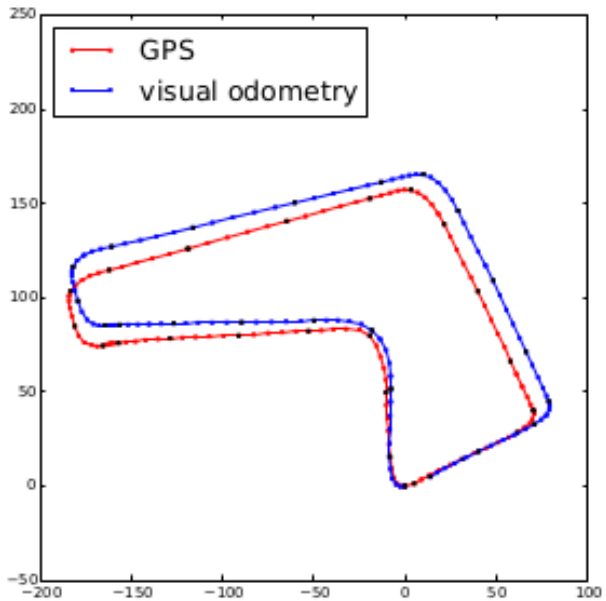
Stereoscopic structure and motion estimation

Siniša Šegvić
(experiments performed by Ivan Krešo)



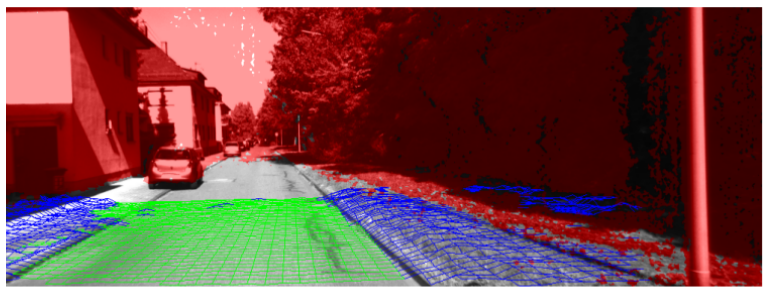


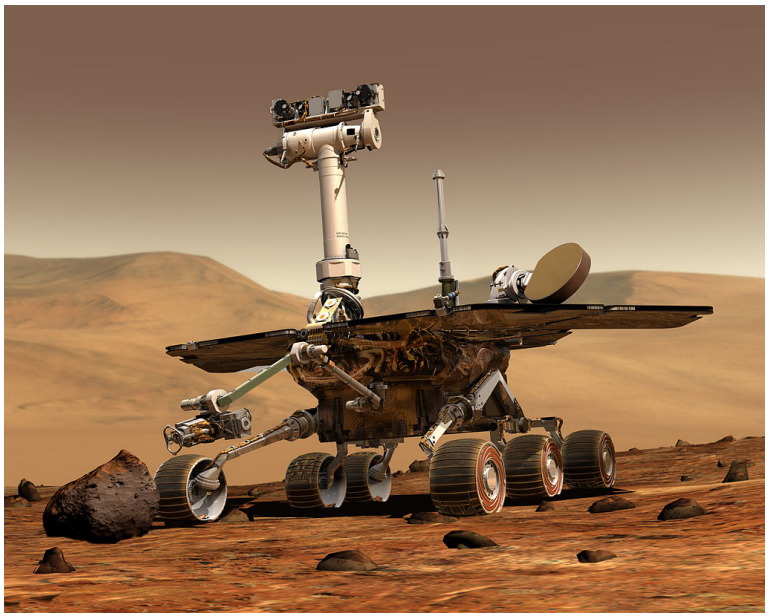
Table of contents

Goal

Structure and motion estimation:

- **recover** the camera motion
- **reconstruct** the scene geometry
- **understand** and **navigate** real environments

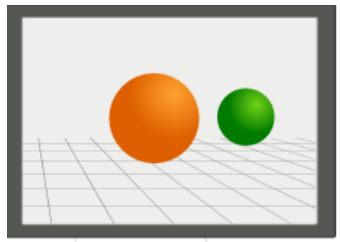
[jurvetson2012]



[NASA/JPL-Caltech]

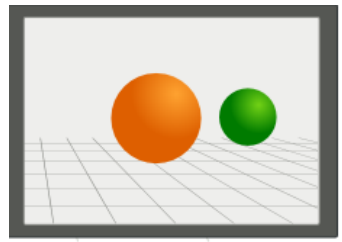
Two cameras are better than one!

Which of the following two spheres is closer to the camera?

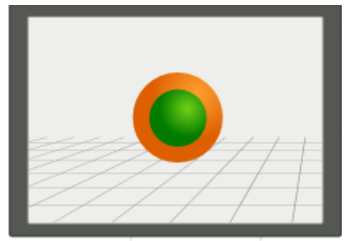


Two cameras are better than one!

Which of the following two spheres is closer to the camera?

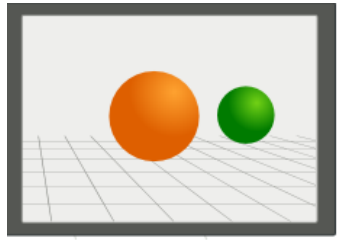


Consider now a different view onto the same scene:

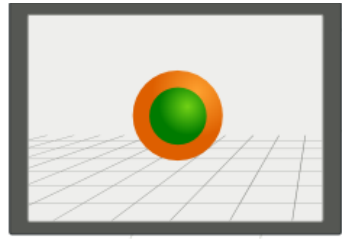


Two cameras are better than one!

