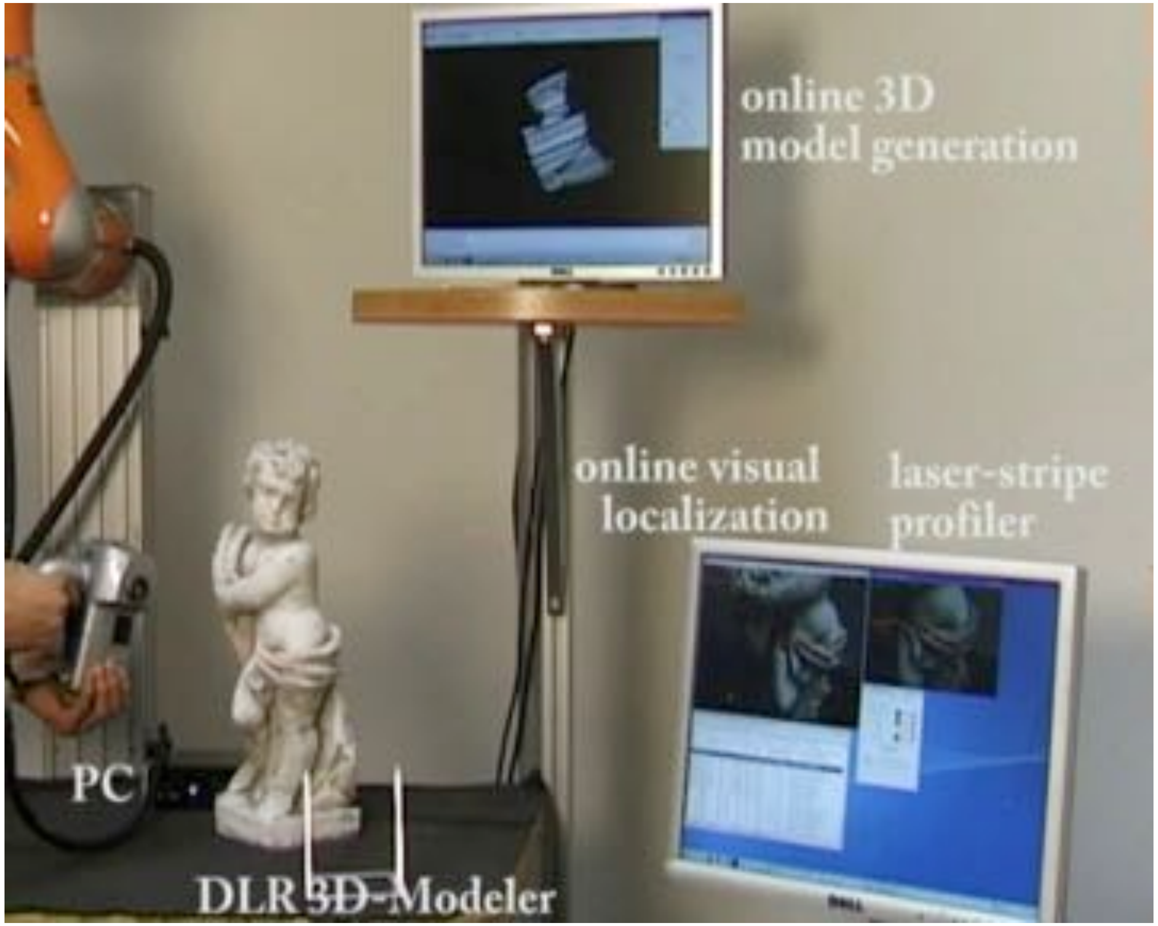
High Accuracy at Example of Light Section 3D Reconstruction







Accuracy of the system - Construction of 3D models (2008)



Camera localization accuracy allows direct stiching of the line responses from the light-section system

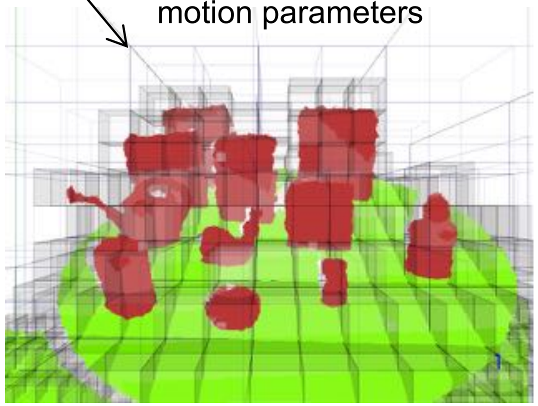
120fps Monocular Navigation from Sparse Optical Flow

GPU implementation of sparse flow (feature-based OpenCV) system
using only 10% of the resources

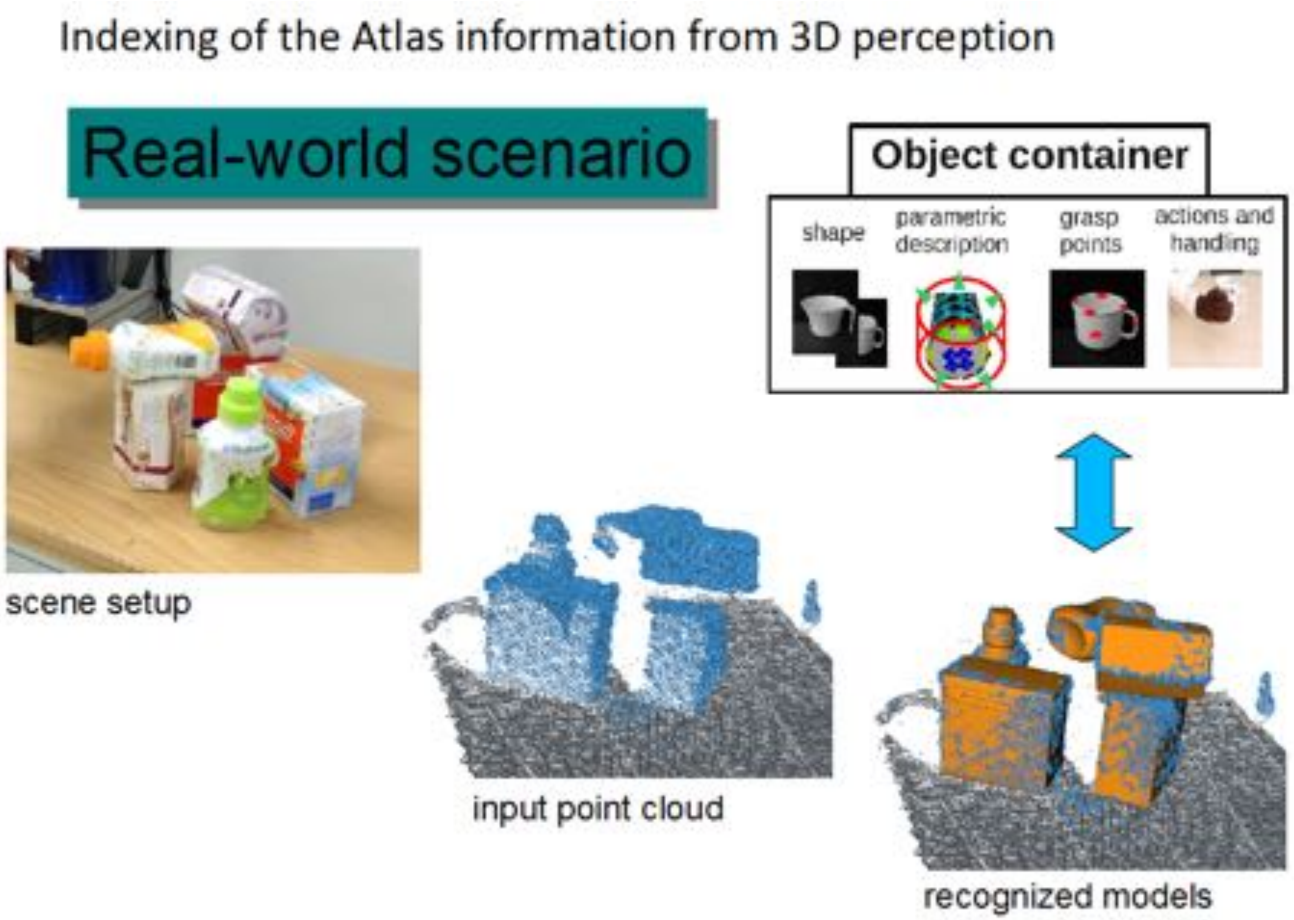
What can we do with the 3D PointClouds?

