Low-level Vision Tasks

Stereo

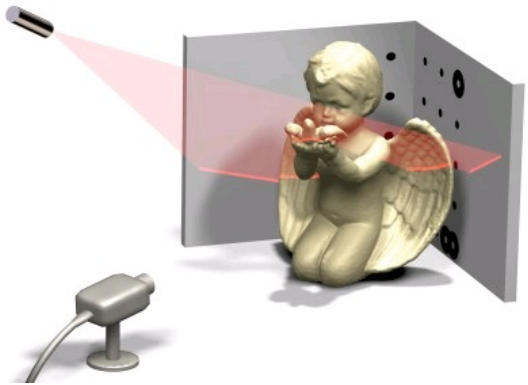
Kuk-Jin Yoon

Visual Intelligence Lab.
Department of Mechanical Engineering

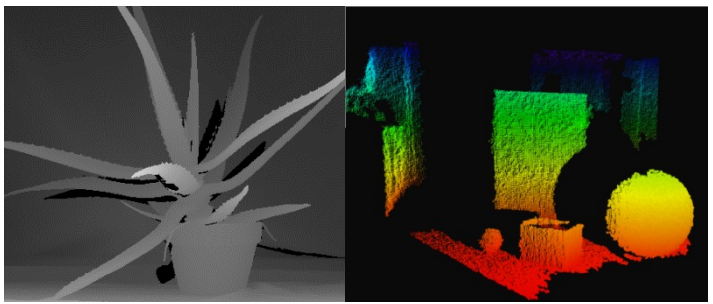
KAIST

3D Sensing

- Estimating depth or distance from a sensor to the scene surface, or complete 3D shape (structure) of the scene based on the geometrical and photometrical properties
  - has long been an important topic in metrology and computer vision.



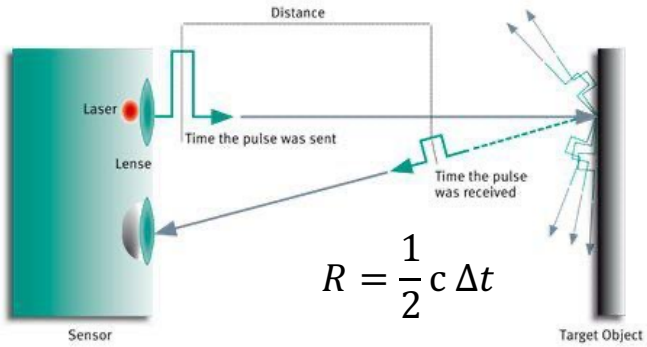
3d sensing with laser scanner



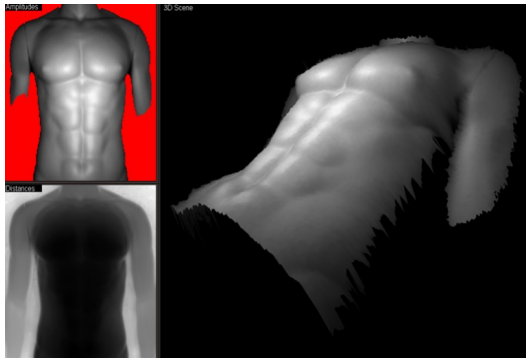
3D sensing results using stereo vision

3D Sensing Methods in Metrology

- Time-of-flight

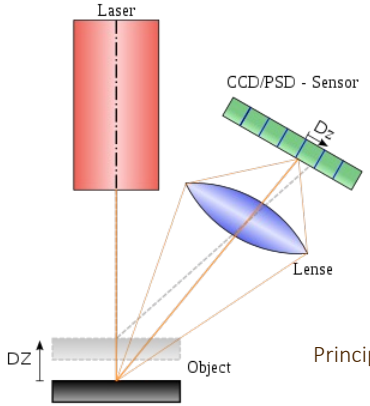


$$R = \frac{1}{2} c \, \Delta t$$

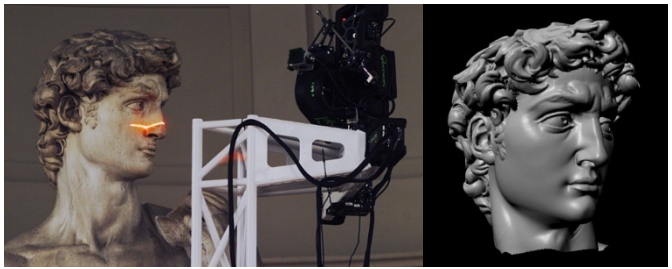


Principle of ToF sensors and acquired 3D data

- Laser scanning



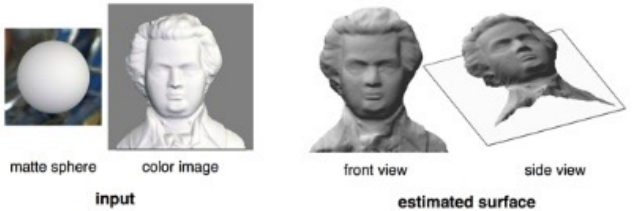
Principle of Laser scanning



Digital Michelangelo Project
http://graphics.stanford.edu/projects/mich

3D Sensing Methods in Computer Vision

- Shape(Structure)-from-X
  - 'X' can be any visual cue that can be extracted from images such as
    - shading
    - silhouette
    - motion
    - stereo
    - …

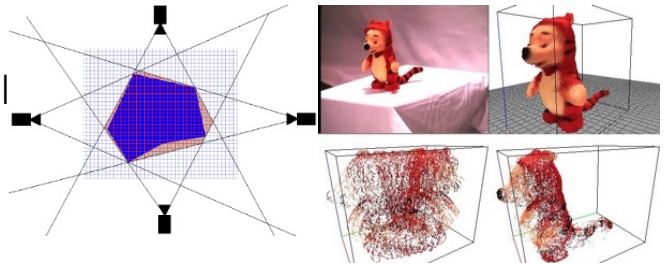


Shape estimation in natural illumination, cvpr 2011

Skeletal Graphs for Efficient Structure from Motion, cvpr 2008

uncontroll

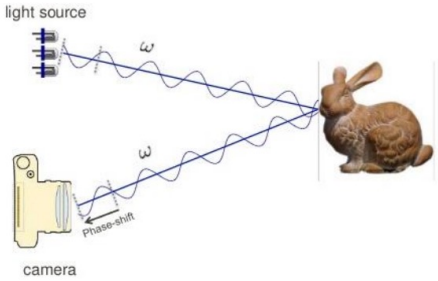
thods in metrology



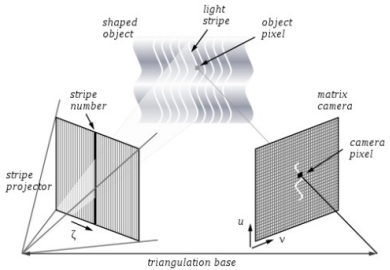
Visual Hull Alignment and Refinement Across Time: A 3D Reconstruction Algorithm Combining Shape-From-Silhouette with Stereo, cvpr 2003

Depth Estimation

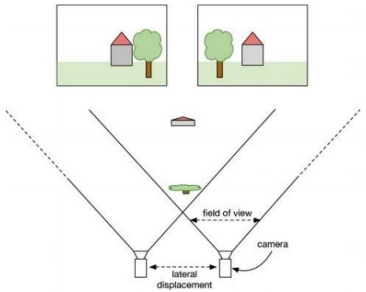
Multiple ways to estimate depth



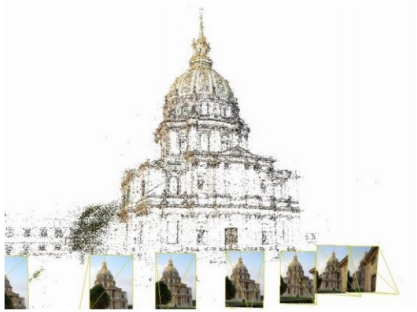
Time of flight



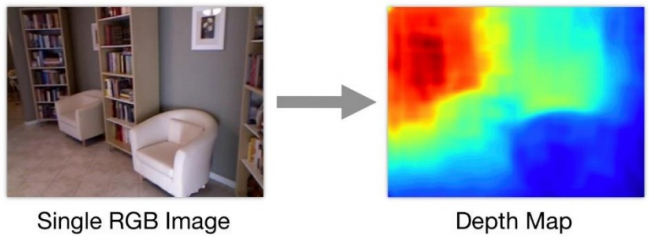
Structure light



Stereo matching



Structure from motion



Single Image Depth Estimation

Most cost-effective approaches

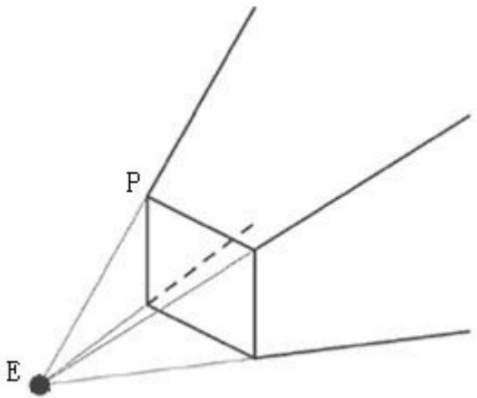
Main focus of this talk

KAIST

Single Image Depth Estimation

Estimate depth from a single image

- Monocular depth estimation: $D = F(I)$
- Highly-ill posed, infinite configurations
- Deep learning to rescue!

