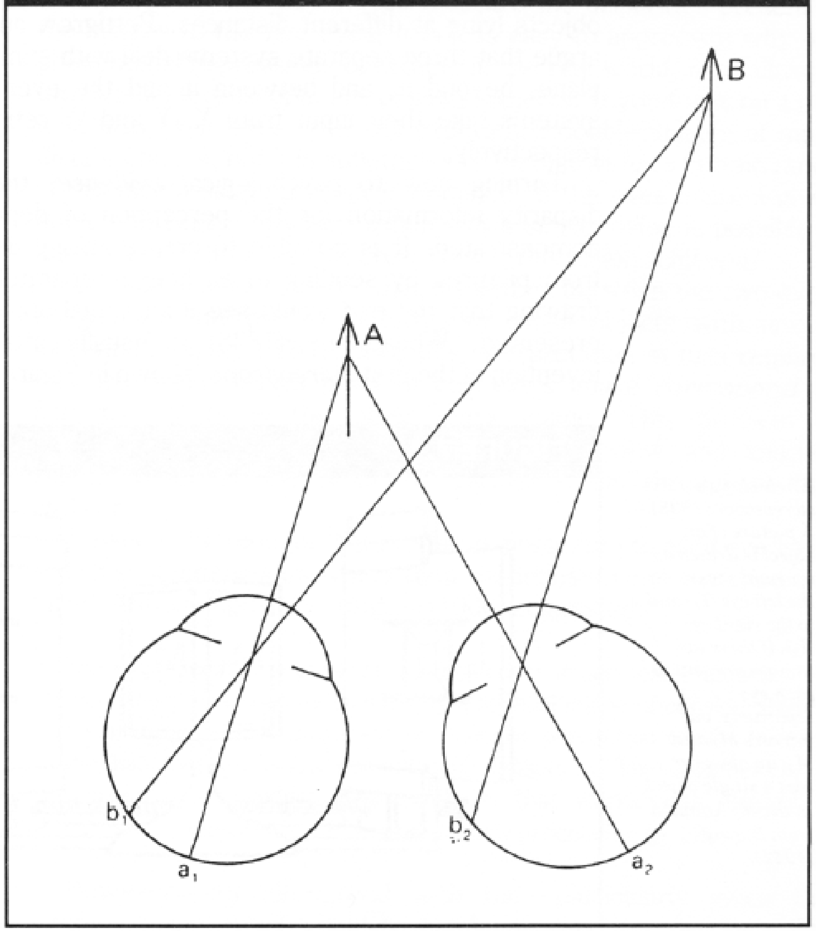
Stereopsis

CS 554 – Computer Vision

Pinar Duygulu
Bilkent University

Disparity

FIGURE 7.3



From Bruce and Green, Visual Perception, Physiology, Psychology and Ecology

Disparity occurs when
Eyes verge on one object;
Others appear at different
Visual angles

Adapted from David Forsyth, UC Berkeley

Disparity

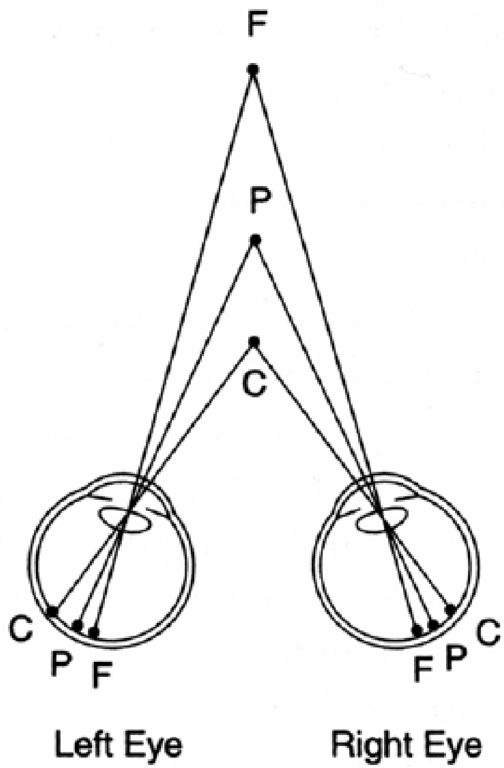


Figure 5.3.2 Crossed versus uncrossed binocular disparity. When a point P is fixated, closer points (such as C) are displaced outwardly in crossed disparity, whereas farther points (such as F) are displaced inwardly in uncrossed disparity.

Adapted from David Forsyth, UC Berkeley

From Palmer, "Vision Science", MIT Press

Disparity

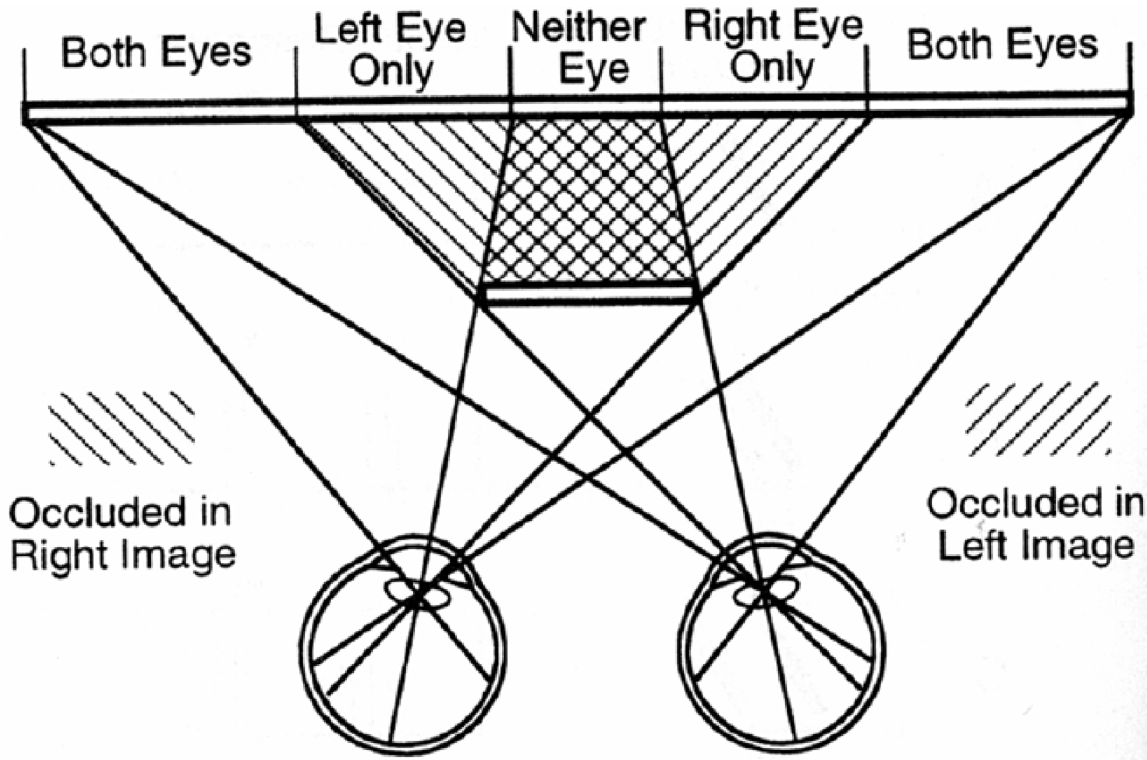


Figure 5.3.23 Da Vinci stereopsis. Depth information also arises from the fact that certain parts of one retinal image have no corresponding parts in the other image. (See text for details.)

From Palmer, “Vision Science”, MIT Press

Adapted from David Forsyth, UC Berkeley

Stereo Vision

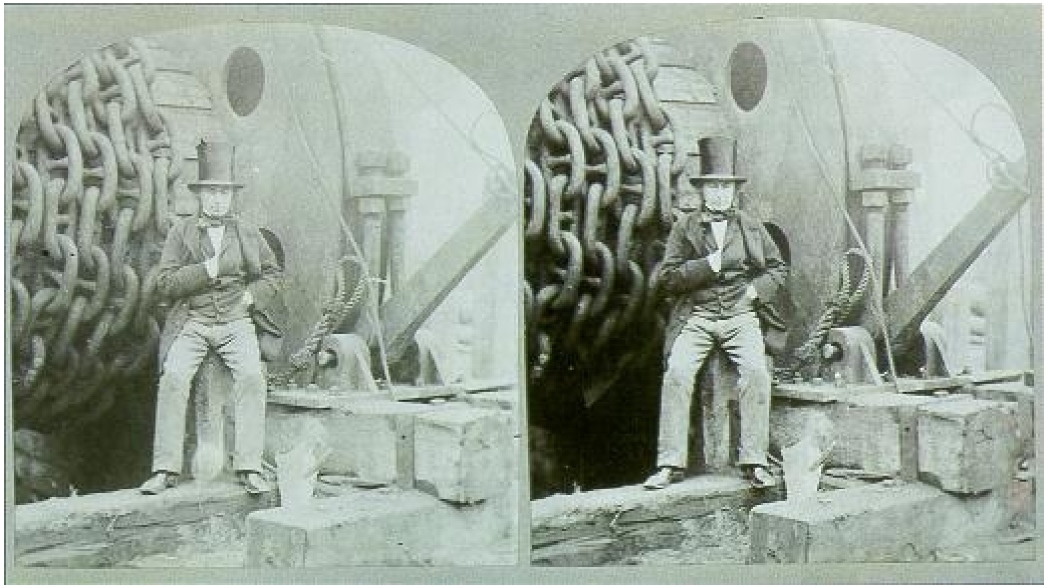
- The whole process is called **stereo vision** and it is derived from the Greek word **`*stereos*'** which means form or solid i.e. having three dimensions.

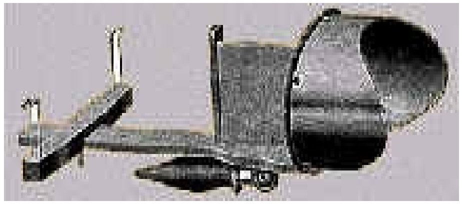
- **Stereoscopy** is the science by which two photographs of the same object taken at slightly different angles are viewed together, giving an impression of depth and solidity as in ordinary human vision.

- **Stereo photography** is the art of taking two pictures of the same subject from two slightly different viewpoints and displaying them in such a way that each eye sees only one of the images.

http://www.photostuff.co.uk/stereo.htm

Stereo photography





- Capturing the image on film requires the photographer to take two pictures from slightly different viewpoints.

- In order to view the captured photographs, the images have to be displayed in such a way that each of the viewer's eyes sees only one image.

http://www.photostuff.co.uk/stereo.htm

Anaglyph

\+

=

Left Eye Image
(Red channel only)

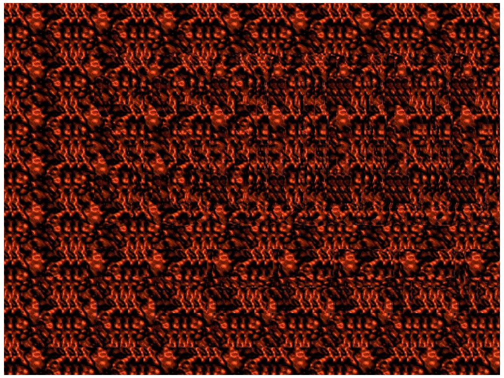
Right Eye Image
(Red channel removed)

Anaglyph
(Left & Right
images overlaid)

- Requires the viewer to wear glasses with red and green/cyan lenses.
- The left image has the blue and green colour channels removed to leave a purely red picture while the right image has the red channel removed.
- The two images are superimposed into one picture which produces a picture very like the original with a red and cyan fringes around objects where the stereo separation produces differences in the original images.
- The red and cyan lenses in the glasses let the eyes separate the two superimposed images into their individual components which the brain then combines to form a 3D-image.

http://www.photostuff.co.uk/stereo.htm

Freeview



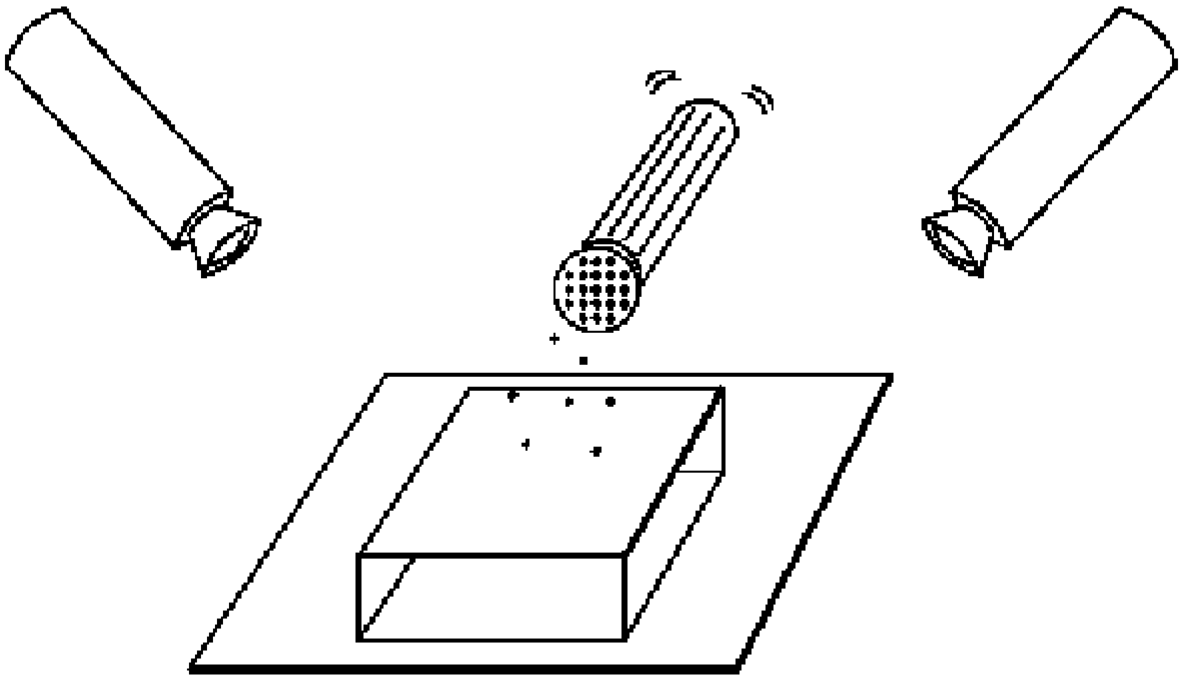
- Free Viewing, the eyes should not converge but look parallel as if the image being looked at is in the distance.
- The brain is fooled into thinking that it has two separate images and creates a 3-D visualisation.

- Single Image Random Dots Stereogram (SIRDS)
- Single Image Stereogram (SIS)

- "Magic Eye" pictures are created by computer and rely on the fact that the brain depends on matching vertical edges to synchronise the left and right images.
- The picture is made up of columns of patterns, which vary slightly across the picture.
- The brain interprets the columns as left and right pairs and the slight differences between each column define the subject e.g. the fish.

