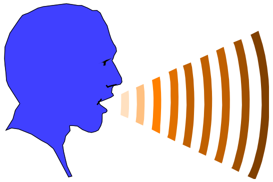


Image and Video Processing

Stereo and Multiview Video Processing

Yao Wang
Tandon School of Engineering, New York University

Outline

- Depth perception
- Depth from disparity
- Disparity estimation
- Intermediate view synthesis
- Stereo and multiview video compression
- Depth sensors
- Stereo and multiview video display
- Other emerging cameras

Perception of Depth

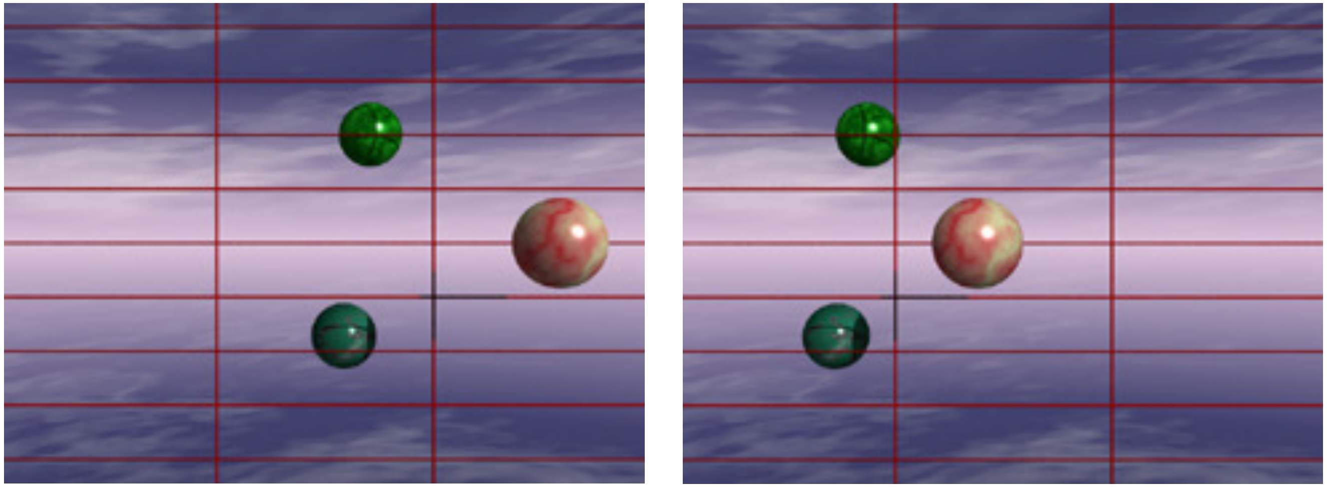
- Monocular cues:
  - Shape/size
  - Occlusion (one object blocks another)
  - Shading and texture
  - Linear perspective (think railroad tracks)
  - Relative height (with respect to the horizon)
  - Motion parallax
  - Aerial haze (blueness on the horizon)
- Motion cues
  - motion parallax
- Binocular cue: Stereopsis
  - The use of two images (or their disparity) to form a sense of depth

From Amy Reibman

Depth Perception by Stereopsis

- Human eye perceives depth by having two eyes with slightly shifted views
  - The shift is called “disparity”
  - Perceived depth depends on the “disparity”
  - Such depth perception is called “stereopsis”…

A Visual Experiment



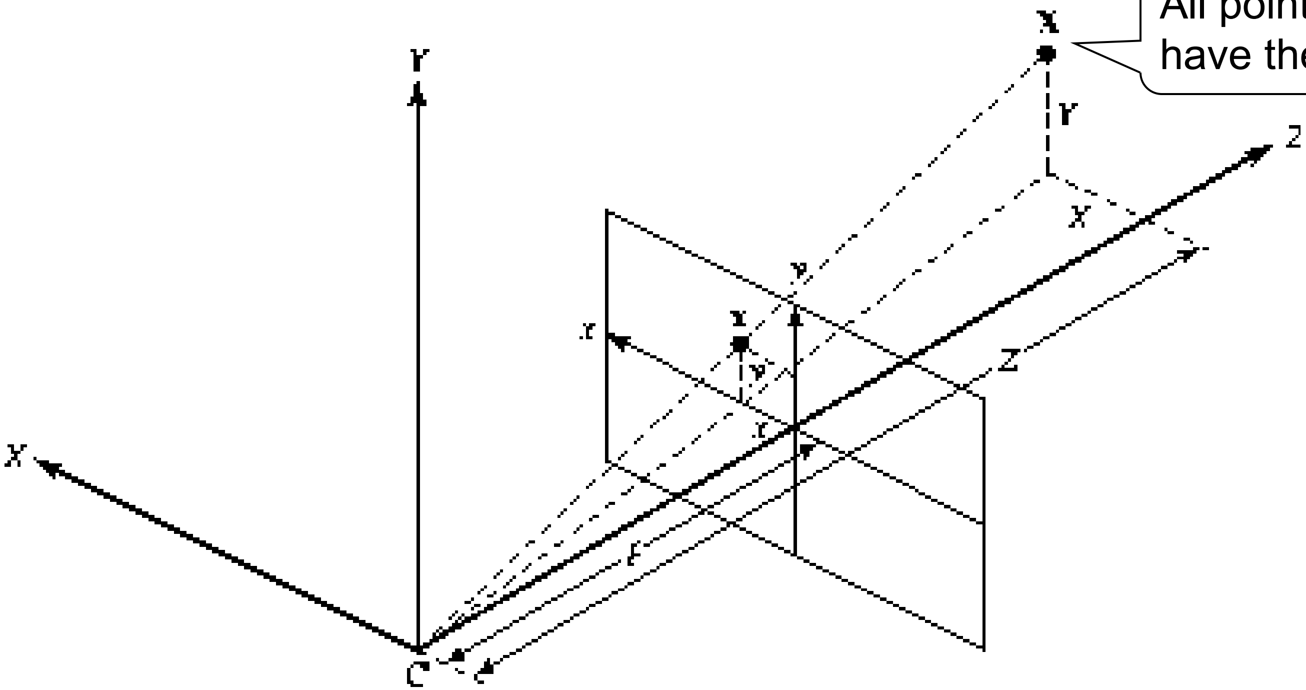
Try to look at the left and right images using your left and right eyes separately while try to merge the two images into one. Can you tell which ball is closer?

Pictures generated by ray-tracing. Courtesy of Rushang Wang

How do we deduce depth from stereo views?

- Depth from disparity

Perspective Projection Revisited



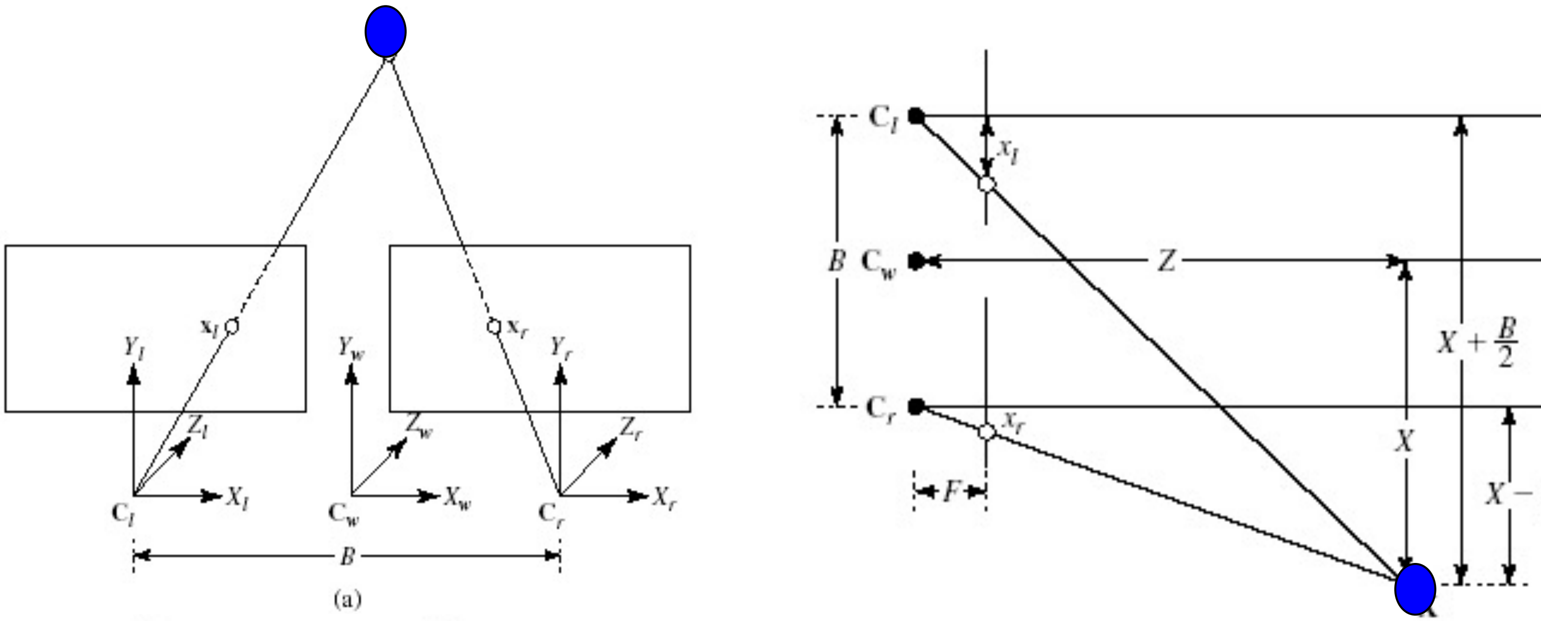
All points in this ray will have the same image

(a)

$$\frac{x}{F}=\frac{X}{Z},\frac{y}{F}=\frac{Y}{Z}\Rightarrow x=F\frac{X}{Z},y=F\frac{Y}{Z}$$

x, y are inversely related to Z

Parallel Camera Configuration: Relation between Depth and Disparity



(a)

(b)

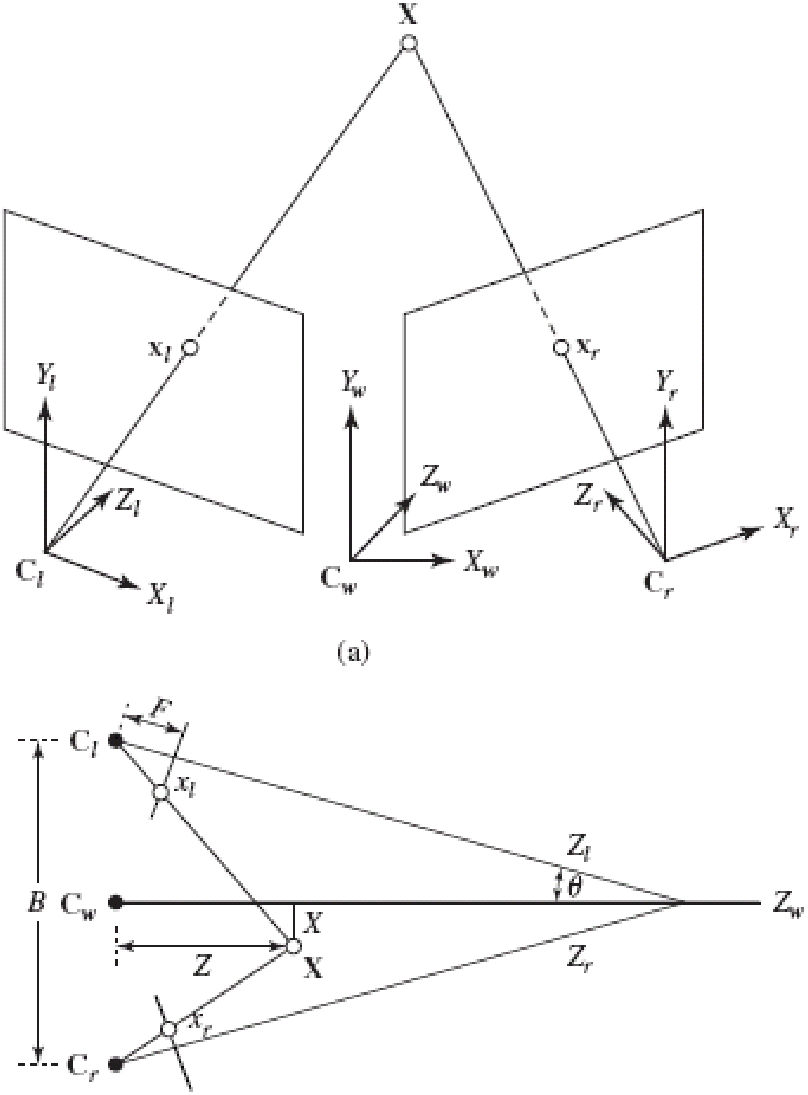
$$X_l = X + \frac{B}{2}, \quad X_r = X - \frac{B}{2}, \quad Y_l = Y_r = Y, \quad Z_l = Z_r = Z;$$

$$x_l = F\frac{X + B/2}{Z}, \quad x_r = F\frac{X - B/2}{Z}, \quad y_l = y_r = y = F\frac{Y}{Z}.$$

$$\boxed{d_x = x_l - x_r = \frac{FB}{Z}.}$$

i) Only horizontal disparity
ii) Disparity is inversely proportional to Z
iii) Range of disparity increases with B

Convergent Camera Configuration



(a)

(b)

$$x_l = F\frac{\cos\theta(X+B/2)-\sin\theta Z}{\sin\theta(X+B/2)+\cos\theta Z},$$

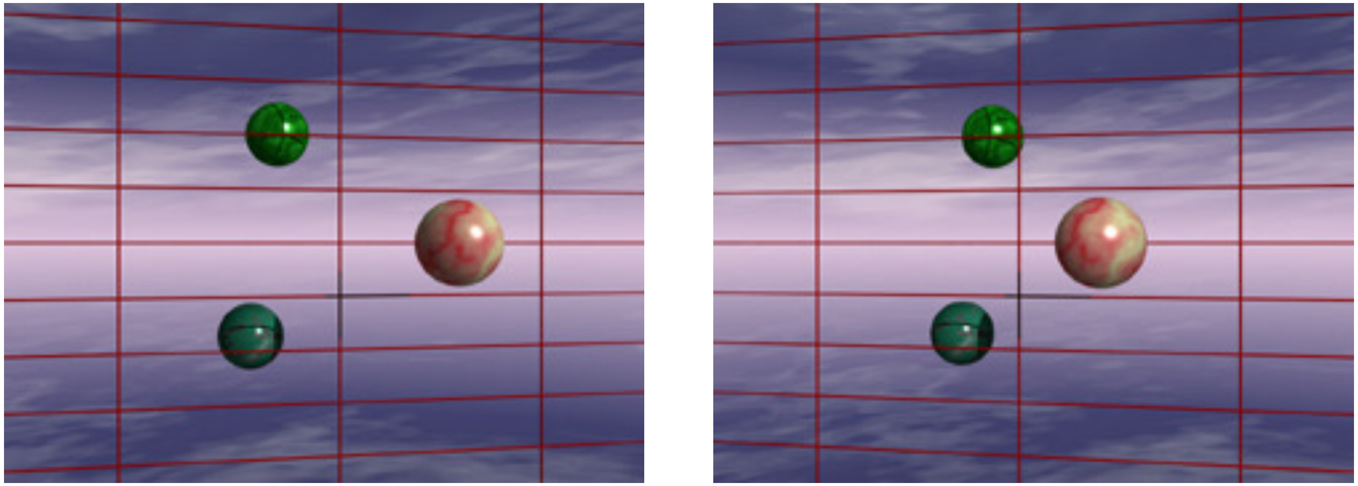
$$x_r = F\frac{\cos\theta(X-B/2)+\sin\theta Z}{-\sin\theta(X-B/2)+\cos\theta Z},$$

$$y_l = F\frac{Y}{\sin\theta(X+B/2)+\cos\theta Z},$$

$$y_r = F\frac{Y}{-\sin\theta(X-B/2)+\cos\theta Z}.$$

both horizontal and vertical disparity

Example Images (Converging Camera)



Notice the keystone effect

Can get a better depth perception of objects closer to the camera than with the parallel set up

But when displayed using a parallel projection system and viewed by the human eye, the vertical disparity causes perceptual discomfort. Geometric correction is needed before displaying these images.