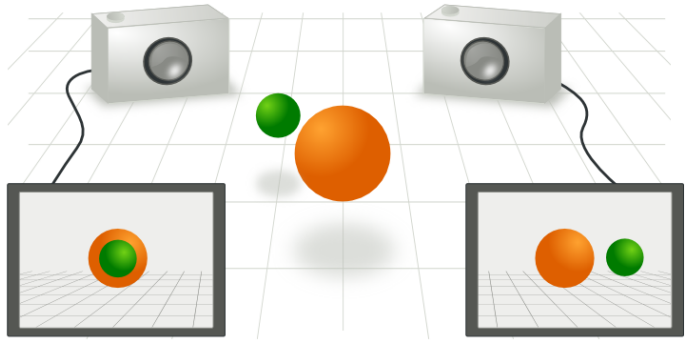
Which of the following two spheres is closer to the camera?



Consider now a different view onto the same scene:



Stereo lets us reason about depth of the image objects!



[nordmann2012]

About images

From the ancient times images have been created as 2D windows to inexistent 3D worlds

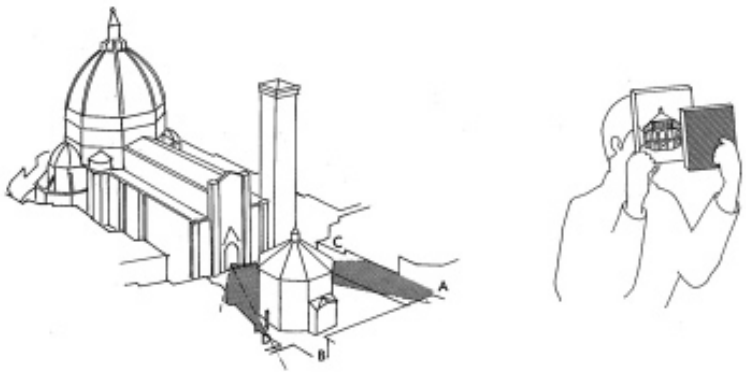
Artists have tried hard to produce images that would appear as real as possible

However, exact laws of image formation have not been widely known before 14th century

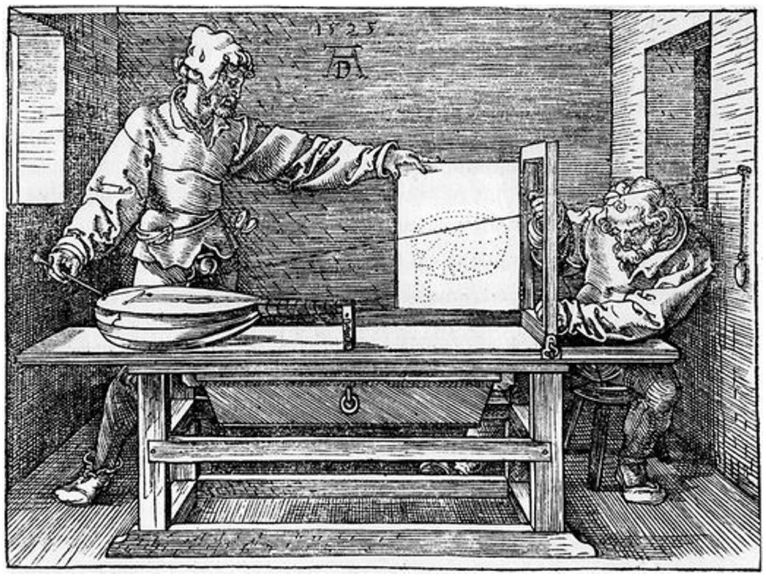
Explaining the perspective

Bruneleschi's device ("peep show") allowed to debug the painting by comparing it with a reflection of the real subject

Duerer's woodcut shows a painful but working approach to produce an exact perspective sketch



[Brockelman08]



[Duerer]

Perspective for artisans

Laws of perspective entered painting handbooks by 17th century

Devices have been built in order to make it easy to guess the perspective even for amateurs and less gifted professionals