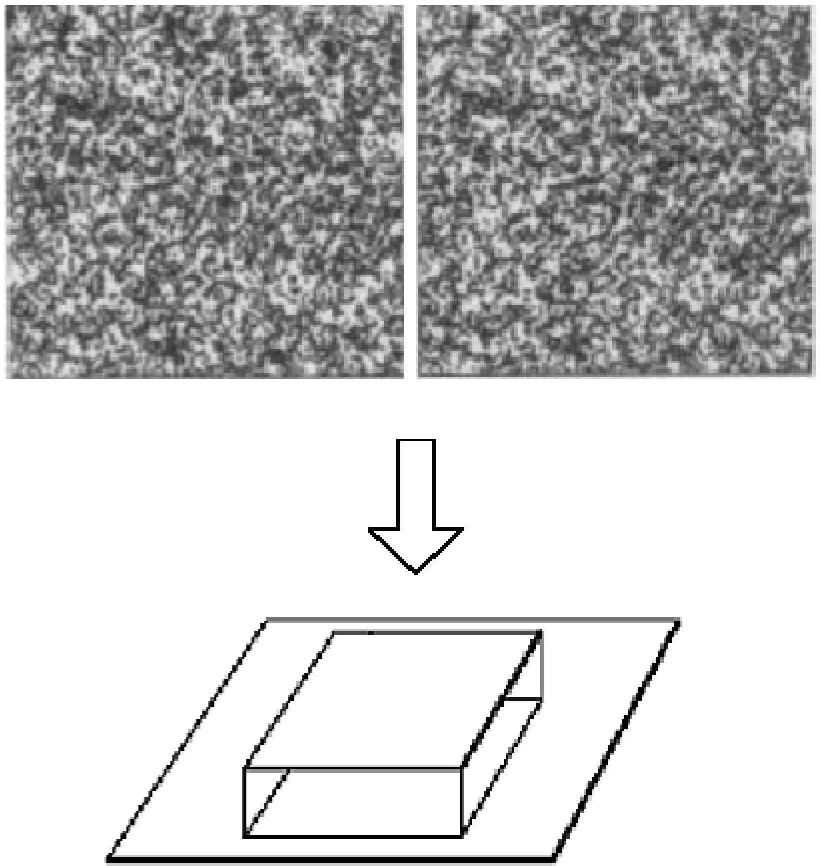
http://www.photostuff.co.uk/stereo.htm

Random dot stereograms



Adapted from David Forsyth, UC Berkeley

Random dot stereograms



Adapted from Trevor Darrell, MIT

Random dot stereograms

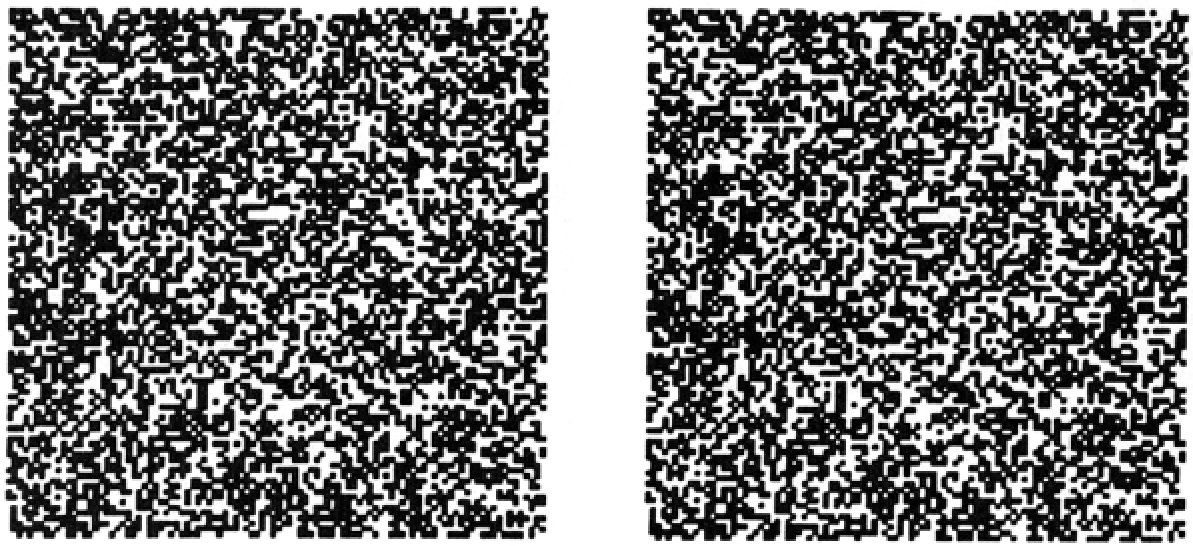


Figure 5.3.8 A random dot stereogram. These two images are derived from a single array of randomly placed squares by laterally displacing a region of them as described in the text. When they are viewed with crossed disparity (by crossing the eyes) so that the right eye's view of the left image is combined with the left eye's view of the right image, a square will be perceived to float above the page. (See pages 210–211 for instructions on fusing stereograms.)

square

Adapted from David Forsyth, UC Berkeley

Random dot stereograms

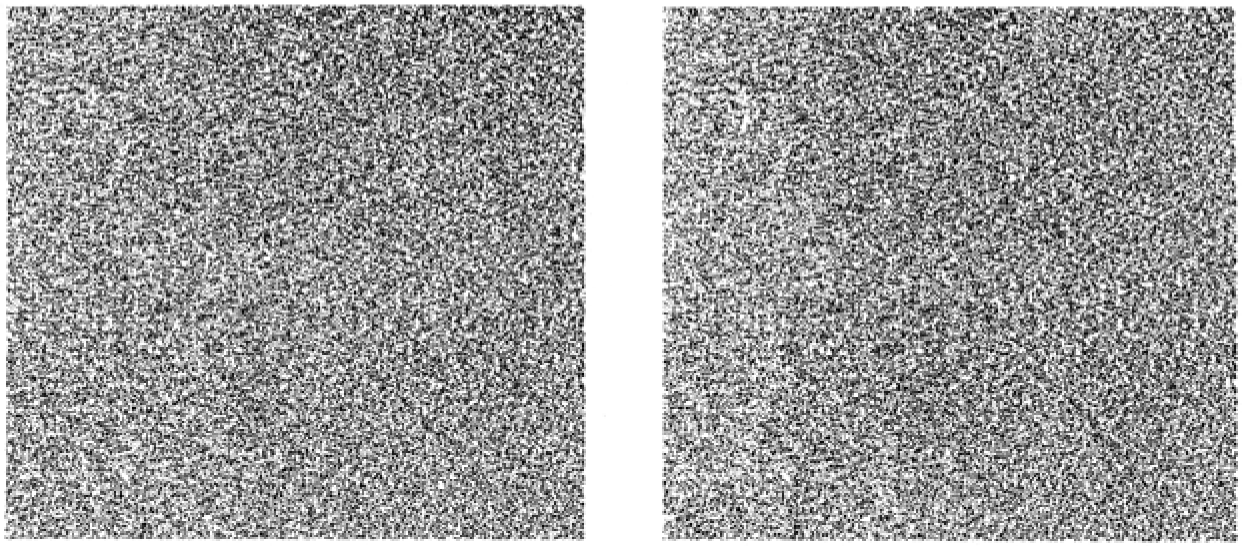


Figure 5.3.9 A random dot stereogram of a spiral surface. If these two images are fused with crossed convergence (see text on pages 210–211 for instructions), they can be perceived as a spiral ramp coming out of the page toward your face. This perception arises from the small lateral displacements of thousands of tiny dots. (From Julesz, 1971.)

Spiral ramp

Adapted from David Forsyth, UC Berkeley

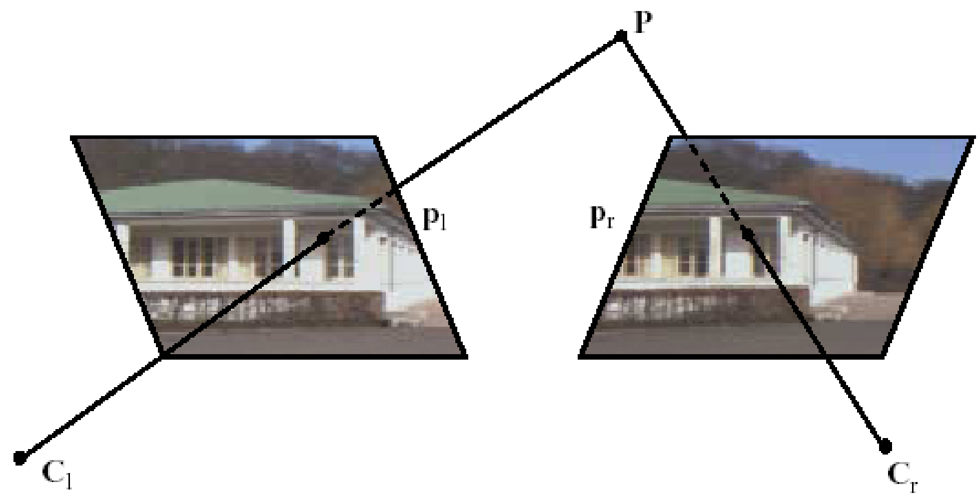
Random dot stereograms

Human binocular fusion cannot be explained by peripheral processes directly associated with the physical retinas.

Instead, it must involve the central nervous system and an imaginary *cyclopean retina* that combines the left and right image stimuli as a single unit

Adapted from David Forsyth, UC Berkeley

Stereo vision = correspondences + reconstruction



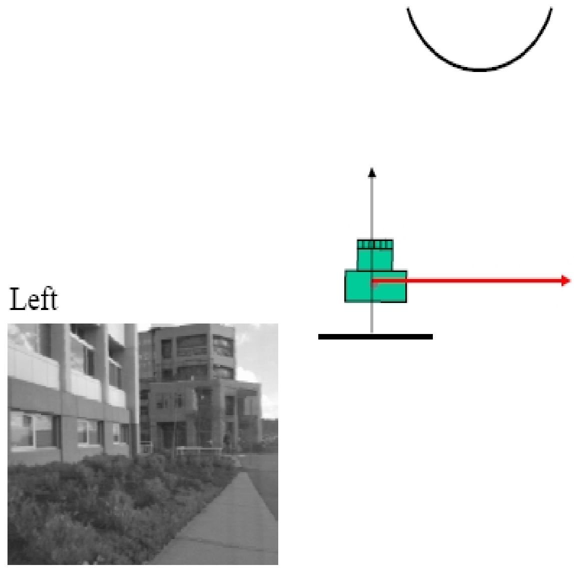
Stereovision involves two problems:

Correspondence : Given a point p_l in one image, find the corresponding point in the other image

Reconstruction: Given a correspondence (p_l, p_r) compute the 3D coordinates of the corresponding point in space, P