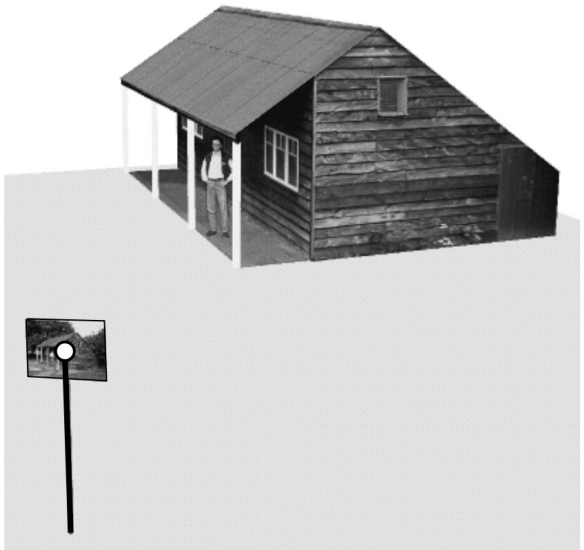


Needs semantic info to understand the scene

Single View Metrology

a

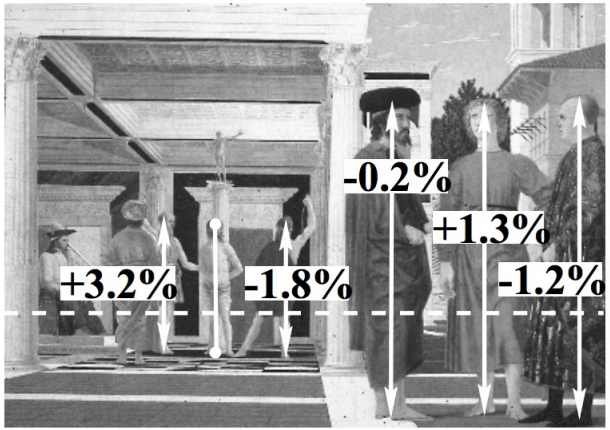
b



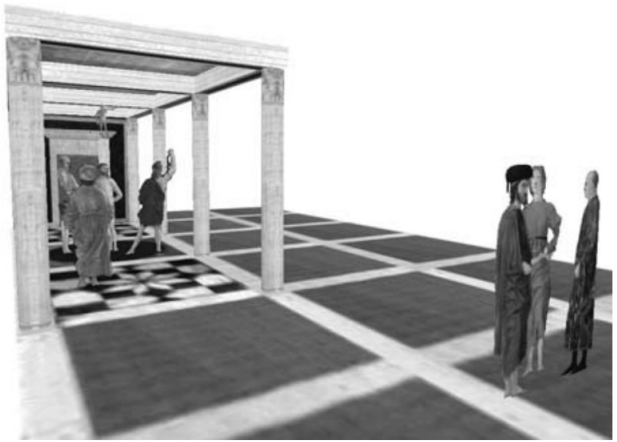
c

Single View Metrology

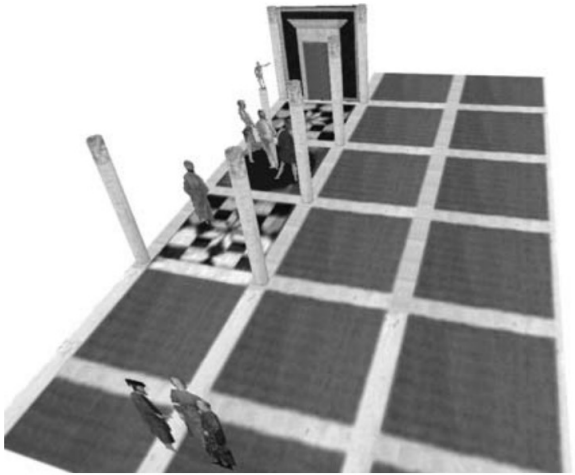
a



b



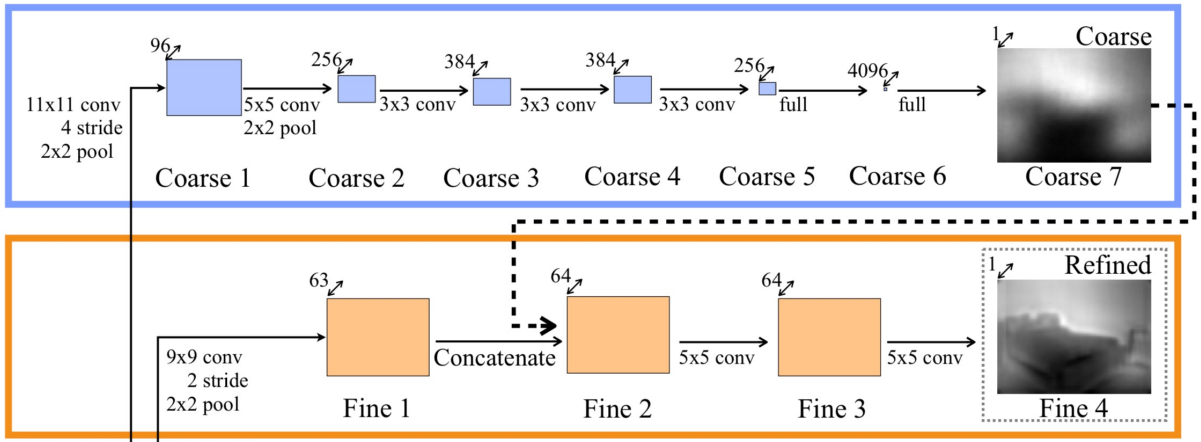
c



d

Deep-learning-based Monocular Depth Estimation (1)

- Depth Map Prediction from a Single Image using a Multi-Scale Deep Network (Eigen et. al, NIPS 2014)

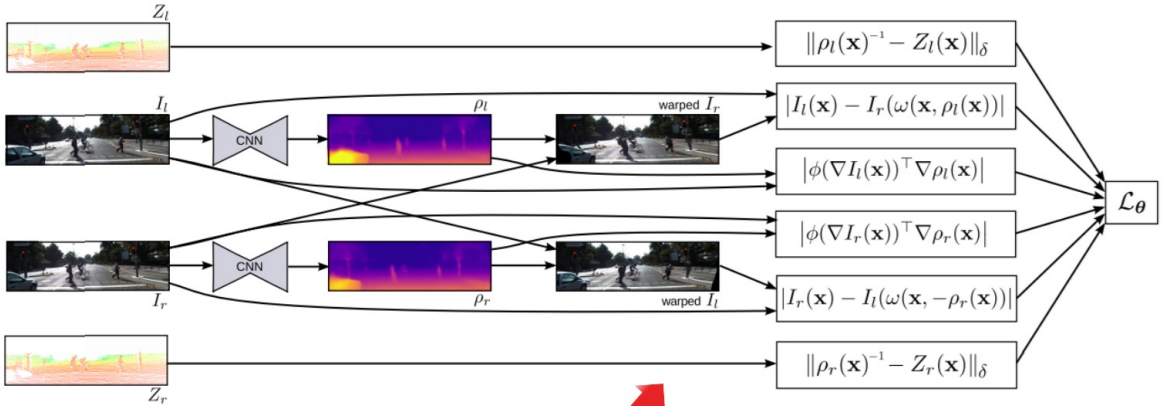


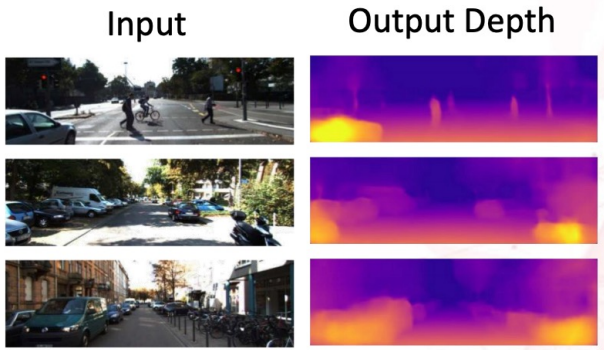
		Coarse					Fine
Layer	input	1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4

Deep-learning-based Monocular Depth Estimation (2)

Semi-supervised/Unsupervised schemes

- Use stereo pairs for training
- Based on left-right consistency and smoothness of depth





[8] defines a series of loss functions based on left-right consistency and depth smoothness

[7] C. Godard, O. Mac Aodha, G.J. Brostow"Unsupervised monocular depth estimation with left-right consistency," In *Proc. CVPR*, 2017.

[8] Y Kuznietsov, J Stückler and L. Bastian, "Semi-supervised deep learning for monocular depth map prediction," In *Proc. CVPR*, 2017.

Deep-learning-based Monocular Depth Estimation (3)

- Digging Into Self-Supervised Monocular Depth Estimation, ICCV 2019

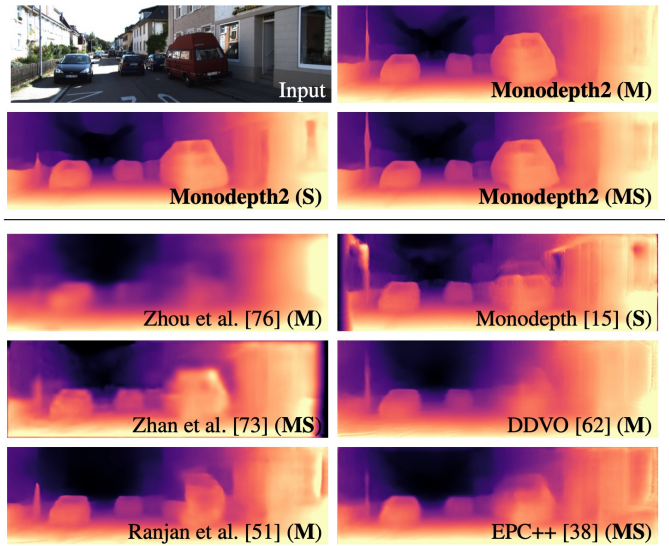


Figure 1. **Depth from a single image.** Our self-supervised model, **Monodepth2**, produces sharp, high quality depth maps, whether trained with monocular (M), stereo (S), or joint (MS) supervision.

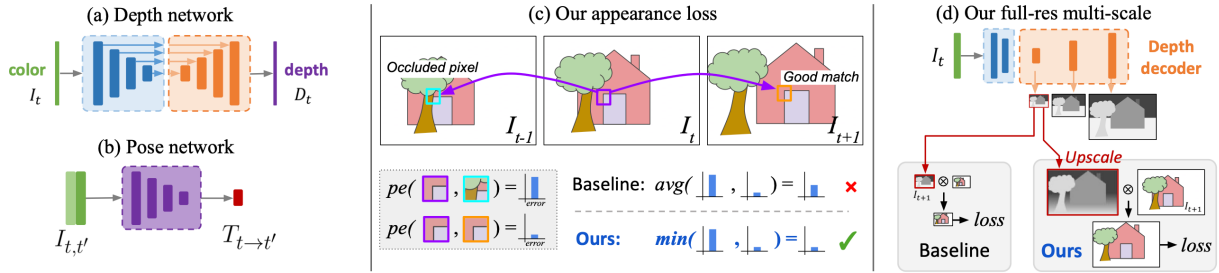


Figure 3. **Overview. (a) Depth network:** We use a standard, fully convolutional, U-Net to predict depth. **(b) Pose network:** Pose between a pair of frames is predicted with a separate pose network. **(c) Per-pixel minimum reprojection:** When correspondences are *good*, the reprojection loss should be *low*. However, occlusions and disocclusions result in pixels from the current time step not appearing in both the previous and next frames. The baseline *average* loss forces the network to match occluded pixels, whereas our *minimum reprojection* loss only matches each pixel to the view in which it is visible, leading to sharper results. **(d) Full-resolution multi-scale:** We upsample depth predictions at intermediate layers and compute all losses at the input resolution, reducing texture-copy artifacts.

- Combining the depth network with the pose network to handle occlusions and moving objects

Deep-learning-based Monocular Depth Estimation (4)

- Enforcing geometric constraints of virtual normal for depth prediction, ICCV 2019
  - enforcing a high- order geometric constraint in the 3D space for the depth prediction task.