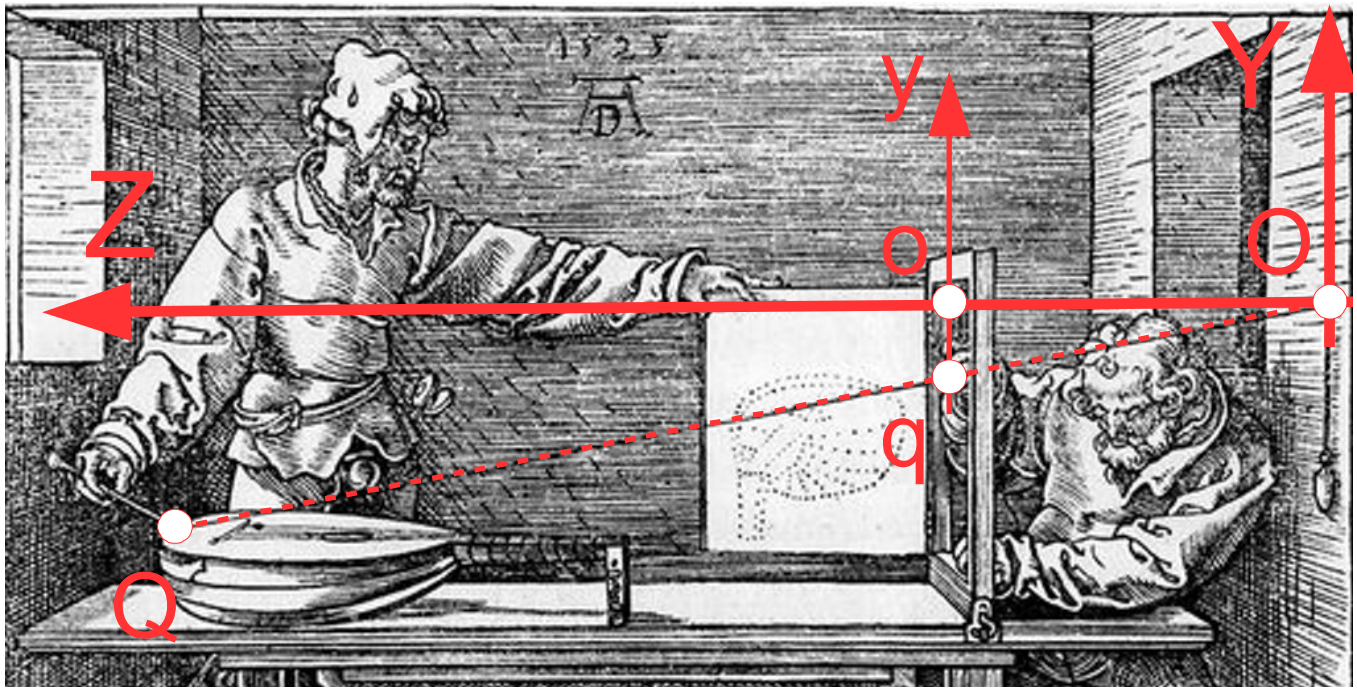
[Bosse]

[Duerer]

Hence the painters declared the perspective as solved and turned their attention to impressionism, expressionism etc

Laws of image formation

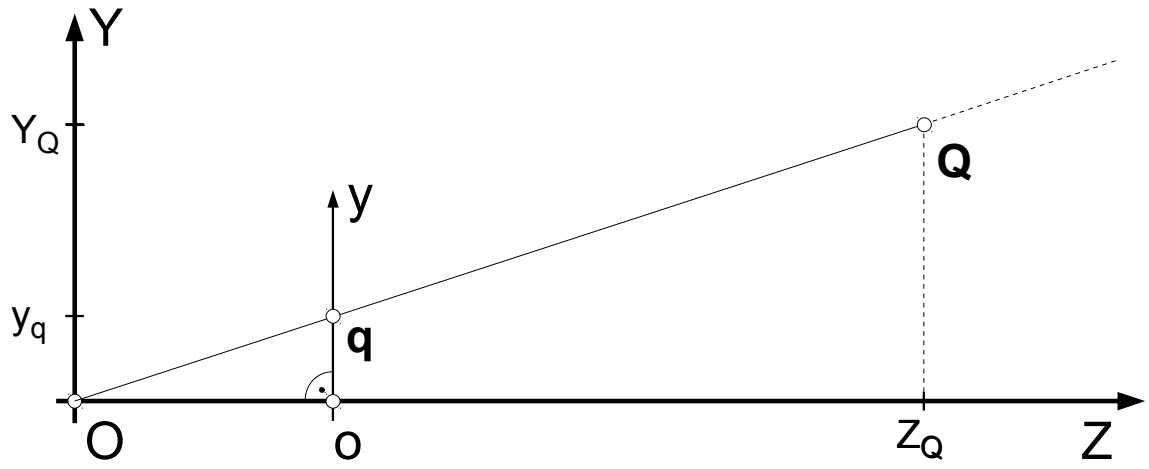
In the Duerer's woodcut, the perspective projects 3D point $Q(X_Q, Y_Q, Z_Q)$ onto the image point $q(x_q, y_q)$



In order to derive equations for x_q and y_q we assume:

- world coordinates are centered in the principal point $\mathbf{O}$
- the Z axis of the world is perpendicular to the image plane
- image coordinates are centered in $\mathbf{o}$, vec$(\mathbf{O},\mathbf{o}) \parallel Z$
- the Y axis of the world is aligned with the image coordinate y

For simplicity, we only look at the Y coordinate:



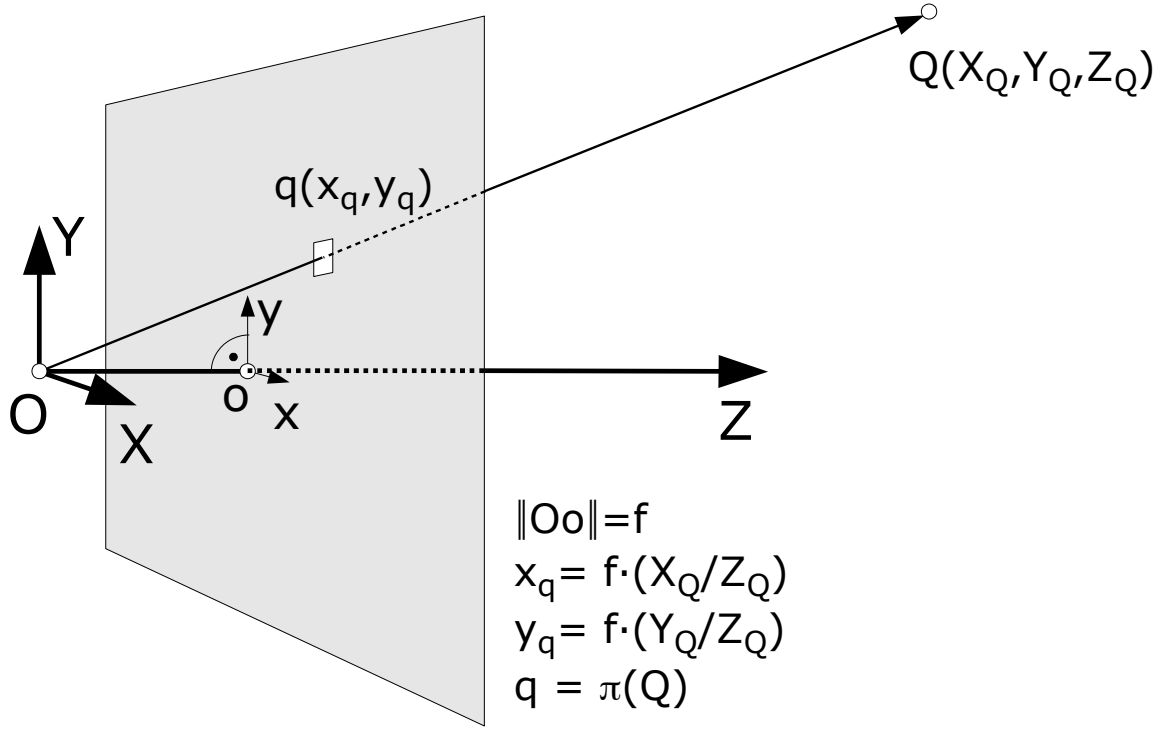
From the triangle similarity we easily derive:

$$y_q = |Oo| \cdot \frac{Y_Q}{Z_Q} = f \cdot \frac{Y_Q}{Z_Q}$$

In an analogous manner we would obtain the equation for x_q:

$$x_q = f \cdot \frac{X_Q}{Z_Q}$$

We are now ready to describe image formation in an ideal camera:



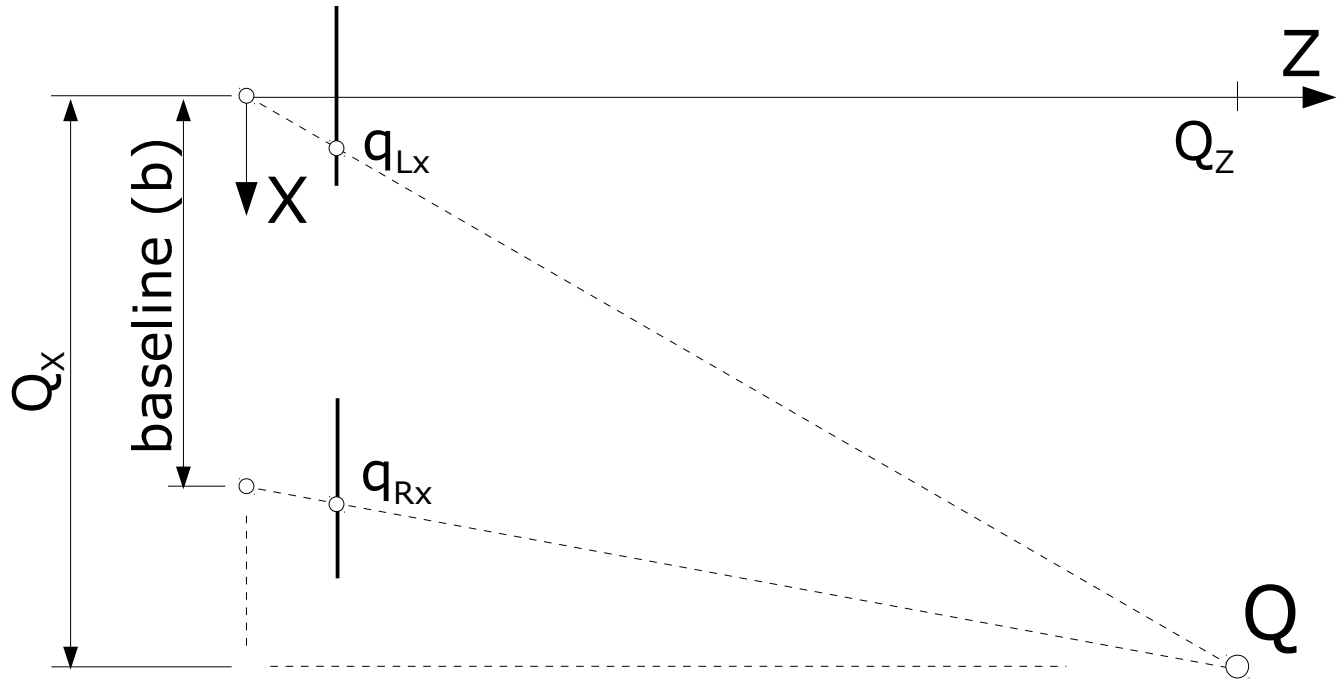
This model is able to describe what happens in real cameras too:

- no surprise: manufacturers try hard to mimic exact perspective
- f is a camera parameter which needs to be calibrated (there are other parameters, but we won't go into details)
- if we wish to express x_q in pixels, f equals the lens focal length divided by the pixel size of the imaging sensor

Geometry of a rectified stereo rig

We have seen that from single images we can only recover 3D directions.