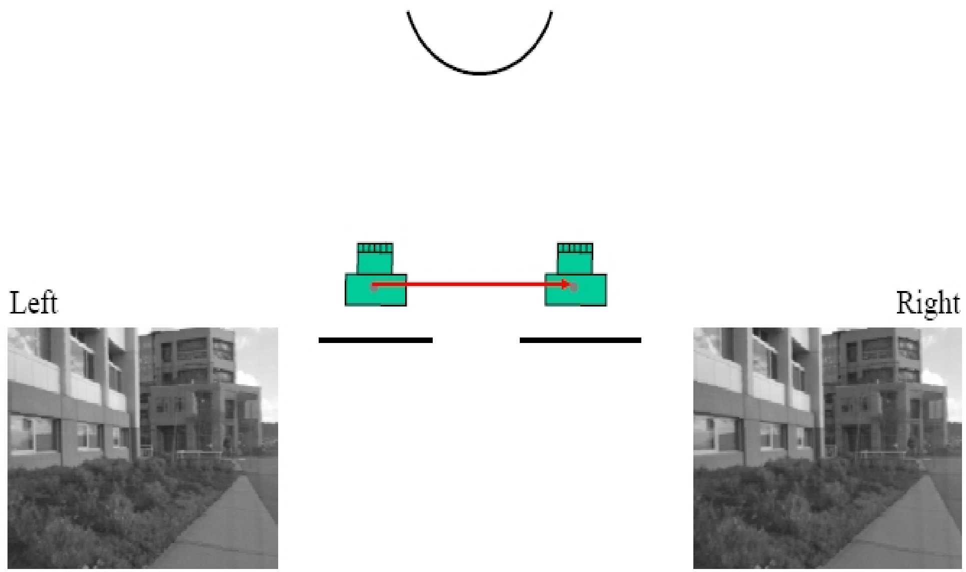
Adapted from Martial Hebert, CMU

Binocular Stereo



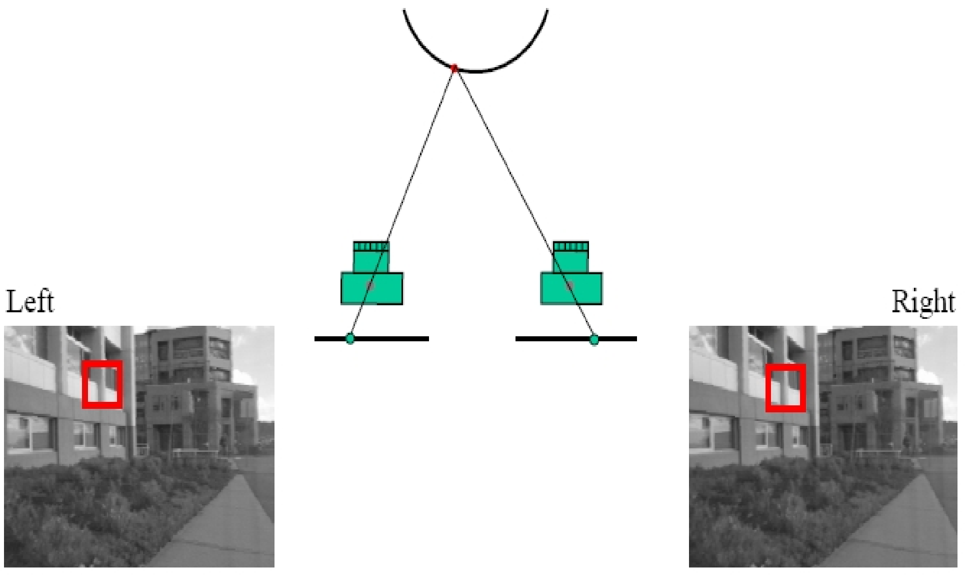
Adapted from Michael Black

Binocular Stereo



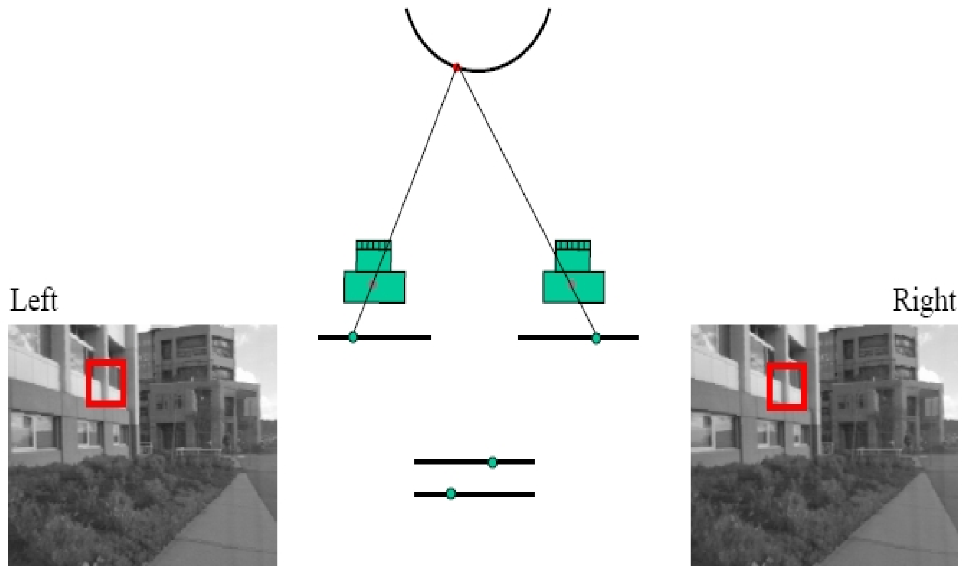
Adapted from Michael Black

Binocular Stereo



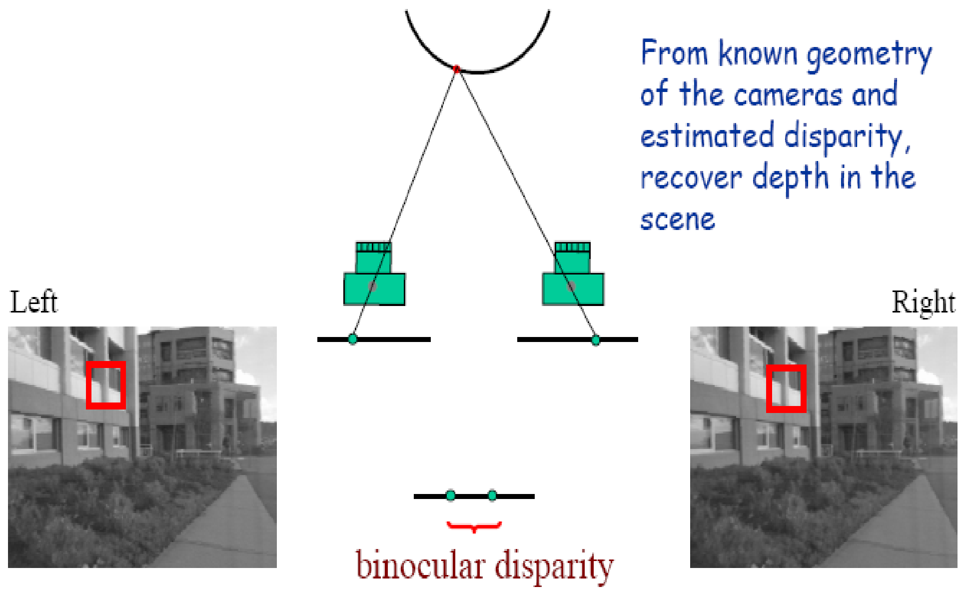
Adapted from Michael Black

Binocular Stereo



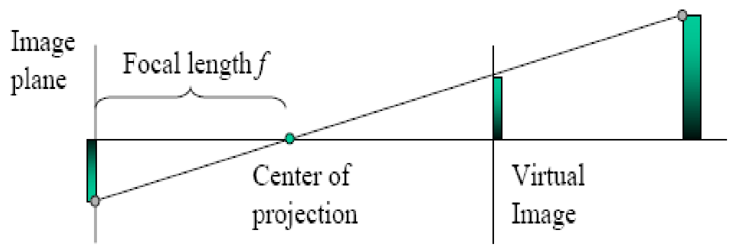
Adapted from Michael Black

Binocular Stereo



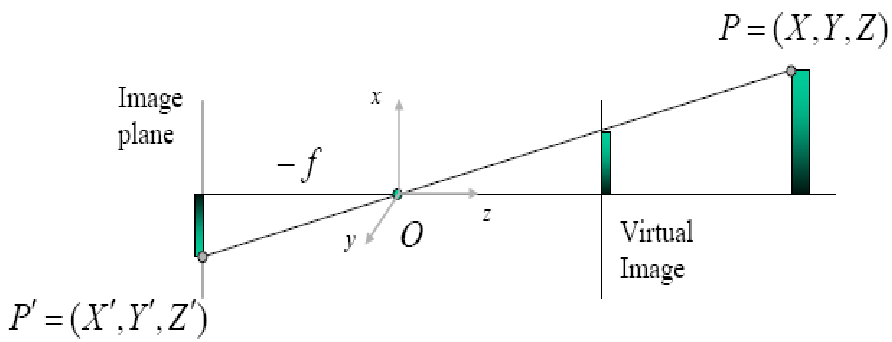
Adapted from Michael Black

Depth Estimation



Adapted from Michael Black

Depth Estimation



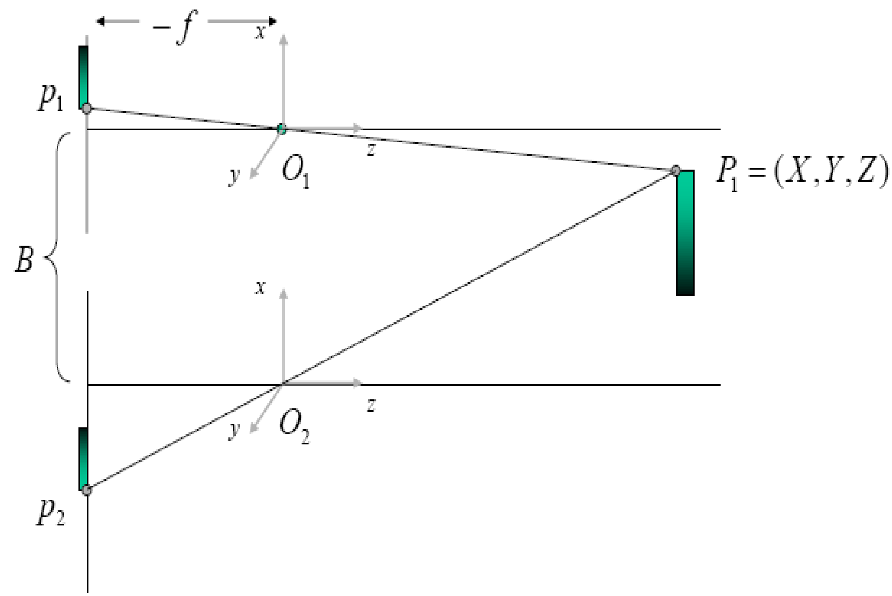
$$Z' = -f, \quad X' = -f\frac{X}{Z}, \quad Y' = -f\frac{Y}{Z}$$

$$x = -X', \quad y = -Y'$$

$$(X, Y, Z) \rightarrow (x, y, 1) = (f\frac{X}{Z}, f\frac{Y}{Z}, 1)$$

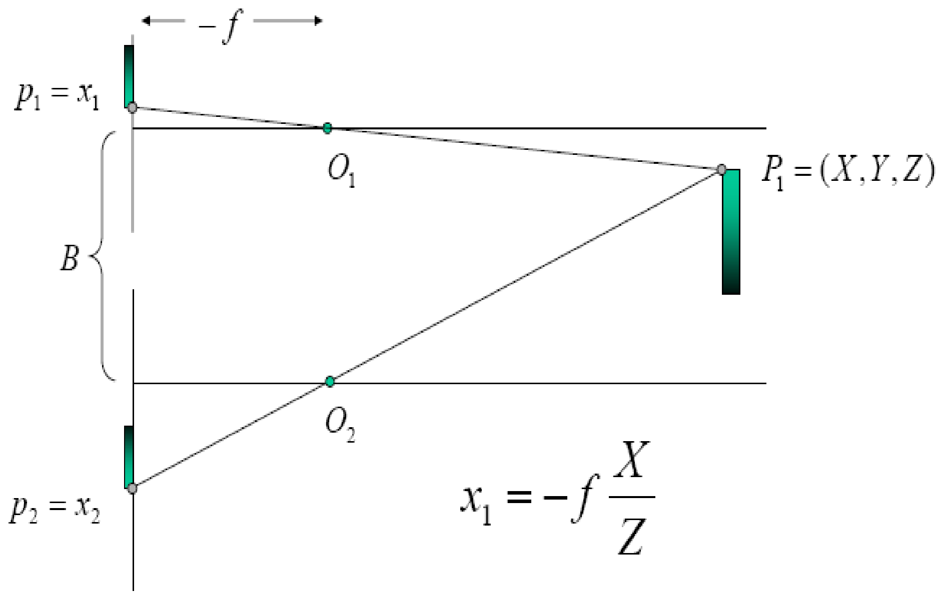
Adapted from Michael Black

Depth Estimation



Adapted from Michael Black

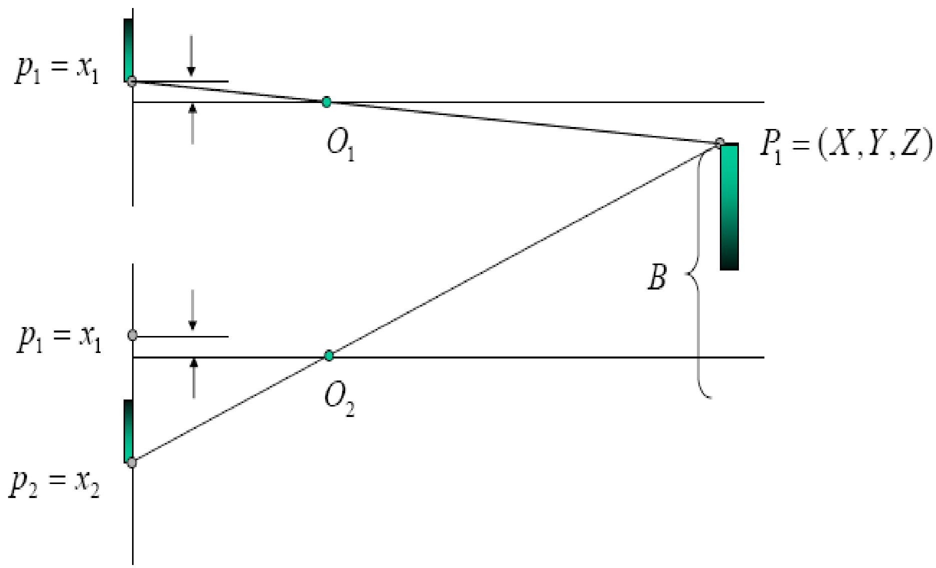
Depth Estimation



$$x_1 = -f\frac{X}{Z}$$

Adapted from Michael Black

Depth Estimation

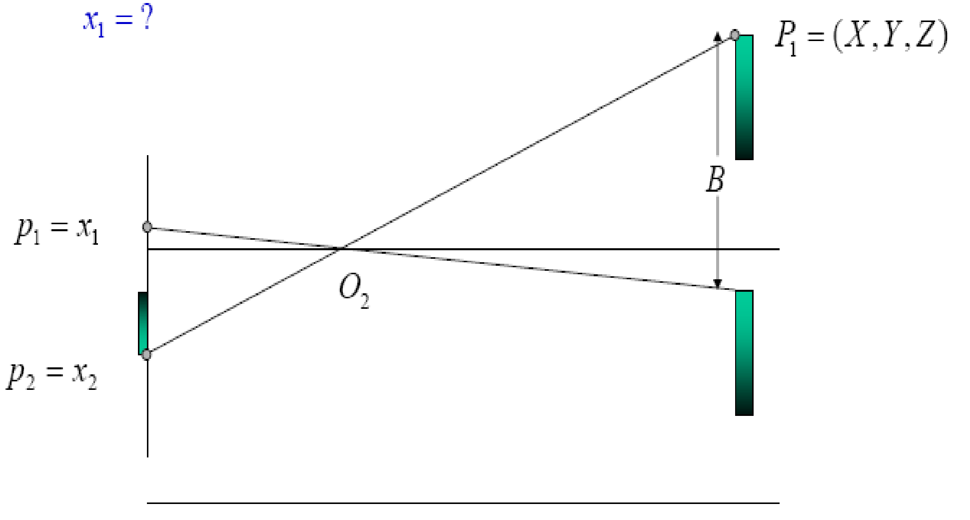


Adapted from Michael Black

Depth Estimation

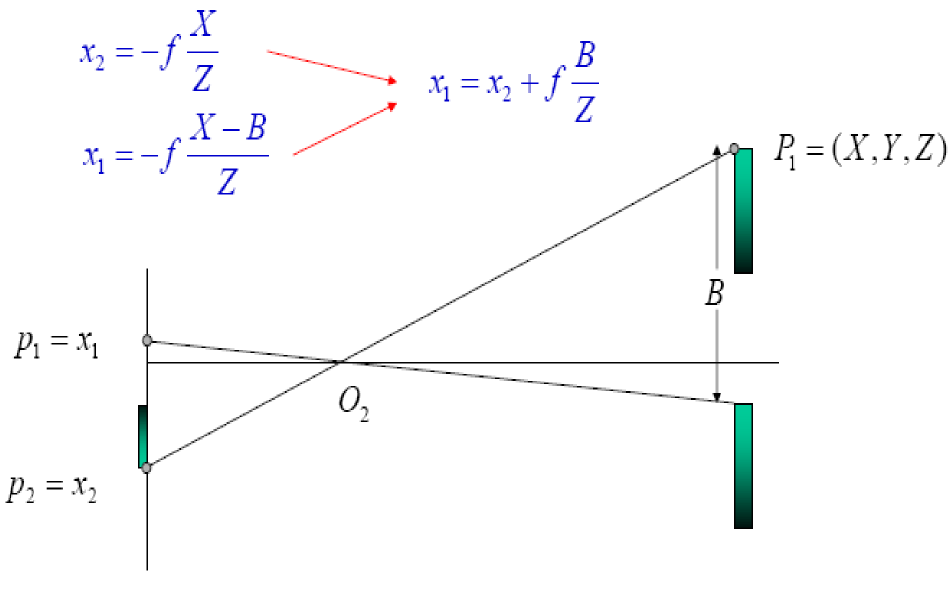
$$x_2 = -f\frac{X}{Z}$$

$$x_1 = ?$$



Adapted from Michael Black

Depth Estimation

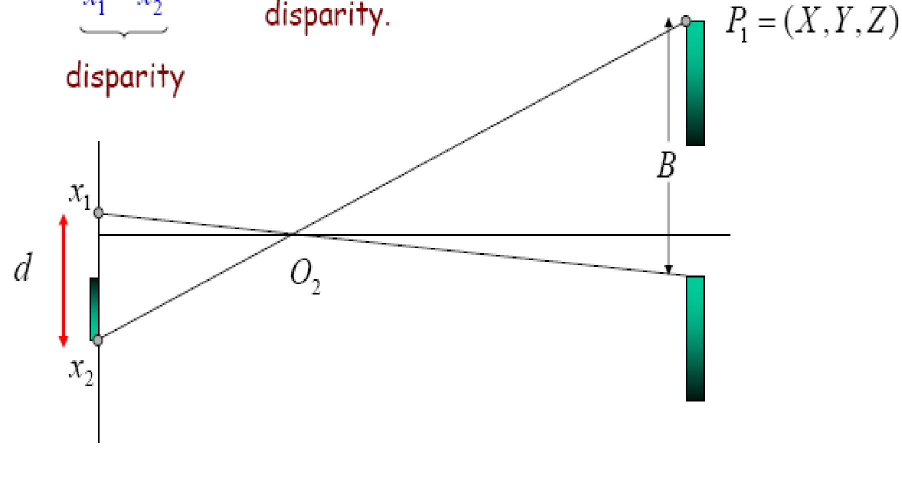


Adapted from Michael Black

Depth Estimation

For a calibrated camera, we know f and the baseline B.
Then depth can be computed from disparity.

$$Z = \frac{fB}{\underbrace{x_1 - x_2}_{\text{disparity}}}$$



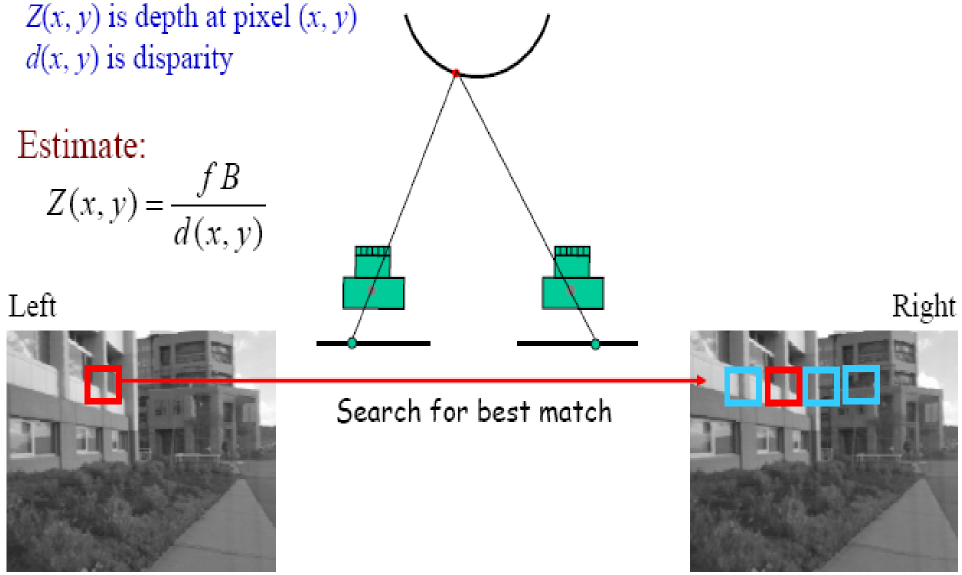
Adapted from Michael Black

Correspondence

$Z(x, y)$ is depth at pixel (x, y)
$d(x, y)$ is disparity

Estimate:

$$Z(x, y) = \frac{f B}{d(x, y)}$$



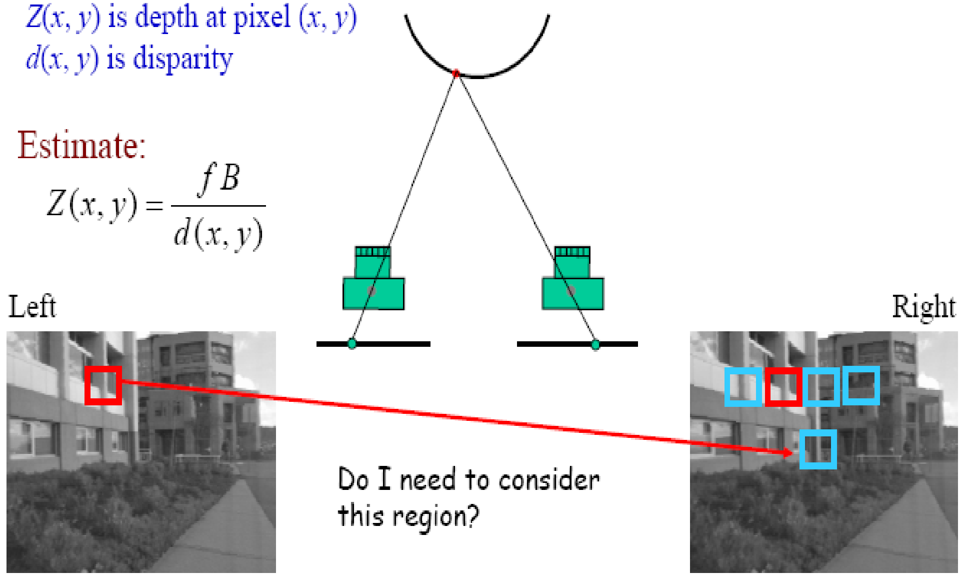
Adapted from Michael Black

Correspondence

$Z(x, y)$ is depth at pixel (x, y)
$d(x, y)$ is disparity

Estimate:

