3D Computer Vision

Radim Šára   Martin Matoušek

Center for Machine Perception
Department of Cybernetics
Faculty of Electrical Engineering
Czech Technical University in Prague

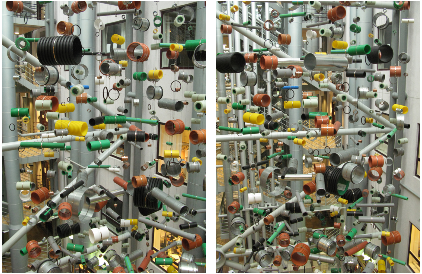
https://cw.fel.cvut.cz/wiki/courses/tdv/start

http://cmp.felk.cvut.cz
mailto:sara@cmp.felk.cvut.cz
phone ext. 7203

rev. December 6, 2022

Open Informatics Master's Course

How Difficult Is Stereo?



Centrum för teknikstudier at Malmö Högskola, Sweden

The Vyšehrad Fortress, Prague

- top: easy interpretation from even a single image
- bottom left: we have no help from image interpretation
- bottom right: ambiguous interpretation due to a combination of missing texture and occlusion

A Summary of Our Observations and an Outlook

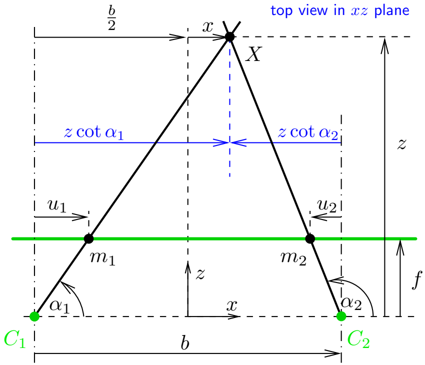
1. simple matching algorithms do not work
   - the success of a model-free stereo matching is unlikely →153
   - without scene recognition or use high-level constraints the problem seems difficult
2. stereopsis requires image interpretation in sufficiently complex scenes — or another-modality measurement

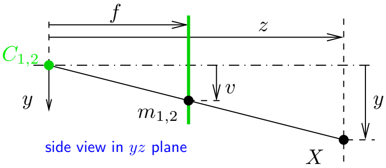
we have a tradeoff: model strength ↔ universality

Outlook:

1. represent the occlusion constraint: — correspondences are not independent due to occlusions
   - disparity in rectified images
   - uniqueness as an occlusion constraint
2. represent piecewise continuity — the weakest of interpretations; piecewise: object boundaries
   - ordering as a weak continuity model
3. use a consistent framework
   - finding the most probable solution (MAP)

▶Binocular Disparity in a Standard Stereo Pair





- Assumptions: single image line, standard camera pair

$$b = z \cot \alpha_1 - z \cot \alpha_2 \qquad b = \frac{b}{2} + x - z \cot \alpha_2$$

$$u_1 = f \cot \alpha_1 \qquad u_2 = f \cot \alpha_2$$

- eliminate α_1, α_2 and obtain:
  $X = (x, y, z)$ from **disparity** $d = u_1 - u_2$:

$$\boxed{z = \frac{b\,f}{d}}, \quad x = \frac{b}{d}\,\frac{u_1 + u_2}{2}, \quad y = \frac{b\,v}{d}$$

f, d, u, v in pixels, b, x, y, z in meters

Observations

- constant disparity surface is a frontoparallel plane
- distant points have small disparity
- relative error in z is large for small disparity

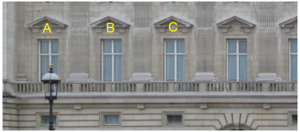
$$\frac{1}{z}\,\frac{\mathrm{d}z}{\mathrm{d}d} = -\frac{1}{d}$$

- increasing the baseline or the focal length increases disparity and reduces the error

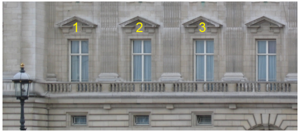
Structural Ambiguity in Stereovision

- suppose we can recognize local matches independently but have no scene model
- lack of an occlusion model
- lack of a continuity model