labeling

motion parameters



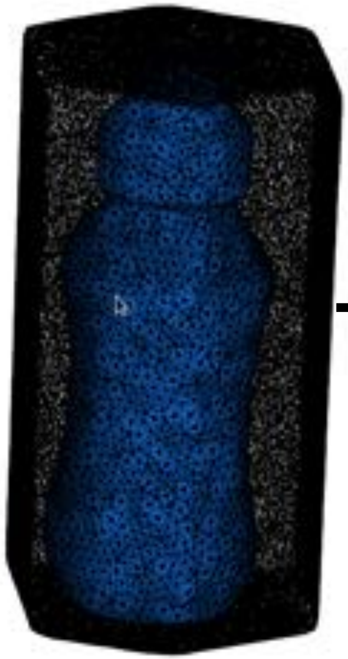
What is in the scene? (labeling)



ObjectRANSAC system fitting 3D models into cluttered scenes (Papazov et al. 2010)

Deformable Registration from Generic Models

(special issue SGP'11 Papazov et al.)



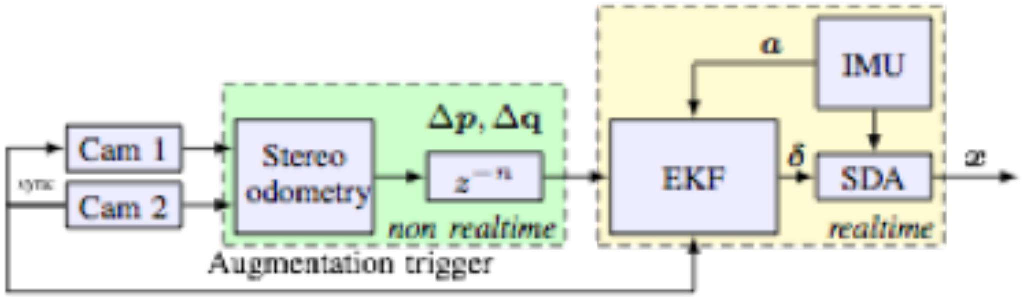
Matching of a detailed shape to a primitive prior

The manipulation “heat map” from the generic model gets propagated

Deformation of the original model generates a deformation heat-map showing the similarities of object regions to the model.

Navigation for Control
VINS filter design

[Schmid et al. IROS 2012]



- Synchronization of real-time and non realtime modules by sensor hardware trigger

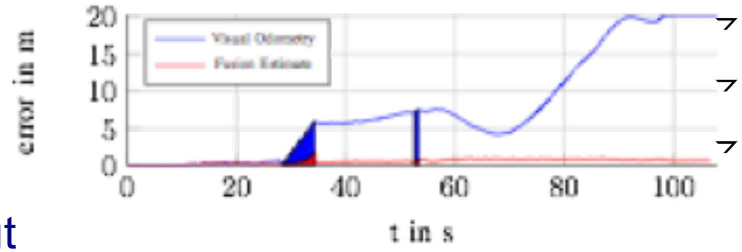
- Direct system state: $\boldsymbol{x} = \left(\boldsymbol{p}_{ob}^{o,T} \quad \boldsymbol{v}_{ob}^{n,T} \quad \boldsymbol{q}_{b}^{o,T} \quad \boldsymbol{b}_{a}^{b,T} \quad \boldsymbol{b}_{\omega}^{b,T}\right)^T$
- High rate calculation by „Strap Down Algorithm“ (SDA)

- Indirect system state: $\boldsymbol{\delta} = \left(\boldsymbol{\delta}_{p}^{o,T} \quad \boldsymbol{\delta}_{v}^{o,T} \quad \boldsymbol{\delta}_{\omega}^{o,T} \quad \boldsymbol{\delta}_{b_a}^{b,T} \quad \boldsymbol{\delta}_{b_\omega}^{b,T}\right)^T$
- Estimation by indirect Extended Kalman Filter (EKF)

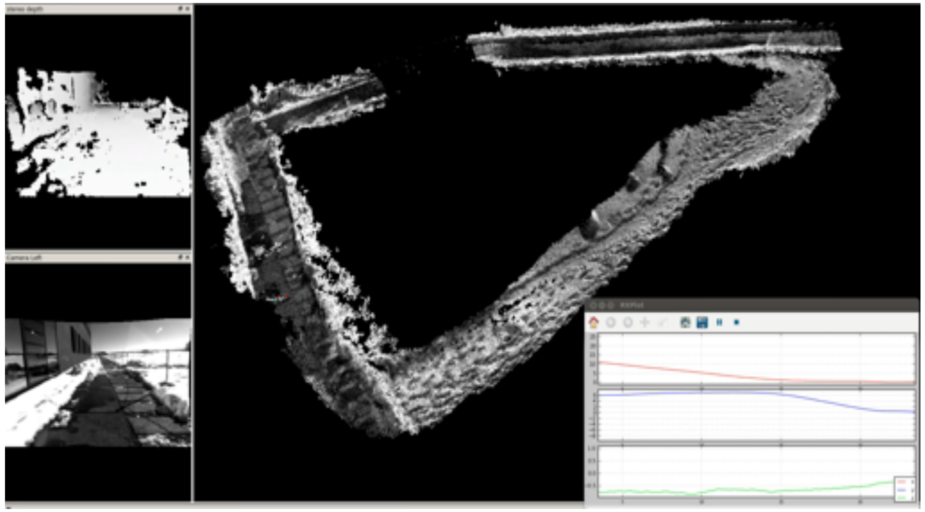
VINS-Systems

Fusion of heterogeneous data with varying latencies (with DLR)

- 70 m trajectory
- Ground truth by tachymeter
- 5 s forced vision drop out with translational motion
- 1 s forced vision drop out with rotational motion

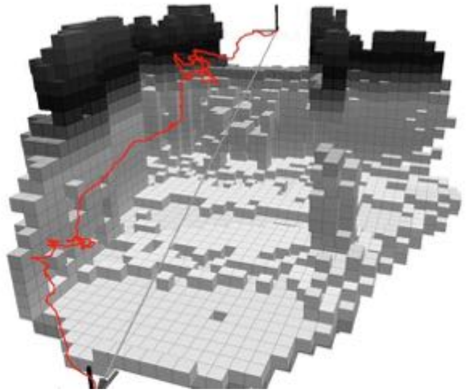


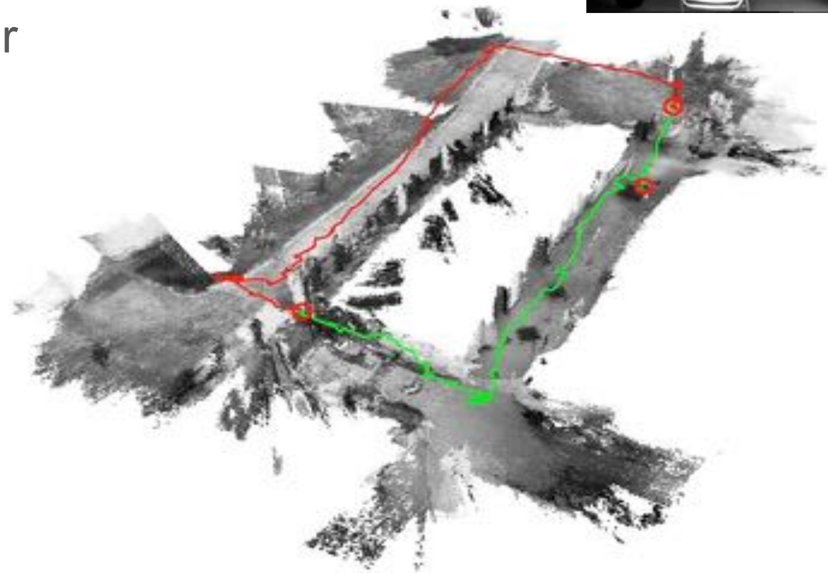
- Estimation error < 1.2 m
- Odometry error < 25.9 m
- Results comparable to runs without vision drop outs