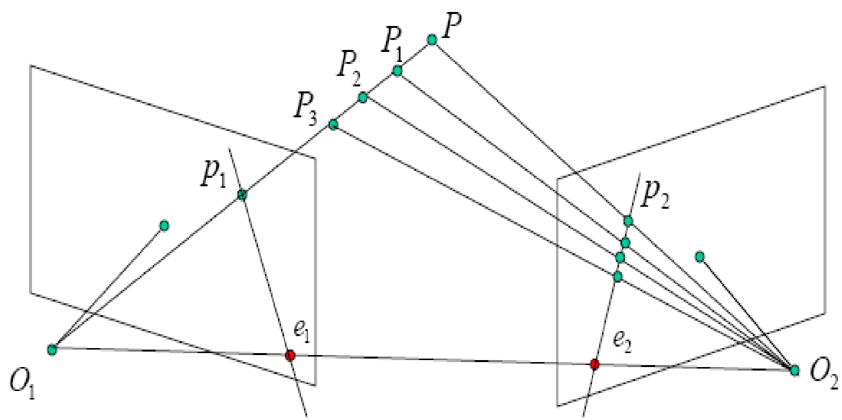
$$Z(x,y) = \frac{fB}{d(x,y)}$$



Adapted from Michael Black

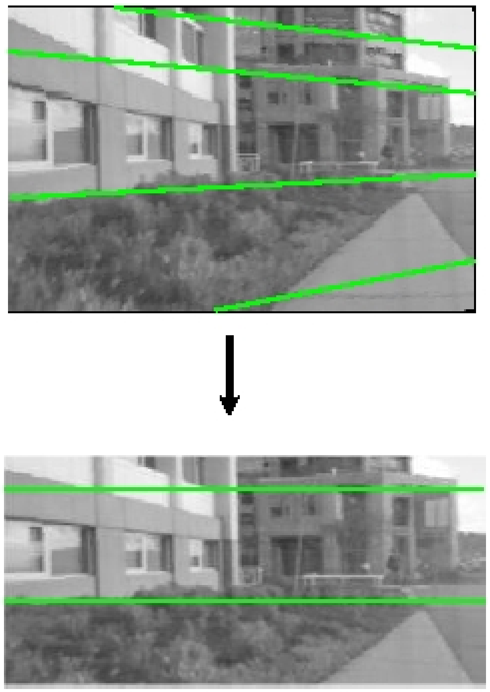
Correspondence

Possible matches for p_1 are constrained to lie along the epipolar line in the other image



Adapted from Michael Black

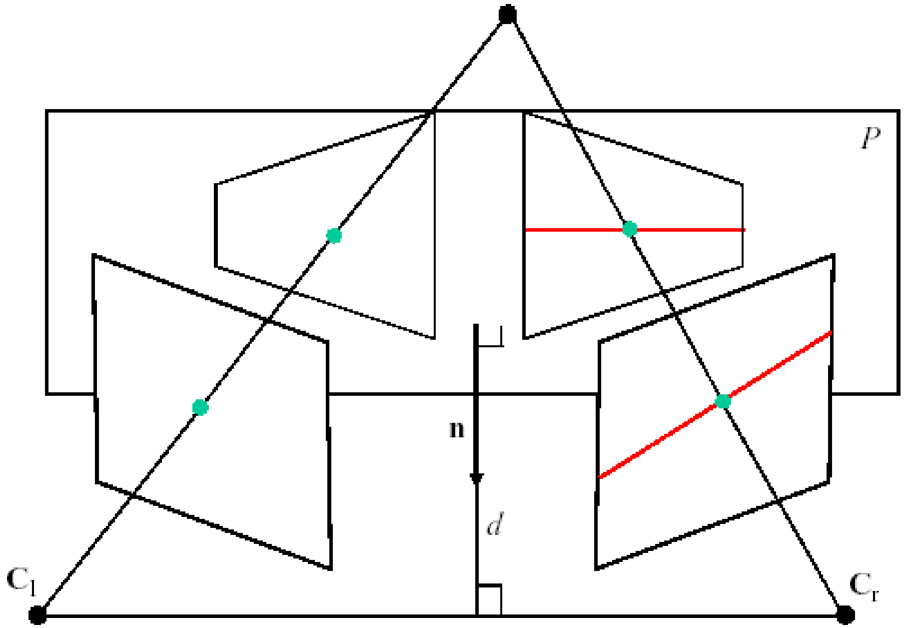
Rectification



Searching along epipolar lines at arbitrary orientation is intuitively expensive. It would be nice to be able to always search along the rows of the right image. Fortunately, given the epipolar geometry of the setereo pair, there always exists a transformation that maps the images into a pair of images with the epipolar lines parallel to the rows of the image. This transformation is called *rectification*. Images are almost always rectified before searching for correspondences in order to simplify the search.

Adapted from Martial Hebert, CMU

Rectification

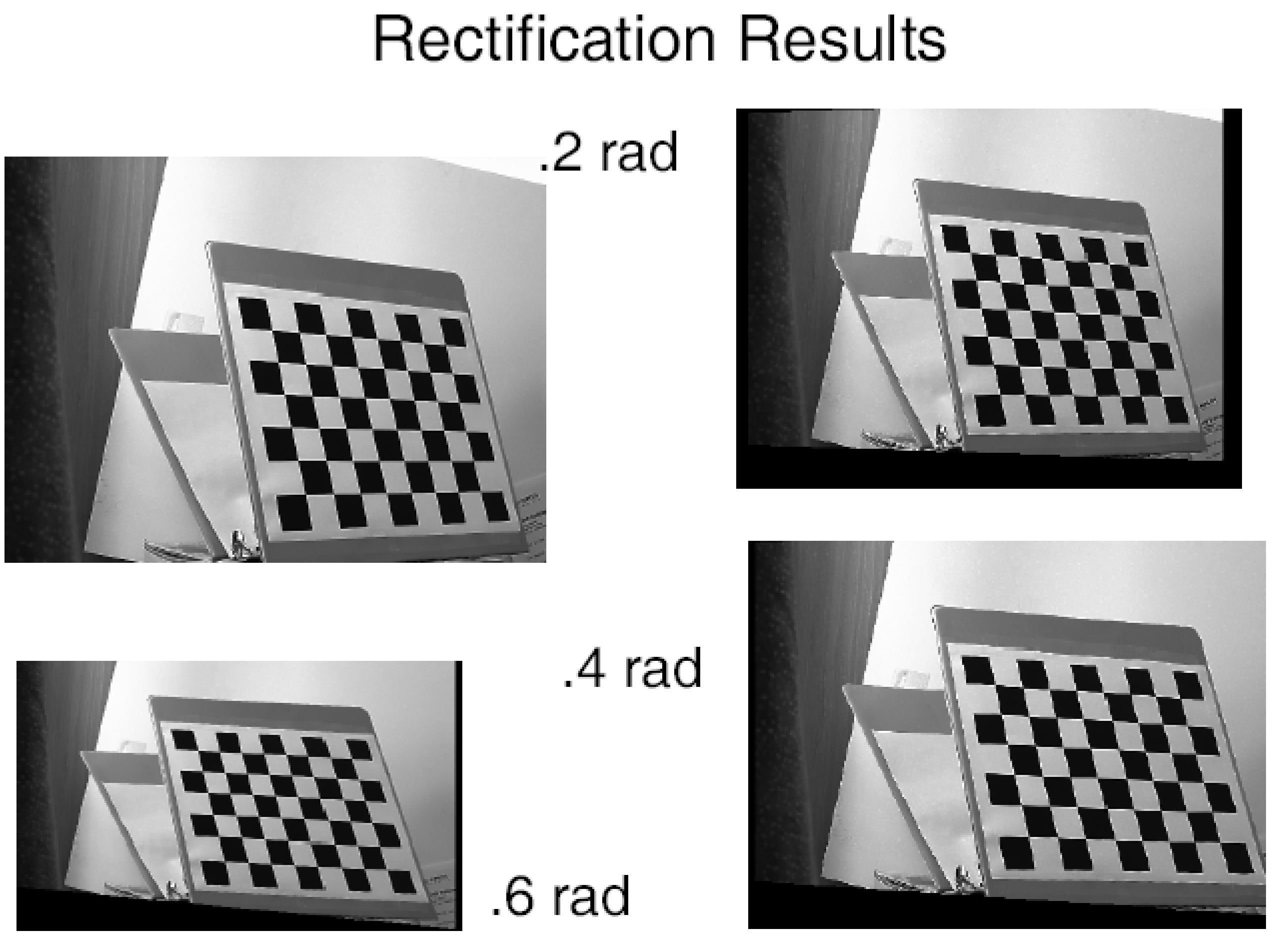


We know that, given a plane **P** in space, there exists two homographies $\mathbf{H}_l$ and $\mathbf{H}_r$ that map each image plane onto **P**. That is, if $\mathbf{p}_l$ is a point in the left image, then the corresponding point in **P** is **Hp** (in homogeneous coordinates). If we map both images to a common plane **P** such that P is parallel to the line $\mathbf{C}_l\mathbf{C}_r$, then the pair of virtual (rectified) images is such that the epipolar lines are parallel. With proper choice of the coordinate system, the epipolar lines are parallel to the rows of the image.

The algorithm for rectification is then:

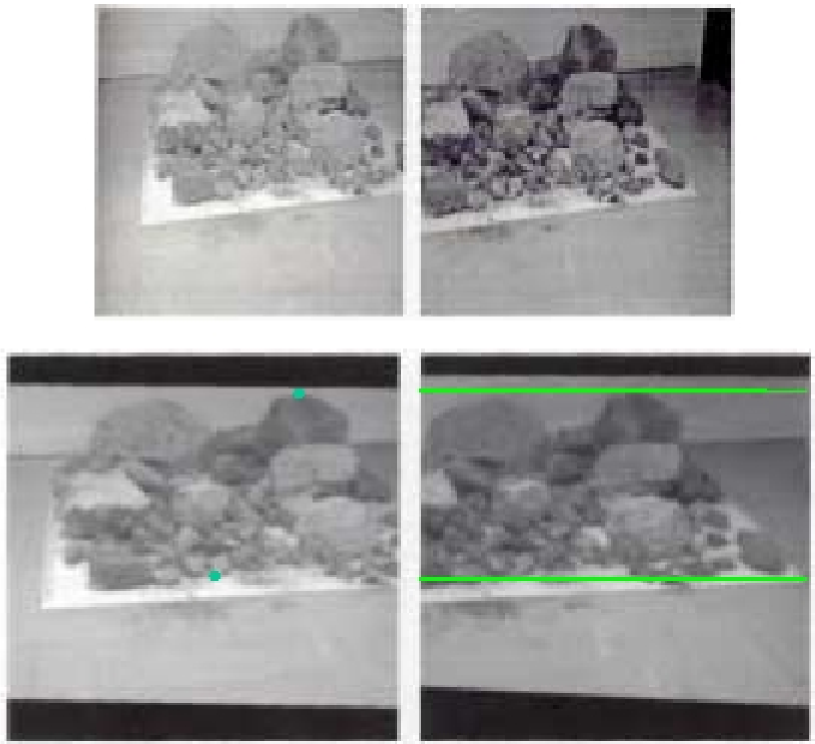
- Select a plane **P** parallel to $\mathbf{C}_r\mathbf{C}_l$
- Define the left and right image coordinate systems on **P**
- Construct the rectification matrices $\mathbf{H}_l$ and $\mathbf{H}_r$ from **P** and the virtual image's coordinate systems.

Rectification Results



Adapted from G. Hager, JHU

Rectification

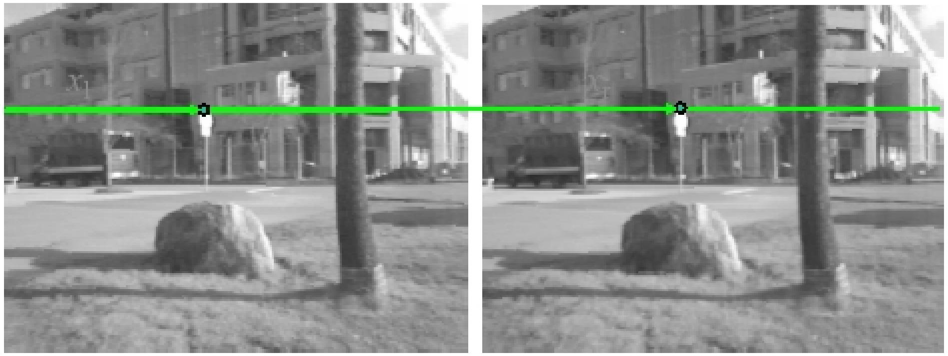


Adapted from Martial Hebert, CMU

Disparity

Assuming that images are rectified to simplify things, given two corresponding points $\mathbf{p}_l$ and $\mathbf{p}_r$, the difference of their coordinates along the epipolar line x_l-x_r is called the disparity d. The disparity is the quantity that is directly measured from the correspondence.

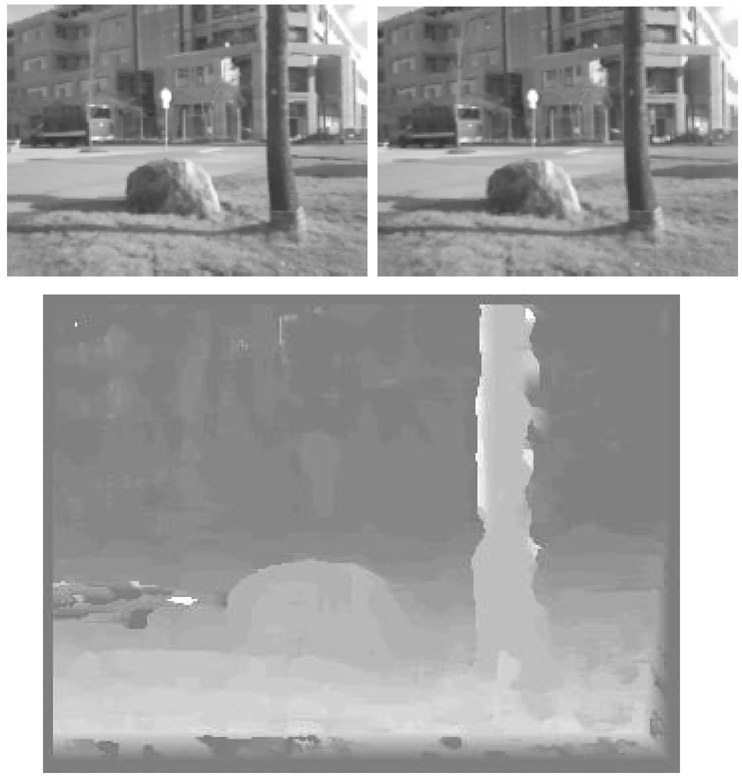
It turns out that the position of the corresponding 3-D point $\mathbf{P}$ can be computed from $\mathbf{p}_l$ and d, assuming that the camera parameters are known.