With a rectified camera pair, the depth can be elegently recovered:



The projections of the 3D point Q onto two images are:

$$q_{Lx} = f \cdot \frac{Q_x}{Q_Z} \quad q_{Rx} = f \cdot \frac{Q_x - b}{Q_Z}$$

Recovering depth with stereo

By subtracting the image coordinates we obtain the disparity d:

$$d = q_{Lx} - q_{Rx} = f \cdot \frac{b}{Q_Z}$$

The disparity uniquely determines the depth Q_Z:

$$Q_Z(d) = f \cdot \frac{b}{d}$$

We finally note that if the baseline is perfectly aligned with the x axes of the two images the y coordinates are identical (q_{Ly}=q_{Ry}!)

Error analysis

The disparity can not be estimated with infinite precision.

It is therefore instructive to see how errors in the disparity affect the accuracy of the reconstructed depth.

The rate of the error propagation in univariate functions is governed by the function derivative:

$$\Delta Q_Z(d) = Q'_Z(d)\Delta d$$

After some easy development we obtain:

$$\Delta Q_Z(d) = -\frac{Q_Z^2}{f \cdot b}\Delta d$$

Error analysis (2)

The propagation of disparity error to the depth error:

$$\Delta Q_Z(d) = K\Delta d,\ K = -\frac{Q_Z^2}{f \cdot b}$$

We note the following:

- the error propagation constant depends quadratically on Q_Z (stereo is useless at extremely long range)
- large baseline can help,but it also implies a smaller stereo coverage in the close range
- large f also helps (small pixels, telephoto lenses), while also increasing the noise and compromising the close range performance

Error analysis (3)

Case study A (Q_Z=10m):

- wide field of view: f=4mm/5μm
- baseline b=10cm
- each pixel in disparity error $\Rightarrow$ K=1.25m of the depth error

Case study B (Q_Z=10m):

- narrow field of view f=16mm/5μm
- baseline b=100cm
- each pixel in disparity error $\Rightarrow$ K=0.31m of the depth error
- if we reduce disparity error to 0.1, the error in depth shall be 3 cm;

Stereo calibration

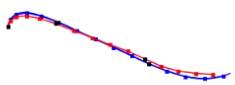
Real camera pairs are never perfectly aligned, however, with some manual work, they can be calibrated.

By using calibration parameters, stereo images can be transformed so that they appear as if they were acquired by a rectified pair

This process is denoted as rectification and is often performed as the first processing step in stereo analysis.

Overview of motion estimation

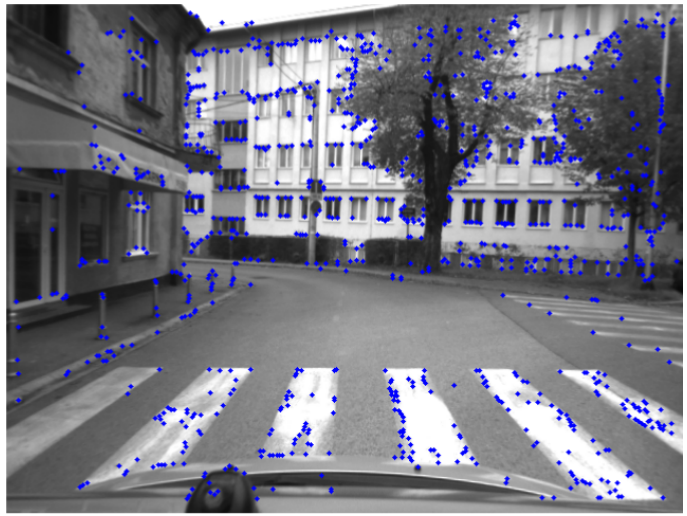
We usually start from inter-frame point correspondences:



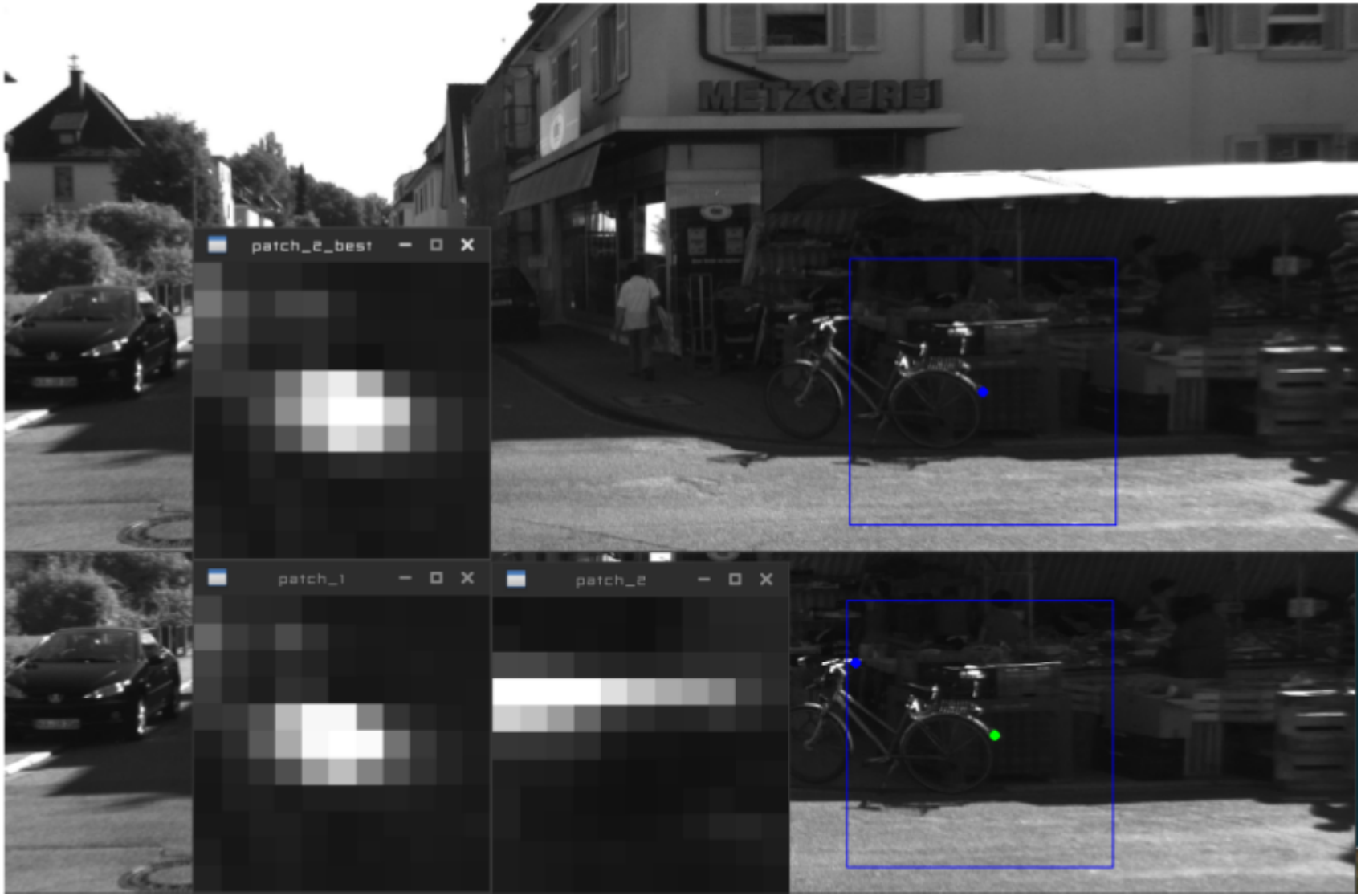
The camera motion is recovered under the constraint that it must explain the inter-frame change of the corresponding 3D points.

Point feature tracking

One possible approach: independent detection (e.g. Harris) plus descriptor matching (e.g. patch appearance, NCC).



NCC matching of patch descriptors at Harris keypoints:



NCC1=0.9545, NCC2=0.0047, Threshold = 0.9

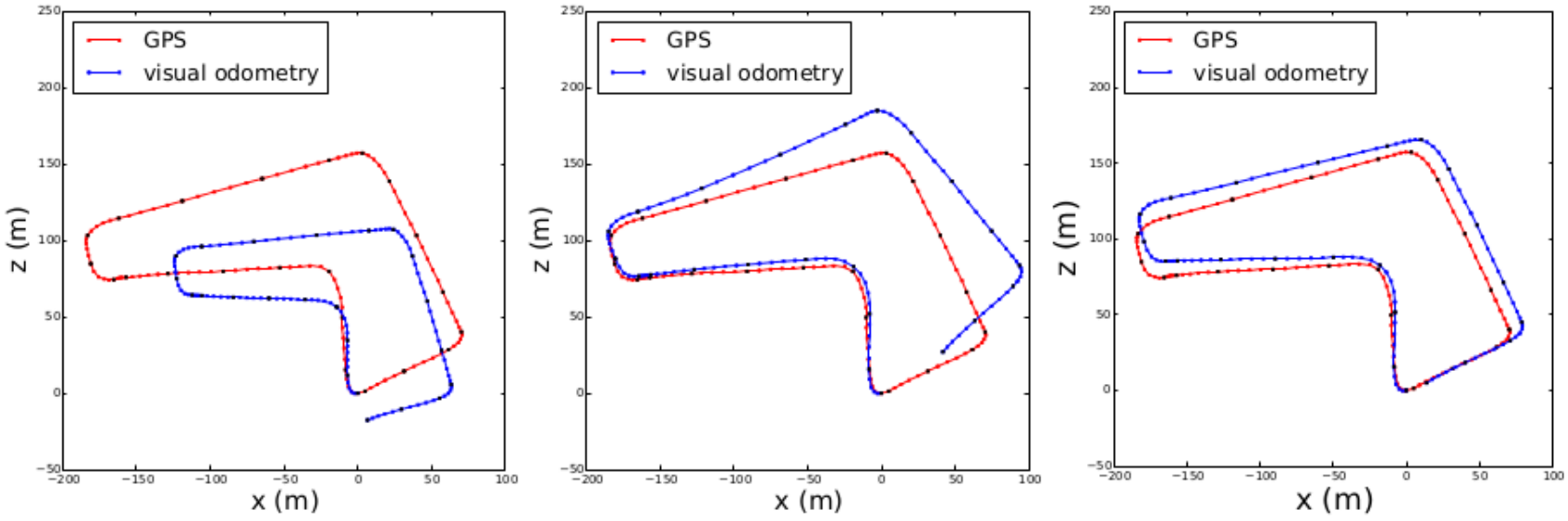
Two-frame motion recovery

1. reconstruct the tracked features $Q^M_{i,j}$ in the previous frame (j)
2. if the camera motion is ($\mathbf{R}$, $\mathbf{T}$), their estimated current image locations are:
$$q^E_{i,j+1} = \pi(\mathbf{R} \cdot Q^M_{i,j} + \mathbf{T})$$
3. now recover the motion parameters ($\mathbf{R}$, $\mathbf{T}$) by requesting that estimated feature locations $q^E_{i,j+1}$ be as similar as possible to the tracked features $q^M_{i,j+1}$:
$$(\mathbf{R}, \mathbf{T}) = \arg\min_{\mathbf{R},\mathbf{T}} \sum_i \| q^M_{i,j+1} - q^E_{i,j+1} \|^2$$

The previous procedure is first performed on random correspondence triples (RANSAC style).