Epipolar Geometry: Arbitrary Case (advanced material, not required)

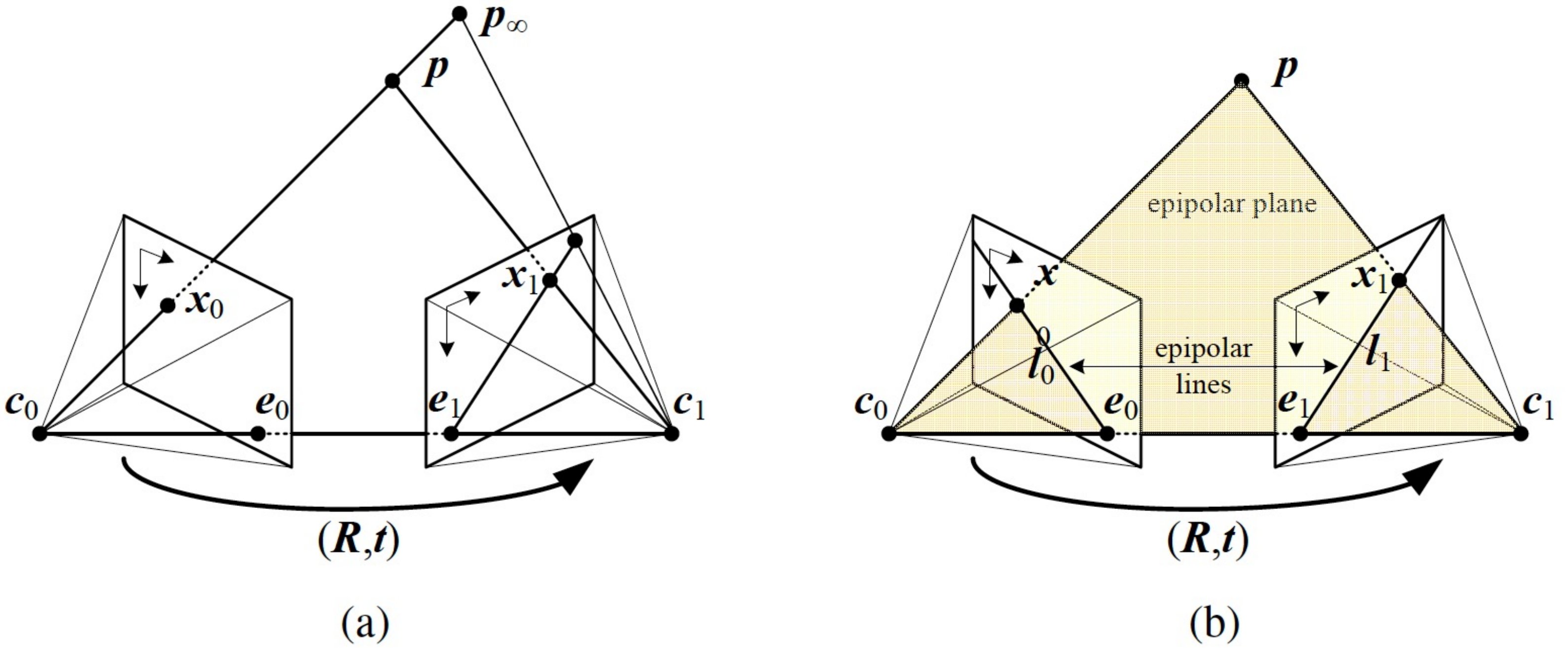


Figure 11.3 Epipolar geometry: (a) epipolar line segment corresponding to one ray; (b) corresponding set of epipolar lines and their epipolar plane.

From [Szeliski2012]

Epipolar Geometry: Parallel Case (not required)

- Epipolar constraint: the corresponding left and right image points should be on the same horizontal line (only horizontal disparity exists)

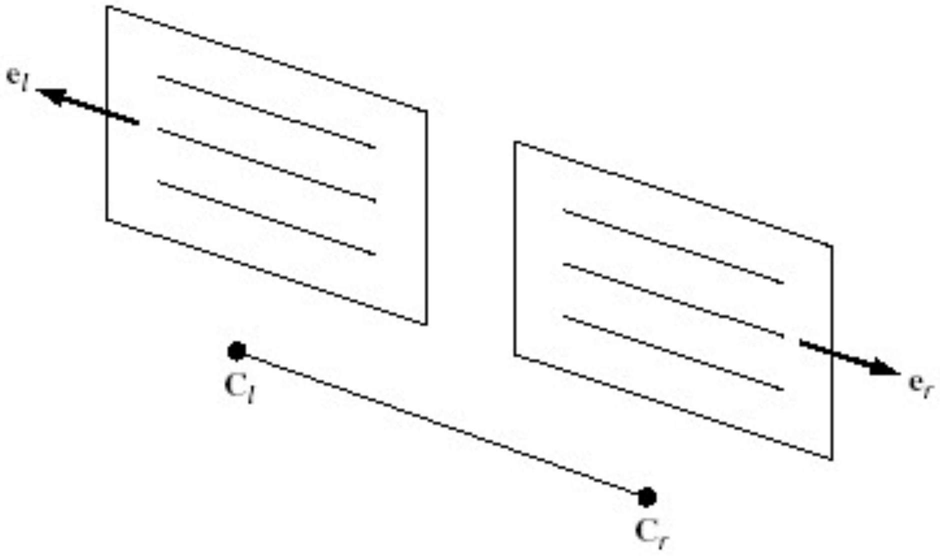
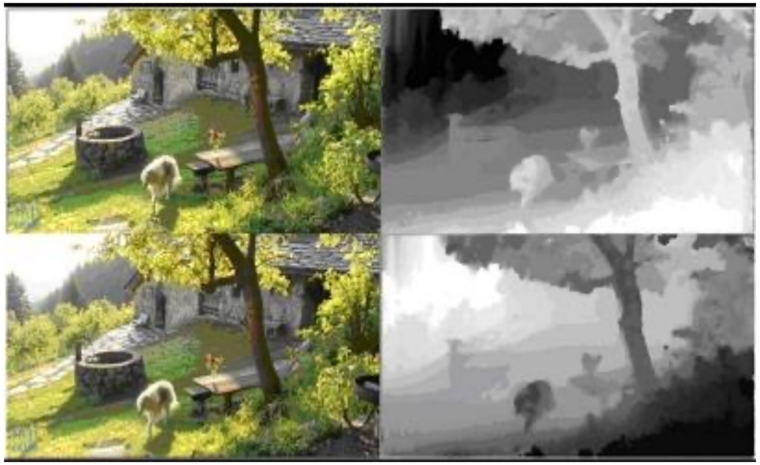


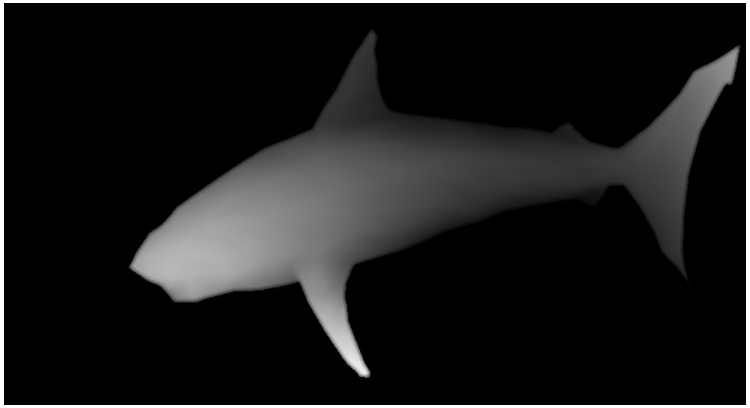
Figure 12.8 Epipolar geometry for a parallel camera configuration: epipoles are at infinity, and epipolar lines are parallel. Adapted from O. Faugeras, *Three-Dimensional Computer Vision—A Geometric Viewpoint*, Cambridge, MA: MIT Press, 1993. Copyright 1993 MIT Press.

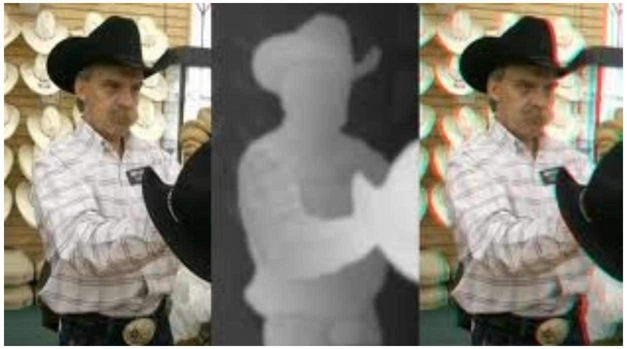
Rectification: creation of images as if acquired from parallel cameras

Examples of Disparity and Depth Maps

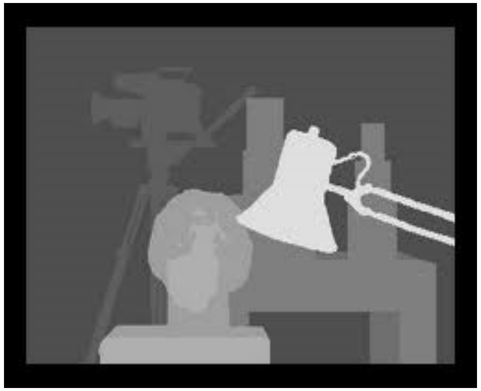
- Brighter objects closer, dark objects further away