$$d = x_l - x_r$$

Adapted from Martial Hebert, CMU

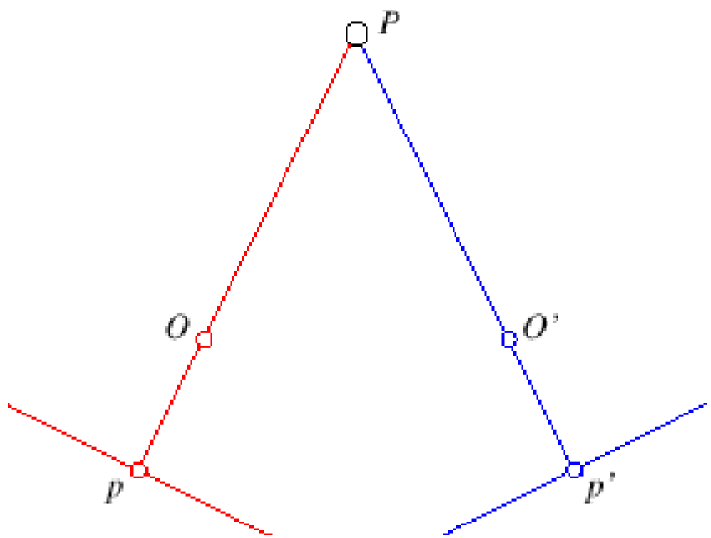
Disparity

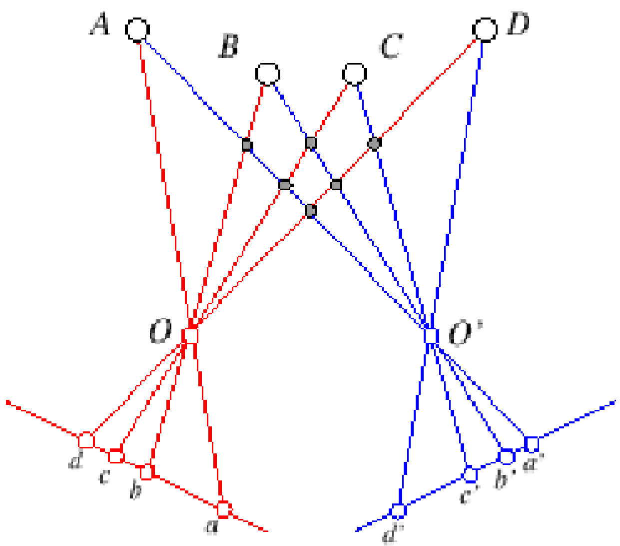


Larger disparity → closer to camera

Adapted from Martial Hebert, CMU

Stereopsis



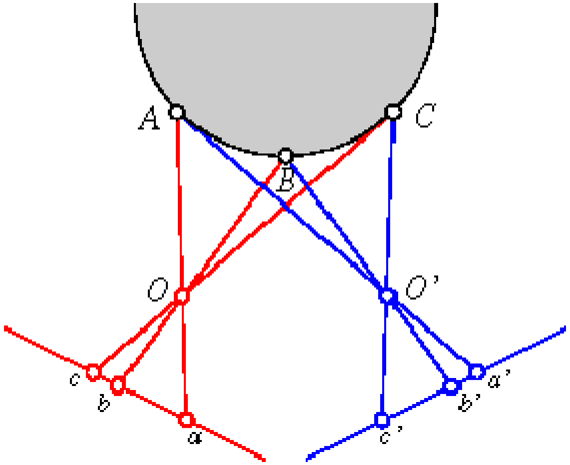


If a single image point is observed at any given time Stereo vision is easy

However, each picture consists of hundreds/thousands of pixels, therefore it is very hard to find the correct correspondences

Adapted from David Forsyth, UC Berkeley

Ordering constraint

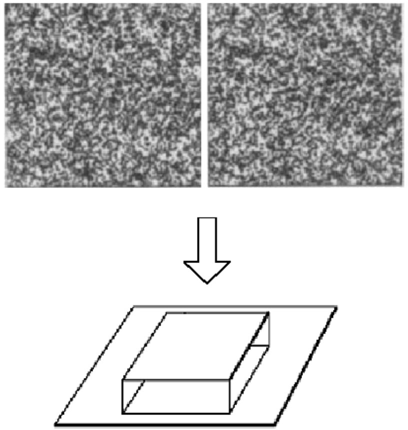


"It is reasonable to assume that the order of matching image features along a pair of epipolar lines is the inverse of the order of the corresponding surface attributes along the curve where the epipolar plane intersects the observed object's boundary."

This is the so-called *ordering constraint* introduced by [Baker and Binford, 1981; Ohta and Kanade, 1985].

Adapted from Trevor Darrell, MIT

Correspondence is ambiguous (Marr & Poggio)



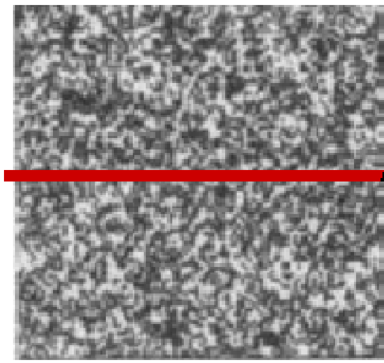
Three constraints :

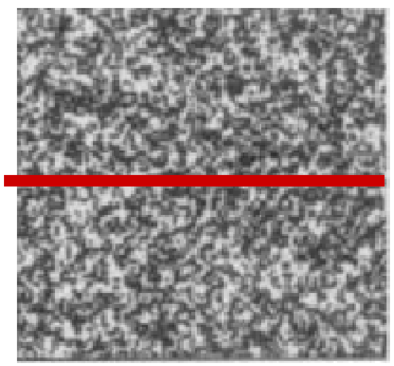
Compatibility : black dots can only match black dots , or more generally, two image features can only match if they have possibly arisen from the same physical marking

Uniqness : a black dot in one image matches at most one black dot in another image

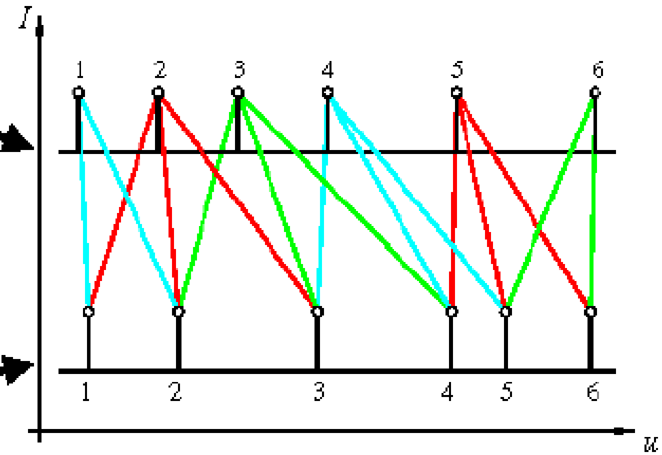
Continuity : the disparity of matches varies smoothly almost everywhere in the image

Correspondence is ambiguous



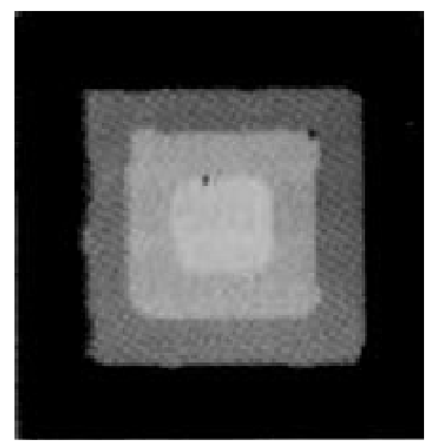


Marr-Poggio (1976)



Three constraints:

- ***compatibility***
- ***uniqueness***
- ***continuity***



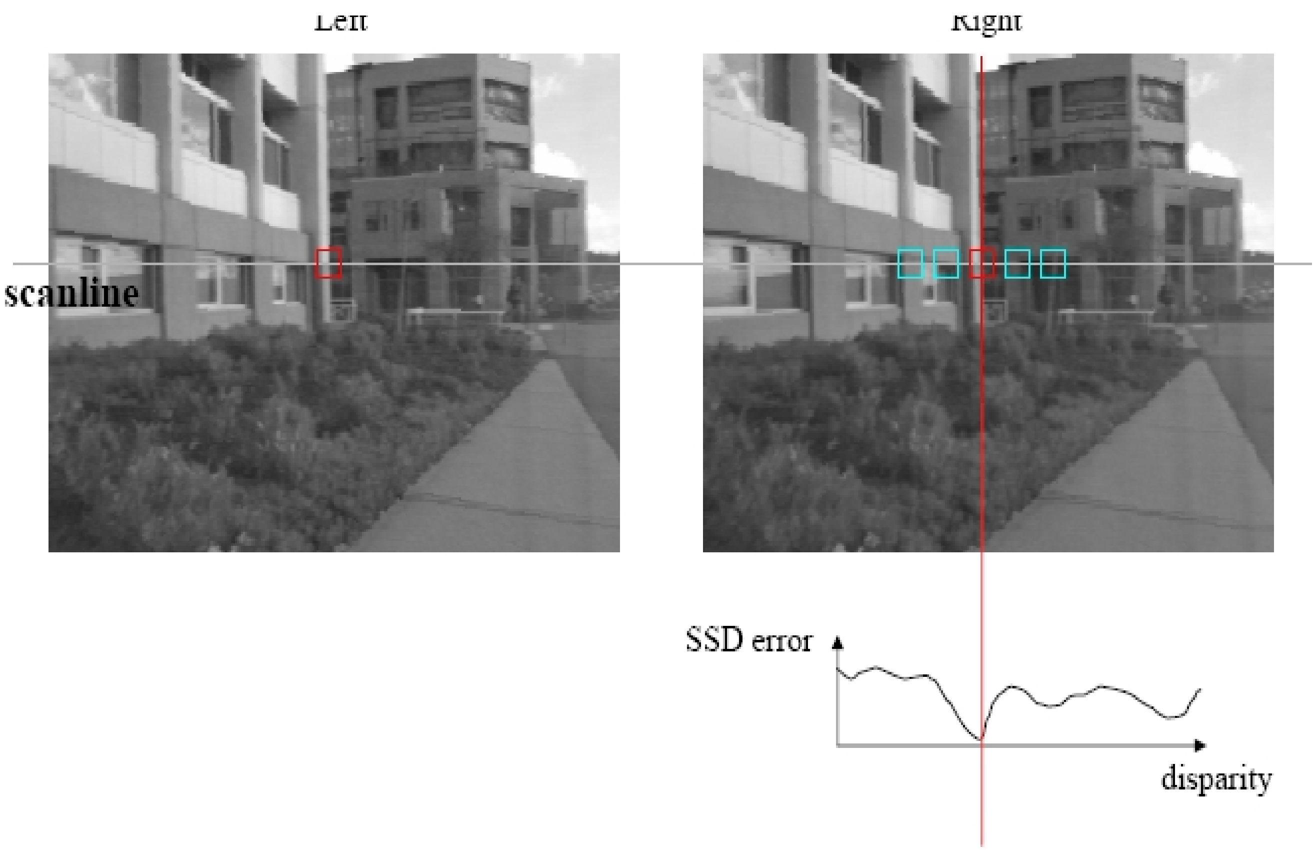
Works well on RDS….but not so well on natural images…

Adapted from Trevor Darrell, MIT

Correspondence using window matching

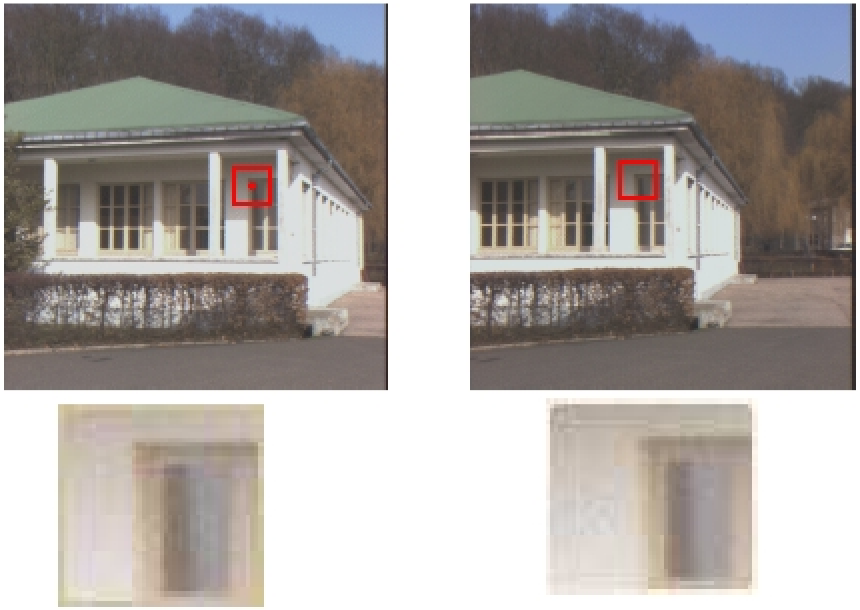
Points are highly individually ambiguous…

More unique matches are possible with small regions of image.



Adapted from Michael Black

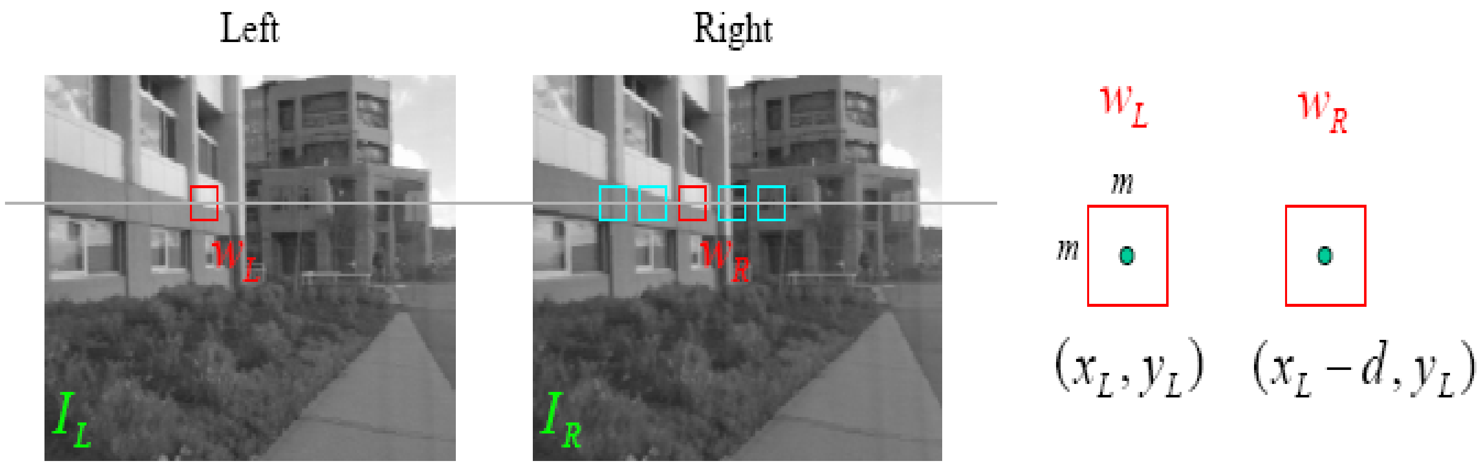
Finding Correspondences



$W(\mathbf{p}_l)$ $W(\mathbf{p}_r)$

Adapted from Martial Hebert, CMU

Sum of squared distances



w_L and w_R are corresponding m by m windows of pixels.