Navigation under strong illumination changes

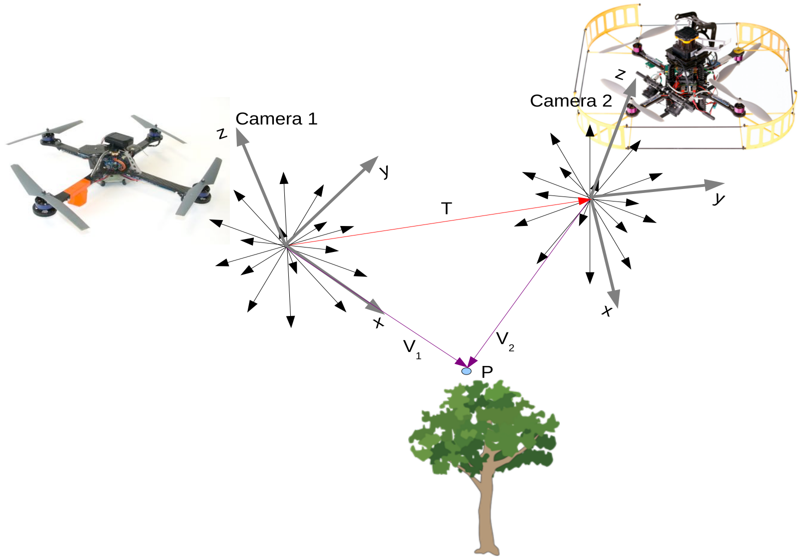
- Autonomous indoor/outdoor flight of 60m
- Mapping resolution: 0.1m
- Leaving through a window
- Returning through door

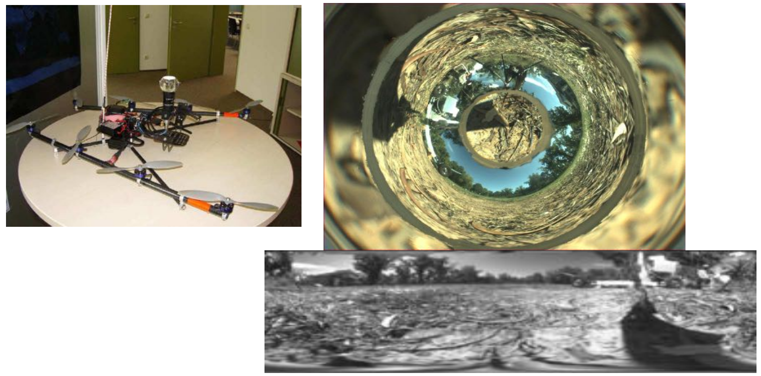




Collaborative Reconstruction with Self-Localization (CVP2008)

Vision in Action: Efficient straegies for
cognitive agents in complex envirnments)





$$V_2 = \mathbf{R} * (V_1 + T)$$

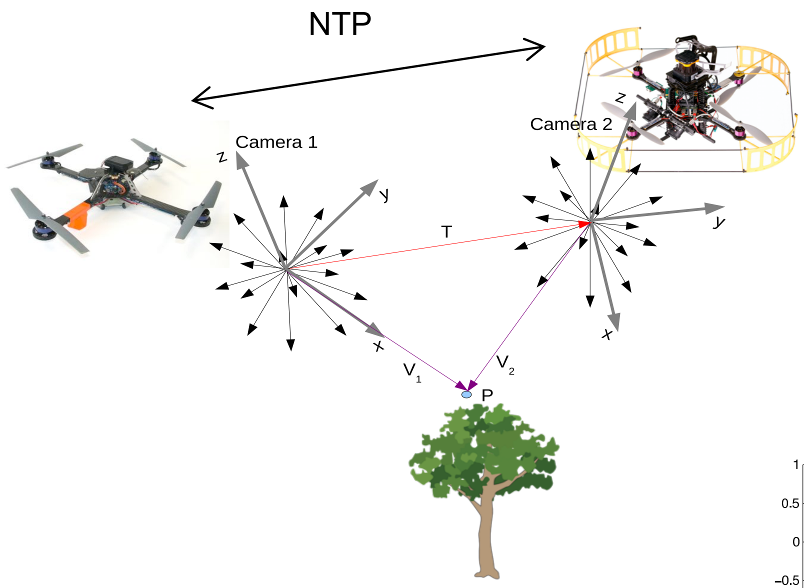
$$\lambda_2 n_2 = \mathbf{R} * (\lambda_1 n_1 + T).$$

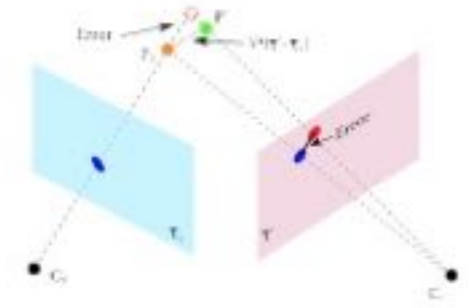
$$(-\mathbf{R}n_1,\, n_2)\begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix} = \mathbf{R}\cdot T$$

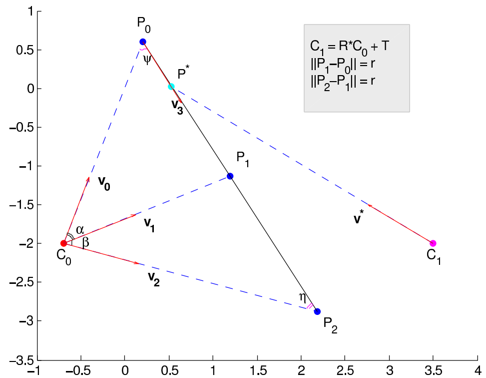
$$\begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix} = (-\mathbf{R}n_1,\, n_2)^{-*}\cdot\mathbf{R}\cdot T = \mathbf{D}^{-*}\cdot\mathbf{R}\cdot T$$

$$\mathbf{D}^{-*} = (\mathbf{D}^{\mathbf{T}}\cdot\mathbf{D})^{-1}\cdot\mathbf{D}^{\mathbf{T}}.$$