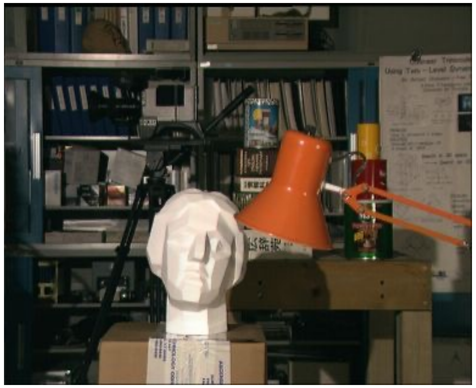


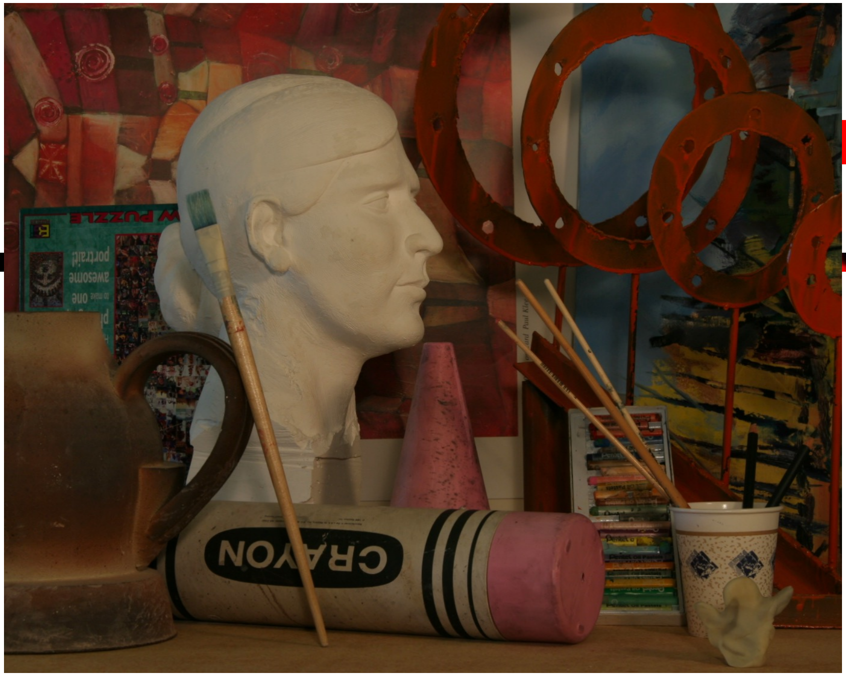


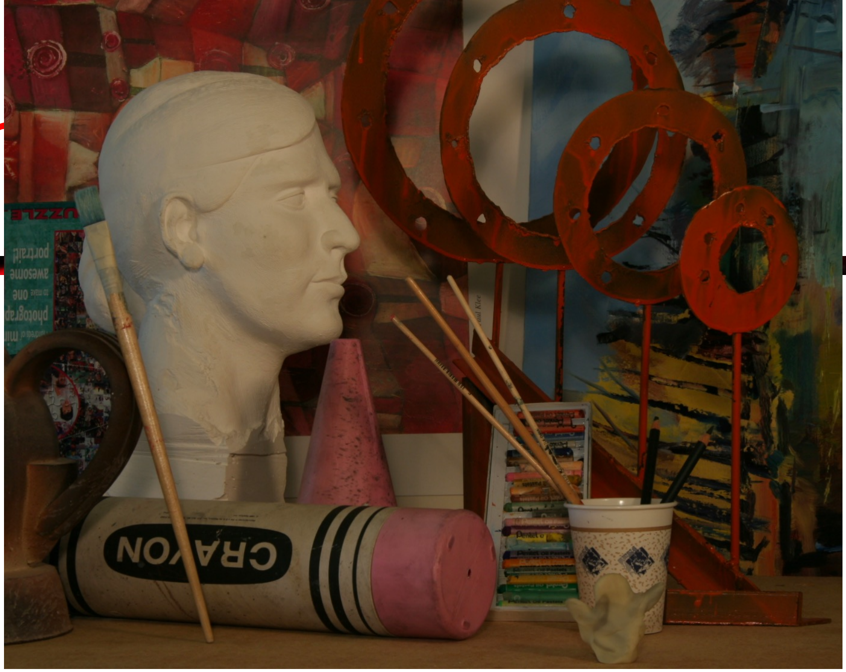


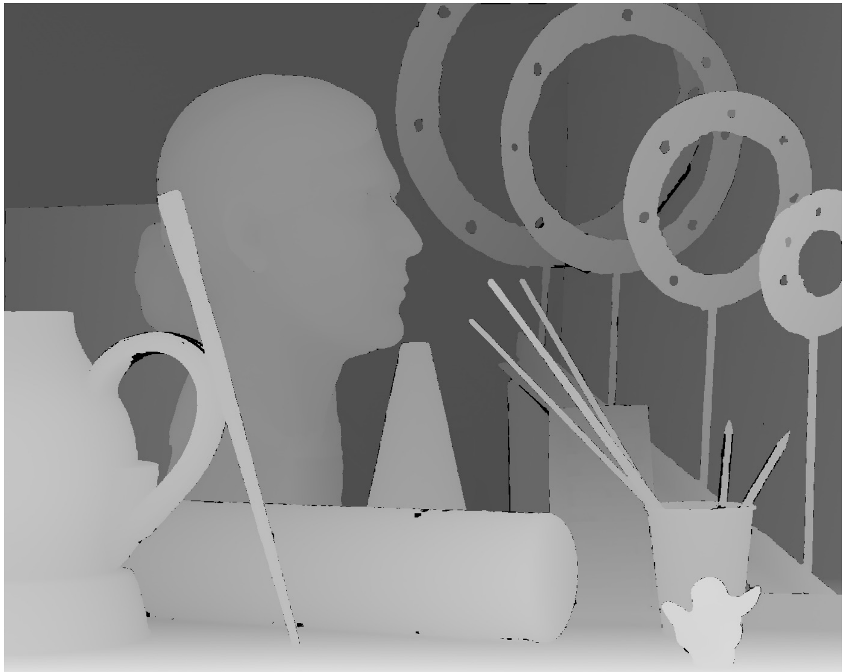
Courtesy of Jill Boyce, 2011











Ground truth disparity map
anchored in the left view

http://vision.middlebury.edu/stereo/data/scenes2005/

Disparity Estimation

- Depth at every pixel can be determined from disparity at that pixel!
- Disparity Estimation Problem:
  - For each point in one (anchor) image, find its corresponding point in the other (target) image – Similar to motion estimation problem
  - Parallel configuration: only horizontal disparity needs to be estimated.
  - Difficulty: disparity range may be very large for objects nearby (up to 40-50 pixels); occlusion areas are prevalent (area appearing only in one view); depth discontinuity are prevalent (must apply smoothness criterion judiciously!); intensity matching does not work in flat surface area.

- Constraints for disparity estimation
- Dense disparity estimation
- Block-based disparity estimation
- Mesh-based disparity estimation
- Line-by-line estimation using dynamic programing
- 3D structure estimation

Constraints for Disparity Estimation

- Epipolar constraints:
  - For parallel set up: two corresponding points are in the same line, only horizontal disparity need to be searched
  - For an arbitrary camera set up: given x_r, possible x_l sits on a line (epipolar line)
- Ordering constraint (for points at the same depth):
  - If two points in the right image are such that $x_{r,1} < x_{r,2}$
  - Then the corresponding two points in the left image satisfy $x_{l,1} < x_{l,2}$
- Models for disparity functions

Models for Disparity Functions

- Affine model for plane surface, parallel set-up:
  - If an imaged object has a plane surface

$$Z(X, Y) = aX + bY + c.$$

then the disparity function for points on the surface satisfies affine model:

$$d(x_l, y) = x_r - x_l = \frac{F - a x_l - b y}{\frac{a}{2} - \frac{c}{B}}$$

(Proof: HW!)

- For an arbitrary scene, we can divide the reference (right) image into small blocks so that the object surface corresponding to each block is approximately flat. Then the disparity function over each block can be modeled as affine.
- Using similar approach, can derive models for curved surfaces (higher order polynomial)

Dense Disparity Estimation

- Very similar to dense motion estimation:
  - Also makes use of constant intensity assumption
  - For one pixel in one image, find its corresponding pixel in the other image so that the intensity/color at that pixel (or its surrounding pixels) matches
  - Optical flow equation still applies if the expected disparity is small!
    - Temporal gradient -> difference between left and right images
- Three methods discussed for dense motion estimation still apply
  - Lucas-Kanade method: based on optical flow equation
    - Only valid for small disparity or far away scene
  - Block matching
  - Regularization: minimizing matching error+disparity smoothness
  - Only needs to solve for horizontal disparity for parallel set up!

Block-Based Disparity Estimation

- Following the method for block-based motion estimation
  - Divide the anchor image (e.g. right image) into regular blocks
  - Assume disparity function over each block is constant or an affine function
  - Determine the constant or the affine parameters to minimize the matching error between corresponding pixels
  - For parallel set up: Only1 horizontal disparity or 3 affine parameters
- Difference from motion estimation
  - Constant disparity model is less effective than constant motion model even over small blocks, only true for a flat surface parallel to the camera plane
  - Affine model is quite good
  - Need a large search range to account for large disparities for objects nearby
  - Occlusion is more prevalent and need to be handled effectively

Imposing constraints between estimated disparity in adjacent blocks/pixels

- Independent estimation at each block/pixel may lead to conflicting estimates
- Smoothness constraint: similar to motion estimation, may add a penalty term to the cost function to discourage significant difference between disparity of nearby pixels

$$E(d) = E_{data}(d) + \lambda E_{smooth}(d)$$

$$E_{data}(d) = \sum C(x, y, d(x, y)),$$

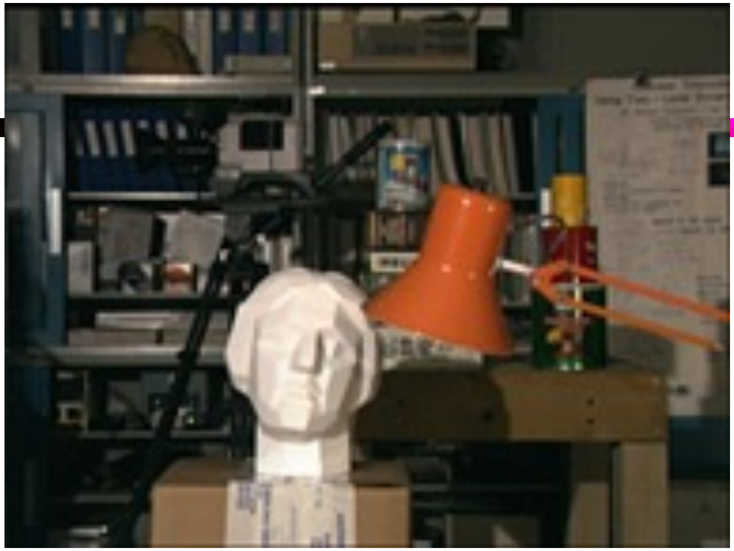
$$c(x, y, d) = \varphi(I_l(x + d, y) - I_r(x, y))$$

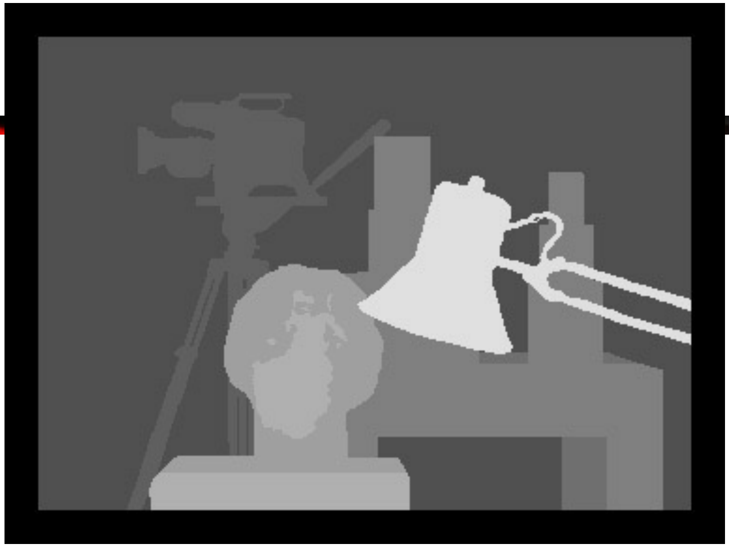
$$E_{smooth}(d) = \sum_{(x,y)} \rho(d(x, y) - d(x+1, y)) + \rho(d(x, y) - d(x, y+1)),$$

To relax smoothness constraint near image edges (where depth discontinuity is more likely):

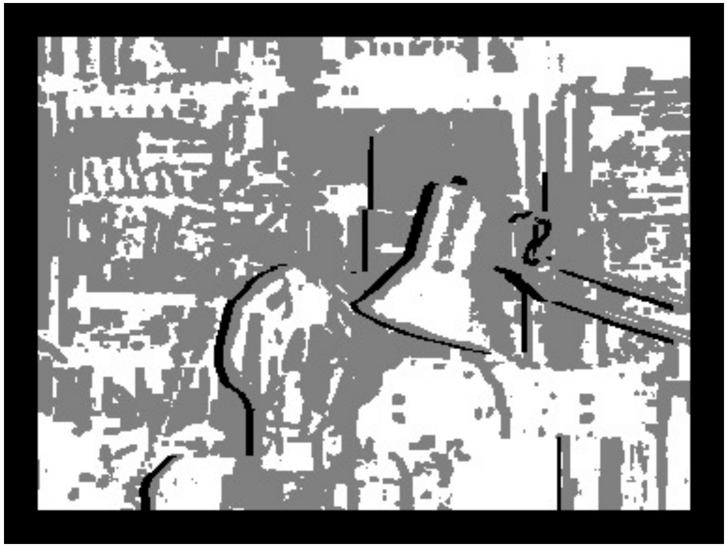
$$\rho_d(d(x, y) - d(x+1, y)) \cdot \rho_I(\|I(x, y) - I(x+1, y)\|)$$

Challenges of disparity estimation

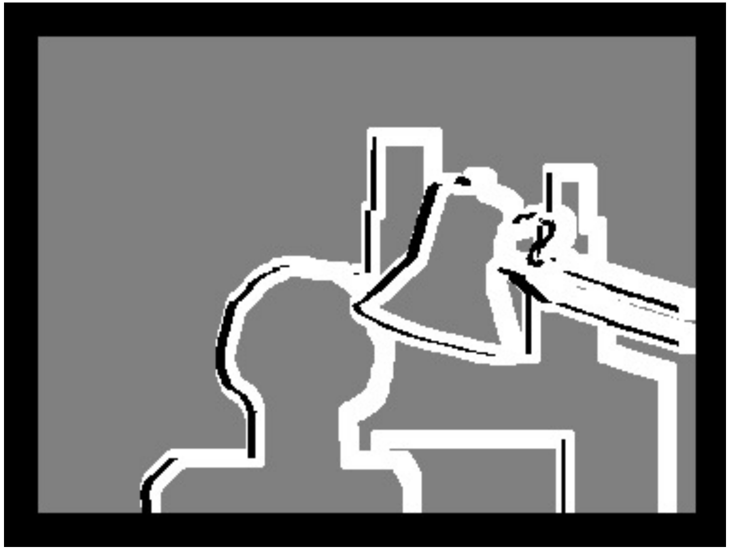




Ground truth depth image



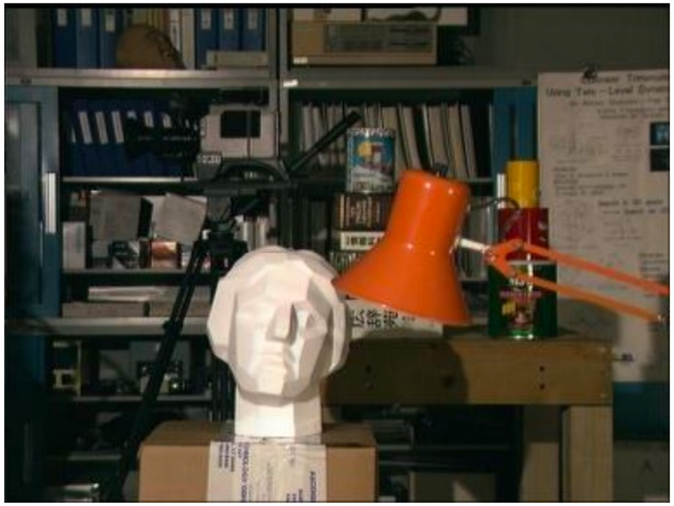
White: flat texture
Black: occluded region in another view