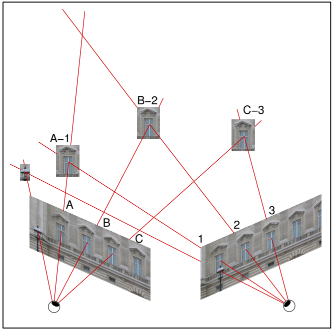
$\Rightarrow$ structural ambiguity in the presence of repetitions (or lack of texture)



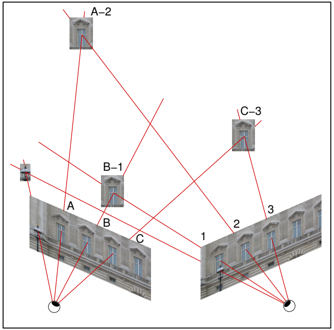
left image



right image

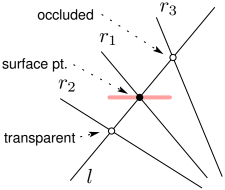


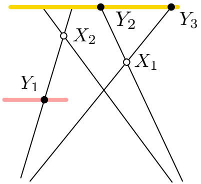
interpretation 1



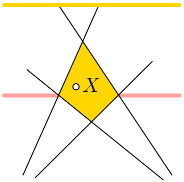
interpretation 2

▶Understanding Basic Occlusion Types





half occlusion



mutual occlusion

- surface point at the intersection of rays l and r_1 occludes a world point at the intersection (l, r_3) and implies the world point (l, r_2) is transparent, therefore

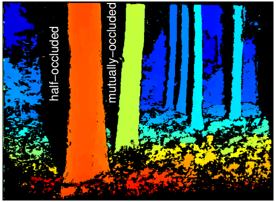
$$(l, r_3) \text{ and } (l, r_2) \text{ are } \underline{\text{excluded}} \text{ by } (l, r_1)$$

- in half-occlusion, every 3D point such as X_1 or X_2 is excluded by a binocularly visible surface point such as Y_1, Y_2, Y_3

  ⇒ decisions on correspondences are not independent

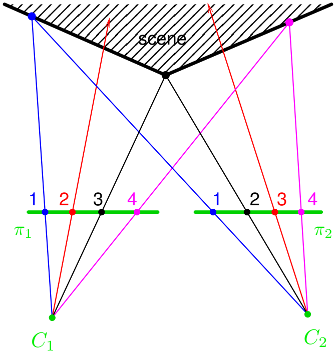
- in mutual occlusion this is no longer the case: any X in the yellow zone above is not excluded

  ⇒ decisions inside the zone are independent on the rest

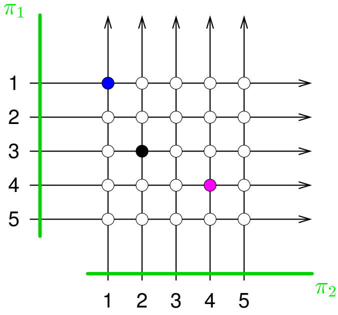


▶Matching Table

Based on scene opacity and the observation on mutual exclusion we expect each pixel to match at most once.



rays in epipolar plane



matching table T

matching table

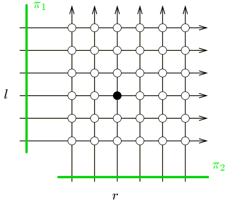
- rows and columns represent optical rays
- nodes: possible correspondence pairs
- full nodes: matches
- numerical values associated with nodes: descriptor similarities

see next

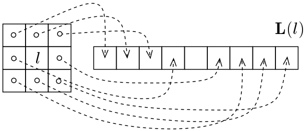
►Constructing An Image Similarity Cost

- let $p_i = (l, r)$ and $\mathbf{L}(l)$, $\mathbf{R}(r)$ be (left, right) image descriptors (vectors) constructed from local image neighborhood windows

in matching table T:



'block' in the left image $\mapsto$ 'a set of random-variable samples':



