Computer Vision Systems Programming VO

3D Vision Applications

Christopher Pramerdorfer

Computer Vision Lab, Vienna University of Technology

Topics

Image formation

3D data acquisition

Kinect applications

Images from wikipedia.org, Shotton et al. 2011, Newcombe et al. 2011

Motivation

CV is about infering information about the world from images

- Knowledge of scene geometry beneficial

This lecture covers

- How scene geometry and images are related
- How this relation can be “inverted”
- CV applications utilizing scene geometry

Image Formation
Pinhole Camera Model

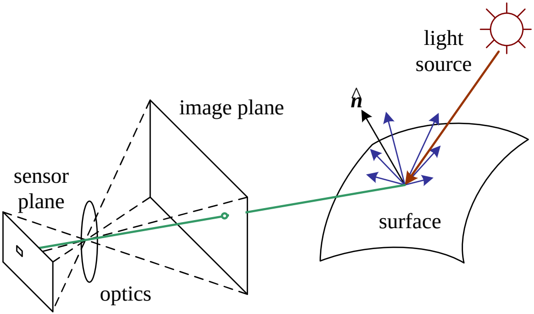


Image from Szeliski 2010

Image Formation
Pinhole Camera Model

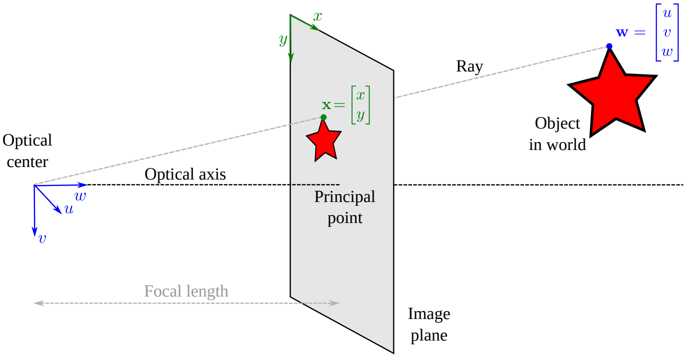


Image adapted from Prince 2012

We obtain $x = fu/w + p_x$, $y = fv/w + p_y$

- f : focal length in pixels
- p_x, p_y : image coordinate of the principal point

This mapping is linear in **homogeneous coordinates**

$$\lambda \tilde{\mathbf{x}} = \begin{pmatrix} \mathbf{\Lambda} & \mathbf{0} \end{pmatrix} \tilde{\mathbf{w}}$$

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} f & 0 & p_x & 0 \\ 0 & f & p_y & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} u \\ v \\ w \\ 1 \end{pmatrix}$$

World and camera coordinate systems usually differ

- Transform $\mathbf{w}$ to camera coordinates before projection

$$
\begin{aligned}
\mathbf{w}' &= \boldsymbol{\Omega}\mathbf{w} + \boldsymbol{\tau} \\
\begin{pmatrix} u' \\ v' \\ w' \end{pmatrix} &= \begin{pmatrix} \omega_{11} & \omega_{12} & \omega_{13} \\ \omega_{21} & \omega_{22} & \omega_{23} \\ \omega_{31} & \omega_{32} & \omega_{33} \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} + \begin{pmatrix} \tau_x \\ \tau_y \\ \tau_z \end{pmatrix}
\end{aligned}
$$

We obtain the full **pinhole camera model**

$$\lambda \tilde{\mathbf{x}} = \begin{pmatrix} \mathbf{\Lambda} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{\Omega} & \boldsymbol{\tau} \\ \mathbf{0}^\top & 1 \end{pmatrix} \tilde{\mathbf{w}}$$

Standard camera model in CV

- Usually together with radial distortion correction

Approximation to actual image formation

- In practice $\mathbf{w}$ is not mapped to a single $\mathbf{x}$

Computing Scene Geometry

We could obtain $\mathbf{w}$ by inverting the pinhole camera model

- But we don't know w

To this end, we must

- Utilize information from multiple images
- Use sensors that capture w (depth sensors)

Image by John Kratz / flickr

In **stereo reconstruction** we have

- n point correspondences $\{(\mathbf{x}_1, \mathbf{x}_2)\}$ in two images
- Taken with calibrated cameras (known $\boldsymbol{\Lambda}, \boldsymbol{\Omega}, \boldsymbol{\tau}$)