The SSD cost measures the intensity difference as a function of disparity :

$$SSD_r(x, y, d) = \sum_{(x',y') \in W_m(x,y)} (I_L(x', y') - I_R(x'-d, y'))^2$$

Adapted from Michael Black

Image Normalization

- Even when the cameras are identical models, there can be differences in gain and sensitivity.
- The cameras do not see exactly the same surfaces, so their overall light levels can differ.
- For these reasons and more, it is a good idea to normalize the pixels in each window:

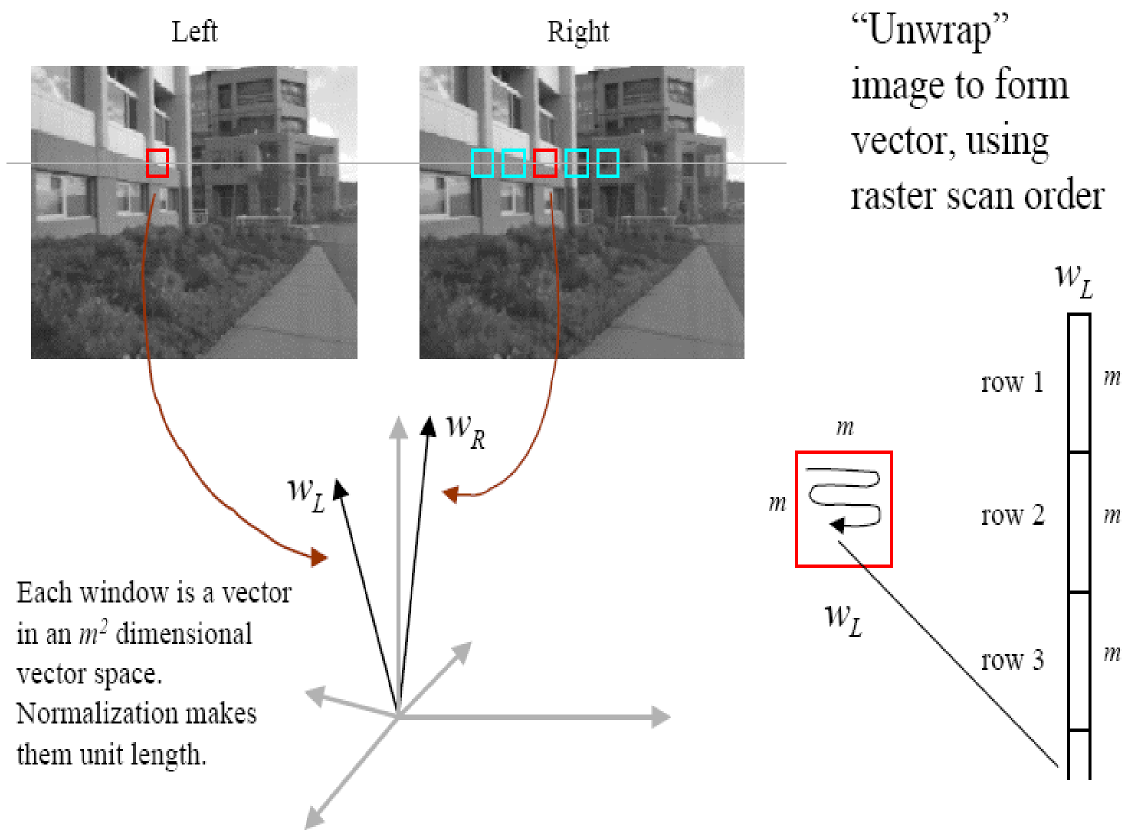
$$\bar{I} = \frac{1}{|W_m(x,y)|} \sum_{(u,v) \in W_m(x,y)} I(u,v)$$ Average pixel

$$\|I\|_{W_m(x,y)} = \sqrt{\sum_{(u,v) \in W_m(x,y)} [I(u,v)]^2}$$ Window magnitude

$$\hat{I}(x,y) = \frac{I(x,y) - \bar{I}}{\|I - \bar{I}\|_{W_m(x,y)}}$$ Normalized pixel

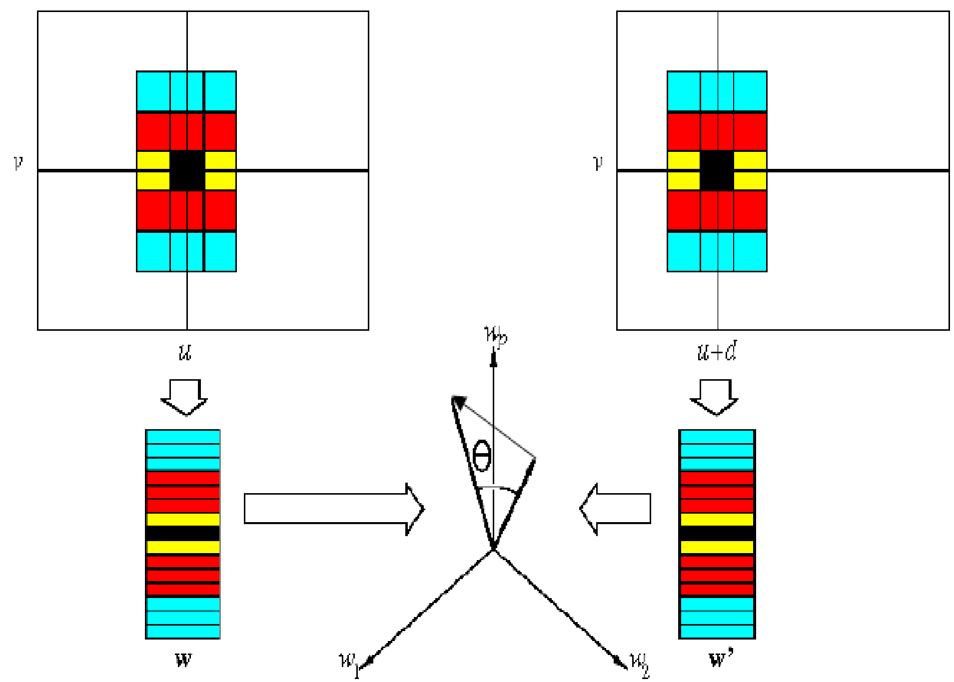
Adapted from Darrell

Images as vectors



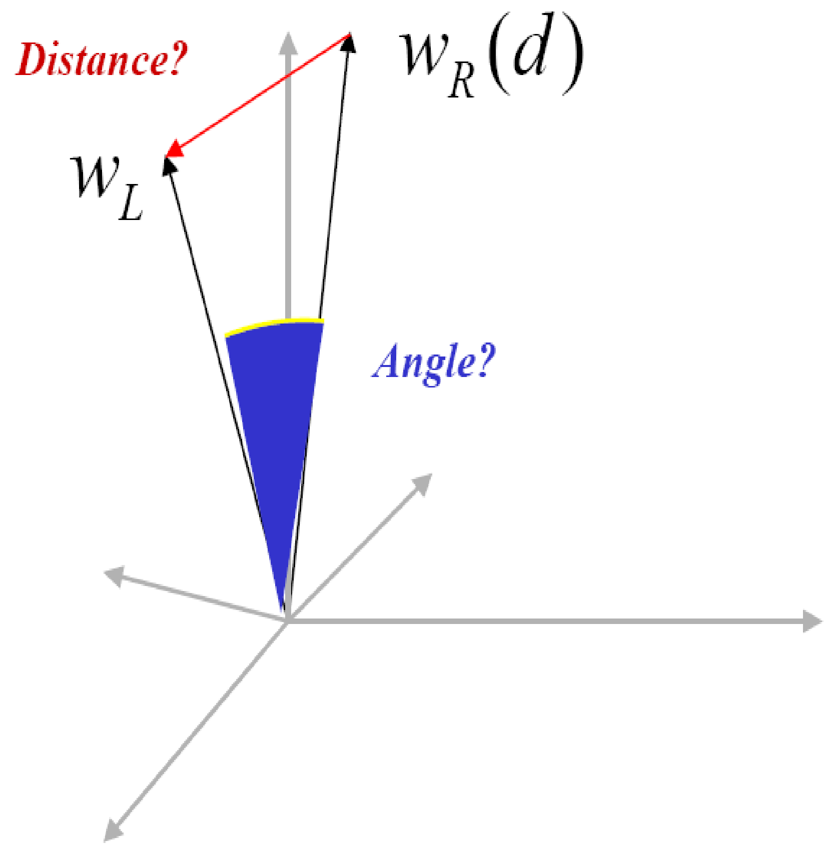
Adapted from Darrell

Images as vectors



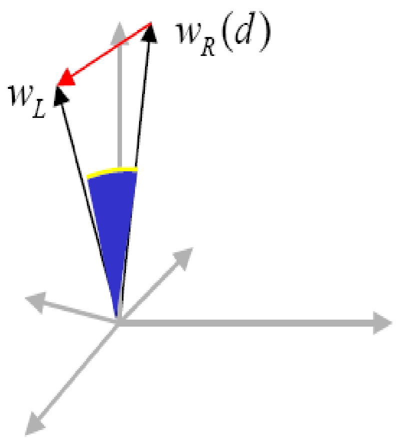
Adapted from Darrell

Possible metric



Adapted from Darrell

Possible metric



(Normalized) Sum of Squared Differences

$$C_{\mathrm{SSD}}(d) = \sum_{(u,v)\in W_m(x,y)} [\hat{I}_L(u,v) - \hat{I}_R(u-d,v)]^2$$

$$= \left\| w_L - w_R(d) \right\|^2$$

Normalized Correlation

$$C_{\mathrm{NC}}(d) = \sum_{(u,v)\in W_m(x,y)} \hat{I}_L(u,v)\hat{I}_R(u-d,v)$$

$$= w_L \cdot w_R(d) = \cos\theta$$

$$\boxed{d^* = \arg\min_d \left\| w_L - w_R(d) \right\|^2 = \arg\max_d w_L \cdot w_R(d)}$$

Adapted from Darrell

Matching using correlation

Left

Disparity Map

Images courtesy of Point Grey Research

Adapted from Michael Black

Matching using correlation

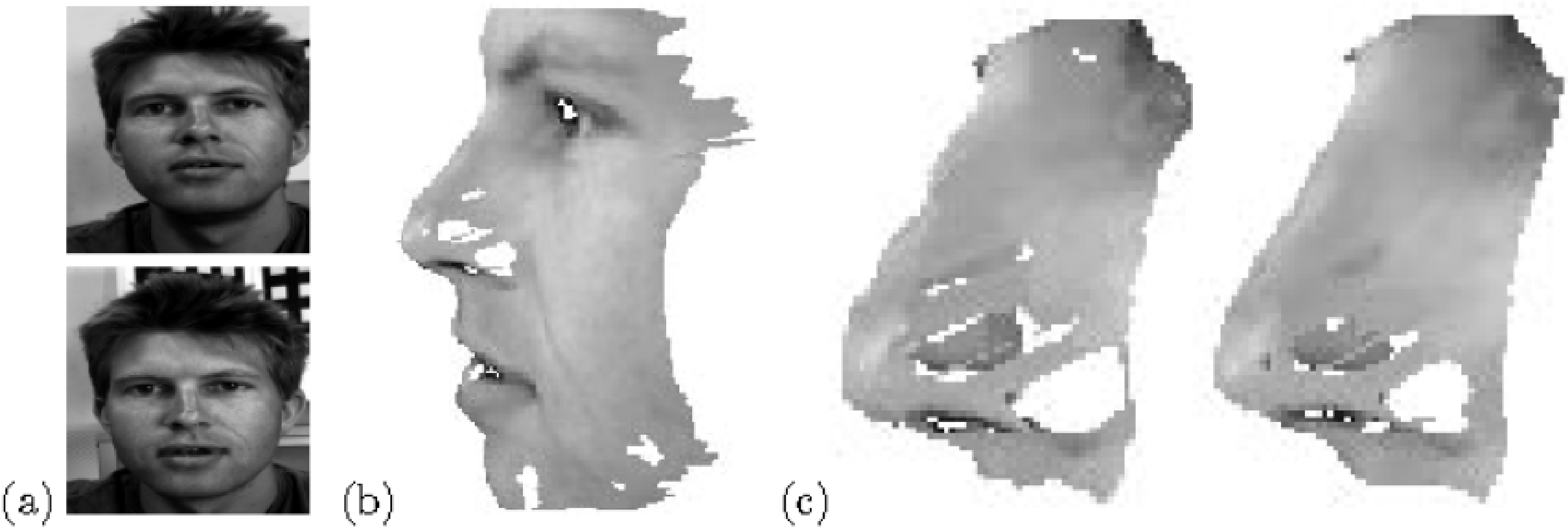


FIGURE 12.13: Correlation-based stereo matching: (a) a pair of stereo pictures; (b) a texture-mapped view of the reconstructed surface; (c) comparison of the regular (left) and refined (right) correlation methods in the nose region. Reprinted from [Devernay and Faugeras, 1994], Figures 5, 8 and 9.

Adapted from Darrell

Problems with window methods

Patch too small?

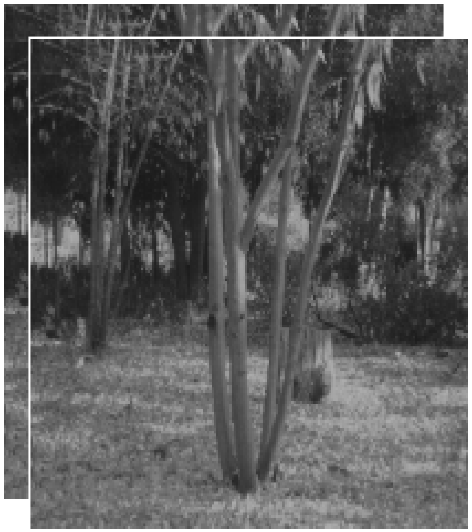
Patch too large?

Can try variable patch size [Okutomi and Kanade], or arbitrary window shapes [Veksler and Zabih]

Should match between physically meaningful quanties, and at multiple scales [Marr]…

Adapted from Trevor Darrell, MIT

Window size



W = 3

W = 20

Better results with *adaptive window*

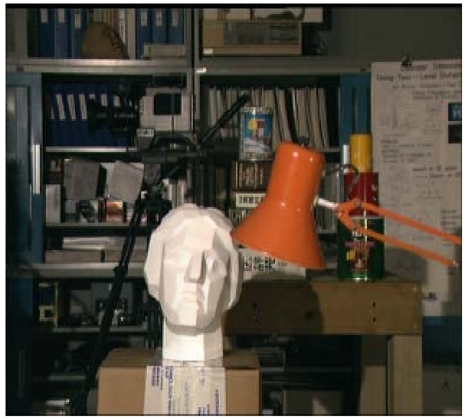
- T. Kanade and M. Okutomi, A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment,, Proc. International Conference on Robotics and Automation, 1991.
- D. Scharstein and R. Szeliski. Stereo matching with nonlinear diffusion. International Journal of Computer Vision, 28(2):155-174, July 1998

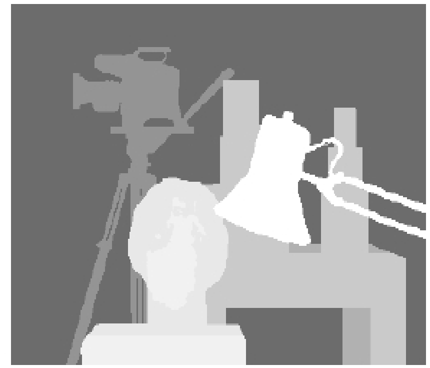
(Seitz)

Adapted from Michael Black

Stereo Results

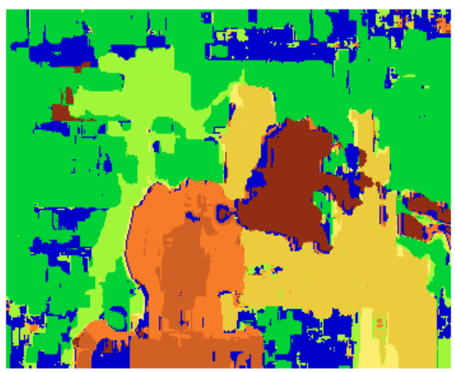
– Data from University of Tsukuba





Scene

Ground truth



Window-based matching
(best window size)

Ground truth

(Seitz)

Adapted from Michael Black

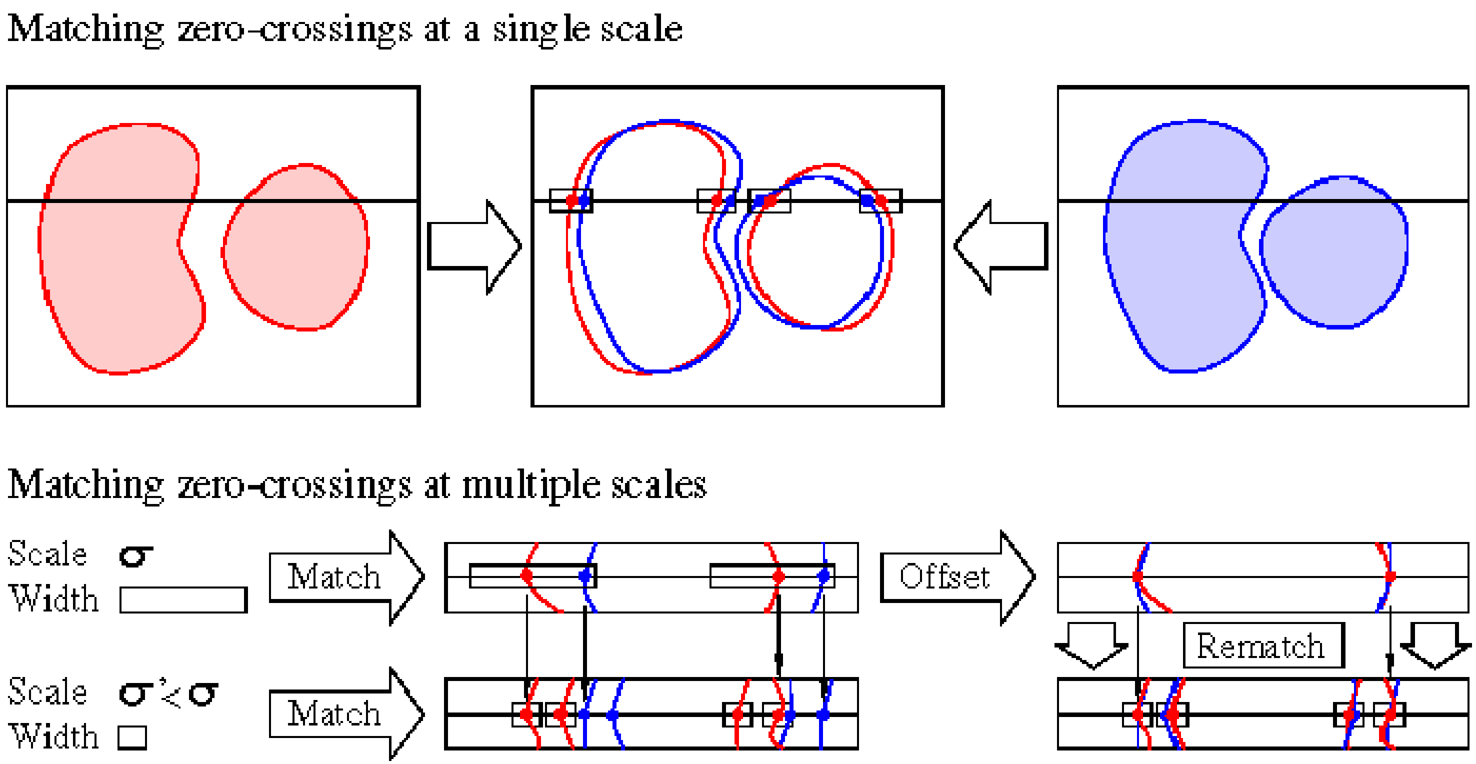
Multi – scale edge matching

1. Convolve the two (rectified) images with $\nabla^2 G_\sigma$ filters of increasing standard deviations $\sigma_1 < \sigma_2 < \sigma_3 < \sigma_4$.
2. Find zero crossings of the Laplacian along horizontal scanlines of the filtered images.
3. For each filter scale σ, match zero crossings with the same parity and roughly equal orientations in a $[-w_\sigma, +w_\sigma]$ disparity range, with $w_\sigma = 2\sqrt{2}\sigma$.
4. Use the disparities found at larger scales to control eye vergence and cause unmatched regions at smaller scales to come into correspondence.