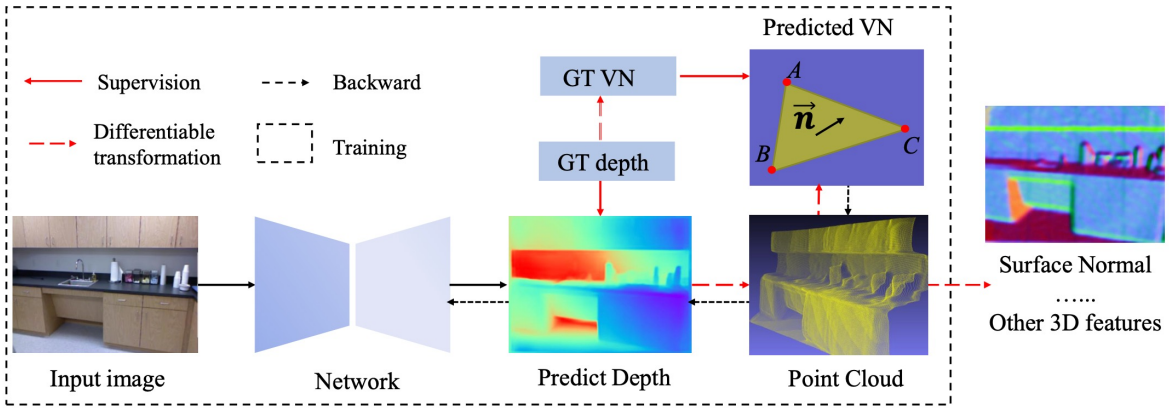


Figure 2. Illustration of the pipeline of our method. An encoder-decoder network is employed to predict the depth, from which the point cloud can be reconstructed. A pixel-wise depth supervision is firstly enforced on the predicted depth, while a geometric supervision, virtual normal constraint, is enforced in 3D space. With the well trained model, other 3D features, such as the surface normal, can be directly recovered from the reconstructed 3D point cloud in the inference.

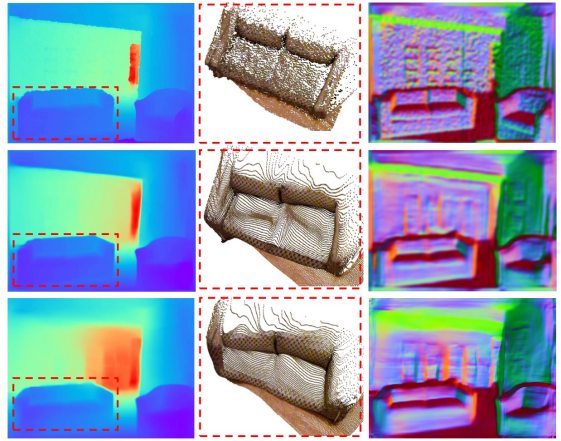
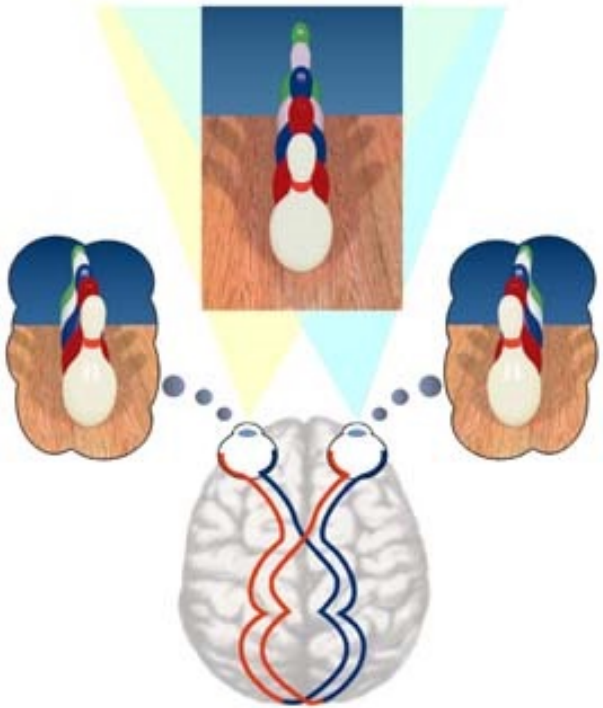


Figure 1. Example results of ground truth (the first row), our method (the second row) and Hu *et al.* [18] (the third row). By enforcing the geometric constraints of virtual normals, our reconstructed 3D point cloud can represent better shape of sofa (see the left part) and the recovered surface normal has much less errors (see green parts) even though the absolute relative error (rel) of our predicted depth is only slightly better than Hu *et al.* (0.108 *vs.* 0.115).

Stereopsis

- Disparity
  - Informally, the difference between two pictures
  - Allows us to gain a sense of depth
- Stereopsis
  - Ability to perceive depth from disparity
    - Gaining the sense of depth by fusing the images recorded by two or more cameras and exploiting the difference
- Stereo vision
  - Design algorithms that mimic stereopsis
    - can use two-view (binocular) or multi-view



Stereopsis

Applications of Stereo Vision System

Driver Assistant System (DAS)

Stereo Vehicle System for DAS

Detecting and tracking dynamic objects to prevent collision

- Dynamic Object Detection
- Object Tracking and Motion Prediction in 3D Space
- Ego motion Estimation

Stereo Camera

RGB & Depth Image

Feature Detection & Tracking (KLT)

Vehicle Pose Estimation (Ego Motion)

Motion Segmentation & Detection

Multi-Object Tracking & Path Prediction

3D Reconstruction via Stereo Vision

- Camera calibration (skip)
  - Calibrating camera parameters
- Correspondence search
  - Finding all corresponding points
- 3D computation
  - Computing depth or surfaces

Stereo Vision (in computer vision)

- Triangulate on two images of the same point to recover depth.
  - Feature matching across views
  - Calibrated cameras

Matching correlation windows across scan lines

Stereo Vision (in computer vision)

- Triangulate on two images of the same point to recover depth.
  - Feature matching across views
  - Calibrated cameras



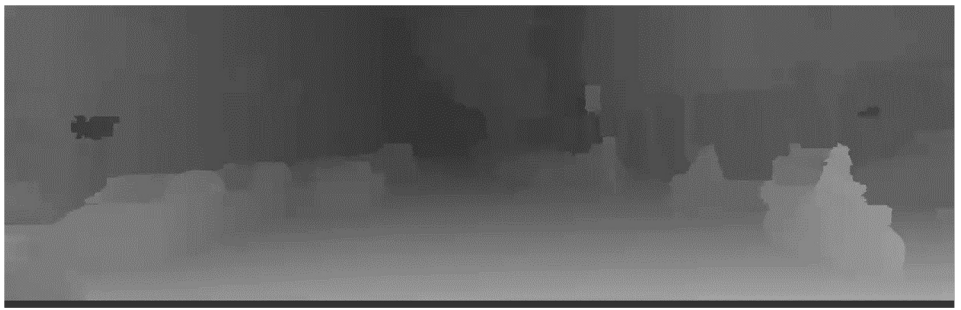
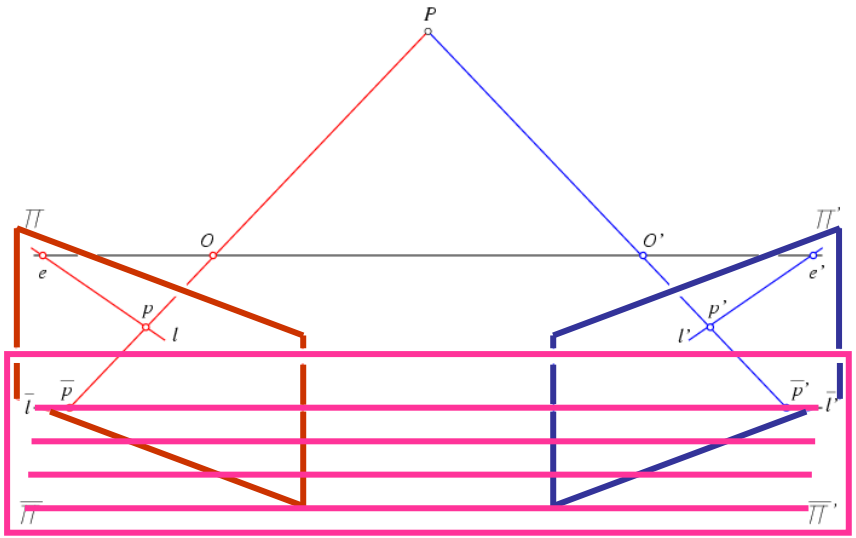



Image Rectification

- Rectification
  - Re-projecting two image planes onto a common plane
    - Simplifying the stereo process



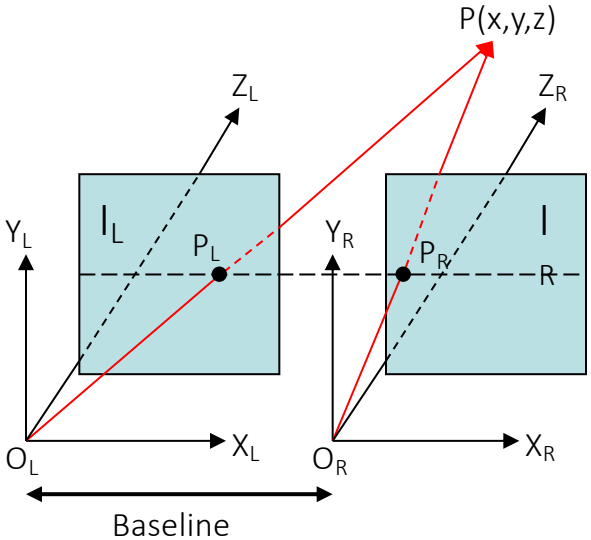
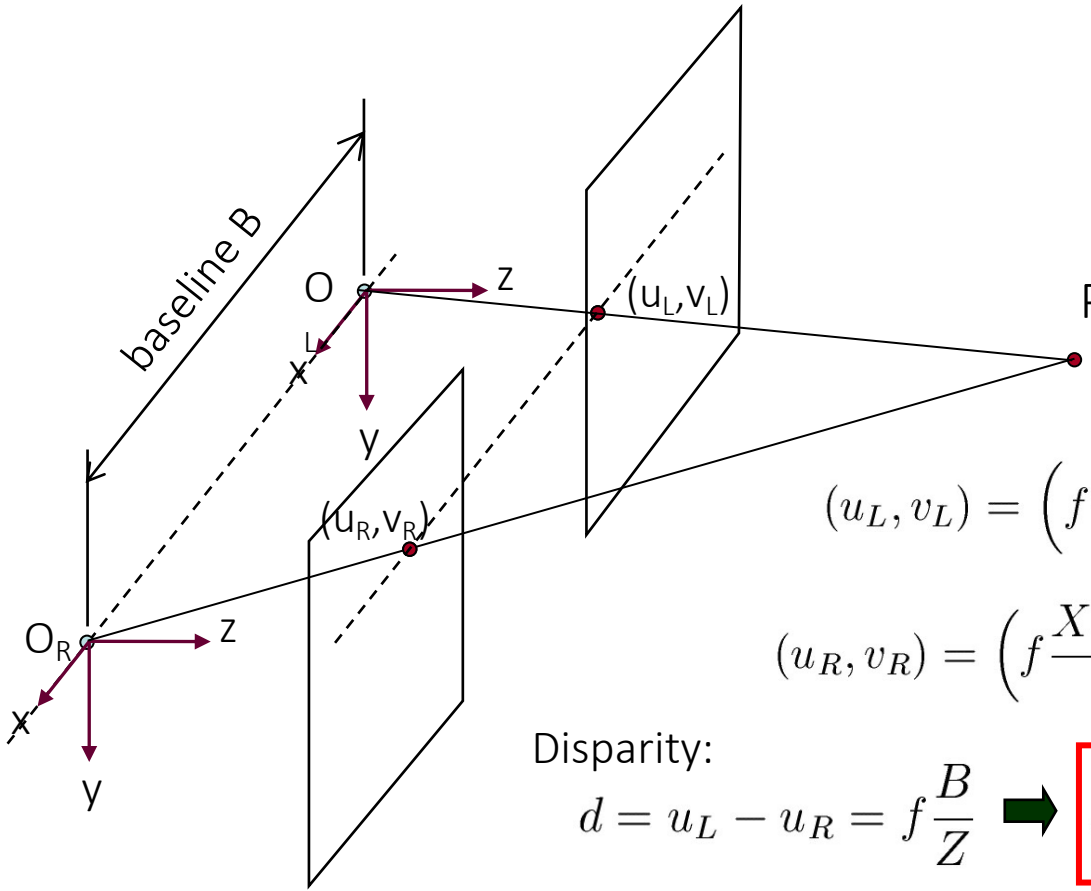