The solution is finally obtained by reestimation on the set of inliers.

Experimental results



Conclusions:

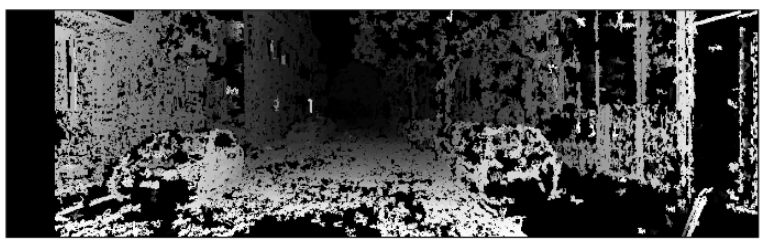
- 12cm baseline able to deliver useful results
- carefully prepared GPS acquisition able to provide useful motion groundtruth
- accurate calibration matters a lot

Overview of dense stereo reconstruction

The main goal: to obtain a dense disparity map (i.e. a dense 3D reconstruction).

Two main approaches:

- local methods (left): evaluate each pixel independently;
- global methods (right): take into account the whole image while evaluating each pixel.



Local methods are related to recovering point correspondences, but much more computationally intensive due to dense output.

Semi global matching

A popular light-weight global approach with good performance.

Cost = pixelwise data cost C(p,d_p) + discontinuity costs P1,P2:

$$E(p,d) = \sum_p (C(p,d_p) + \sum_{q \in N_p} P1 \cdot T(|d_p - d_q| = 1) \quad (1)$$

$$+ \sum_{q \in N_p} P2 \cdot T(|d_p - d_q| > 1)) \quad (2)$$

Problem: the test $T(|d_p - d_q| = 1)$ depends on neighbours (which again depend on their neighbours, etc.)

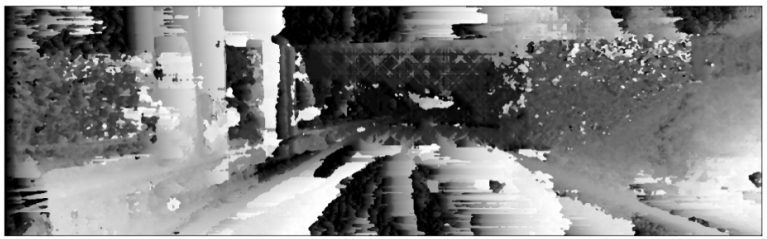
The solution is performed by dynamic programming in 8 image directions.

Popular data costs

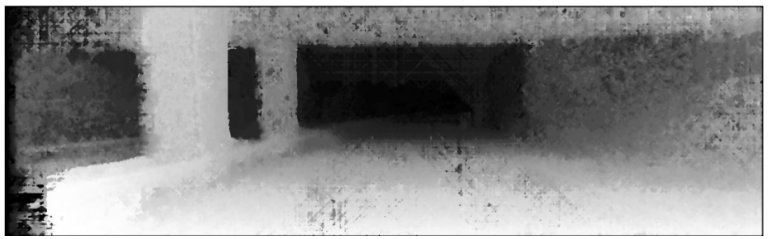
- SAD: $\sum_{i \in W} |p_i - q_i|$
- ZSAD: $\sum_{i \in W} |(p_i - q_i) - (\mu_p - \mu_q)|$
- Census (Hamming distance of binary descriptors):

$$\begin{array}{cccccccc} 32 & 64 & 96 & & 0 & 0 & 1 & \\ 32 & \mathbf{64} & 96 & \rightarrow & 0 & & 1 & \rightarrow (11010110)_2 \rightarrow 214 \\ 32 & 32 & 96 & & 0 & 0 & 1 & \end{array}$$