Forsyth & Ponce

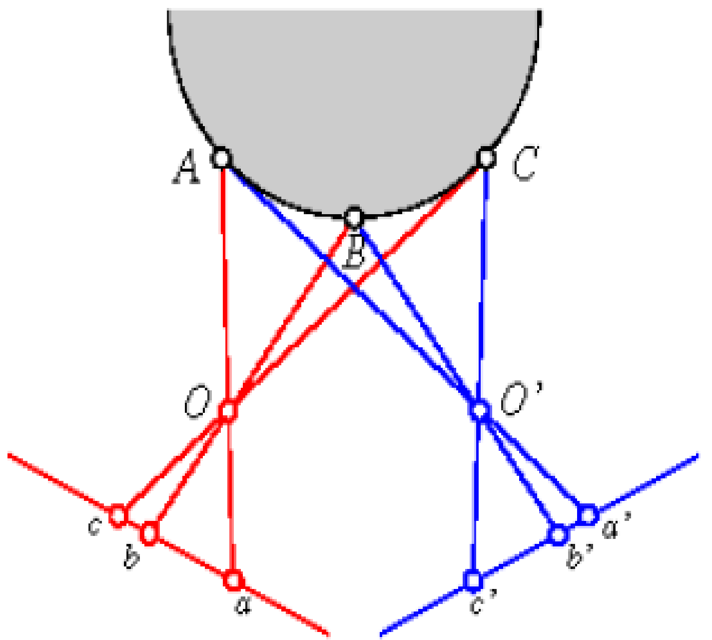
Marr-Poggio Algorithm

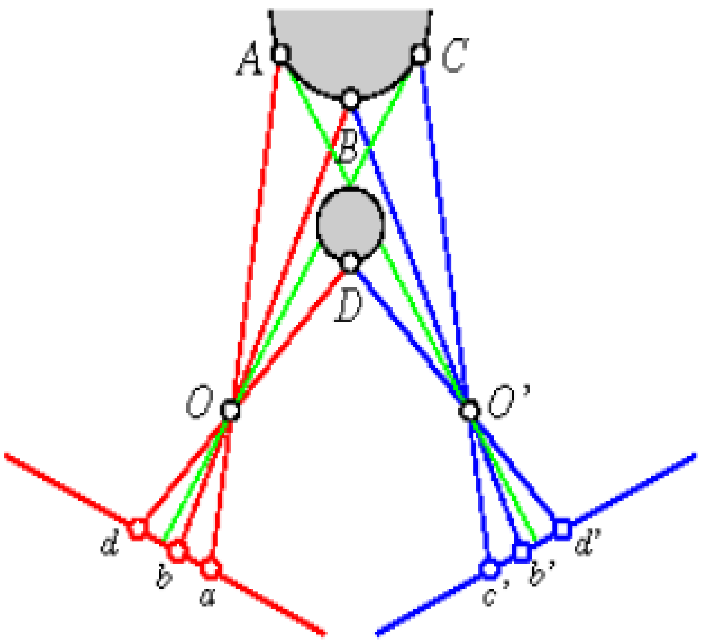
Search for edges, a.k.a. “zero crossings”: (more during edge detection lectures…)



Adapted from Trevor Darrell, MIT

Correspondence

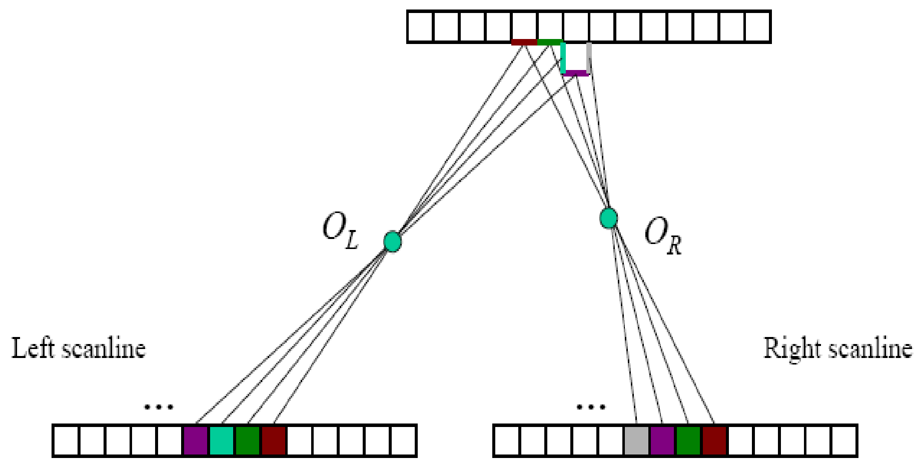




Oops!

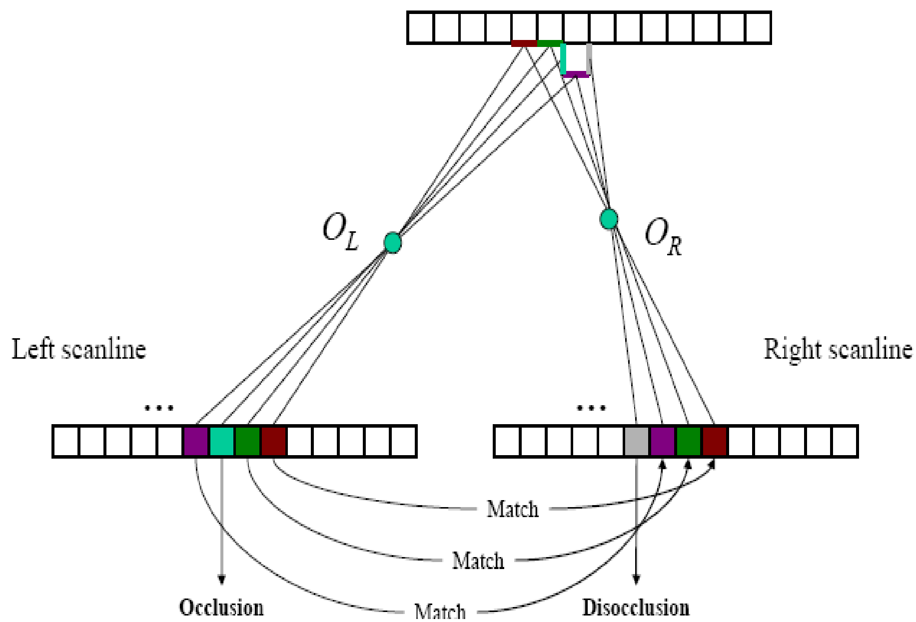
Adapted from Darrel & Freman

Correspondence



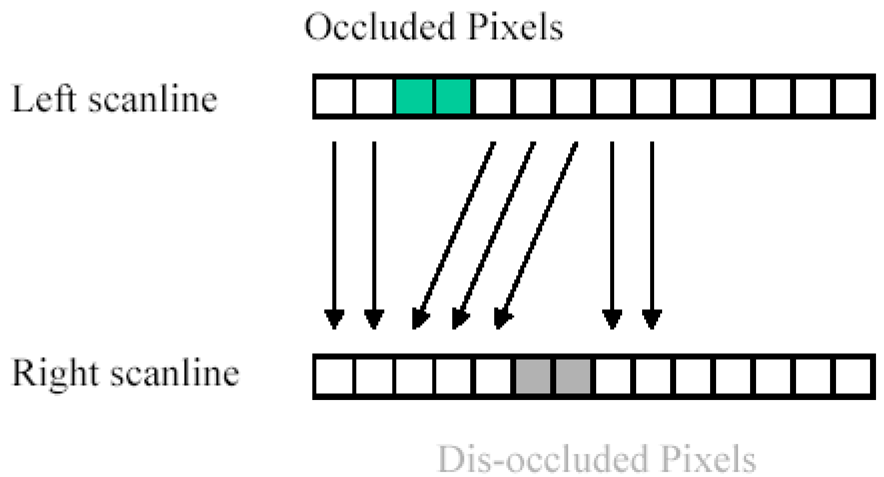
Adapted from Michael Black

Correspondence



Adapted from Michael Black

Search over correspondences

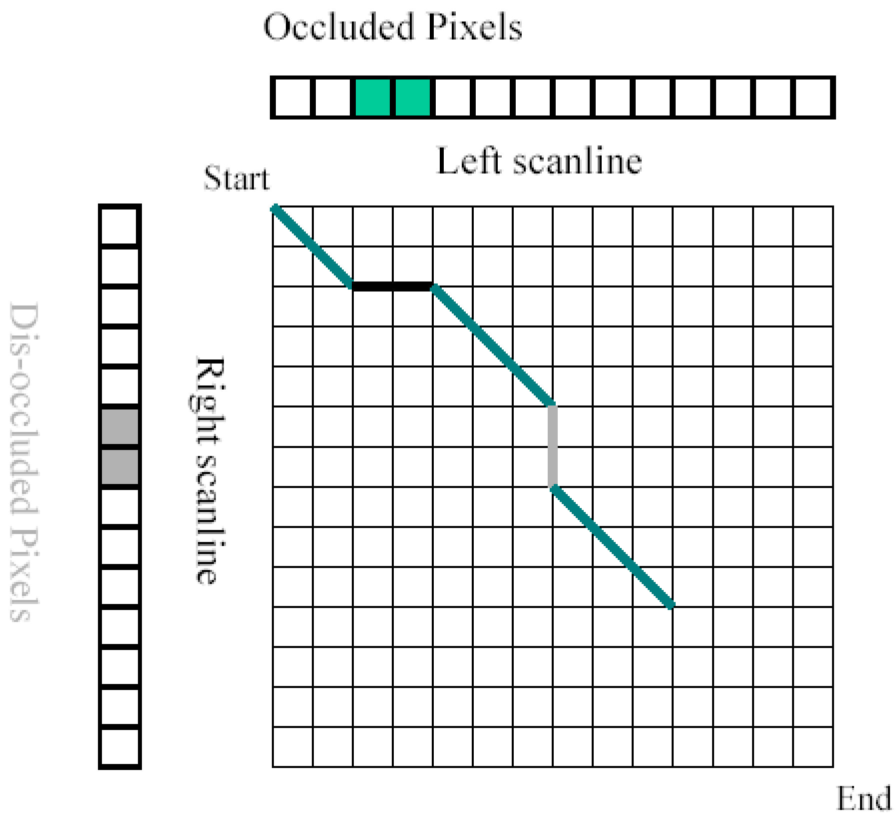


Three cases:

- Sequential – cost of match
- Occluded – cost of no match
- Disoccluded – cost of no match

Adapted from Trevor Darrell, MIT

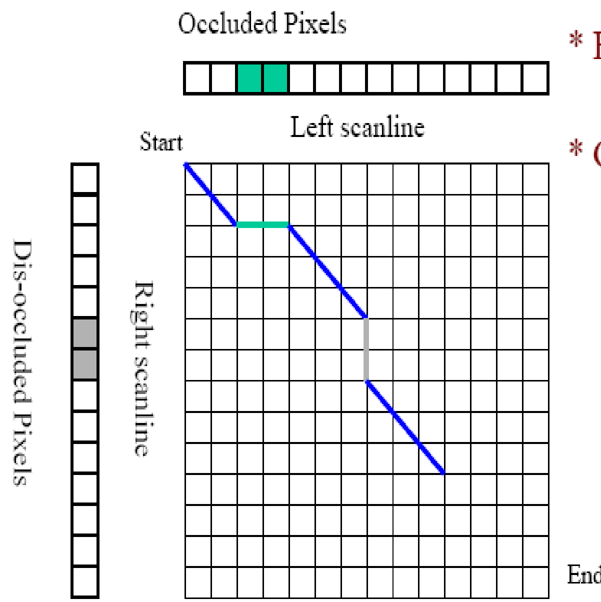
Dynamic programming



Dynamic programming yields the optimal path through grid. This is the best set of matches that satisfy the ordering constraint

Adapted from Trevor Darrell, MIT

Stereo Matching with Dynamic Programming



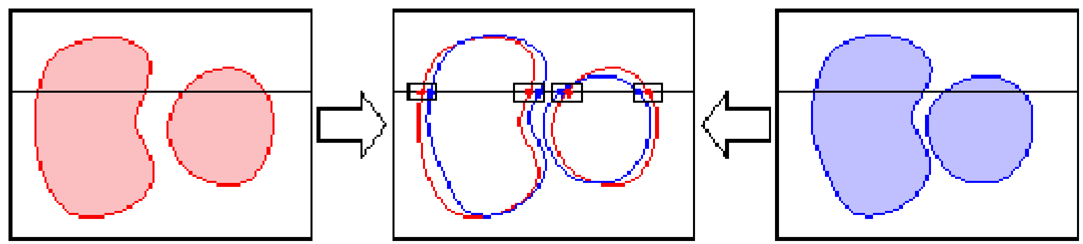
* Enforces ordering constraint.

* Given appropriate cost functions, solves for best path (matches, occlusions, disocclusions).