Image Rectification

- Property of rectified images
  - Focal points are at the same height.
  - Focal lengths are the same.
  - The epipolar lines are scanlines and also parallel to the baseline.
    - Image planes of cameras are parallel.
  - Disparity is defined as the difference of the x-axis position.
    - Possible to use simplified algorithms
    - Possible to improve the efficiency

Basic Derivations with Rectified Images



$$(u_L, v_L) = \left(f\frac{X}{Z}, f\frac{Y}{Z}\right)$$

$$(u_R, v_R) = \left(f\frac{X-B}{Z}, f\frac{Y}{Z}\right)$$

Disparity:

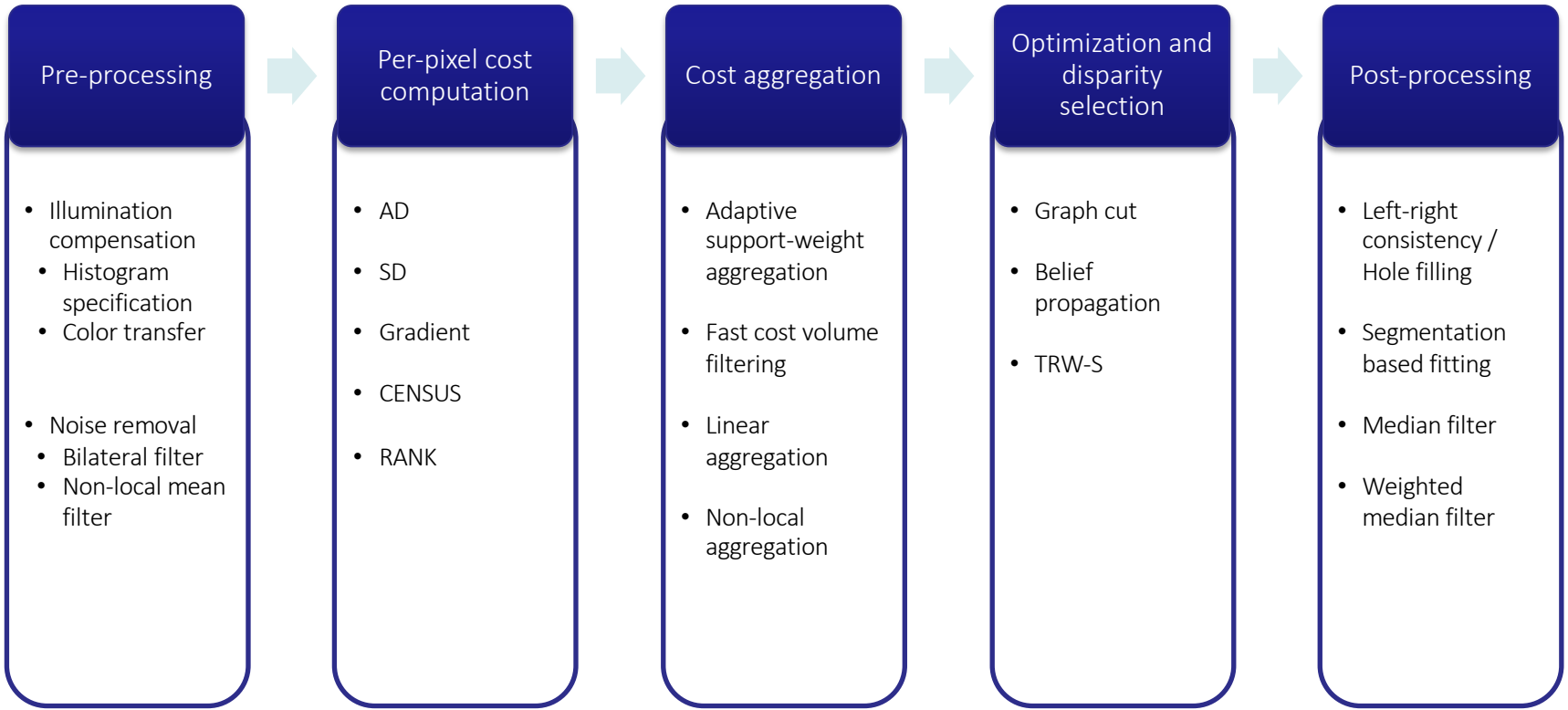
$$d = u_L - u_R = f\frac{B}{Z} \Rightarrow \boxed{Z = f\frac{B}{d}}$$

Figure from D. Lowe

Steps in Area-based Stereo

- Pre-processing
  - e.g. noise reduction, illumination compensation, etc.
- Per-pixel raw cost computation
  - Pixel-by-pixel matching cost (or (dis-)similarity) computation
- Cost aggregation
  - Aggregating pixel-by-pixel matching cost for reducing image ambiguity
- Disparity selection
  - Determining the optimal disparity of each pixel based on the aggregated matching cost
- Post-processing
  - Removing or correcting false matches

Solving the Stereo Matching Problem

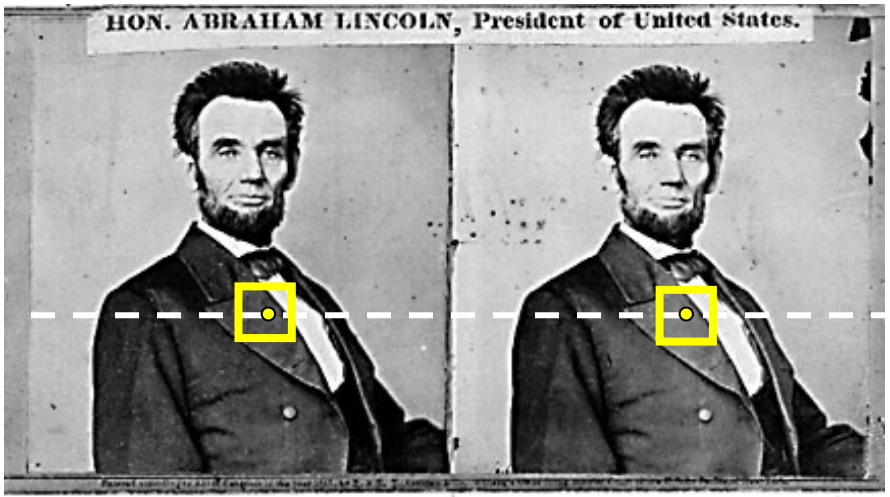


Pixel Matching

- Applying feature matching criterion at all pixels simultaneously and then finding the most similar pixel

- Similarity between pixels
  - Photometric constraint: intensity conservation assumption (ICA)
    - Correspondences have the same intensity in all images.
    - True for Lambertian surfaces
      - A Lambertian surface has a brightness that is independent of viewing angle
    - Violations
      - Noise
      - Specularity
      - Non-Lambertian materials
      - Pixels that contain multiple surfaces

Pixel Matching



For each epipolar line

For each pixel in the left image

- compare with every pixel on the same epipolar line in the right image
- pick one pixel with the minimum match cost

This leaves too much ambiguity, so:

Improvement: match windows

Similarity Measures

- Similarity measures with local windows
  - SSD(Sum of Squared Differences)

$$SSD(d) = \sum_i \sum_j \{I_L(x+d+i, y+j) - I_R(x+i, y+j)\}^2$$

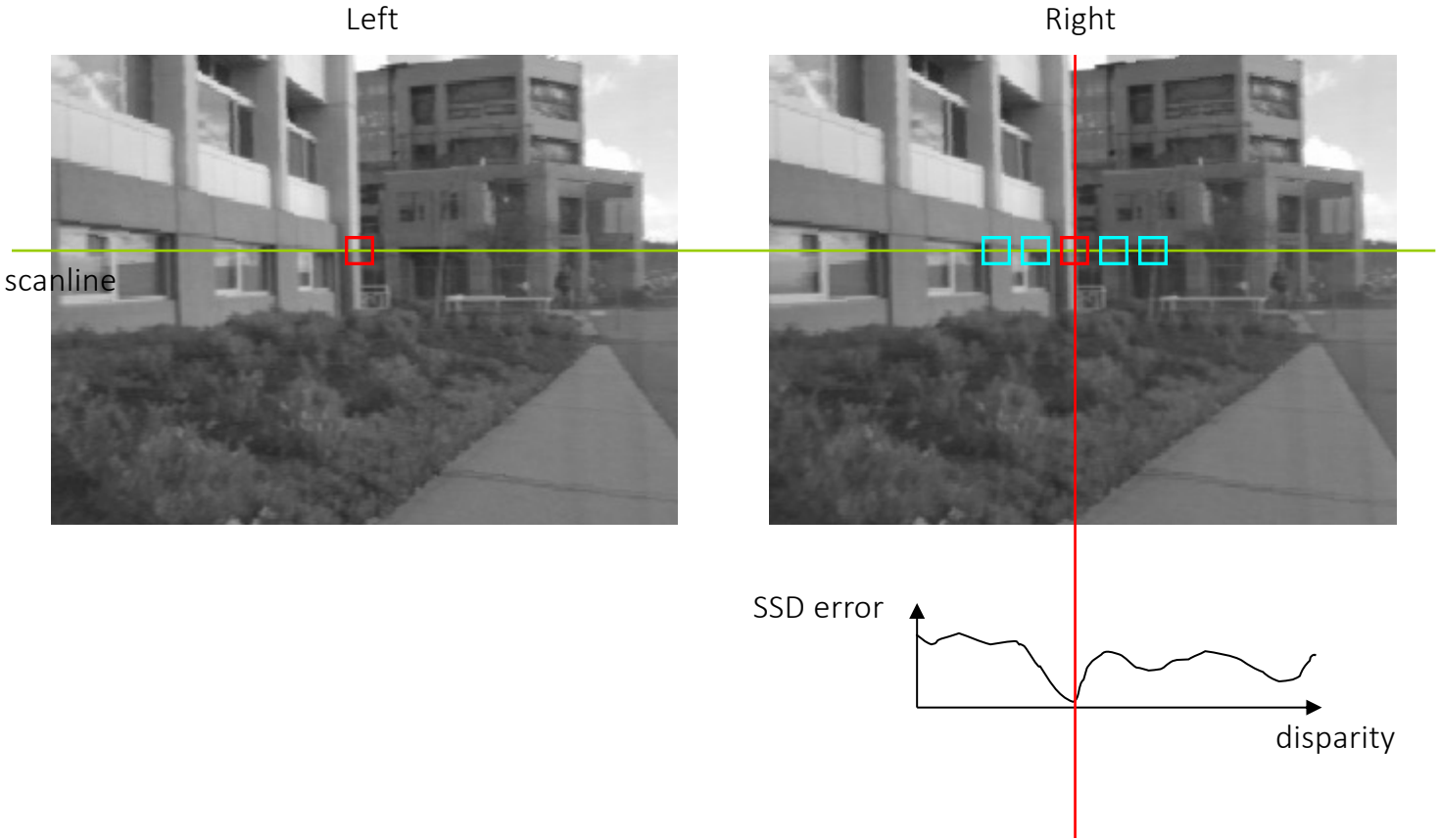
  - SAD(Sum of Absolute Differences)

$$SAD(d) = \sum_i \sum_j | I_L(x+d+i, y+j) - I_R(x+i, y+j) |$$

  - NCC(Normalized Cross Correlation)

$$NCC(d) = \frac{C(I_L, I_R) - \sum_i \sum_j \mu_L \mu_R}{\sum_i \sum_j \sigma_L \sigma_R}$$