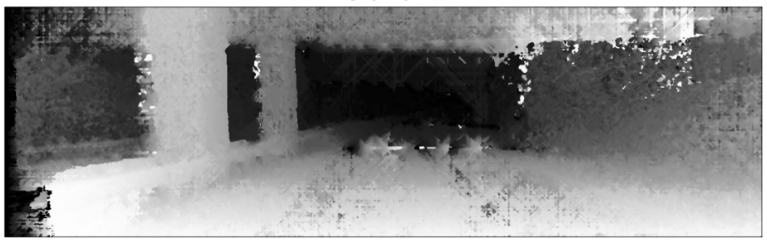
opencv



sad



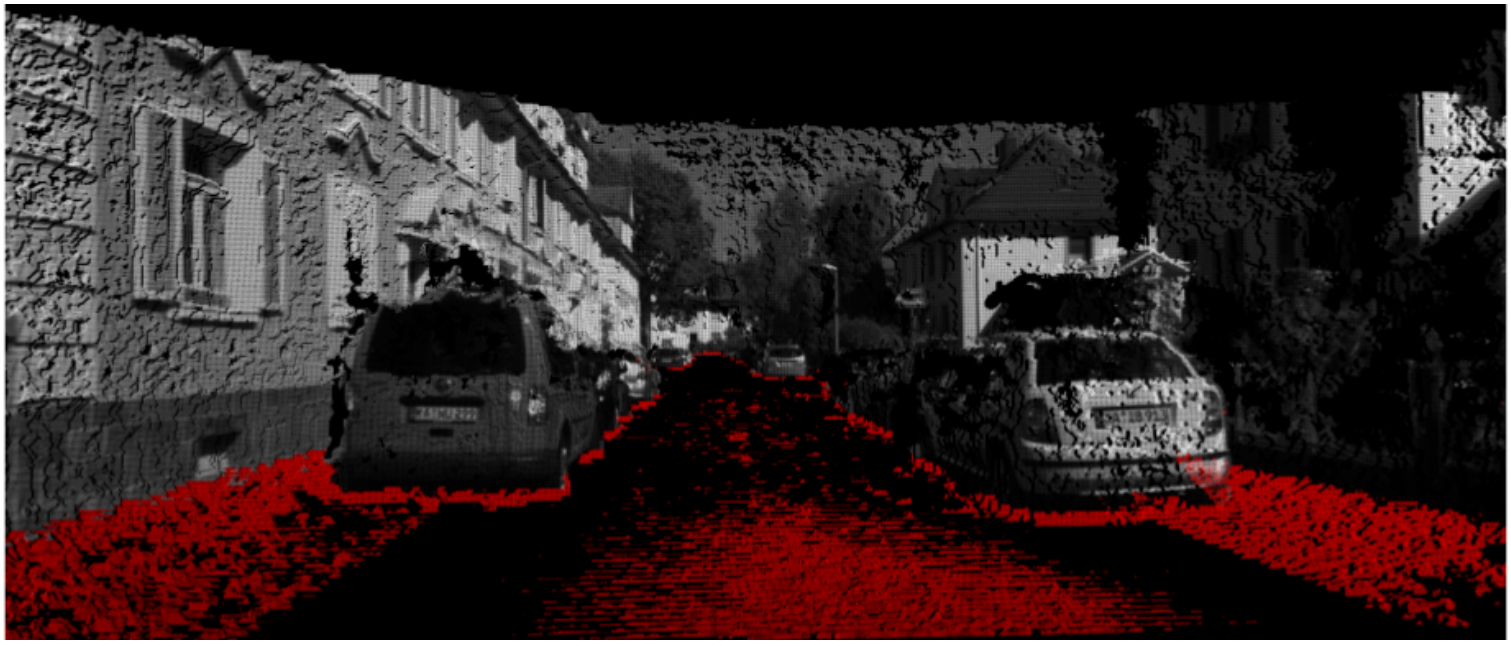
census



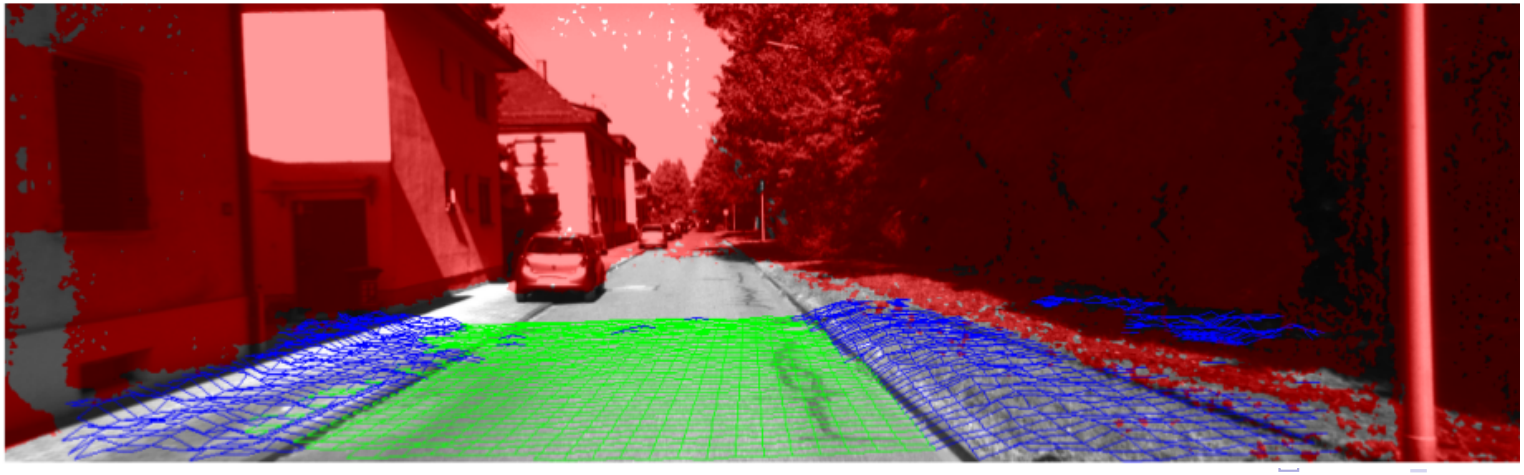
zsad

Applications

Groundplane estimation:



Building a traversability map:



Conclusion

Stereo provides an interesting alternative to laser scanners for structure and motion estimation

Main challenges:

- computational complexity
  - motion: realtime
  - local reconstruction: 4Hz
  - SGM: 1 fps
- large memory requirements (SGM: 1 GB),
- bad correspondences, uneven illumination...

Applications:

- motion recovery,
- object detection (up to 100m distance),
- scene understanding.