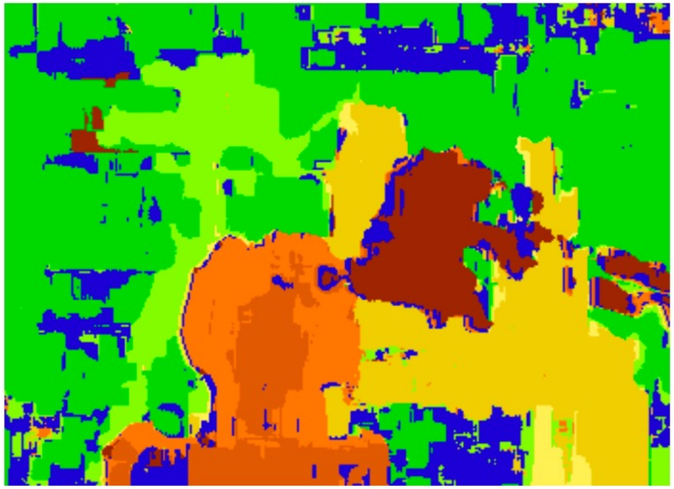
White: depth discontinuity region
Black: occluded region in another view

Sample Results

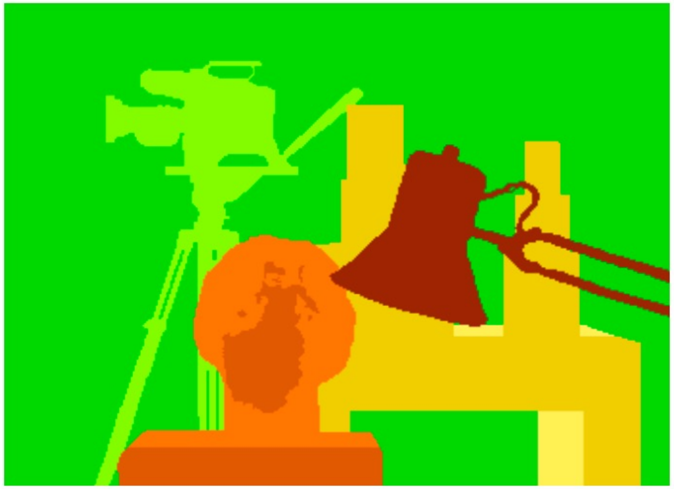
Data



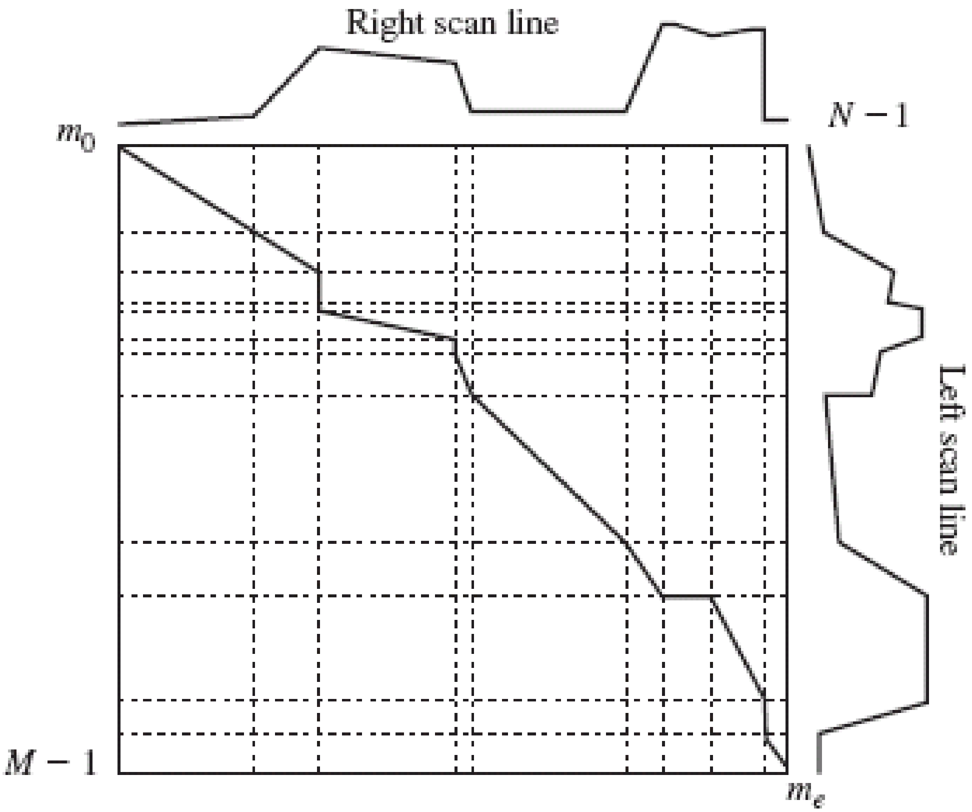
Window-based matching



Ground truth



Intra-line edge matching using dynamic programing (Optional)

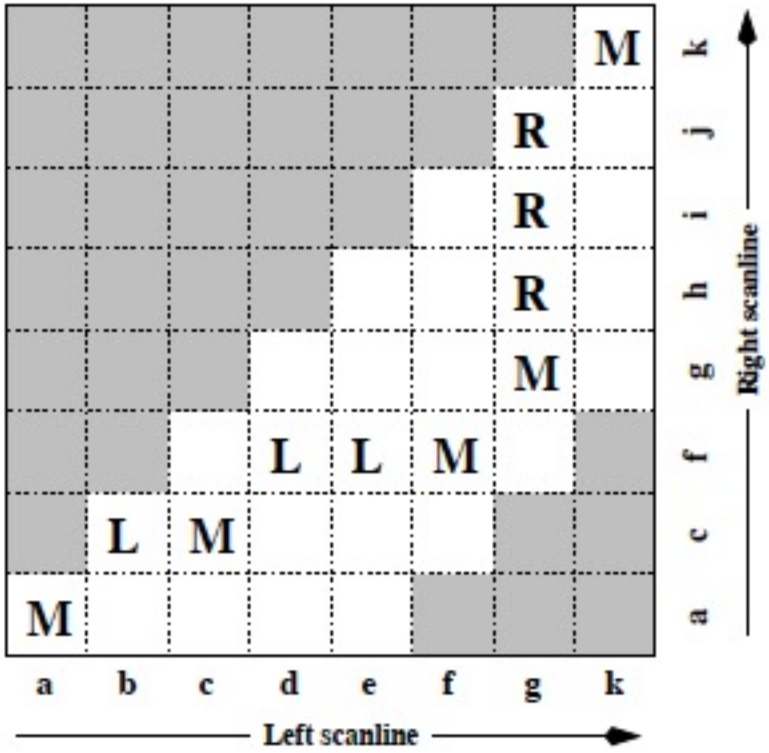


Each point in the graph corresponds to one pair of left edge and right edge, with a corresponding matching cost. The problem is to find a path that has the minimal total cost. Because of the ordering constraint, the path cannot go backwards. The minimal cost path can be determined using dynamic programming.

Figure 12.11 Path-finding analogy of the stereo matching problem. Adapted from O. Faugeras, *Three-Dimensional Computer Vision—A Geometric Viewpoint*, Cambridge, MA: MIT Press, 1993. Copyright 1993 MIT Press.

Y. Ohta and T. Kanade. Stereo by intra- and interscanline search using dynamic programming. IEEE TPAMI, 7(2):139–154, 1985.

Considering Occlusion in the Dynamic Programming Approach (optional)



	a	b	c	d	e	f	g	k
k	▨	▨	▨	▨	▨	▨	▨	M
j	▨	▨	▨	▨	▨	▨	R	
i	▨	▨	▨	▨	▨		R	
h	▨	▨	▨	▨			R	
g	▨	▨	▨				M	
f	▨	▨		L	L	M		▨
c	▨	L	M				▨	▨
a	M					▨	▨	▨

M: matching
L: appearing only in left
R: appearing only in right

From [Scharstein02]

Datasets for Disparity Estimation

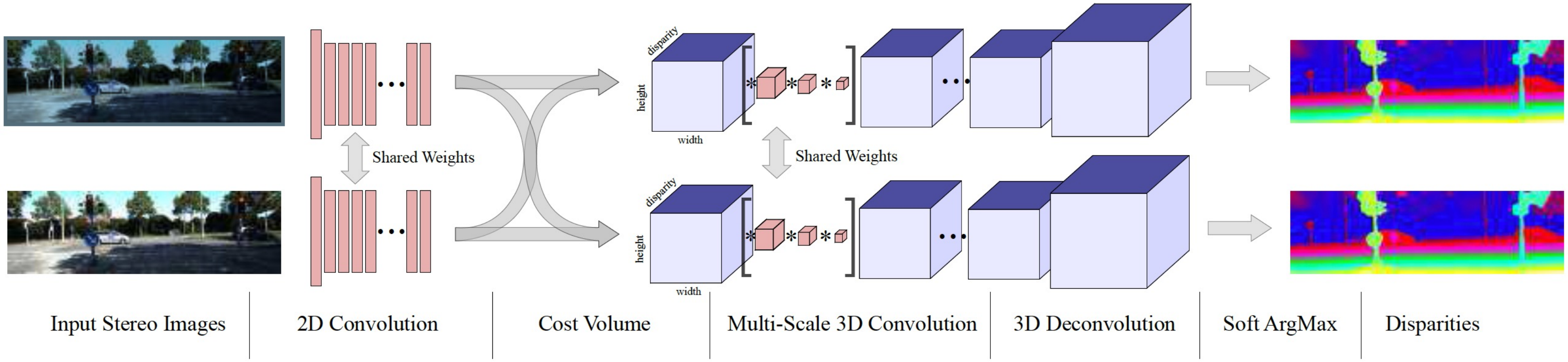
- Middlebury: http://vision.middlebury.edu/stereo/ (2002)
  - A public domain database containing sample pairs of stereo images with ground truth disparity
  - Open to submission of algorithms and estimation results
  - The site evaluates the accuracy against ground truth
- KITTY dataset: http://www.cvlibs.net/datasets/kitti/ (2012, 2015)
- Scene Flow dataset (synthetic) (2015)

https://lmb.informatik.uni-freiburg.de/resources/datasets/SceneFlowDatasets.en.html

Deep-Learning Based Disparity Estimation

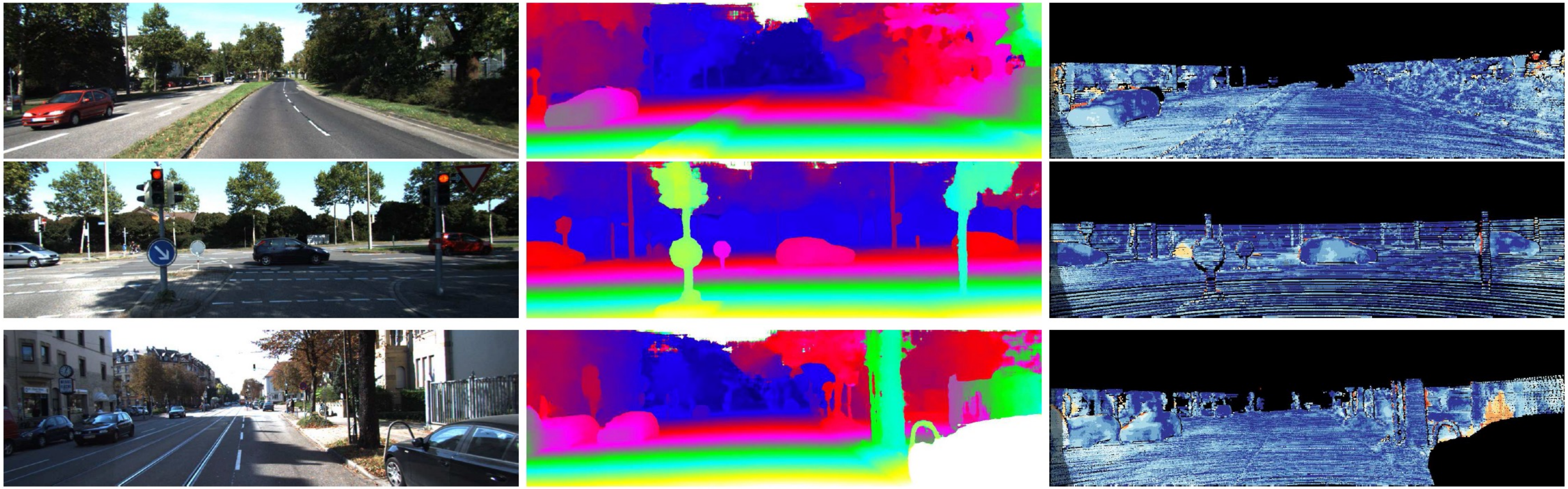
- J. Zbontar and Y. Le Cun. Computing the stereo matching cost with a convolutional neural network. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
- W. Luo, A. G. Schwing, and R. Urtasun. Efficient Deep Learning for Stereo Matching. CVPR, 2016
- Z. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang. A deep visual correspondence embedding model for stereo matching costs. CVPR 2016.
- N. Mayer, E. Ilg, P. H¨ausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. CoRR, abs/1510.0(2002), 2015.
- Kendall, Alex, et al. "End-to-end learning of geometry and context for deep stereo regression." *CVPR*. 2017. (GC-net)
- Liang, Zhengfa, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. "Learning for disparity estimation through feature constancy." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2811-2820. 2018.
- For evaluation of various methods on KITTI dataset:
  - http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo

GC-Net

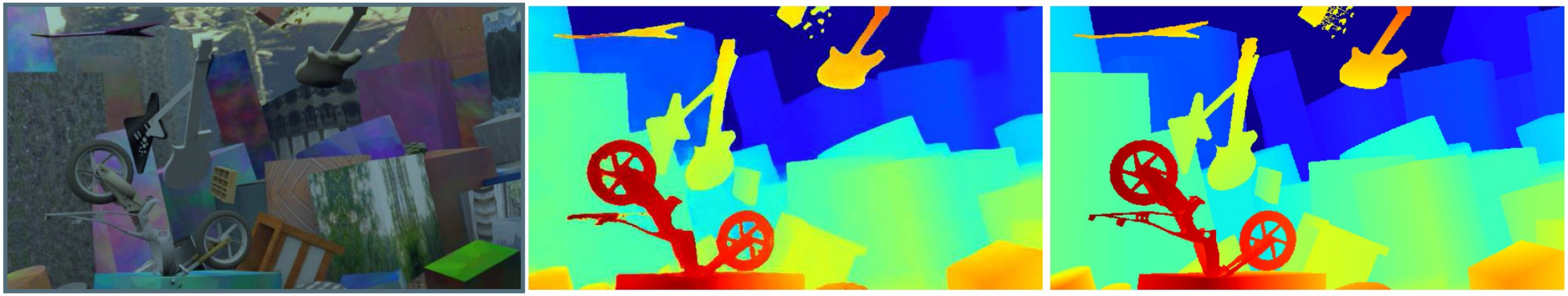


From: Kendall, Alex, et al. "End-to-end learning of geometry and context for deep stereo regression." *CVPR*. 2017.