Correspondence Using Correlation

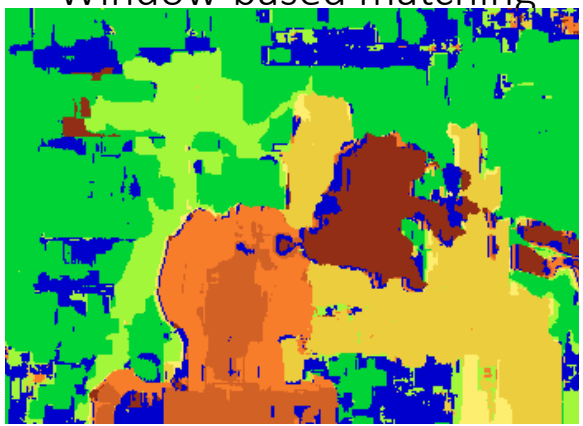


Slide from Lowe

Results with Window Search

Data

Window-based matching



Ground truth

Better Methods Exist...

Graph cuts

Ground truth

Y. Boykov, O. Veksler, and R. Zabih, Fast Approximate Energy Minimization via Graph Cuts, PAMI 2001

For the latest and greatest: http://www.middlebury.edu/stereo/

Global Methods

- Global methods
  - Trying to find a depth surface by minimizing a pre-defined global energy using optimization techniques
  - Most global methods focus on how to minimize E(D|I)

- MRF model for two-view stereo
  - Finding the optimal disparity map D_{opt}

$$D_{opt} = \arg\max_{D} P(D \mid I) = \frac{P(I \mid D)P(D)}{P(I)}$$

  - I : a set of input images
  - D : disparity field of a reference image

$$P(I \mid D) \propto \prod_{p} \exp(-\phi(p, d_p, I))$$

cost function of pixel p with disparity d_p for given observation I

$$P(D) \propto \prod_{p} \prod_{q \in N(p)} \exp(-\psi_c(d_p, d_q))$$

joint clique potential function of d_p and d_q

MRF Model in terms of Cost Functions

$$P(D \mid I) \propto \prod_{p} \exp(-\phi(p, d_p, I)) \prod_{p} \prod_{q \in N(p)} \exp(-\psi_c(d_p, d_q))$$

$$-\ln(P(D \mid I)) \propto \sum_{p} \phi(p, d_p, I)) + \sum_{p} \sum_{q \in N(p)} \psi_c(d_p, d_q)$$

$$E(D \mid I) = \sum_{p} D(p, d_p, I) + \sum_{p} \sum_{q \in N(p)} V(d_p, d_q)$$

- where

$$D(p, d_p, I) = \phi(p, d_p, I) \qquad V(d_p, d_q) = \psi_c(d_p, d_q)$$

Maximizing P(D|I) = Minimizing E(D|I)

Solving the Stereo Matching Problem

- Two key steps
  - Defining cost functions

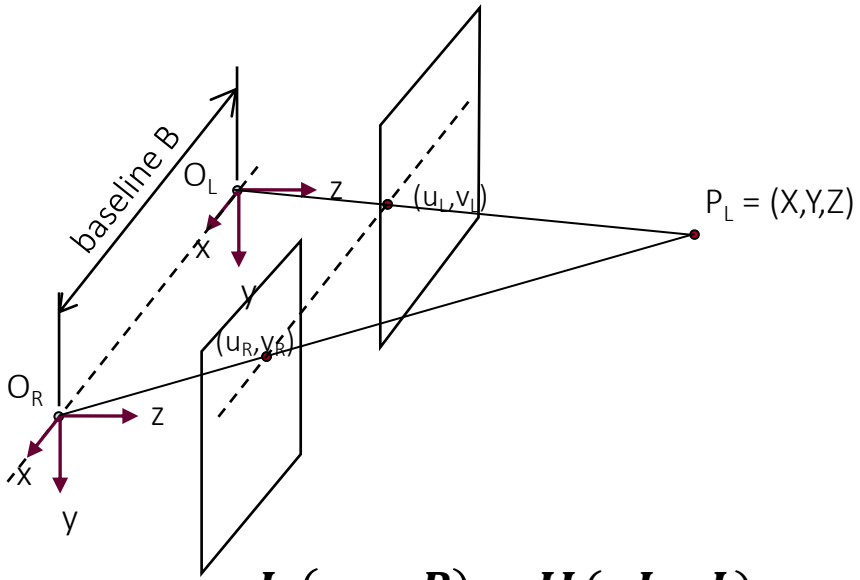
$$E(D \mid I) = \sum_{p} D(p, d_p, I) + \sum_{p} \sum_{q \in N(p)} V(d_p, d_q)$$

    - Should be robust to image ambiguity, occlusion, and depth discontinuities
    - Data cost (likelihood)
      - Measuring (dis-)similarity between image points
    - Discontinuity cost (prior)
      - Making the overall surface smooth while preserving details
  - Minimizing the cost functions to get the optimal disparity maps
    - Using optimization techniques such as graph cut(GC), belief propagation(BP), dynamic programming (DP), simulated annealing(SA), and so on.

Likelihood (Data Cost Function)

$$E(D \mid I) = \sum_{p} D(p, d_p, I) + \sum_{p} \sum_{q \in N(p)} V(d_p, d_q)$$

- Measuring similarity between two pixels with disparity *d*
- Using photometric constraints
  - Intensity Conservation Assumption (ICA)
    - All correspondences have the same intensity (or brightness) across all images.
    - Valid only for Lambertian surfaces



$$I_R(u_R, vR) = IL(uL, vL)$$

Likelihood (Data Cost Function)

- Per-pixel cost computation
  - AD (Absolute difference)
  - SD (Squared difference)
  - Census (Census filter based cost)
  - Rank (Rank transform based cost)
  - …

- Aggregated cost
  - SAD(Sum of Absolute Differences)
  - SSD(Sum of Squared Differences)
  - NCC(Normalized Cross Correlation)
  - Adaptive support weight aggregation (CVPR 2005)
  - Fast cost volume filtering (CVPR 2010)
  - Linear aggregation (ICCV 2011)
  - Non-local aggregation (CVPR 2012)

- When comparing pixels just by using per-pixel raw matching costs, there might be so many matching candidates having the same similarity on the one epipolar line.

- Aggregating per-pixel raw matching cost within predefined neighborhood to reduce the ambiguity and to increase the discriminative power

Prior (Smoothness Cost)

- Enforcing some constraints to get a more reasonable disparity map

$$E(D \mid I) = \sum_{p} D(p, d_p, I) + \sum_{p} \sum_{q \in N(p)} V(d_p, d_q)$$

- Using a smoothness constraint
  - Making an optimal disparity map piece-wise smooth assuming that depth discontinuity generally coincides with color (or intensity) discontinuity.

$$V(d_p, d_q) = \begin{cases} 0 & \text{if } d_p = d_q \\ \rho(\Delta C) & \text{otherwise} \end{cases}$$

$$\rho(\Delta C) = \begin{cases} P \times s & \text{if } \Delta C < T \\ s & \text{otherwise} \end{cases}$$

conventional Potts model

Disparity Selection via Optimization

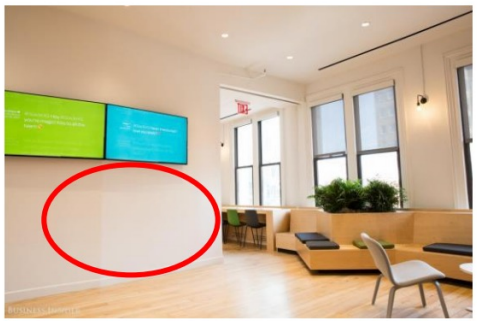
- Obtaining a disparity map that minimizes the pre-defined cost function via optimization techniques such as
  - Bayesian diffusion (Scharstein and Szeliski 1998)
  - Belief propagation (Sun et al. 2003, 2005)
  - Graph cut (Boykov et al. 2001, Kolmogorov and Zabih 2001)
  - Dynamic programming (Bobick and Intille 1999)
  - Tree-reweighted message passing (Kolmogorov 2005)

Challenges in Stereo Matching

Underdetermined (ill-posed)

Photometric variations

Occlusions



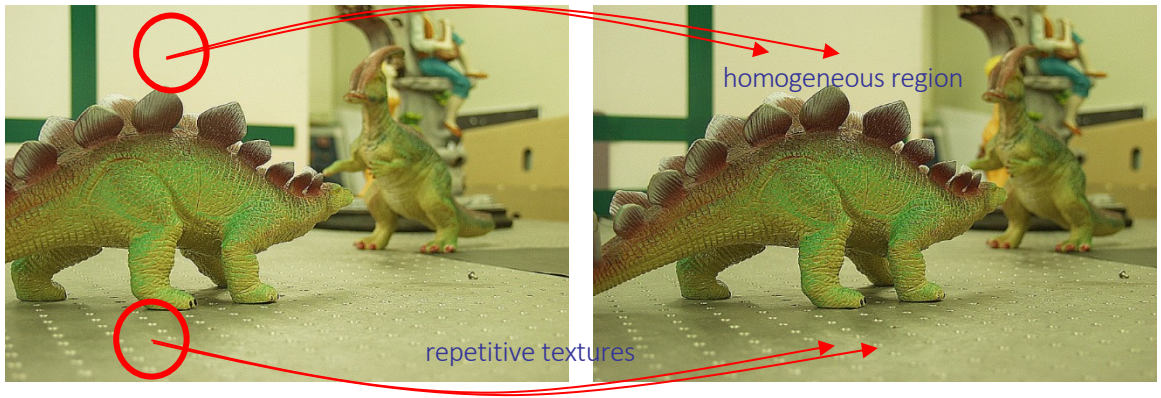
Texture-less Areas

Repetitive patterns

Reflections

Limitations of Passive Stereo Vision

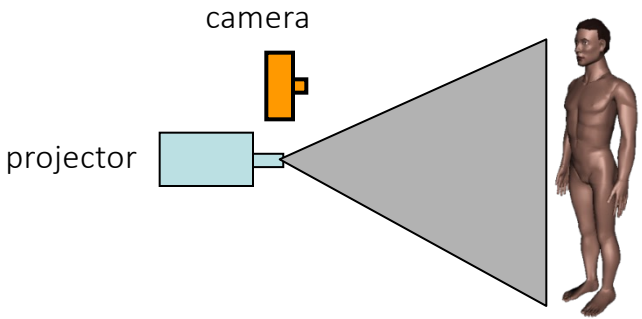
- Stereo matching error due to inherent image ambiguity