Asynchronous Stereo for Dynamic Scenes



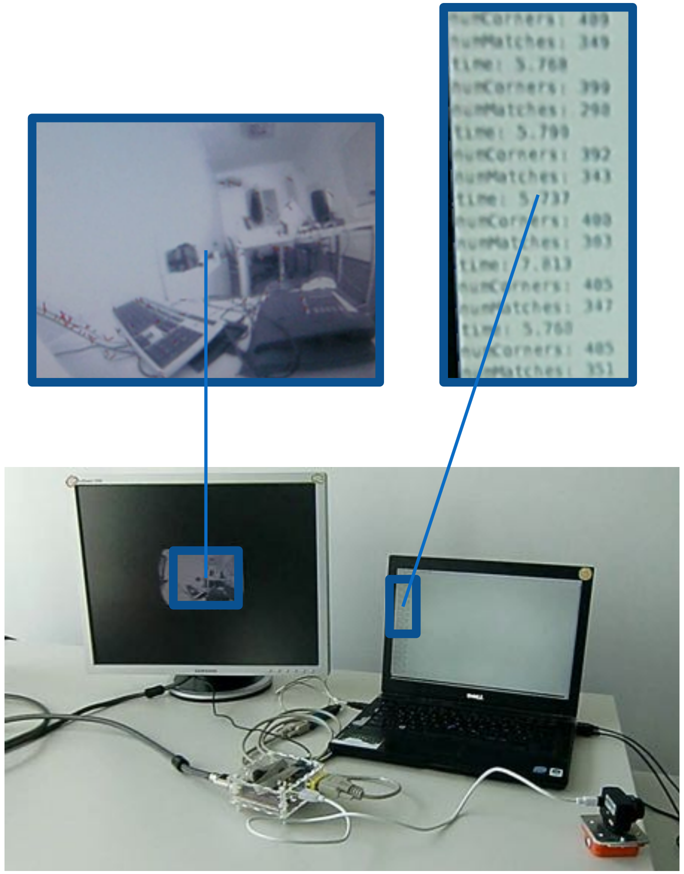




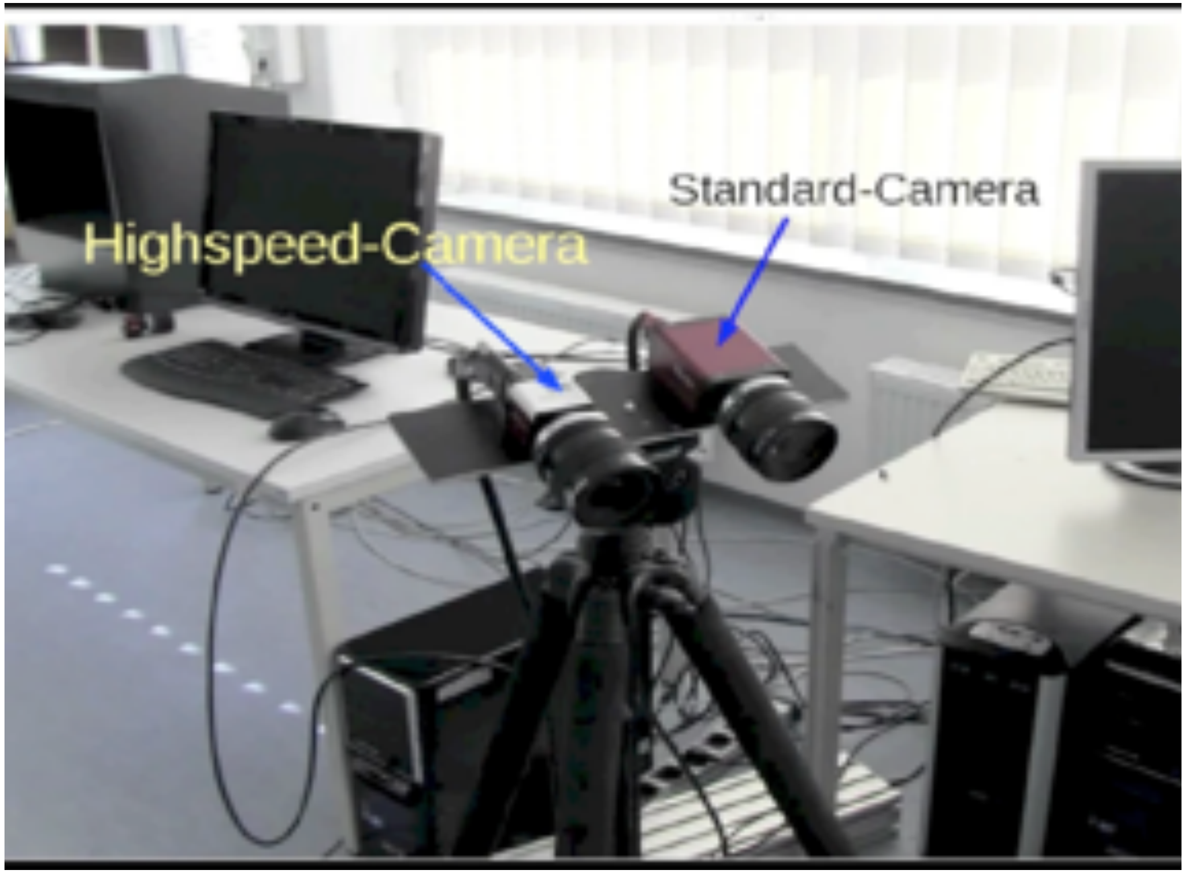
Mkhitaryan, Burschka VISAPP 2014

Back to Autonomous Vehicle Applications - Processing Units: Local Feature Tracking Algorithms (AGAST, fastest keypoint detector part of OpenCV developed by us)

- Image-gradient based → Extended KLT (ExtKLT)
  - patch-based implementation
  - feature propagation
  - corner-binding
  - \+ sub-pixel accuracy
  - algorithm scales bad with number of features
- Tracking-By-Matching → AGAST tracker
  - AGAST corner detector
  - efficient descriptor
  - high frame-rates (hundrets of features in a few milliseconds)
  - \+ algorithm scales well with number of features
  - pixel-accuracy



Hybrid High-Speed Stereo System

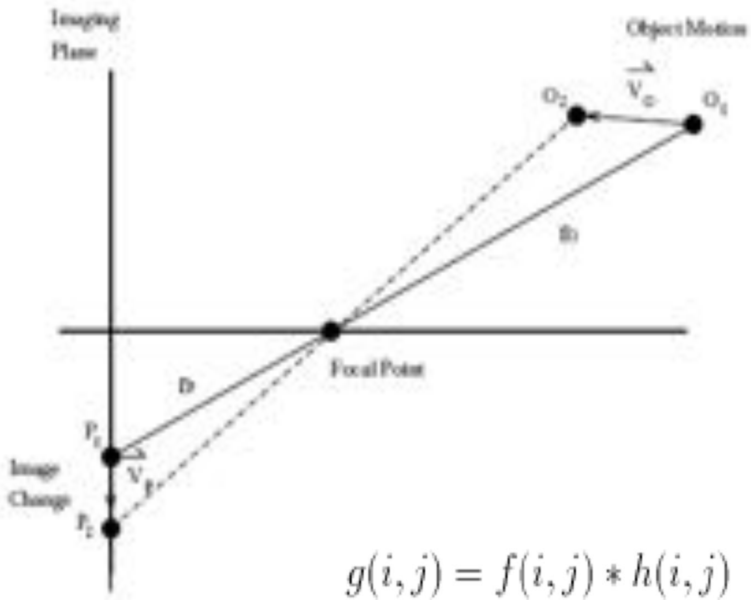


Previous approach – Navigation from Optical Flow between images

Can motion be calculated directly a single image?

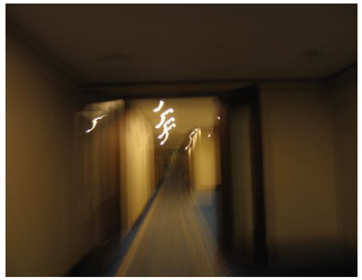
→

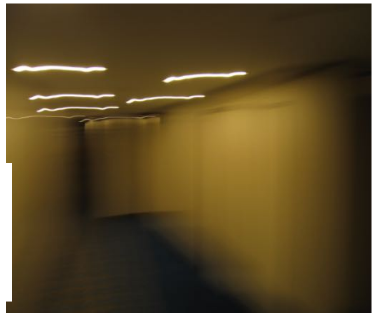
What is the underlying principle?
Point Spread Function (PSF)



$$g(i,j) = f(i,j) * h(i,j)$$

Horizontal motion $h(x,y) = \begin{cases} \frac{1}{d}, & 0 \le |x| \le d \cdot \cos\alpha \wedge y = \sin\alpha \cdot x \\ 0, & \text{otherwise} \end{cases}$