> We form our model by developing differentiable layers representing each major component in traditional stereo pipelines. This allows us to learn the entire model end-to-end while leveraging our geometric knowledge of the stereo problem.



(b) KITTI 2015 test data qualitative results. From left: left stereo input image, disparity prediction, error map.



(c) Scene Flow test set qualitative results. From left: left stereo input image, disparity prediction, ground truth.

From: Kendall, Alex, et al. "End-to-end learning of geometry and context for deep stereo regression." *CVPR*. 2017.

Depth and Structure From Disparity

- One can deduce depth and correspondingly 3D positions (structure) of a point from its disparity
  - Main application of stereo imaging
- Parallel case:

$$X_l = X + \frac{B}{2}, \quad X_r = X - \frac{B}{2}, \quad Y_l = Y_r = Y, \quad Z_l = Z_r = Z;$$

$$x_l = F\frac{X + B/2}{Z}, \quad x_r = F\frac{X - B/2}{Z}, \quad y_l = y_r = y = F\frac{Y}{Z}.$$

$$d_x = x_l - x_r = \frac{FB}{Z}.$$

→

$$\frac{Z}{F} = \frac{B}{d};$$

$$Y = y\frac{Z}{F}$$

$$X = \frac{x_l + x_r}{2}\frac{Z}{F}$$

Depth Estimation from a Single Image (Deep Learning Approaches, Supervised)

- Ming Y, Meng X, Fan C, et al. Deep learning for monocular depth estimation: A review. Neurocomputing, 2021, 438: 14-33.
- Ranftl, René, Alexey Bochkovskiy, and Vladlen Koltun. "Vision transformers for dense prediction." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

- Earlier papers:
- D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. NIPS, pages 1–9, 2014.
- F. Liu, C. Shen, G. Lin, and I. Reid. Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. Pattern Analysis and Machine Intelligence, page 15, 2015.
- R. Garg, V. Kumar BG, and I. Reid. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. ECCV, pages 1–16, 2016.
- Laina, Iro, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. "Deeper depth prediction with fully convolutional residual networks." In *2016 Fourth international conference on 3D vision (3DV)*, pp. 239-248. IEEE, 2016.
- Fu, Huan, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. "Deep ordinal regression network for monocular depth estimation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2002-2011. 2018.
- NYU Depth dataset https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html

Depth Estimation Dataset

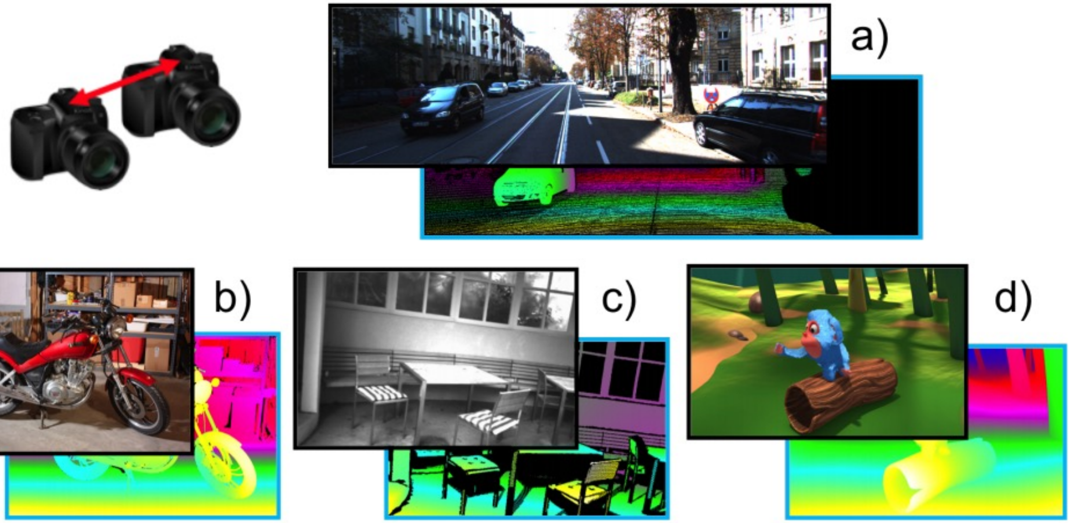


Fig. 2. **Overview of the most popular stereo datasets in literature**, with examples of reference images and associated ground truth disparity. a) KITTI 2015 [9], b) Middlebury 2014 [8], c) ETH3D [10], d) Freiburg SceneFlow [11].

Poggi, M., Tosi, F., Batsos, K., Mordohai, P., and Mattoccia, S. (2020). On the synergies between machine learning and stereo: a survey. arXiv preprint arXiv: 2004.08566.

Auto Encoder for Depth Prediction

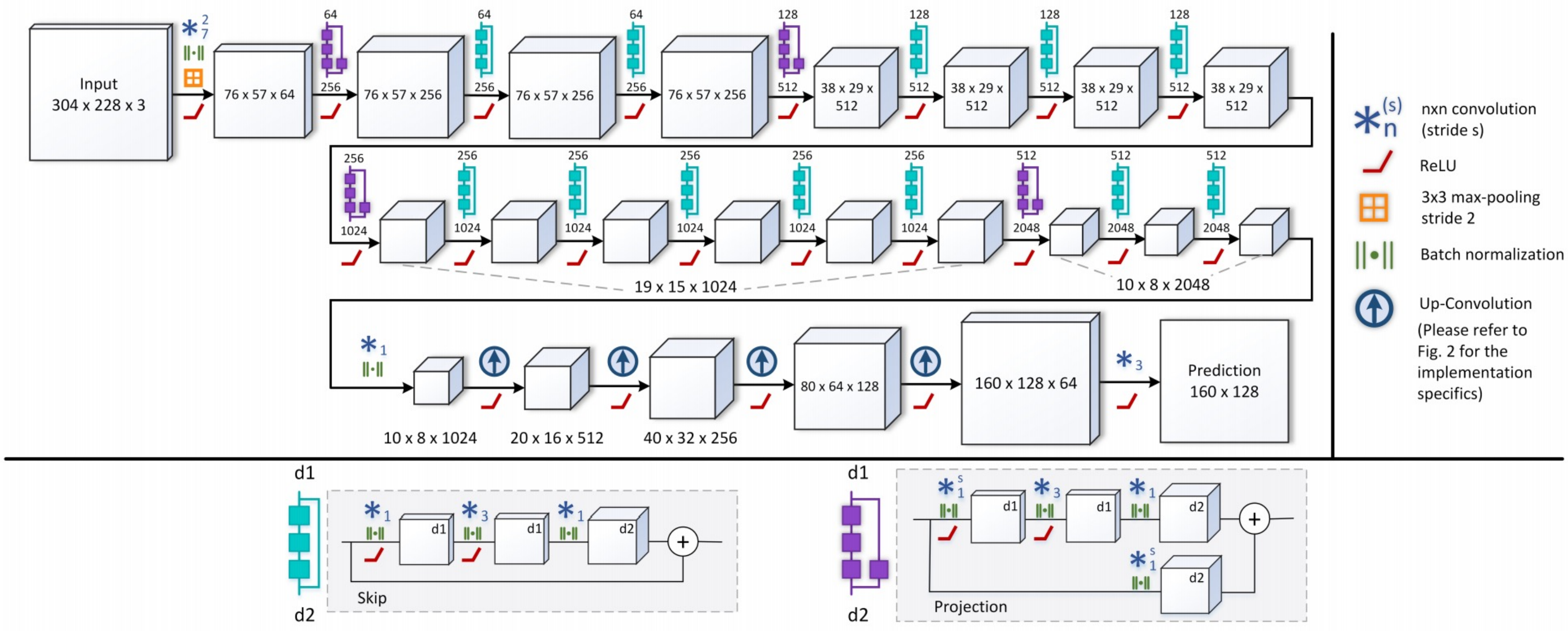


Figure 1. **Network architecture.** The proposed architecture builds upon ResNet-50. We replace the fully-connected layer, which was part of the original architecture, with our novel up-sampling blocks, yielding an output of roughly half the input resolution

Laina, Iro, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. "Deeper depth prediction with fully convolutional residual networks." In *2016 Fourth international conference on 3D vision (3DV)*, pp. 239-248. IEEE, 2016.

Results by using different encoders

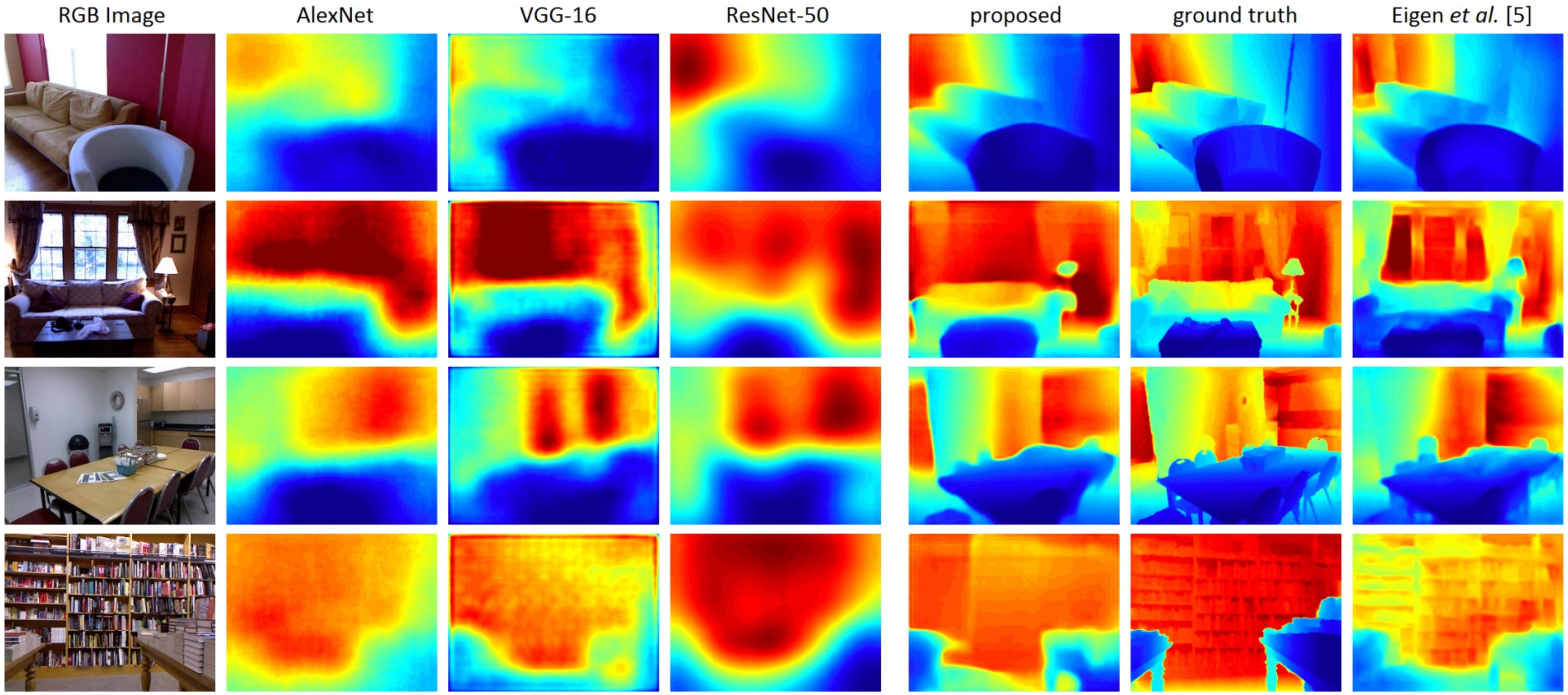


Figure 4. **Depth Prediction on NYU Depth** Qualitative results showing predictions using AlexNet, VGG, and the fully-connected ResNet compared to our model and the predictions of [5]. All colormaps are scaled equally for better comparison

Learning Depth Estimation Without Ground Truth (Self-Supervision)

- Large training sets with ground truth depth information is hard to obtain!
- Wouldn't it be nice if we can just train the network with stereo images or video?
- Self-supervision!
  - Given a stereo pair
    - Left image - > Depth -> generate right image
    - Generated right image should be similar to the actual right image
  - Given a video sequence
    - Take say 3 consecutive images: $f_{t-1,}$ $f_{t,}$ f_{t+1}
    - Middle image f_t -> Depth image D_t
    - Pose estimation to estimate the camera transformation from to left f_t to f_{t-1}, and from f_t to f_{t+1}
    - Pose+depth -> generate the left and right images, respectively
    - Generated image should be similar to the actual f_{t-1} and f_{t+1}

Poggi, M., Tosi, F., Batsos, K., Mordohai, P., and Mattoccia, S. (2020). On the synergies between machine learning and stereo: a survey. arXiv preprint arXiv: 2004.08566.
Laga H, Jospin L V, Boussaid F, et al. A survey on deep learning techniques for stereo-based depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