  - Disparity values of ambiguous pixels are assigned by the constraint used for the prior

Active Stereo with Structured Light

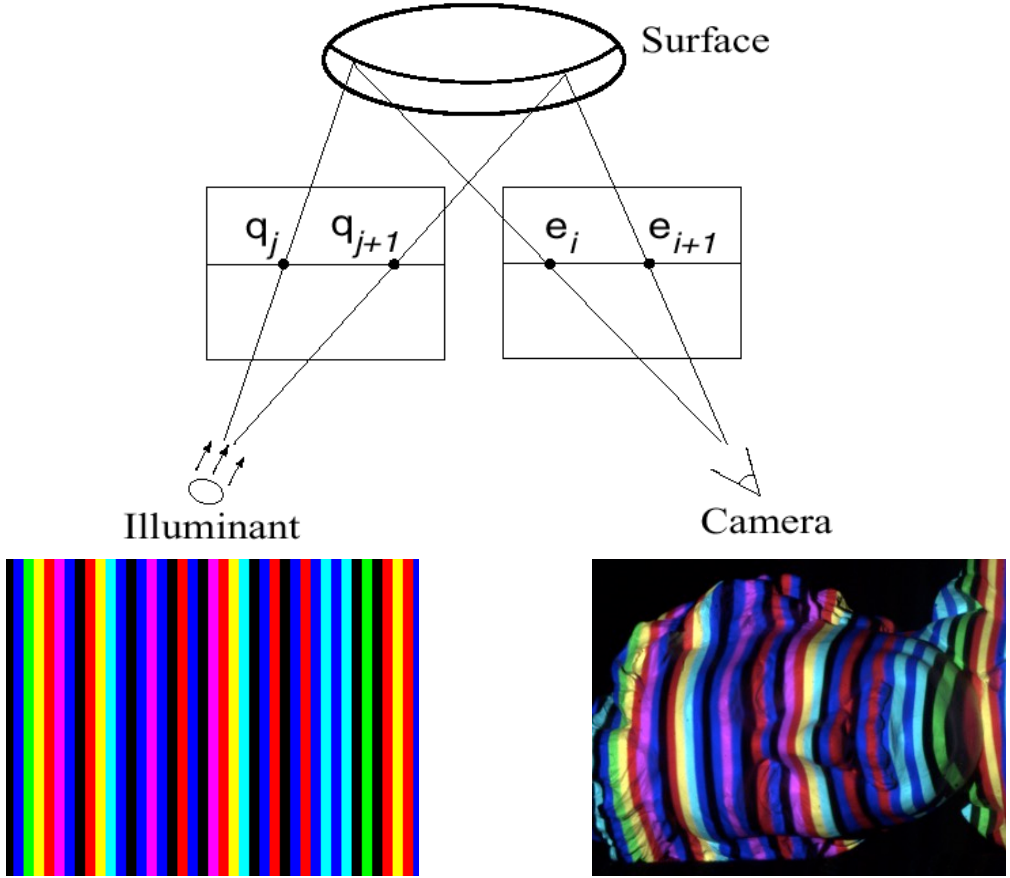
- Project “structured” light patterns onto the object
  - Simplifies the correspondence problem
  - Allows us to use only one camera



L. Zhang, B. Curless, and S. M. Seitz. Rapid Shape Acquisition Using Color Structured Light and Multi-pass Dynamic Programming. 3DPVT 2002

Slide credit: Lazebnik

Active Stereo with Structured Light

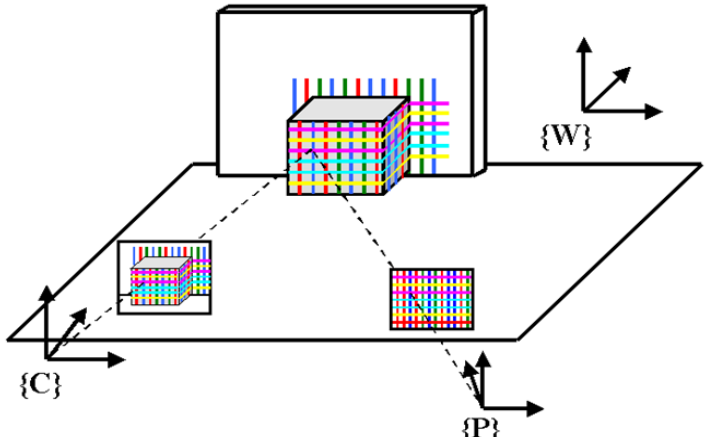


L. Zhang, B. Curless, and S. M. Seitz. Rapid Shape Acquisition Using Color Structured Light and Multi-pass Dynamic Programming. 3DPVT 2002

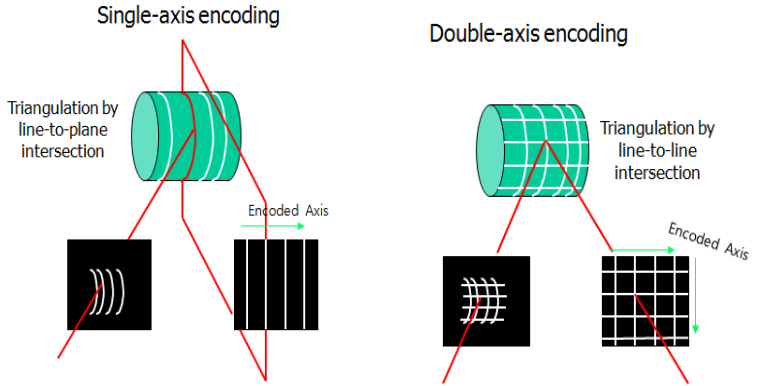
Slide credit: Lazebnik

Active Stereo Vision

- Structured-light-based stereo vision
  - Replacing one camera (in two-view stereo) with a projector
  - Perform stereo matching using a known structured pattern and a captured image

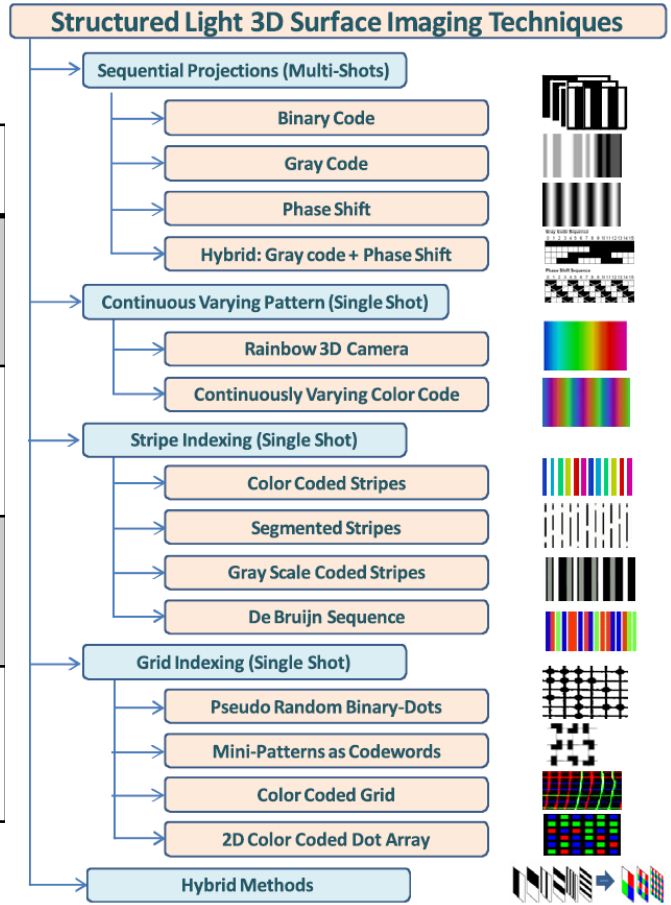


- Structured-light design
  - Consisting of dots, lines, and plai
  - Designed to resolve the inherent ambiguity (homogeneous area, r patterns)



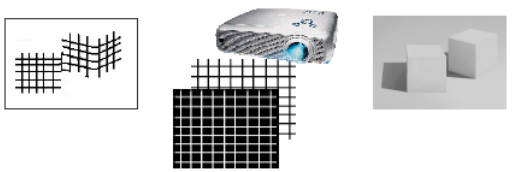
Categories of Structured-light Techniques

Types of techniques		Advantage	Weakness
Discrete patterns	Spatial multiplexing	• A unique pattern is required • Can measure moving objects	• Lower resolution than time multiplexing method • Occlusions problem
	Time multiplexing	• Highest resolution • High accuracy • Easy implementation	• Large number of patterns • Only motionless objects
Continuous patterns	Phase shifting	• High resolution • High accuracy • Few patterns are required	• Very sensitive to image noise • Only motionless objects
	Frequency multiplexing	• Very few patterns are required • Better accuracy than spatial multiplexing	• Frequent shape object problem • Sensitive to image noise

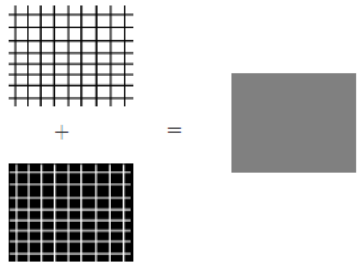


Invisible Structured-light

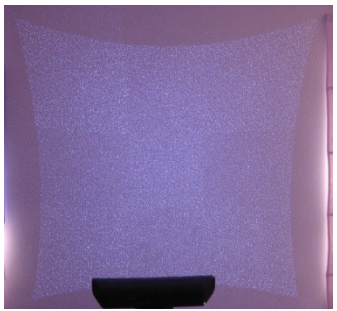
- Using structured-light patterns designed not to be perceived by human
  - Imperceptible Structured-light (ISL)
    - Projecting two reverse patterns alternatively with very high frequency
    - A camera is synchronized with only one pattern projection
    - Because of the response delay of human eye, human perceives not the projected patterns but the averaged gray images.
  - InfraRed Structured-light (IRSL)
    - Projecting structured patterns in IR spectrum range
    - Capturing the pattern-projected image with an IR camera