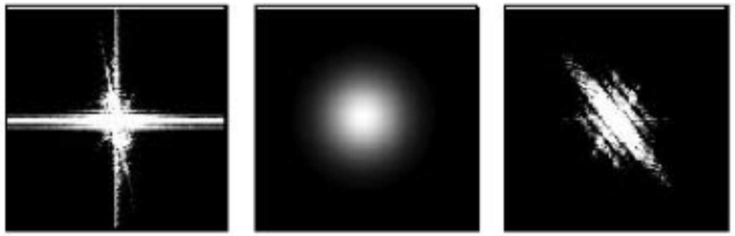
Motion Blurr

$$G(u,v) = I(u,v) \cdot H(u,v),$$

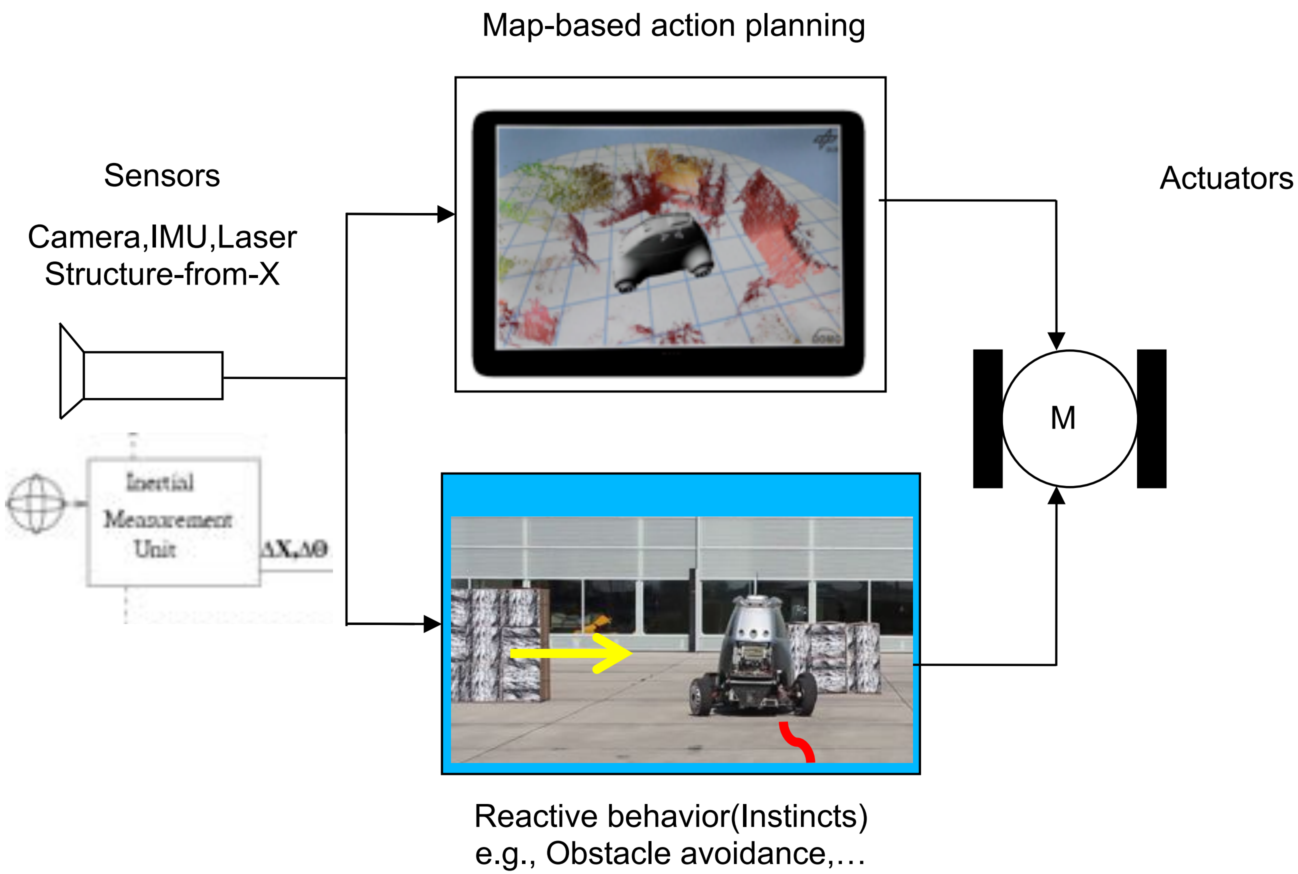
$$h(x,y) = \begin{cases} \frac{1}{d}, & 0 \leq |x| \leq d \cdot \cos\alpha \wedge y = \sin\alpha \cdot x \\ 0, & \text{otherwise} \end{cases}$$

$$H(\omega,\nu) = \frac{\sin \pi d \omega}{\pi d \omega} = \mathrm{sinc}(\pi d \omega)$$

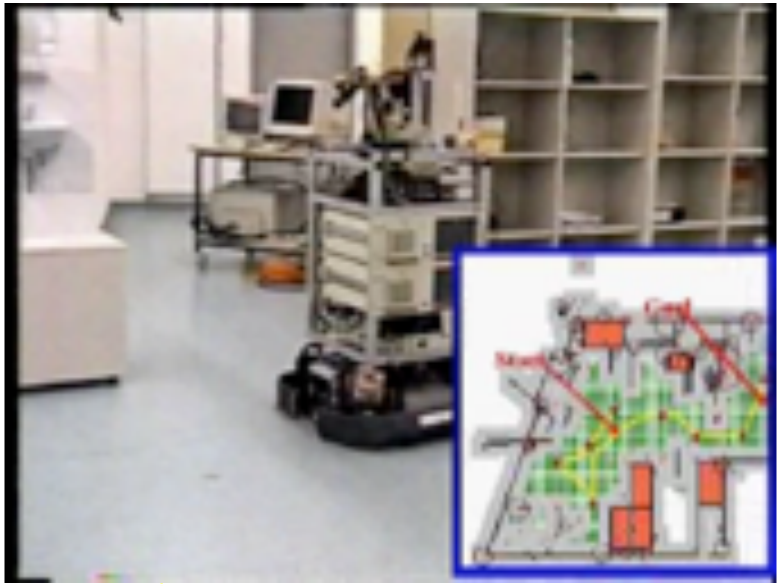
Gaussian window to avoid artifacts in Cepstrum



Coupling Alternatives for Perception Modules



Navigation Strategies (metric vs. non-metric)



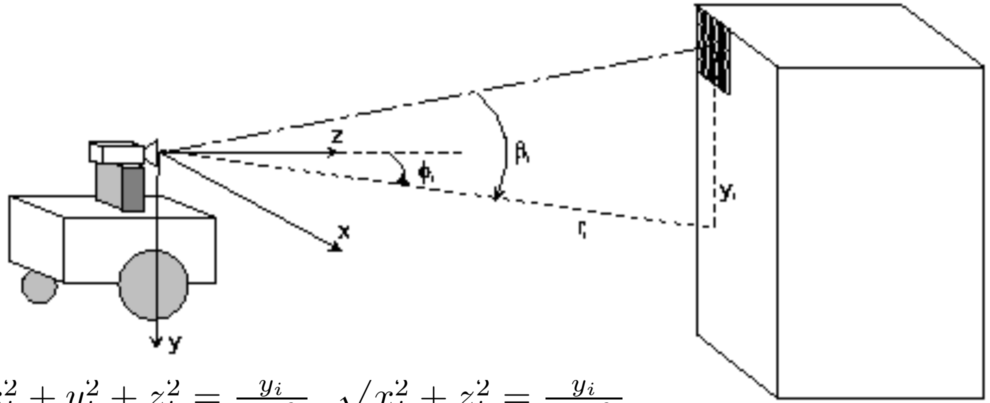
Map-based Navigation

the reconstructed data is stored in 3D maps to be used for obstacle avoidance and mission planning.