Goal is to estimate corresponding world coordinates $\mathbf{w}$

- Accomplished via triangulation

Computing Scene Geometry
Stereo

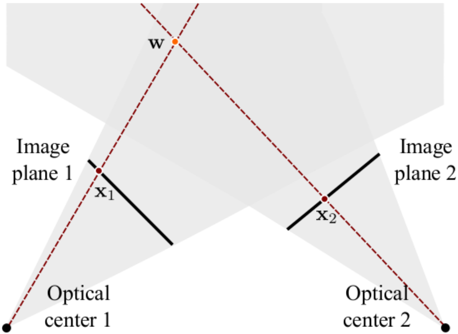


Image from Prince 2012

Computing Scene Geometry
Stereo

The challenge is finding correspondences

We typically want

- Many correspondences to obtain a dense 3D model
- High accuracy and low noise

Usually accomplished via

- Dense feature matching along epipolar lines
- Local or global optimization

$\mathbf{x}_1$ must lie on the **epipolar line**

- Given by $\mathbf{x}_0$ and the camera parameters

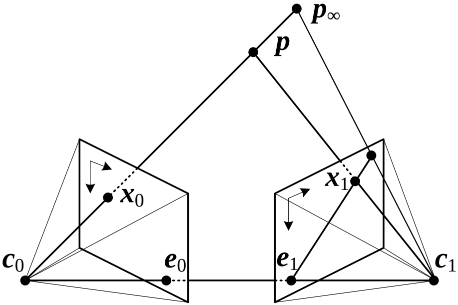


Image from Szeliski 2010

Images are **rectified** before correspondence search

- Relation between x-offset (**disparity** d) and w, $d = fb/w$
- b is the distance between the cameras

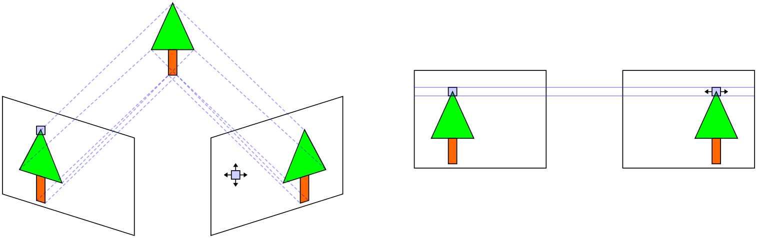


Image adapted from `wikipedia.org`

Computing Scene Geometry
Stereo

Dense matching on rectified images results in a disparity map

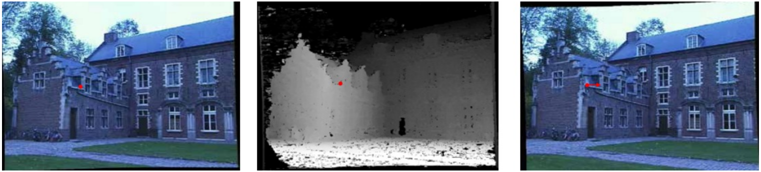


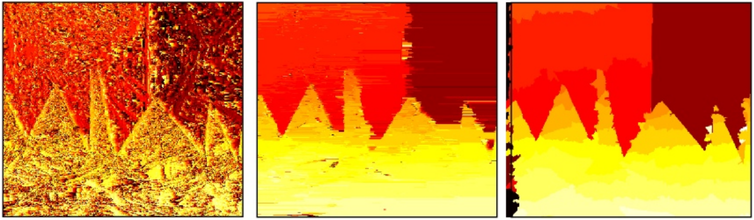
Image from Guido Gerig's slides

Computing Scene Geometry
Stereo

Raw disparity maps are noisy

Quality can be improved by encouraging smoothness

Accomplished via graphical models (e.g. MRFs)



Images from Prince 2012

Limitations of image-based (passive) stereo

- No correspondences in regions without texture
- Relies on proper illumination (no dark living rooms)
- Computational complexity

Alternatively, we can use sensors that capture w directly

- Usually together with brightness or color

These **depth sensors**

- Do not rely on texture
- Are not (significantly) affected by lighting conditions

Released by Microsoft for Xbox 360 in late 2010

Fastest selling consumer electronics device

Image from `wikipedia.org`

Computing Scene Geometry
Depth Sensors – Kinect v1

Image from https://www.youtube.com/watch?v=p2qlHoxPioM

Bibliography

Newcombe, Richard A et al. (2011). **KinectFusion: Real-time dense surface mapping and tracking**. ISMAR.

Prince, S.J.D. (2012). **Computer Vision: Models Learning and Inference**. Cambridge University Press.

Shotton, Jamie et al. (2011). **Real-Time Human Pose Recognition in Parts from a Single Depth Image**. CVPR.

Szeliski, Richard (2010). **Computer vision: algorithms and applications**. Springer.