Principle of the ISL sensor



Resulting and projected patterns



IR pattern of KINECT

KINECT

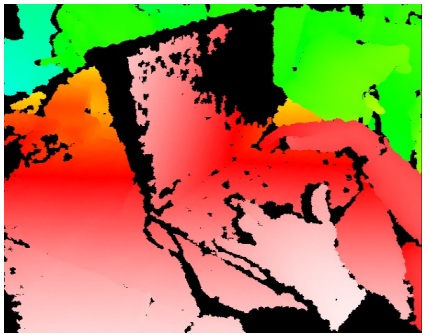
- One success example of stereo vision

- Specification
  - Sensors
    - Color camera, IR camera, IR projector
  - Field of View (FoV)
    - Horizontal FoV: 57 degrees
    - Vertical FoV: 43 degrees
    - Sensing range: 1.2m~3.5m
  - Data Streams
    - 320 x 240 16-bits depth @ 30 frames/sec
    - 640 x 480 32-bits color @ 30 frames/sec

How KINECT Works

- Projecting a binary pattern image via an IR projector
- Capturing a pattern-projected image with an IR camera
- Performing stereo matching using a known pattern image and a captured image and computing 3D

Captured image

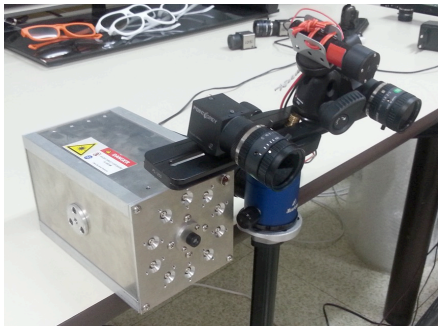


Depth map

  - Pros
    - high accuracy
    - real-time operation
    - relatively high resolution
    - cheap

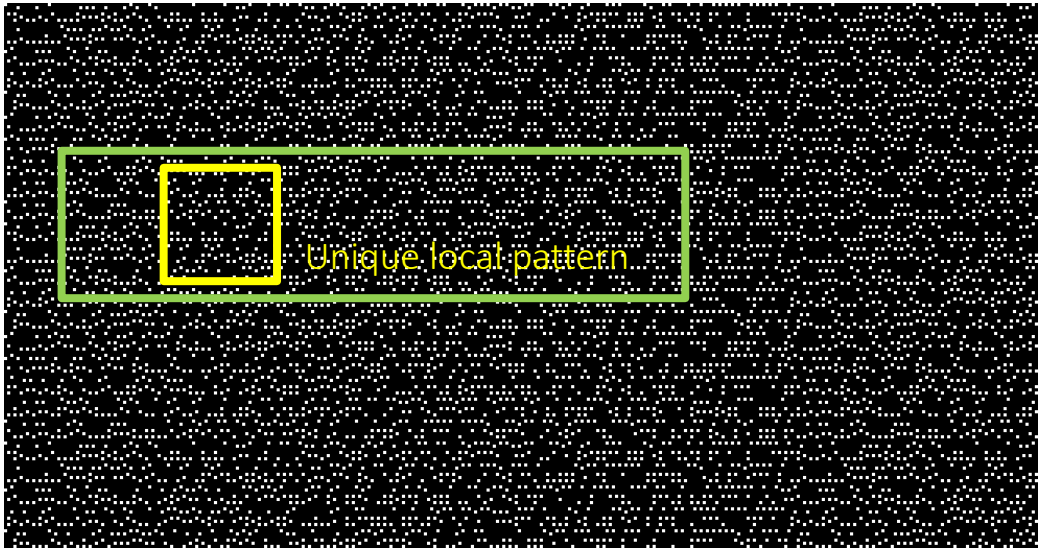
Building KINECT by Yourself

- Possible configuration
  - Two camera + one projector
    - Perform stereo matching with two pattern-projected images
    - Less dependent on the pattern design
  - One camera + one projector
    - Perform stereo matching with a known pattern image and a pattern-projected image
    - More dependent on the pattern design
- Common process
  - Calibration
  - Pattern design
  - Stereo matching



Pattern Design

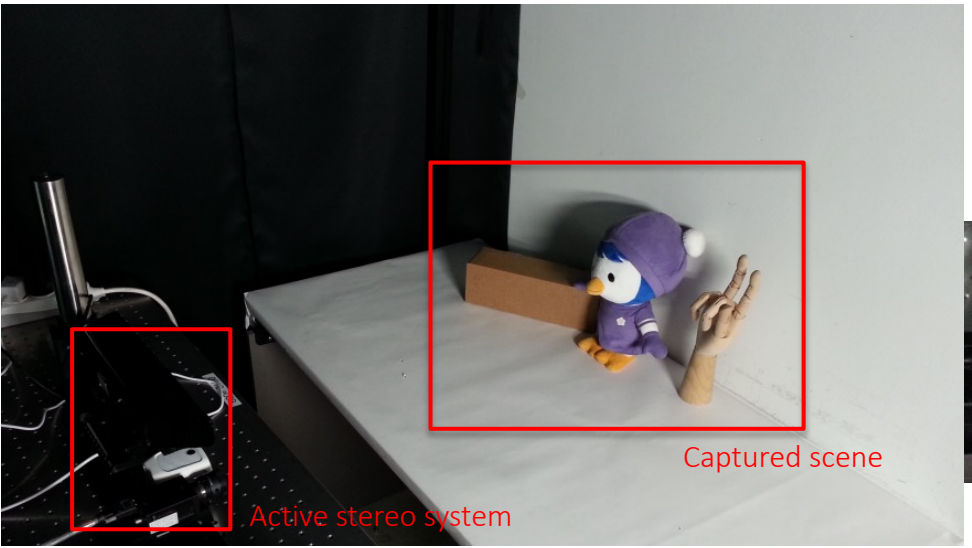
- Requirement of patterns
  - Binary (robust to image noise)
  - Small number of white dots (power consumption)
  - Uniqueness
    - Any local binary pattern should be uniquely appeared in the local neighborhood.



Reducing the inherent ambiguity of stereo matching due to lack of texture or repetitive textures.

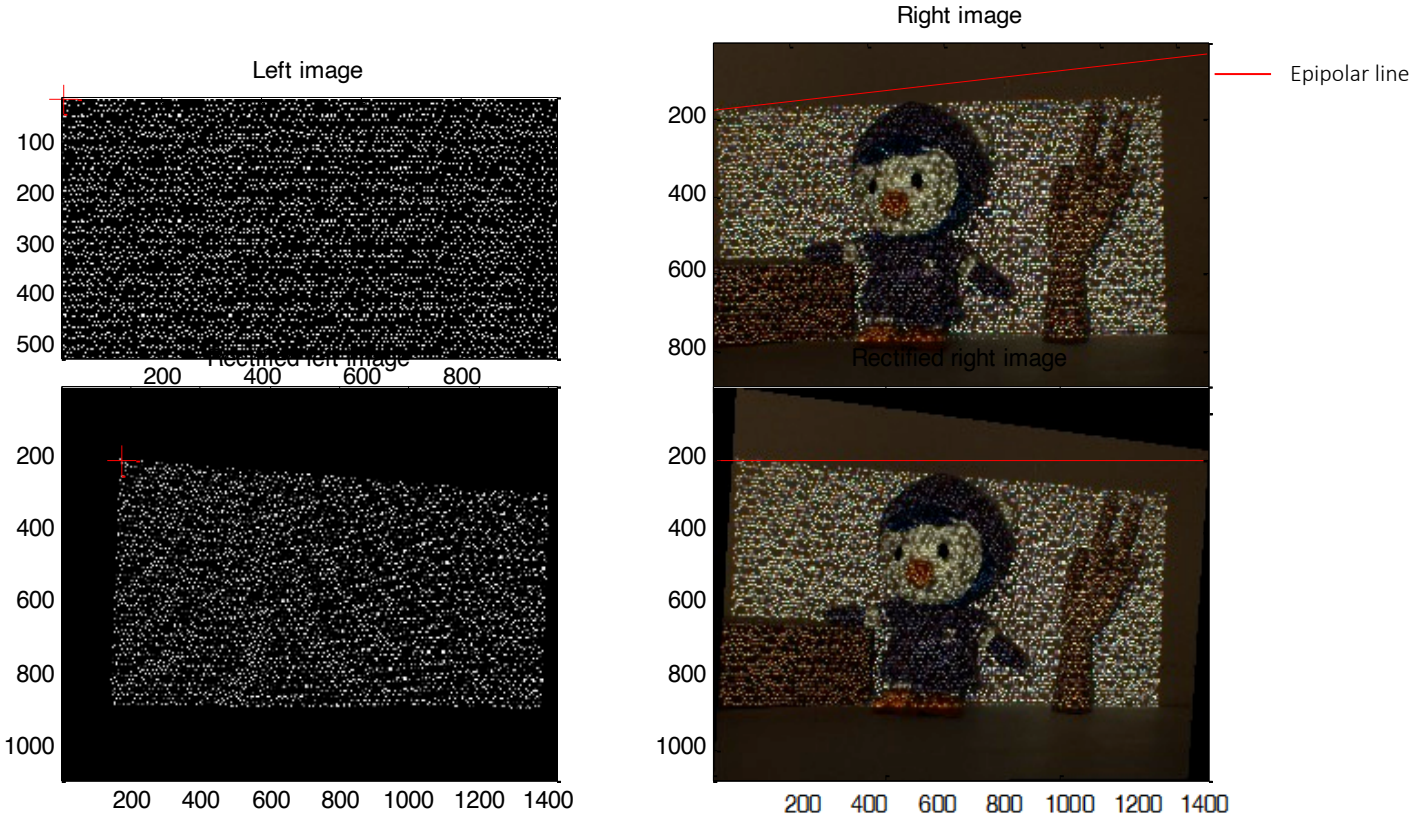
→ Make the stereo matching much easier!