Vision-Based Control

the control signals are generated directly from the sensor perception

Optimal Feature Selection



$$x_i^2 + y_i^2 + z_i^2 = \frac{y_i}{\sin\beta_i}, \quad \sqrt{x_i^2 + z_i^2} = \frac{y_i}{\tan\beta_i},$$

$$= \frac{y_i \cdot \sin\alpha_i}{\tan\beta_i}, \qquad z_i = \frac{y_i \cdot \cos\alpha_i}{\tan\beta_i}$$

$$= \begin{pmatrix} \frac{\tan\beta_i \cdot \cos\alpha_i}{y_i} & -\frac{\tan\beta_i \cdot \sin\alpha_i}{y_i} & -1 \\ -\frac{\sin^2\beta_i \cdot \sin\alpha_i}{y_i} & -\frac{\sin^2\beta_i \cdot \cos\alpha_i}{y_i} & 0 \end{pmatrix}$$

$$\mathcal{J}_i^t = \begin{pmatrix} 0 & -1 \\ -\frac{\sin^2\beta_i}{y_i} & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ -\frac{1}{y_i \cdot \left(1 + \left(\frac{r_i}{y_i}\right)^2\right)} & 0 \end{pmatrix}$$

$$r_i = \sqrt{x_i^2 + z_i^2}$$

Consider the equation system with perturbations in matrix J and vector b:

$$(\mathcal{J} + \epsilon\delta\mathcal{J})x_b = b + \epsilon\delta b \qquad (7)$$

The relative error in the solution caused by perturbations of parameters can be estimated by the following inequality using the condition number κ calculated for J (see [5]):

$$\frac{||x - x_b||}{||x||} \leq \kappa\left(\epsilon\frac{||\delta\mathcal{J}||}{||\mathcal{J}||} + \epsilon\frac{||\delta b||}{||b||}\right) + \mathcal{O}(\epsilon^2) \qquad (8)$$

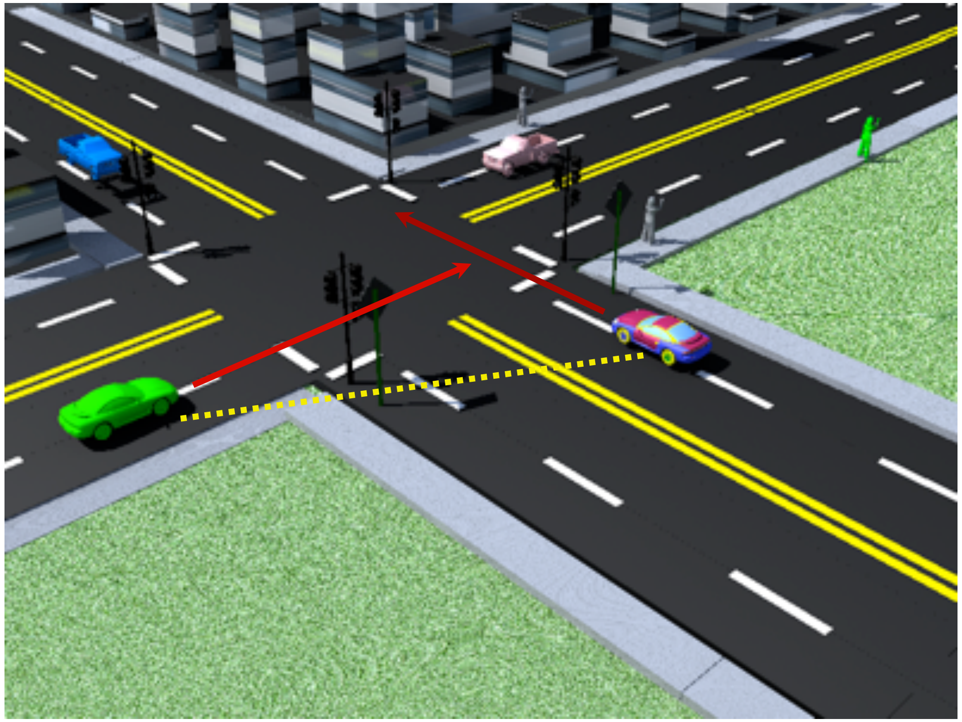
$$\kappa_r = \left[y \cdot \left(1 + \left(\frac{r_i}{y_i}\right)^2\right)\right]^{-1}$$

$$\frac{d\kappa_r}{dy_i} = -\frac{1}{x_i{}^2\left(1 + \frac{r_i{}^2}{y_i{}^2}\right)} + \frac{2 \cdot r_i{}^2}{y_i{}^4\left(1 + \frac{r_i{}^2}{y_i{}^2}\right)^2} = 0$$

$$\Rightarrow \quad y_i = \pm r_i \quad \Rightarrow \beta_i = \arctan\frac{y_i}{r_i} \qquad (12)$$

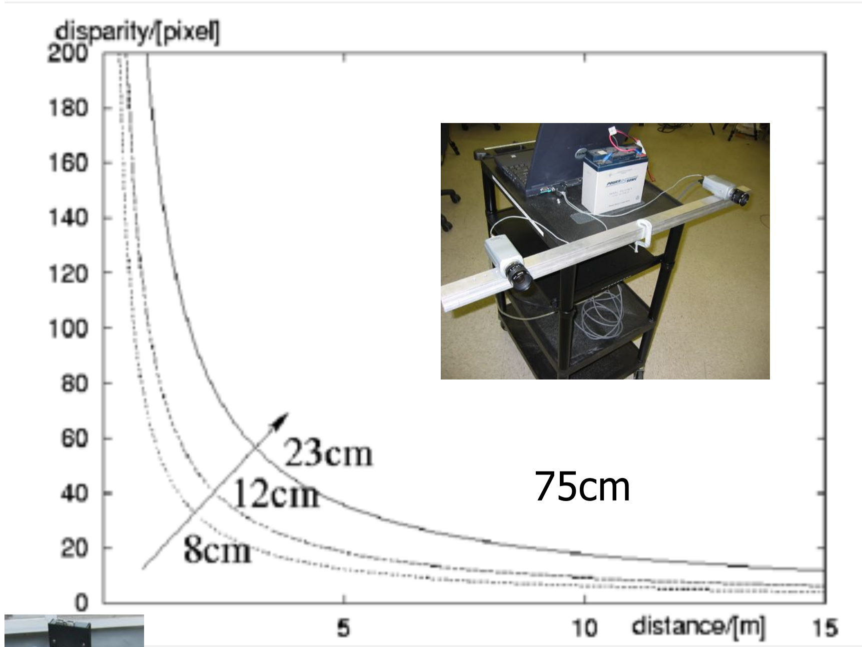
that corresponds to an angle $|\beta_i| = 45°$.

Capturing Motion Properties of Large Dynamic Scenes



Cars in distances over 50m are only a few pixels large

Are lab approaches transferrable to automobile and avionic applications?