MonoDepth 2

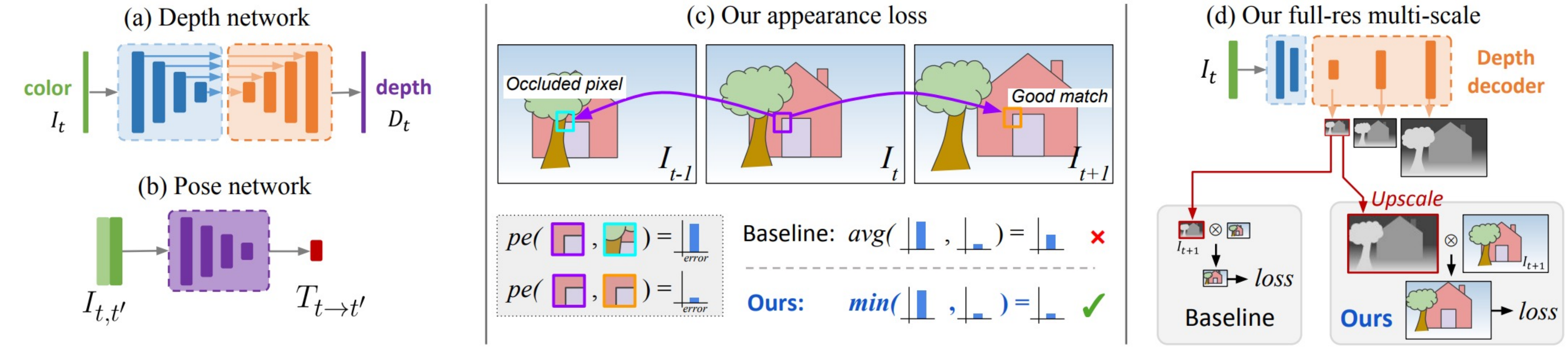


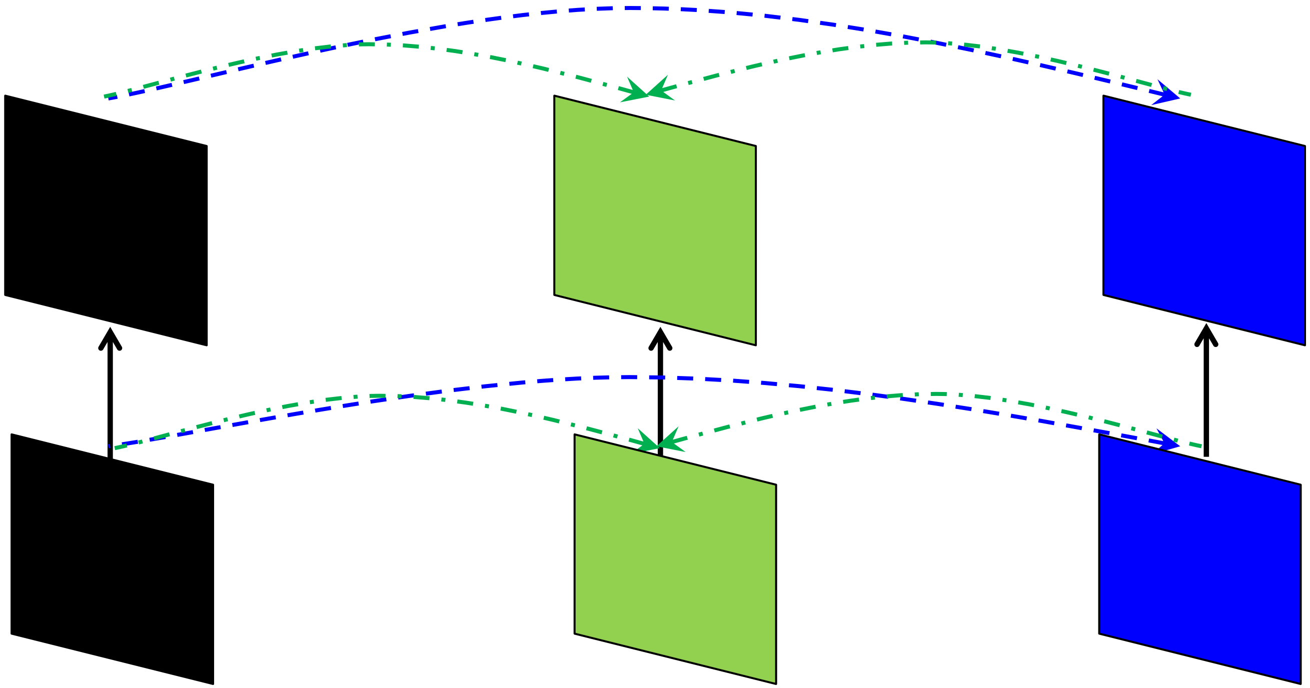
Figure 3. **Overview. (a) Depth network:** We use a standard, fully convolutional, U-Net to predict depth. **(b) Pose network:** Pose between a pair of frames is predicted with a separate pose network. **(c) Per-pixel minimum reprojection:** When correspondences are *good*, the reprojection loss should 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.

Godard, Clément, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. "Digging into self-supervised monocular depth estimation." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3828-3838. 2019. (https://github.com/nianticlabs/monodepth2)

Compression of Stereo and Multiview Video

- Monocular video compression:
  - Key idea: predict current frame from the past frame
- Stereo video:
  - Code the left view as a monocular video
  - For the right view: in addition to predicting from the past frame in the right view, can also predict the corresponding left view
    - Choose the better prediction, or use weighted combination
- Multiview:
  - Code some views independently
  - Remaining views can use both temporal prediction and view interpolation.
- Considered in the video coding standards

Stereo Video Coding



Left view (coded reference to right view)

Right view (coded independently)

Intermediate View Synthesis

- Problem:
  - Interpolate intermediate views from given views
  - Necessary for virtual reality applications
  - Important for multi-view display systems

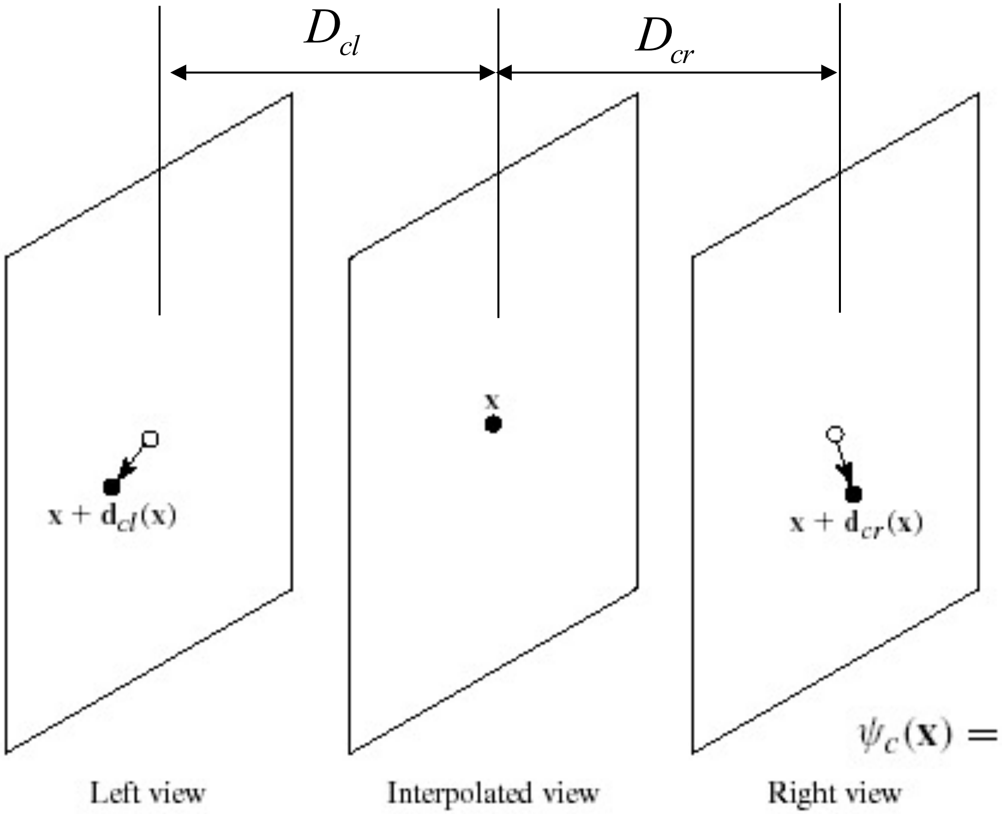
- Linear interpolation: can lead to blurred images

$$\psi_c(\mathbf{x}) = w_l(\mathbf{x})\psi_l(\mathbf{x}) + w_r(\mathbf{x})\psi_r(\mathbf{x}).$$

- Disparity-compensated interpolation

$$\psi_c(\mathbf{x}) = w_l(\mathbf{x})\psi_l(\mathbf{x} + \mathbf{d}_{cl}(\mathbf{x})) + w_r(\mathbf{x})\psi_r(\mathbf{x} + \mathbf{d}_{cr}(\mathbf{x})).$$

Disparity Compensated View Synthesis


D_{cl}
D_{cr}
$\mathbf{x}$
$\mathbf{x}+\mathbf{d}_{cl}(\mathbf{x})$
$\mathbf{x}+\mathbf{d}_{cr}(\mathbf{x})$
Left view
Interpolated view
Right view

Baseline distances (distance between camera centers): Dcl and Dcr

$$\psi_c(\mathbf{x}) = w_l(\mathbf{x})\psi_l(\mathbf{x}+\mathbf{d}_{cl}(\mathbf{x})) + w_r(\mathbf{x})\psi_r(\mathbf{x}+\mathbf{d}_{cr}(\mathbf{x})).$$

Figure 12.13 Disparity-compensated interpolation: $\mathbf{x}$ is interpolated from $\mathbf{x}+\mathbf{d}_{cl}(\mathbf{x})$ in the left view and $\mathbf{x}+\mathbf{d}_{cr}(\mathbf{x})$ in the right view.

$$w_l(\mathbf{x}) = \begin{cases} \frac{D_{cr}}{D_{cl}+D_{cr}}, & \text{if } \mathbf{x} \text{ is visible in both views,} \\ 1, & \text{if } \mathbf{x} \text{ is visible only in the left view,} \\ 0, & \text{if } \mathbf{x} \text{ is visible only in the right view.} \end{cases}$$

How to determine disparity from the central (unknown) view?

- One approach:
  - First determine disparity between left and right for every pixel in the left d_lr(x_l)
  - Then determine disparity between left and central based on distance, d_lc(x_l)=B_cl/(B_cl+B_cr) d_lr(x_l)
  - (B_cl and B_cr are camera center distances between center and left, and center and right cameras)
- For every point x_l in left, find corresponding point in central x_c=x_l+d_lc(x_l)
- But the central point may not be an integer pixel!
- Need to interpolate the integer pixel values from these non-integer pixels
- When using block-based method for estimating d_lr, there may be uncovered points in the central view or multiple-covered points; a dense depth field is better

Mesh-Based Interpolation of Disparity Field

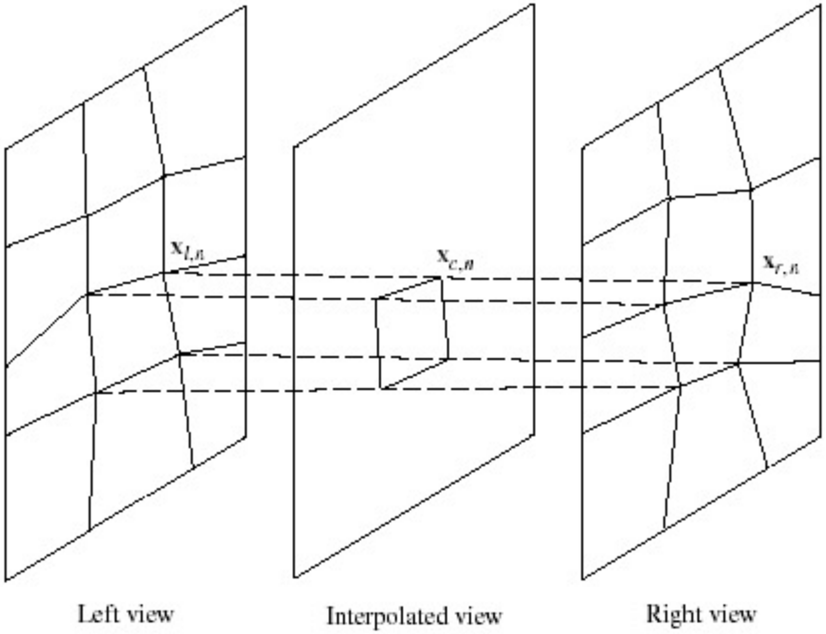


Figure 12.14 The mesh for the central view is generated by linearly interpolating the nodal positions in the left and right views.

$$\mathbf{x}_{c,n} = \frac{D_{cr}}{D_{cl} + D_{cr}} \mathbf{x}_{l,n} + \frac{D_{cl}}{D_{cl} + D_{cr}} \mathbf{x}_{r,n}.$$

No uncovered or multiple covered pixels in central view!

Mesh-Based Disparity Estimation

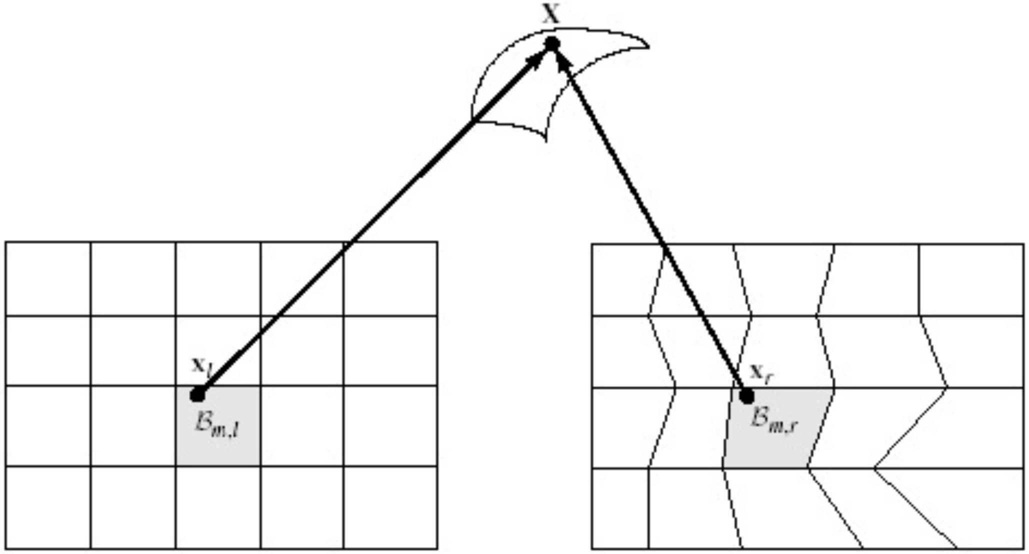


Figure 12.9 Correspondence between 3-D and 2-D meshes. Reprinted from R. Wang and Y. Wang, Multiview video sequence analysis, compression, and virtual viewpoint synthesis, *IEEE Trans. Circuits Syst. for Video Technology* (April 2000), 10(3):397–410. Copyright 2000 IEEE.