Image Acquisition



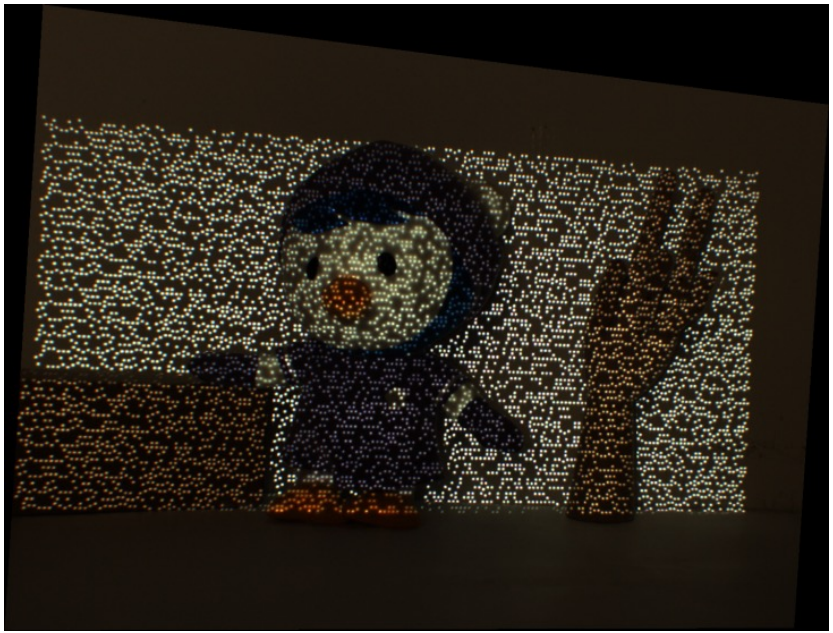
Rectification

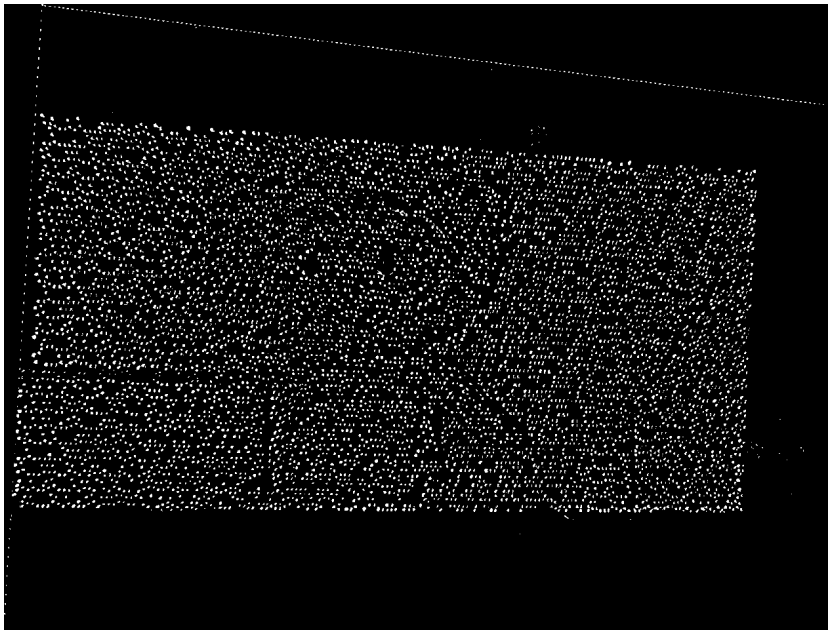
- Rectifying a known pattern image and a captured image so that the epipolar lines coincide with scanlines in an image.



Binarization

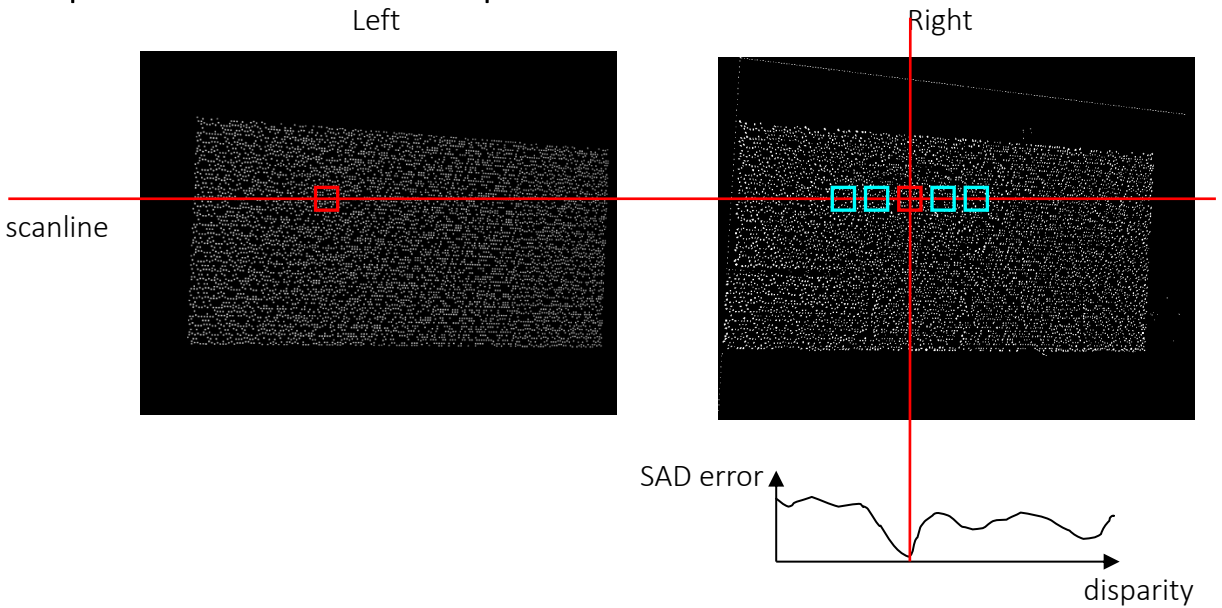
- Extracting structured-light patterns from a captured image using an adaptive thresholding technique



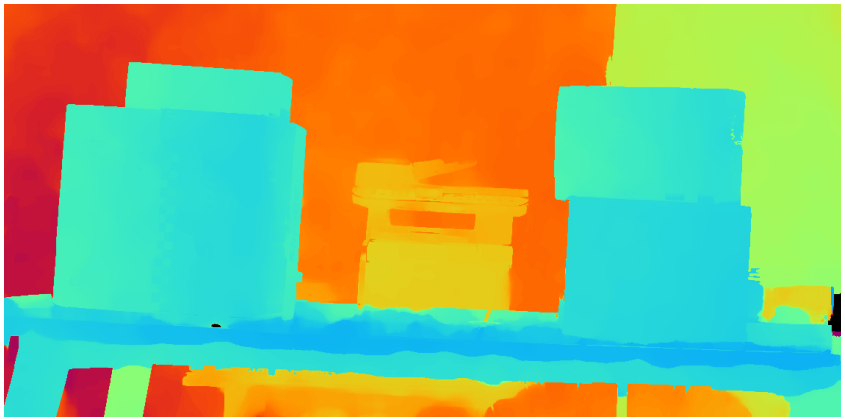


Stereo Matching

- Performing stereo matching with any stereo matching methods
  - Since the projected pattern greatly reduces the matching ambiguity, any simple method works quite well.



Depth Estimation Result



computed depth map (with post-processing)

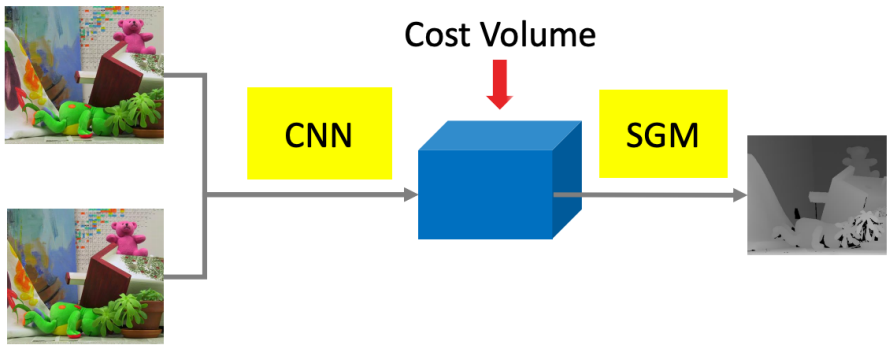
Depth map from KINECT

KAIST

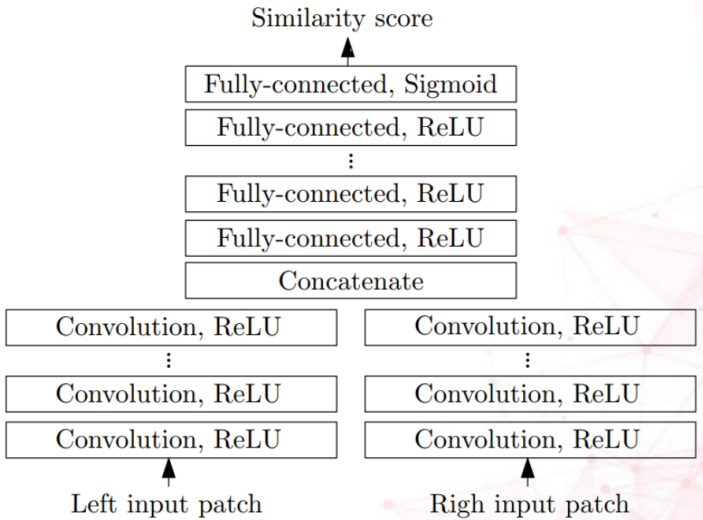
Stereo Matching with Deep Learning (I)

Matching cost learning

- Train a model to classify patches into two classes (similar and not similar)
- A small set of image pairs with ground-truth disparities generate millions of patches
- Depend on the performance of SGM



Overall pipeline



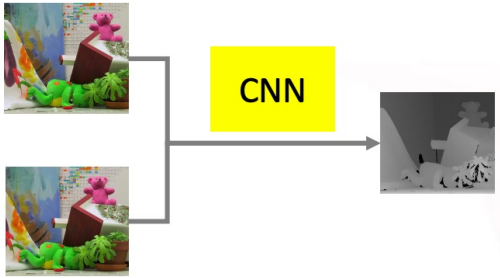
CNN Architecture of [3]

[3] J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 2016.

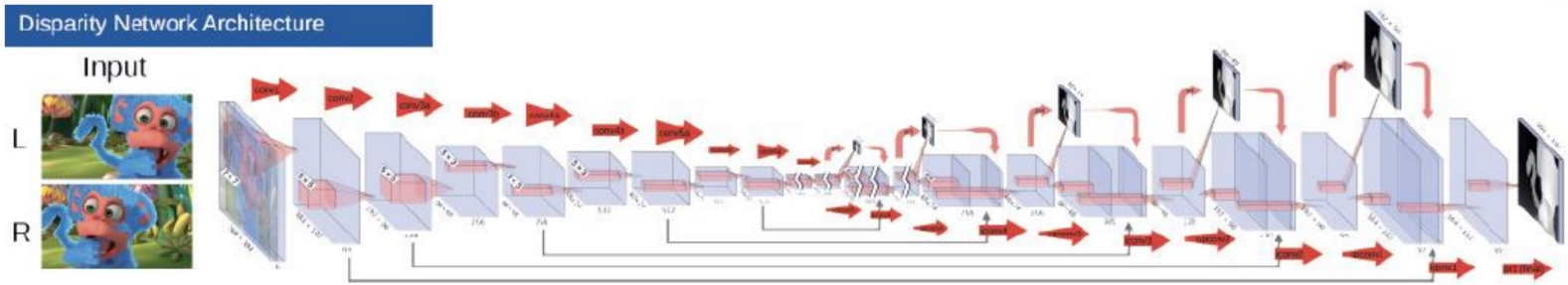
Stereo Matching with Deep Learning (II)

End-to-end Learning

- Train an end-to-end model to regress disparity
- A large-scale dataset is needed to train a good model
- Usually via hourglass structure
- Variations:
  - Correlation layer to compute cost volume, e.g., DispNetC [4]
  - Unsupervised learning with left-right check, e.g., [5]



Overall pipeline



[4] N. Mayer, et al,. "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation." In *Proc. IEEE CVPR*, 2016.

[5] C. Zhou, H. Zhang, X. Shen, and J. Jia. "Unsupervised Learning of Stereo Matching." In *Proc. IEEE CVPR*, 2017.