- a simple block (dis-)similarity is $\mathrm{SAD}(l, r) = \|\mathbf{L}(l) - \mathbf{R}(r)\|_1$ $\quad$ L_1 metric (sum of absolute differences; smaller is better)
- a scaled-descriptor (dis-)similarity is $\quad \mathrm{sim}(l, r) = \dfrac{\|\mathbf{L}(l) - \mathbf{R}(r)\|^2}{\sigma_I^2(l, r)}$ $\quad$ smaller is better
- σ_I^2 – the difference scale; a suitable (plug-in) estimate is $\frac{1}{2}\left[\mathrm{var}\big(\mathbf{L}(l)\big) + \mathrm{var}\big(\mathbf{R}(r)\big)\right]$, giving

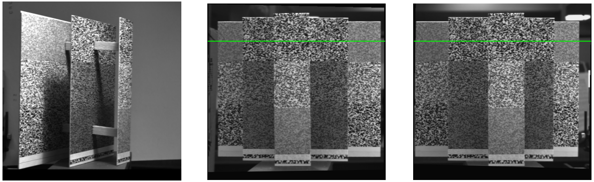
$$\mathrm{sim}(l, r) = 1 - \underbrace{\frac{2\,\mathrm{cov}\big(\mathbf{L}(l), \mathbf{R}(r)\big)}{\mathrm{var}\big(\mathbf{L}(l)\big) + \mathrm{var}\big(\mathbf{R}(r)\big)}}_{\rho\left(\mathbf{L}(l), \mathbf{R}(r)\right)} \tag{36}$$

$\mathrm{var}(\cdot)$, $\mathrm{cov}(\cdot)$ is sample (co-)variance, not invariant to scale difference

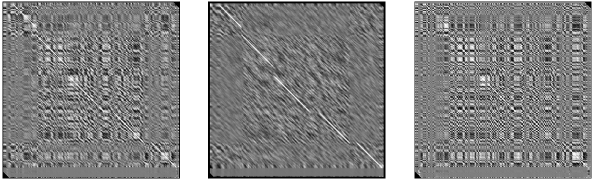
- ρ – MNCC – Moravec's Normalized Cross-Correlation similarity $\quad$ bigger is better $\quad$ [Moravec 1977]

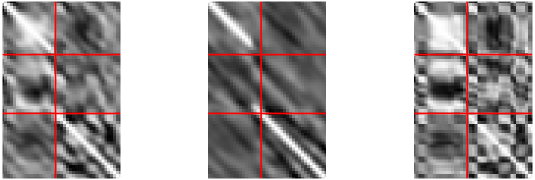
$$\rho^2 \in [0, 1], \qquad \mathrm{sign}\,\rho \sim \text{'phase'}$$

- another successful (dis-)similarity is the Hamming Distance over the Census Transform $\quad$ related to local binary patterns object | left image | right image



5×5 window | 11×11 window | 3×3 window





a good tradeoff | occlusion artefacts | undiscriminable

- MNCC ρ used
  $(\alpha = 1.5,\ \beta = 1)$ →176
- high-similarity structures correspond to scene objects

Things to notice:

constant disparity

- a diagonal in matching table
- zero disparity is the main diagonal
  assuming standard stereopair

depth discontinuity

- horizontal or vertical jump in matching table

large image window

- similarity values have better discriminability
- worse occlusion localization

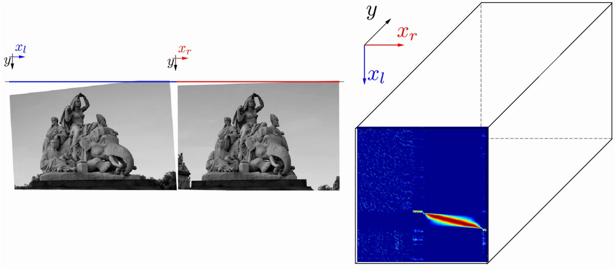
repeated texture

- horizontal and vertical block repetition

Descriptors: Image points are tagged by their (viewpoint-invariant) physical properties:

- texture window [Moravec 77]
- a descriptor like DAISY [Tola et al. 2010]
- learned descriptors
- reflectance profile under a moving illuminant
- photometric ratios [Wolff & Angelopoulou 93-94]
- dual photometric stereo [Ikeuchi 87]
- polarization signature
- . . .