- Estimate the disparity at each node (corner) by minimizing DCP error over 4 blocks attached on this node.
- The disparity within each block modeled by a affine or bilinear function
- Can use a non-regular mesh in the anchor frame to mach with object boundary

Mesh-Based View Synthesis Result

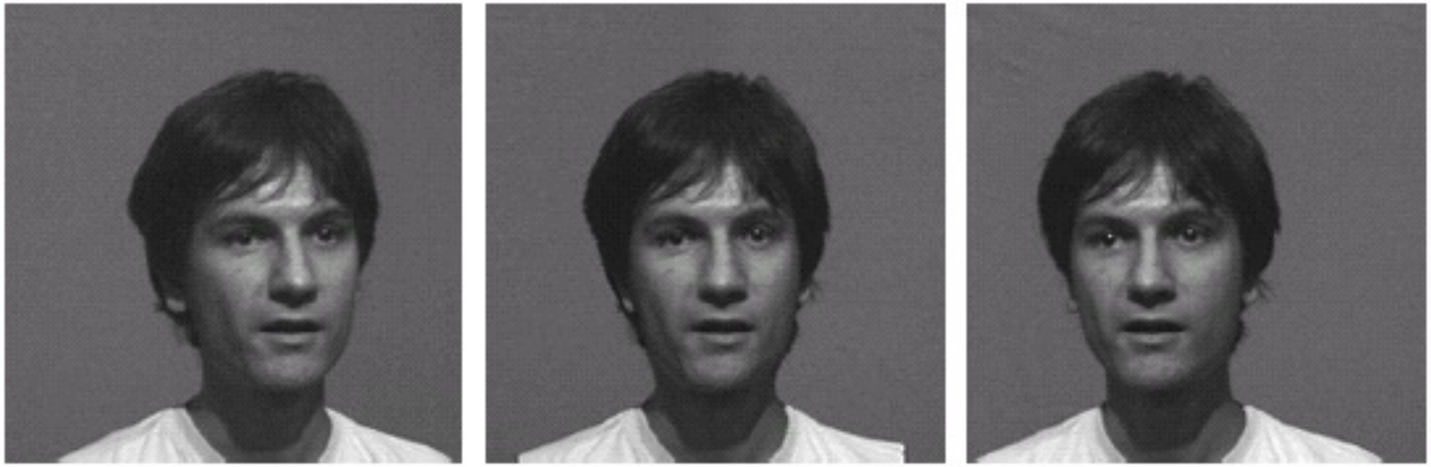


Figure 12.15 An example of intermediate view synthesis: the center image is interpolated from the left and right images. The result is obtained using mesh-based disparity-compensated interpolation, with the disparity map between the left and right views estimated using a mesh-based approach. Reprinted from R. Wang and Y. Wang, Multiview video sequence analysis, compression, and virtual viewpoint synthesis, *IEEE Trans. Circuits Syst. for Video Technology* (April 2000), 10(3):397–410. Copyright 2000 IEEE.

3D Camera / Depth Sensing

- Stereo camera
  - Depth from disparity, possibly with built-in algorithm for depth estimation
- Depth camera
  - Time-of-flight (ToF): Shine a pulsed or modulated laser beam (at ultraviolet, visible or infrared freq) to the imaged object and measure the total time the light travels for each pixel position. From round trip travel time, deduce the distance.
  - LIDAR: a special ToF camera targeted for outdoor long distance observation (e.g. environment, urban mapping), typically by scanning the scene in raster order
- Microsoft Kinect: contain a depth sensor, a color camera, and a microphone array

Stereo cameras in the old days

New 3D cameras: standalone, on laptop, smartphones

https://www.stereolabs.com/zed/

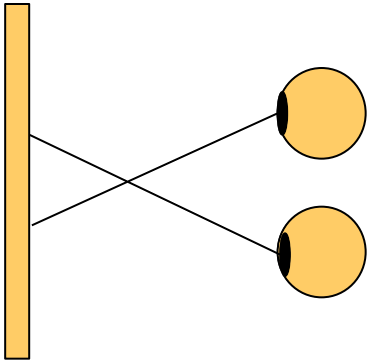
Lidar in commercial applications

http://en.wikipedia.org/wiki/Lidar

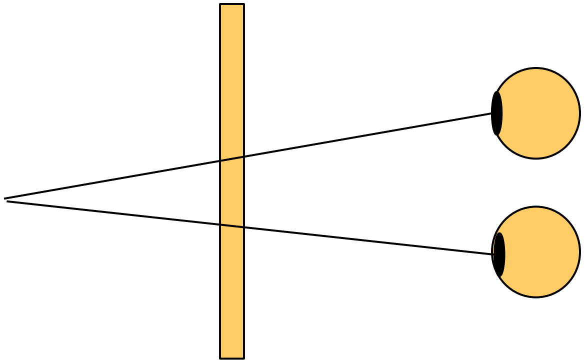
Display of Stereo Images/Sequences

- Principle:
  - Project images for the left and right eyes simultaneously in a way that the two images can be received separately by left and right eyes
- Separation mechanism in stereoscopic display
  - Color filter (Cannot be used for display color stereo images)
  - Polorization
  - Interlace in time the left and right views (Stereographics, Inc.)
  - Head mounted display (HMD: each eye views one image)
  - Viewers need to wear special glasses
- Auto-stereoscopic display
  - Present two or multiple views on the same screen simultaneously
  - A viewer sees different view when looking from different angle
  - Viewers do not need to wear glasses
  - Autostereoscopic lenticular screens

Disparity and Depth



Negative disparity:
Object in front of the screen



Positive disparity
Object behind the screen

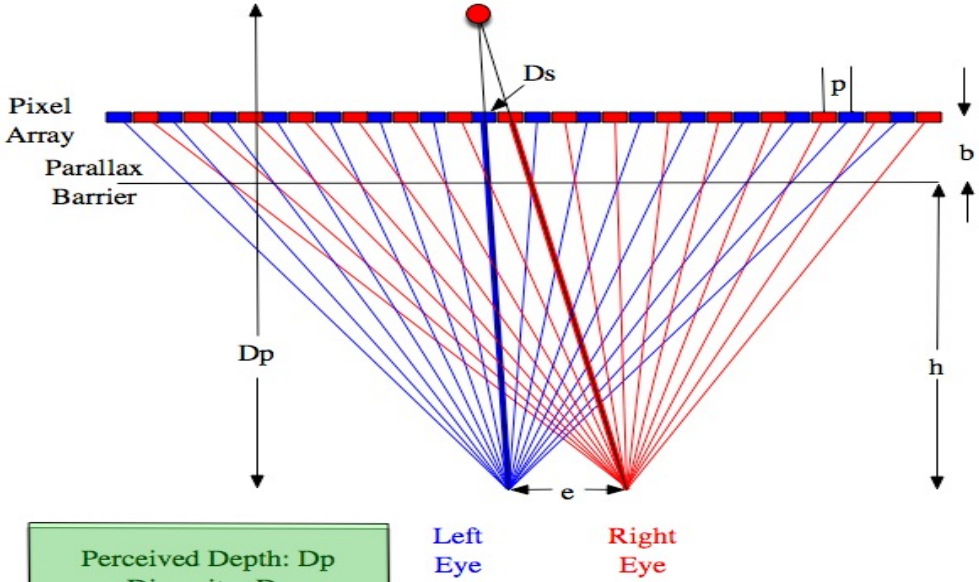
From Amy Reibman

Using glasses

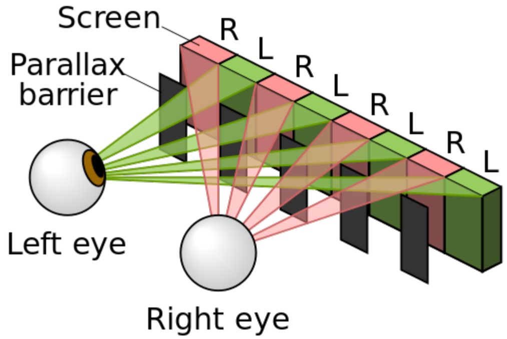
- Anaglyph. Two-color separation of left/right view. Poor color rendition.
- Polarized. For viewing stereo pairs projected through suitable polarizing filters. Better image quality.
- Shutter glasses. Liquid crystal. Expensive. Require high refresh rate. Require synchronization of display and glasses

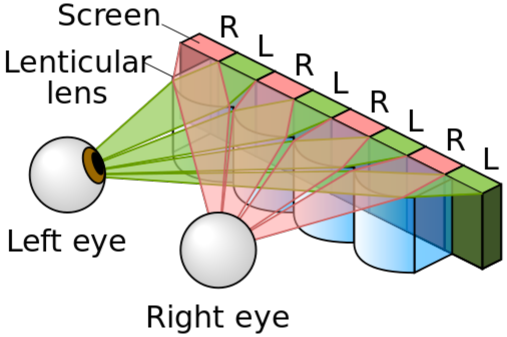
From Amy Reibman

Autostereoscopic display principle



Perceived Depth: Dp
Disparity: Ds
Dp=(b+h)e/(e-Ds)





- The viewer must be positioned in a well-defined spot to experience the 3D effect
- The effective horizontal pixel count viewable for each eye is reduced by one half

http://en.wikipedia.org/wiki/Parallax_barrier

From Amy Reibman

3D-ready consumer TVs

- Display stereo pairs in time-sequential manner
- Active shutter glasses

- Options
  - 3D DLP technology from TI (Samsung & Mitsubishi)
  - 3D plasma (Samsung)

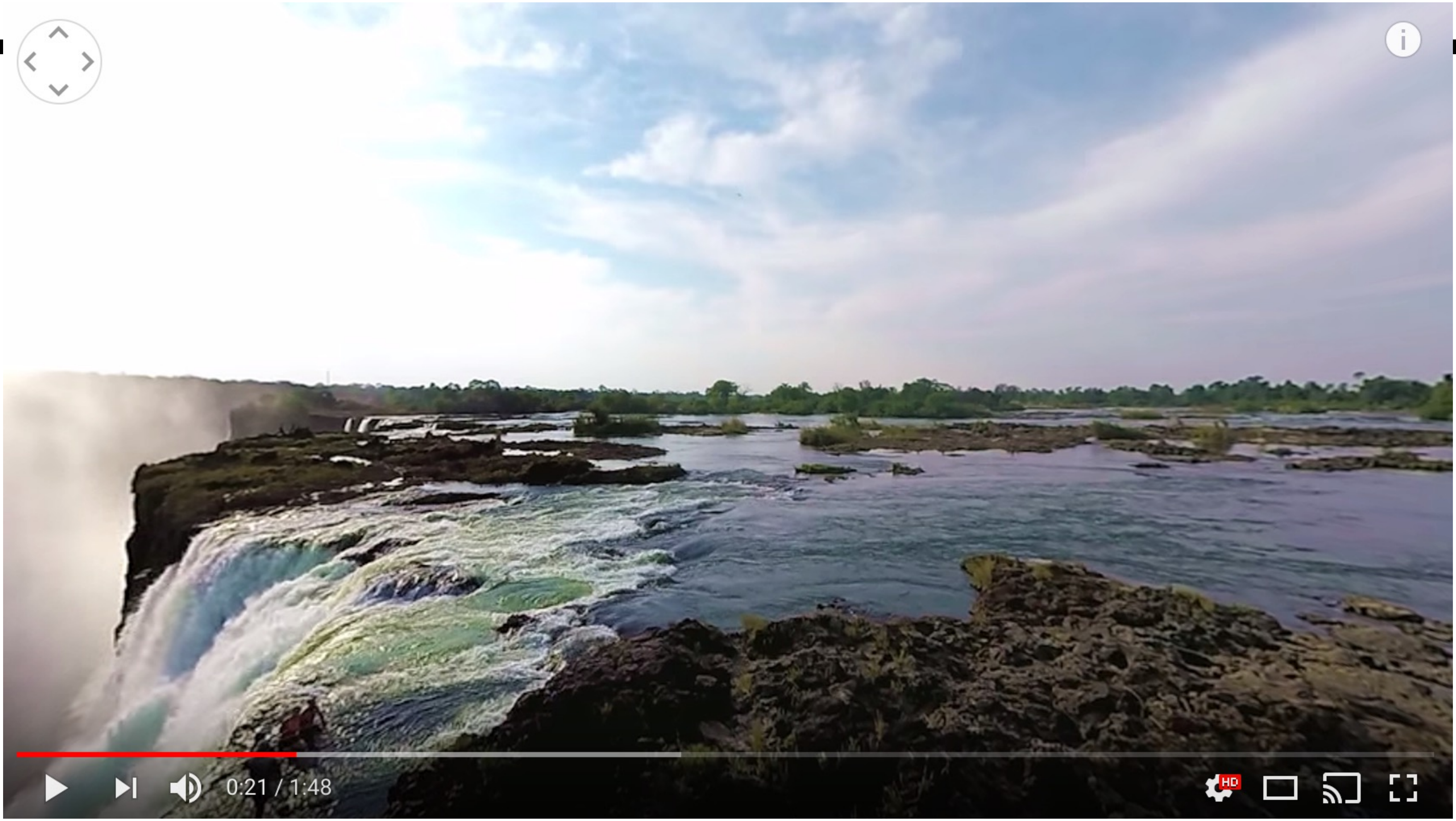
From Amy Reibman

360 degree camera

https://www.digitalcameraworld.com/buying-guides/best-360-cameras

YouTube's Ready To Blow Your Mind With 360-Degree Videos
http://gizmodo.com/youtubes-ready-to-blow-your-mind-with-360-degree-videos-1690989402