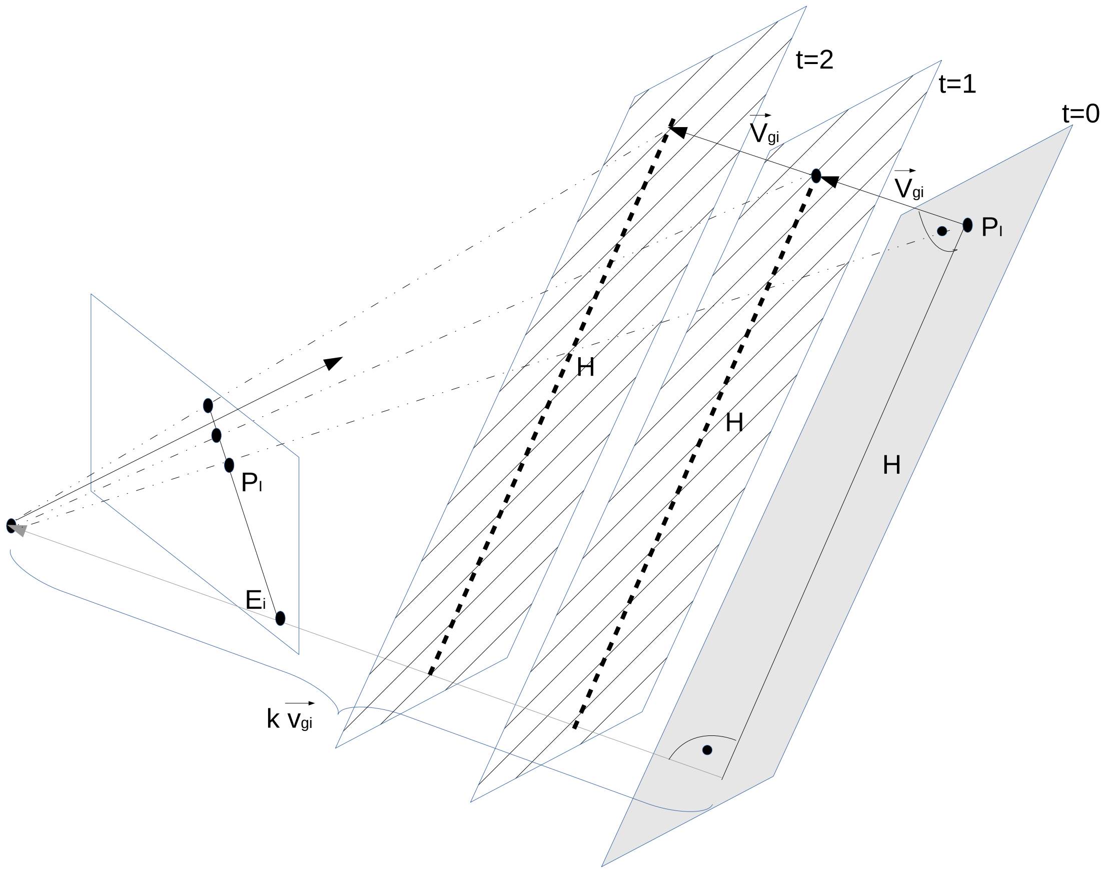


$$d_p = \boxed{\frac{B \cdot f}{p_x}} \cdot \frac{1}{z} [pixel]$$

Sensitivity increase:

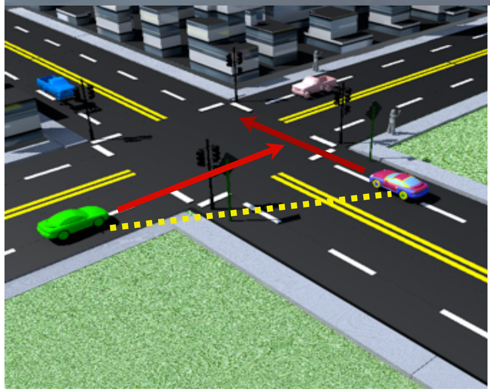
- Larger baseline (B)
- Longer focal length (f) □ field of view
- Smaller pixelsize (p_x) □ "pixel explosion"

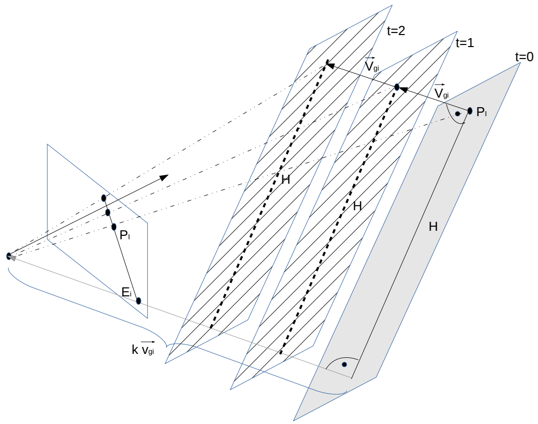
Direct Mapping of Point Motion on Image Observation



Detection of Independent Motion Groups from Optical Flow

- Our goal is a robust detection of **motion direction** and **collision times** from a **monocular uncalibrated** camera sequence.
- Representation of the **dynamic scene ordered by collision times** instead of Cartesian coordinates aenables monocular processing (**no scale necessary**) and better priorisation of collision candidates than in conventional methods
- Independent estimation of motion direction and collision time allows collision categorization in large distances from the camera





Schaub et al., Journal ITSC
Burschka, BMVC 2017

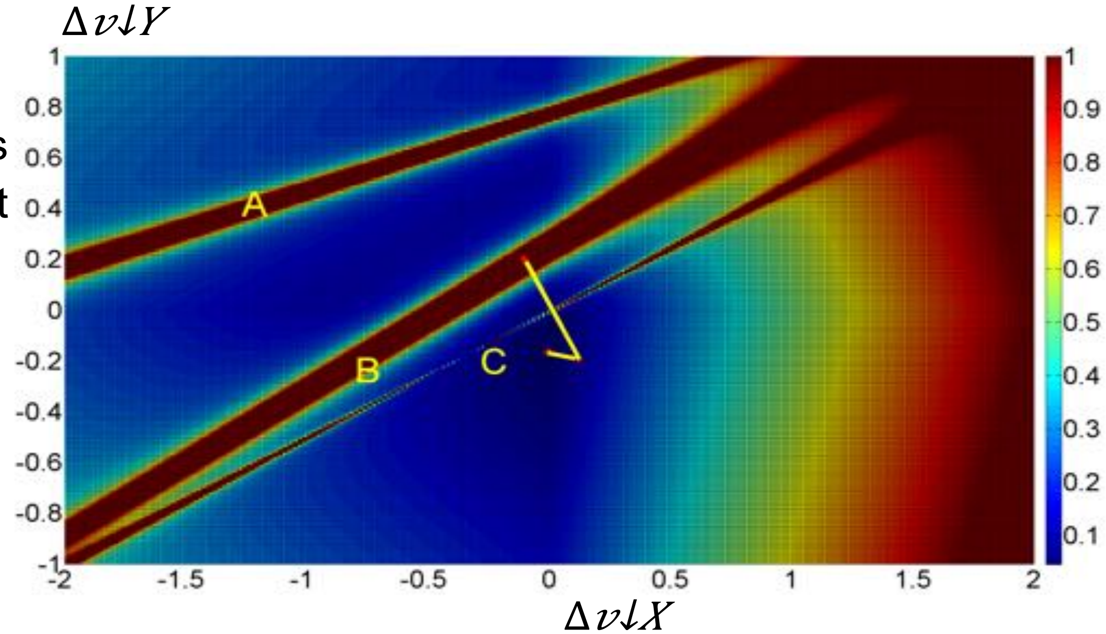
Obstacle Avoidance in Dynamic Spaces

Novel Control Design for Non-metric Control from Monocular

(Schaub&Burschka)

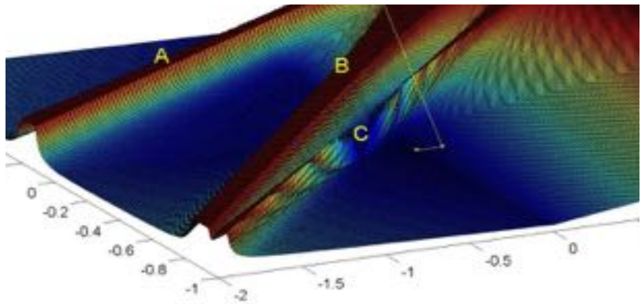
New Controller is necessary for the non-metric input:

- Planning space represented as collision times for different velocities
- Non-Linear Gradient-Descent with an Adaptive Lagrange Interpolation Search (ALIS)
- Weights: $J_d > J_{a_X} > J_{a_Y}$



- Good performance: 2 Steps to reach the optimum

$$J = 0.0349 \qquad J_{fmincon} = 0.3792$$

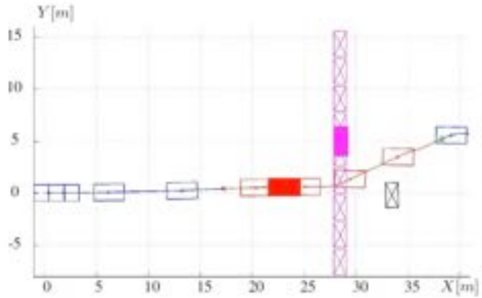


- Realtime implementation with 25 Hz

Navigation based on Pixel-Information from Monocular View

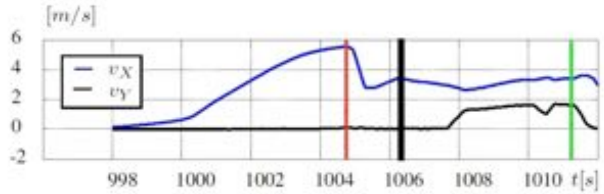
Concept: Shifting the optical flow epipole out of the object's boundaries
→ no collision