- similar points are more likely to match
- image similarity values for all 'match candidates' give the 3D matching table also called: 'disparity volume'

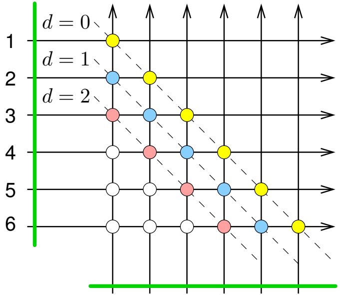


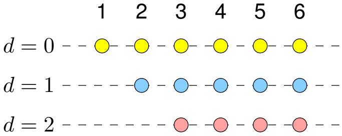
click for video

▶Marroquin's Winner Take All (WTA) Matching Algorithm

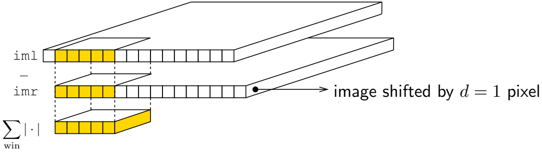
Alg: Per left-image pixel: The most SAD-similar pixel along the right epipolar line →169

1. select disparity range this is a critical weak point
2. represent the matching table diagonals in a compact form





3. use an 'image sliding & cost aggregation algorithm'



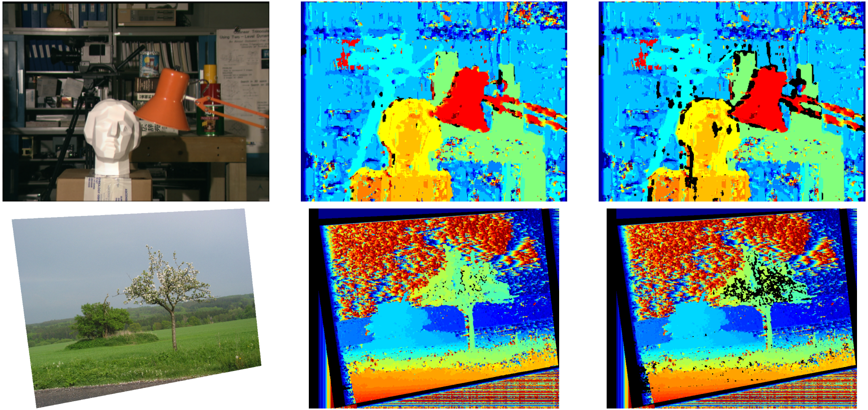
4. take the maximum over disparities d
5. threshold results by maximal allowed SAD dissimilarity

A Matlab Code for WTA

```
function dmap = marroquin(iml, imr, disparityRange)
%       iml, imr - rectified gray-scale images
% disparityRange - non-negative disparity range

% (c) Radim Sara (sara@cmp.felk.cvut.cz) FEE CTU Prague, 10 Dec 12

 thr = 20;                                        % bad match rejection threshold
 r = 2;
 winsize = 2*r+[1 1];                             % 5x5 window (neighborhood) for r=2
 N = boxing(ones(size(iml)), winsize);            % the size of each local patch is
                                                  % N = (2r+1)^2 except for boundary pixels

 % --- compute dissimilarity per pixel and disparity --->

 for d = 0:disparityRange                         % cycle over all disparities
  slice = abs(imr(:,1:end-d) - iml(:,d+1:end));   % pixelwise dissimilarity (unscaled SAD)
  V(:,d+1:end,d+1) = boxing(slice, winsize)./N;   % window aggregation
 end

 % --- collect winners, threshold, output disparity map --->

 [cmap,dmap] = min(V,[],3);                       % collect winners and their dissimilarities
 dmap(cmap > thr) = NaN;                          % mask-out high dissimilarity pixels
end % of marroquin

function c = boxing(im, wsz)
 % if the mex is not found, run this slow version:
 c = conv2(ones(1,wsz(1)), ones(wsz(2),1), im, 'same');
end % of boxing
```