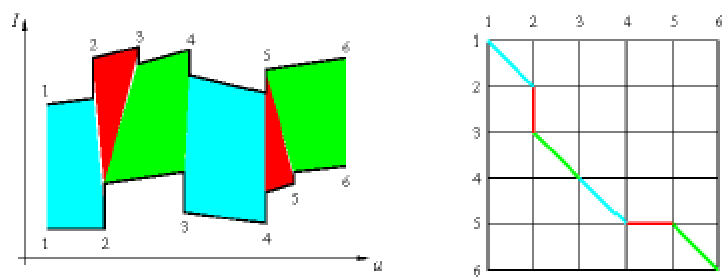
Adapted from Michael Black

DP vs. edges

Edges:



DP:

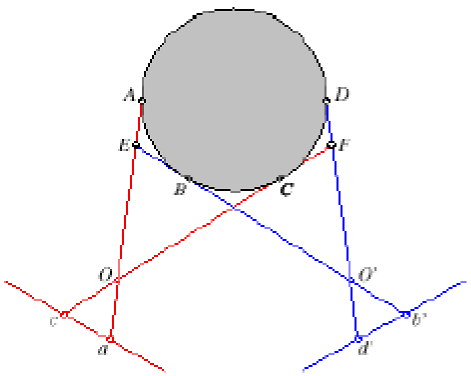


- Which method is better?
  - Edges are more “meaningful” [Marr]…but hard to find!
  - Edges tend to fail in dense texture (outdoors)
  - Correlation tends to fail in smooth featureless areas

Adapted from Trevor Darrell, MIT

Computing correspondences

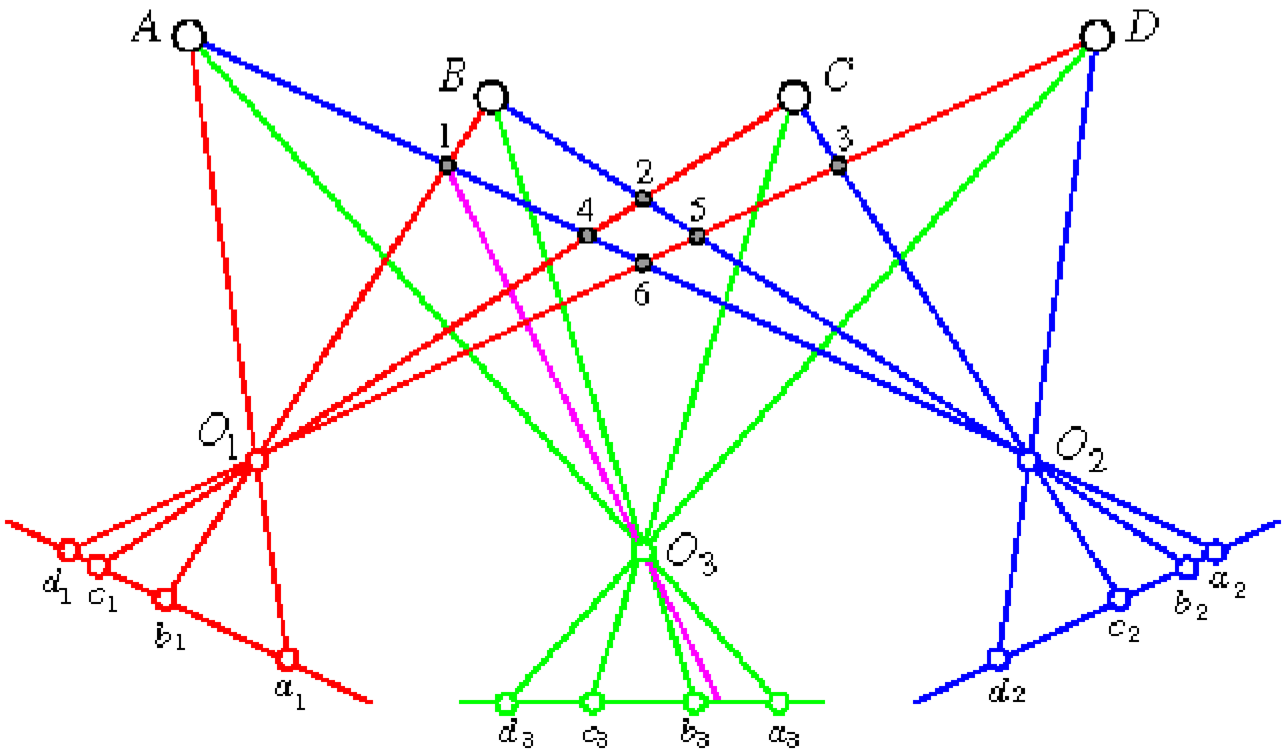
Both methods fail for smooth surfaces



There is currently no good solution to the correspondence problem

Adapted from Trevor Darrell, MIT

Three (calibrated) views



Adding a third camera eliminates the ambiguity inherent in two-view point matching

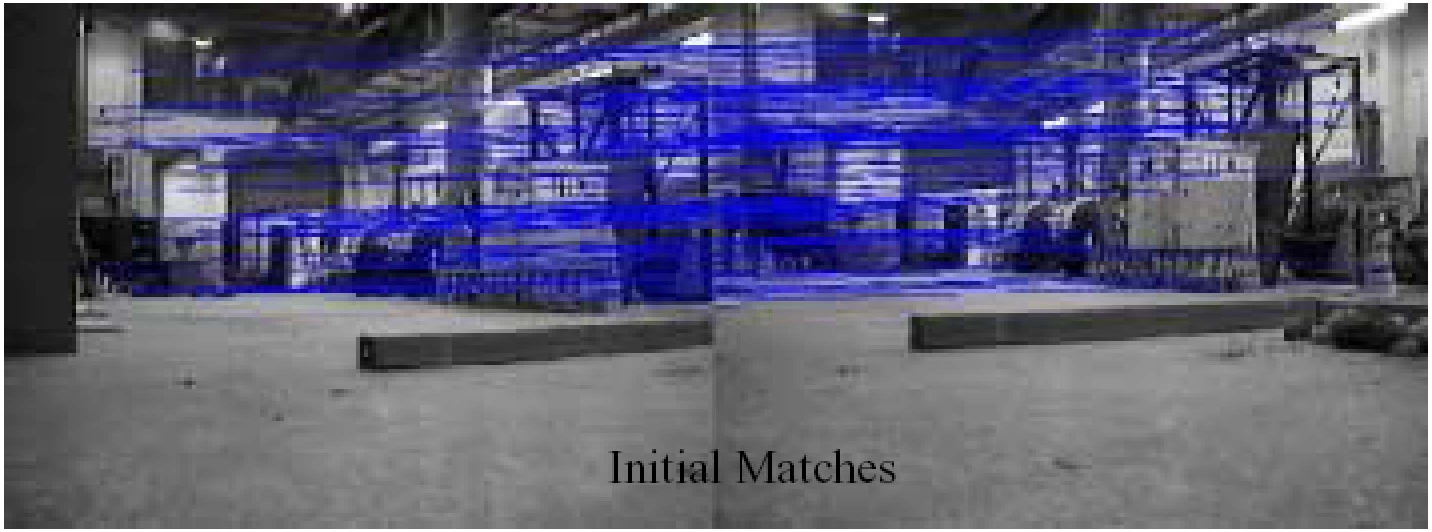
Adapted from Trevor Darrell, MIT

RANSAC

- Do k times:
  - Draw set of 8 correspondences
  - Fit $\boldsymbol{F}$ to the set
  - Count the number d of correspondences that are closer than t to the fitted epipolar lines
  - If $d > d_{\min}$, recompute fit error using all the correspondences
- Return best fit found

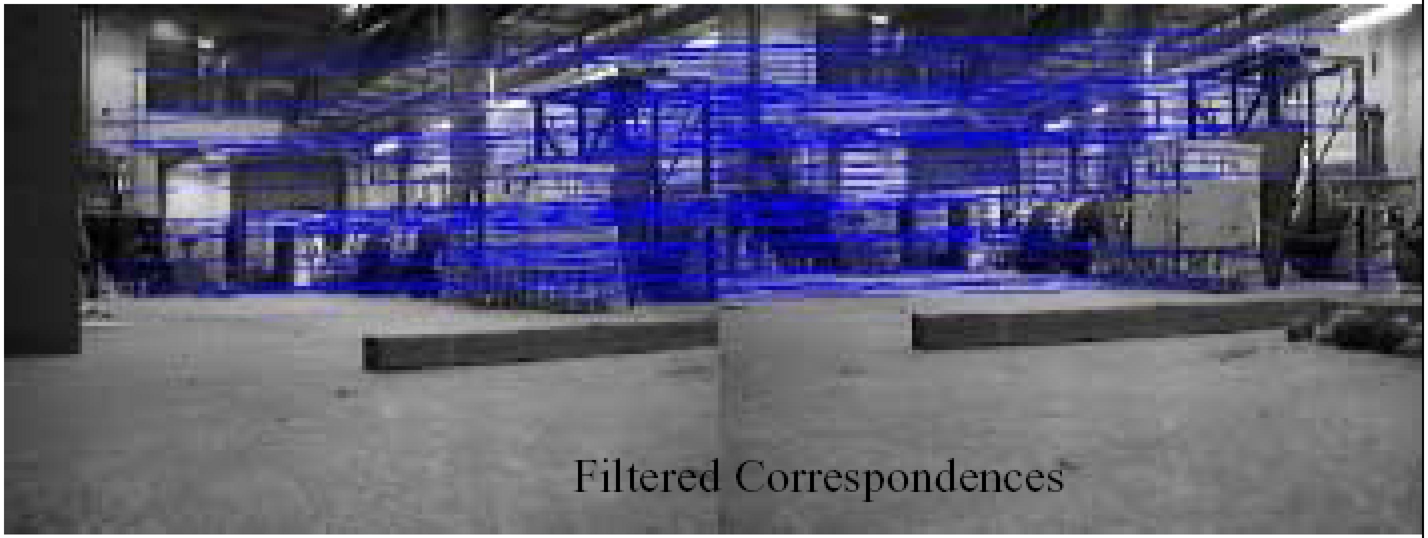
Adapted from Martial Hebert, CMU

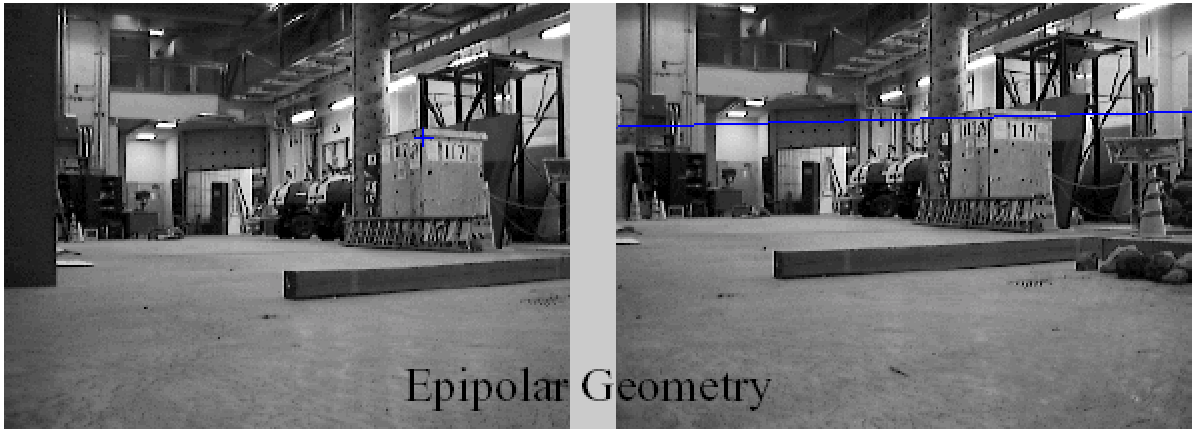
RANSAC



Adapted from Martial Hebert, CMU

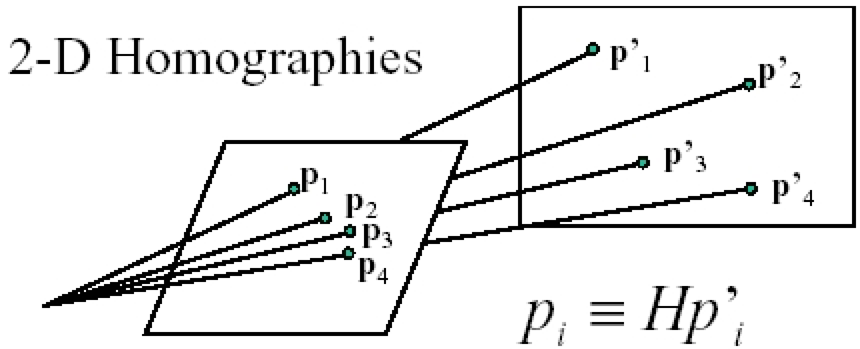
RANSAC

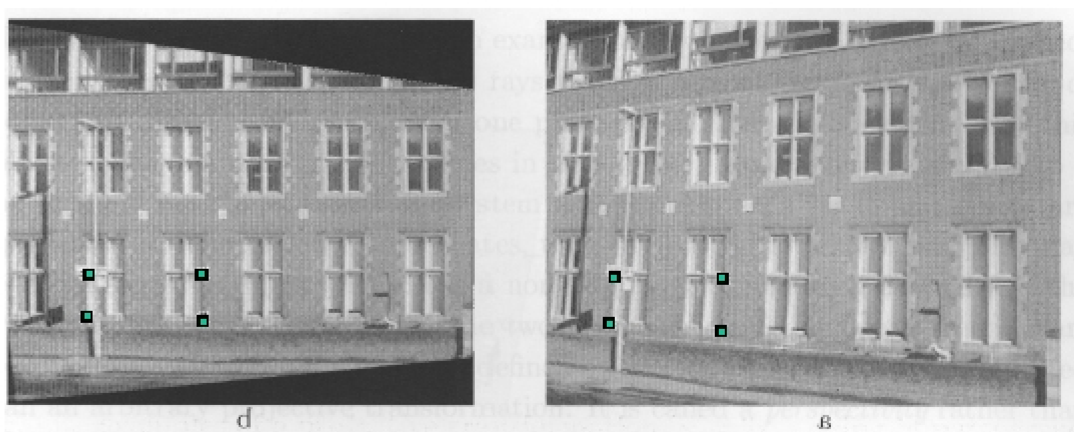




Adapted from Martial Hebert, CMU

2D homographies

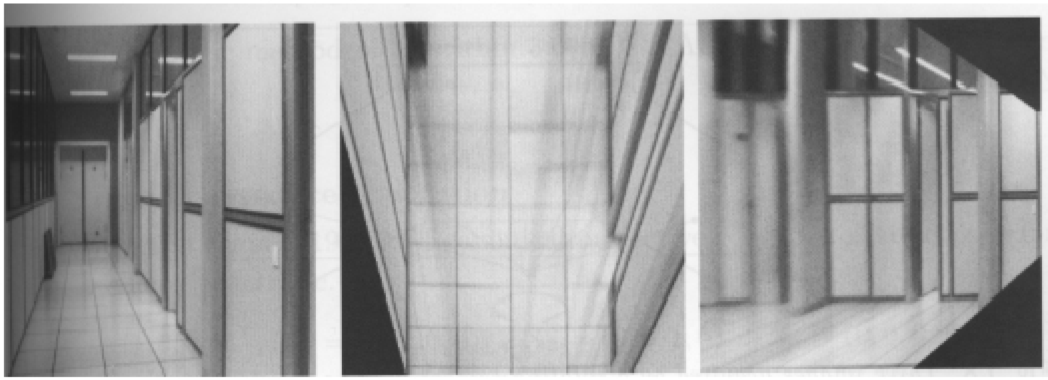


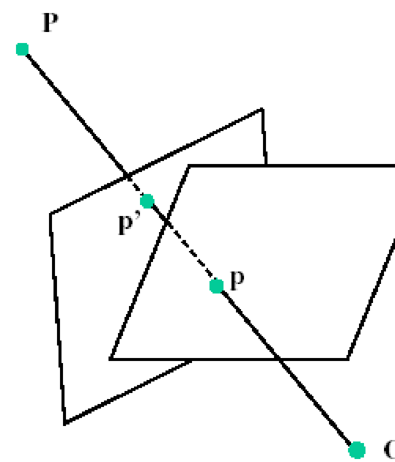


2D homographies transforms points from one plane to another

Adapted from Martial Hebert, CMU

2D homographies

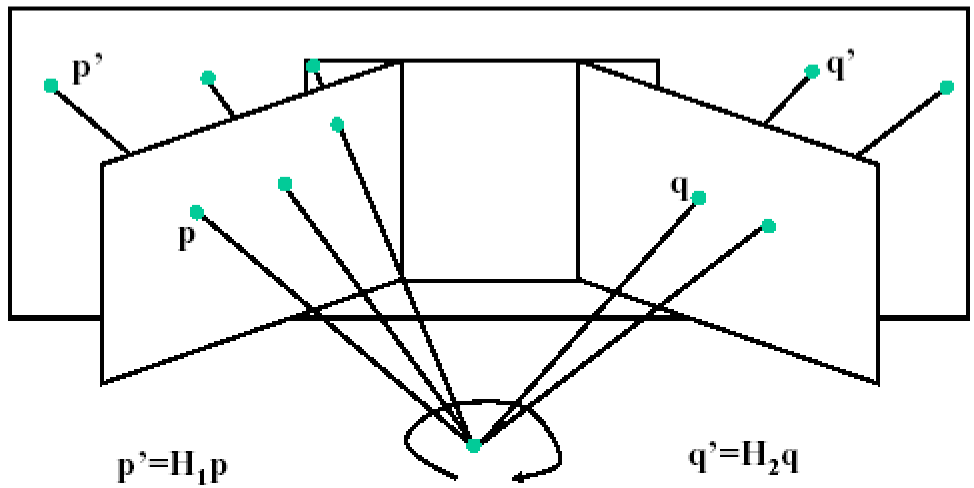




$p' \equiv Hp$

Adapted from Martial Hebert, CMU

2D homographies



If we choose the plane of one of the images from a set of images obtained by rotating the camera around the optical center, for each image, there exists an homography which maps the point p in the original image plane to the reference image plane. If we map all the points from all the images into the reference image plane, we obtain a single image, a mosaic, which contains the data from all the input images.

Adapted from Martial Hebert, CMU

2D homographies

Adapted from Martial Hebert, CMU