Schaub, Burschka ITSCC 2015



Planar motion of the objects $(\mathrm{P} \in \mathbb{R}^3\{X, Y, Z = c\})$

$$E_x = c_x + f s_x \frac{v_Y}{v_X}$$
$$E_y = c_y$$



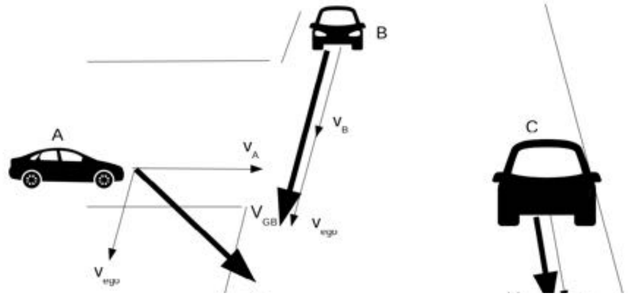
Effect of the relative velocity $v = \{v_X, v_Y\}$ on the Epipole's 2D image position (x, y)
→ find: $\Delta E_x = \mathrm{f}(\Delta v_X, \Delta v_Y)$:

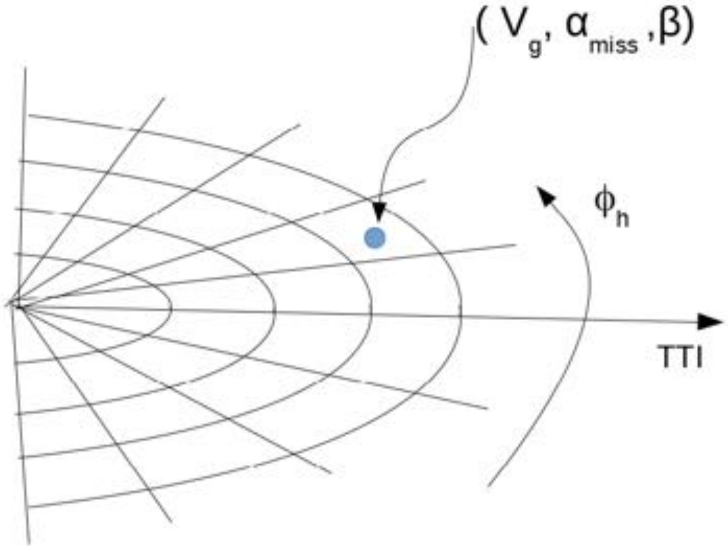
$$\Delta E_x = E_x(v + \Delta v) - E_x(v) = f_x \frac{v_Y \Delta v_X - v_X \Delta v_Y}{(v_X + \Delta v_X) v_X}$$

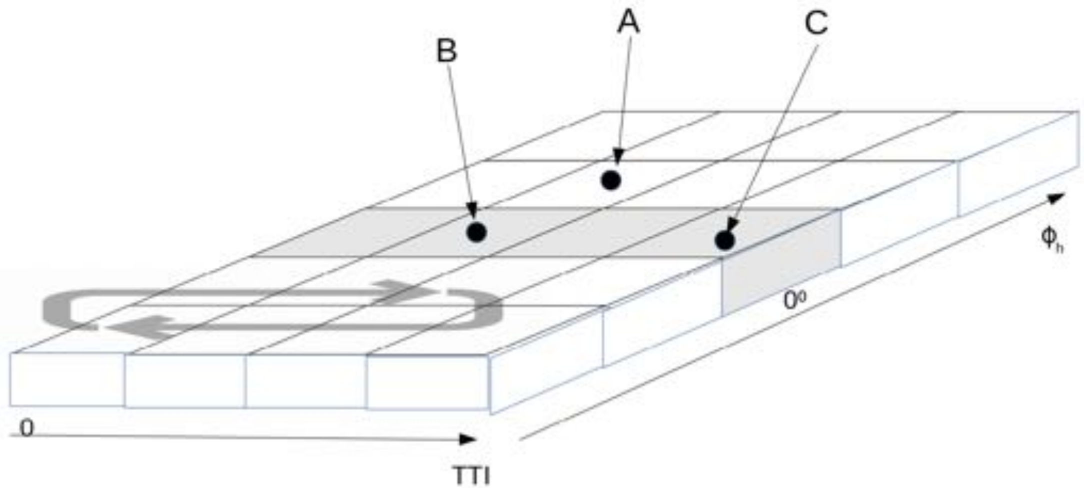
Application in Autonomous Driving (Schaub&Burschka(

Novel Non-Cartesian Map Structure (IROS 2019)

Map is indexed by the azimuth angle of the relative velocity vector and the time until the plane with the velocity vector as a normal passes the focal point TTC







This makes the map content static in dynamic environments. Merely a TTC counter scrolls the grid during the operation but no changes of the grid information is necessary!

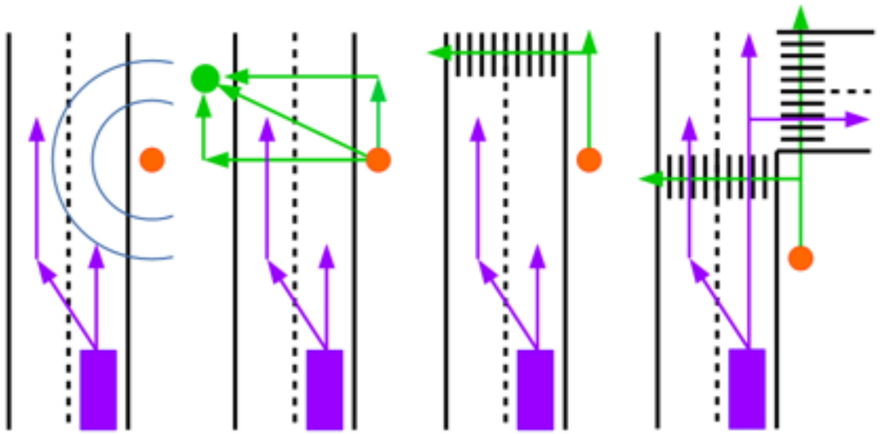
Identification of Interaction Regions in the Scene

Extraction of map-based POI
static

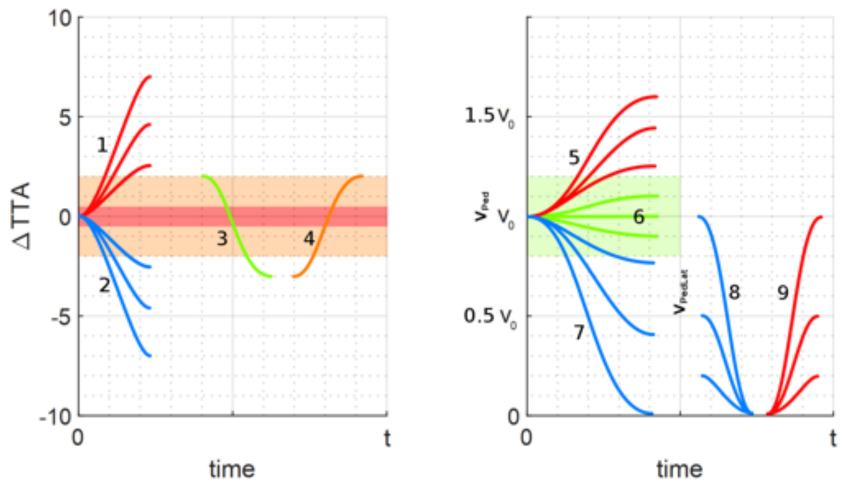
Identification of overlapping ”resource” allocation (competition for POI) dynamic

Estimation of Intention of the Traffic Agents

Changes in the temporal interaction with agents can be used for behavior analysis (IV 2019)



Static and dynamic POI allow a better prediction of intentions



Temporal evolution of TTC at the resource allows to assess passivity of aggressivity of the traffic partner

Conclusions

- Non-metric navigation allows operations directly in camera images without the necessity of metric scale
- Temporal representation helps to assess and predict behaviors
- Learning approaches are based on similarity search and, therefore, built for segmentation and labeling – not for metric measurements
- Scene understanding from single images reconstruct only spatial but no temporal relations
- Early data abstraction loses often important information from the sensor
- Driving can be modelled as resource competition on the road