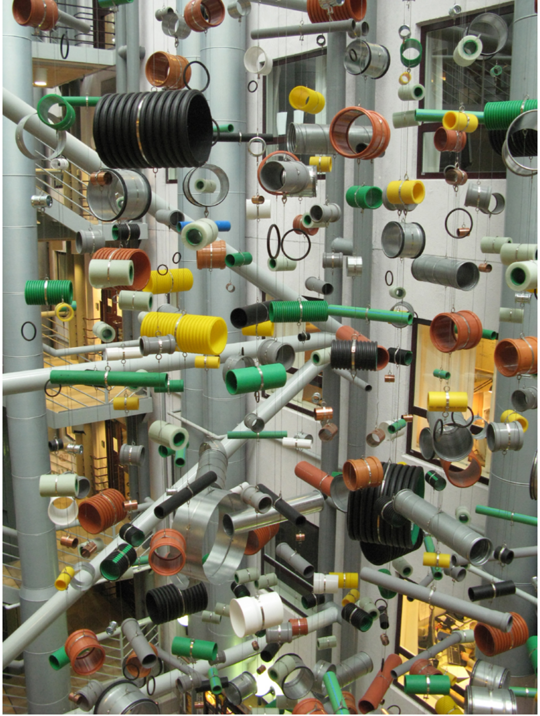
WTA: Some Results

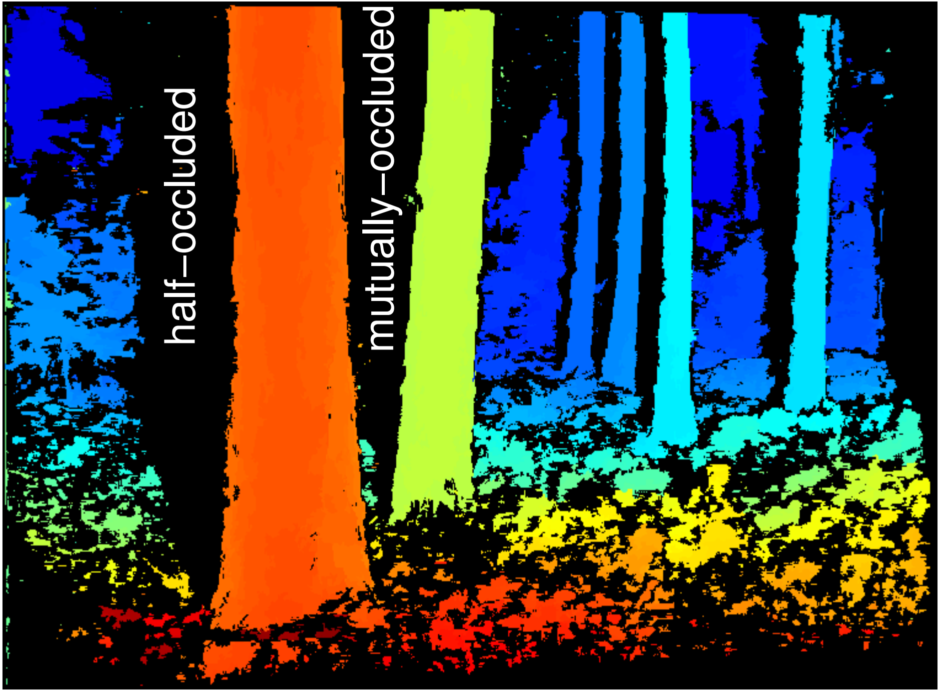
thr $= 20$ thr $= 10$



- results are fairly bad
- false matches in textureless image regions and on repetitive structures (book shelf)
- a more restrictive threshold (thr $= 10$) does not work as expected
- we searched the true disparity range, results get worse if the range is set wider
- chief failure reasons:
  - unnormalized image dissimilarity does not work well
  - no occlusion model (it just ignores the occlusion structure we have discussed $\rightarrow$167)

Thank You




half–occluded
mutually–occluded