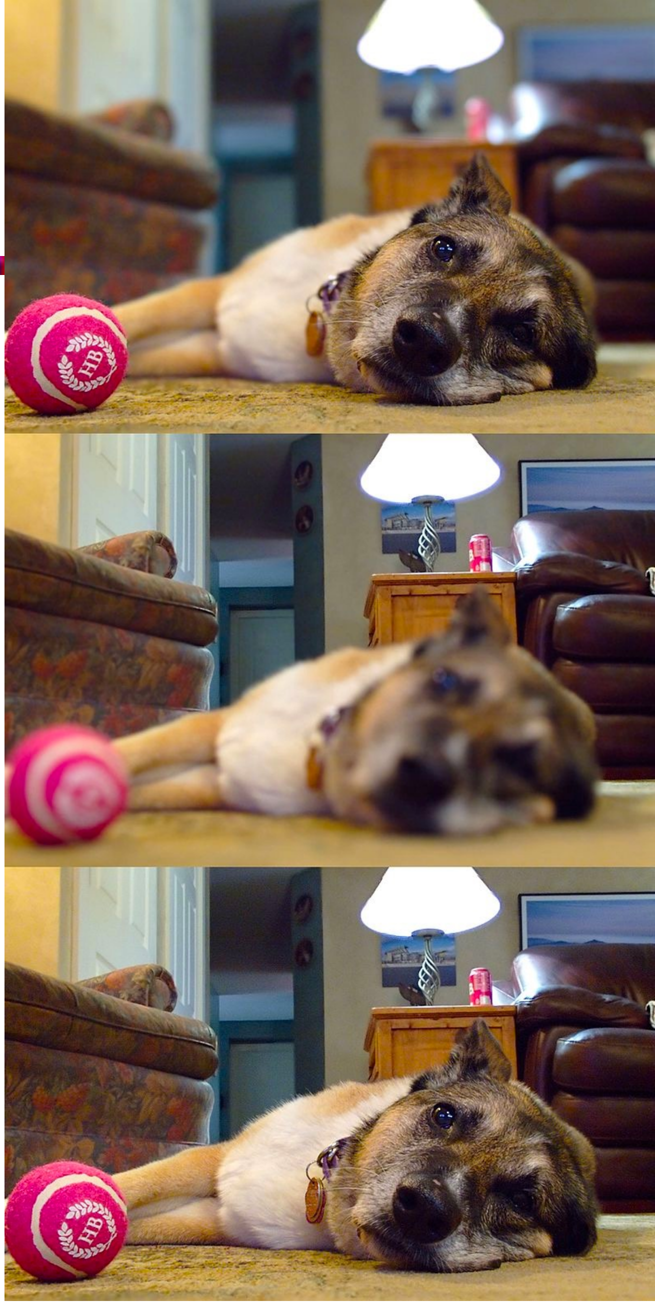
360 Degree Video Streaming



https://www.youtube.com/watch?v=WsMjBMxpUTc

Lightfield Camera

- Captures information about the light field emanating from a scene; that is, the intensity of light in a scene, and also the direction that the light rays are traveling in space.
- From the captured information, can generate the scene from different view angles, and also vary the focus plane
- Lytro: first commercial camera



By Doodybutch - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=56546526

Summary

- Human perception of depth
- Principle in stereo imaging:
  - Relation between depth and disparity for parallel set-up and other more general camera set-ups.
  - Epipolar constraint for an arbitrary set-up
- Disparity estimation:
  - Formulation as an optimization problem similar to motion estimation
  - Block-based approach
  - Mesh-based approach: regular mesh vs. adaptive mesh (not required)
  - Dynamic programming: not required
- 3D cameras: different ways of depth sensing
- Intermediate view synthesis
- Stereo image/video display
- Other emerging cameras …
- VR / AR …

Homework

- Review questions
  - Describe briefly how the human being deduce depth from disparity. A far away object has small or large disparity?
  - Describe how a stereo camera deduce depth
    - Dense disparity estimation algorithms: intensity based, edge based
  - Describe different ways of depth sensing for 3D imaging
  - Describe the general principle of intermediate view synthesis
- Written HW
  - See next slide

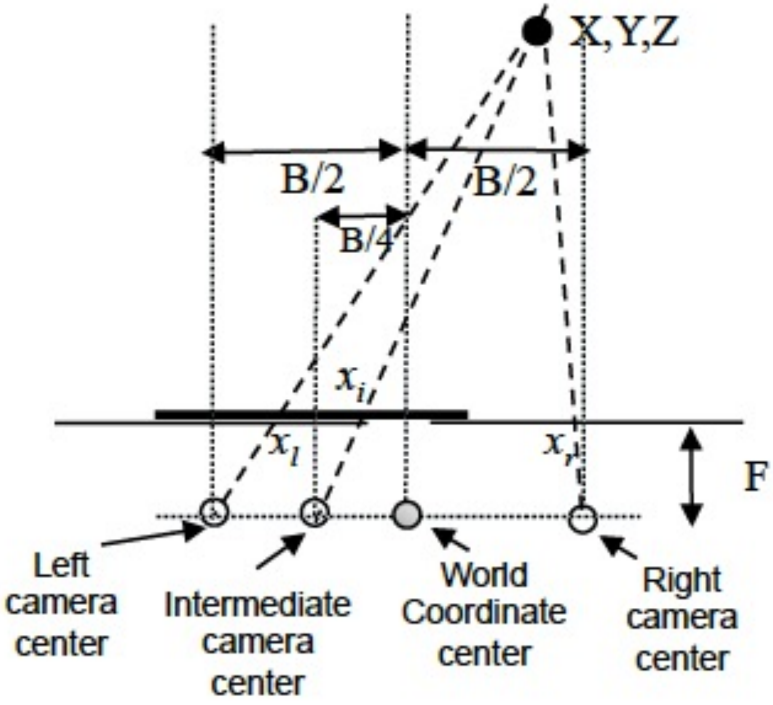
Consider a parallel stereo imaging system with baseline distance *B* and focus length *F*. Suppose that for an object point at world coordinate *(X,Y,Z)*, its image position in the left and right view are (x_l, y) and (x_r, y), respectively.

a. If the observed object has a flat surface described by $Z = aX + bY + c$, show that the disparity function seen from the left image can be modeled by an affine function,

$$d_l(x_l, y) = x_r(x_l) - x_l = (F - ax_l - by)/(\tfrac{c}{B} - \tfrac{a}{2})$$

correction: $d=(F-ax_l-by)/(a/2-c/B)$

b. Suppose we want to generate an intermediate view, whose camera center has a distance of *B*/4 away from the world coordinate origin, as shown below. How would you determine the image coordinate (x_i, y_i) for the same 3D point in this intermediate view? Express x_i, y_i in terms of x_l, x_r, y.



Computer Assignment (Optional)

- Write a program that can estimate the horizontal disparity map between two stereo images captured using parallel set up. To estimate the disparity at each pixel, apply EBMA over a block centered at this pixel. Apply your program to any stereo image pair you can download (e.g. from Middlebury stereo database).
- From the estimated disparity map, generate a depth map, by assuming an reasonable constants for B and F.
- Prob. 12.9 in [Wang2002]
- Prob. 12.11 in [Wang2002]

Reading Assignments

- [Wang2002] Wang, et al, Digital video processing and communications. Chap 12. (Sec. 12.1-12.4 required)
- [Szeliski2021] Richard Szeliski, Computer Vision: Algorithms and Applications. 2021. Chap. 12 (Sec. 12.3-12.5 required).
- Optional readings:
- Depth estimation from stereo images:
  - D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. IJCV 2002.

http://vision.middlebury.edu/stereo/taxonomy-IJCV.pdf

http://vision.middlebury.edu/stereo/ (an excellent website)

  - Tombari, S. Mattoccia, L. D. Stefano, and E. Addimanda. Classification and evaluation of cost aggregation methods for stereo correspondence. 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2008.
  - H. Hirschm¨uller. Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2):328–341, 2008.
- Kinect
  - Derek Hoiem, Lecture note on Kinect, courses.engr.illinois.edu/cs498dh/fa2011/lectures/Lecture%2025%20-%20How%20the%20Kinect%20Works%20-%20CP%20Fall%202011.pdf

Recommended Readings

- Stereo and autostereoscopic display
  - A. Woods, T. Docherty, and R. Koch, "Image distortions in stereoscopic video systems", Proceedings SPIE Stereoscopic Displays and Applications IV, vol. 1915, San Jose, CA, Feb. 1993.
  - Dodgson, N.A., "Autostereoscopic 3D Displays," IEEE Computer Magazine, vol.38, no.8, pp.31,36, Aug. 2005
  - Lecture note by Neil Dodgson, multi-view autostereoscopic display http://www.cl.cam.ac.uk/~nad10/pubs/Stanford3D-2011.pdf

Optional Material

Screen geometry

- Each technology has its own screen size and resolution
- IMAX
  - 48-foot screen; 2048x1080: aspect ratio 1.4
  - Typically all seats are within one screen height
- Real-D XLS:
  - 20-foot screen; 2048x858 per view; aspect ratio 1.85
  - Typically seats are within {single digit} screen heights
- Home TV
  - Typically 8 feet viewing distance
- Screen parallax (i.e. disparity) is affected by the size of the display screen

From Amy Reibman

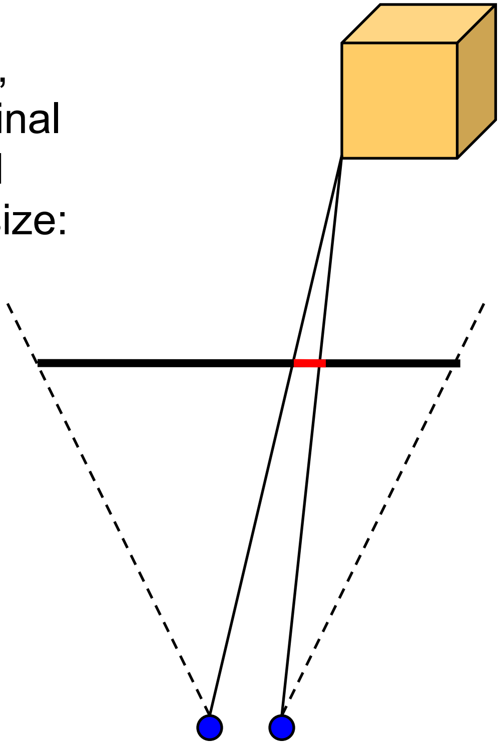
Depth-based adaptation of 3D content

- Perceived depth is affected by the screen size and and viewing distance
- To display for different screens, need to adjust stereo disparity
- Limit maximum disparity to avoid too much eye strain
- Shifting/offsetting one image has only limited success
- Ideally, for a given viewer distance and viewer location, generate an intermediate view for that viewer: Intermediate view synthesis

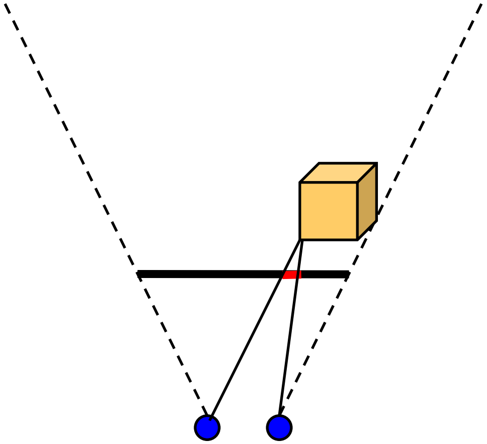
From Amy Reibman

Mismatch in screen sizes and viewing distances (movie theater vs. home)

Real-life,
and original
intended
screen size:



Same screen angle, smaller disparity.
Result: different object size and distance

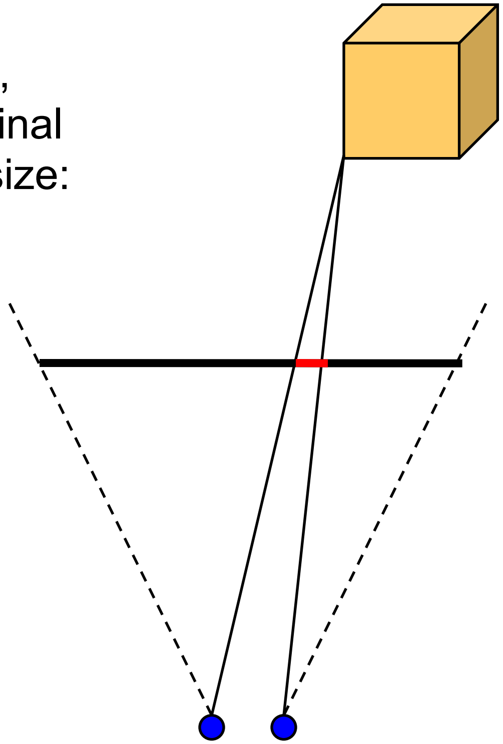


Objects appear to
be in a puppet theater:
small and close
together

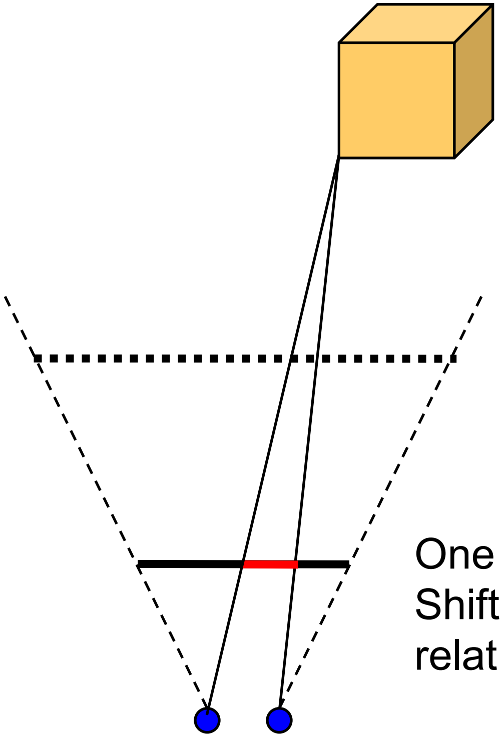
From Amy Reibman

Adaptation of disparity for different screen sizes

Real-life,
and original
screen size:



Same screen angle,
shifted disparity.
Result: same object size



One way to “fix”:
Shift one image
relative to the other

Now, the object appears the correct size. However, objects are almost always behind the screen. Causing conflict of converging and accommodation

From Amy Reibman

Barriers to mass market

- Data and delivery format
- Quality of 3D video production
  - Content creators must be aware of 3D videography
  - Re-purposing of 3D content from cinema into the homes
  - 2D->3D conversion: converting old 2D movie to 3D
- Human factors
  - Stereoscopic glasses are no fun
  - Auto-stereoscopic has its own issues
  - Avoid objectionable 3D effects

From Amy Reibman

Additional perceptual issues

- Both too *much* depth and too many *fast changes* of depth cause visual fatigue
- Conflicting depth information causes visual fatigue
  - Accommodation and vergence are linked when scanning the scene (but can be decoupled over time)
  - Compression, aliasing, other impairments (like keystoning) can make fusing more difficult
  - Screen or glasses scratch or dust
- Cross-talk
  - Left eye sees some of what Right eye should see
  - Stronger in high-contrast and large-disparity areas
  - (But fusing is easier in high-contrast areas)

From Amy Reibman