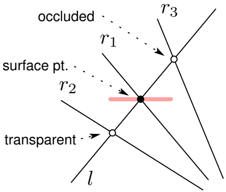
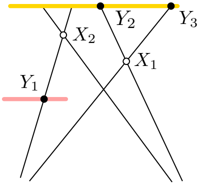
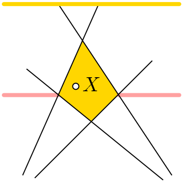
# ▶Understanding Basic Occlusion Types




half occlusion




mutual occlusion

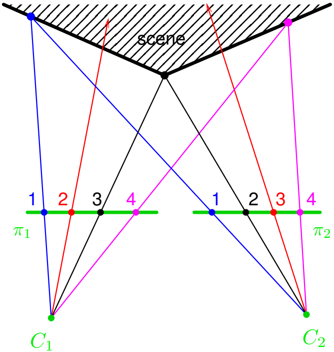
- surface point at the intersection of rays $l$ and $r_1$ occludes a world point at the intersection $(l, r_3)$ and implies the world point $(l, r_2)$ is transparent, therefore

> $(l, r_3)$ and $(l, r_2)$ are excluded by $(l, r_1)$

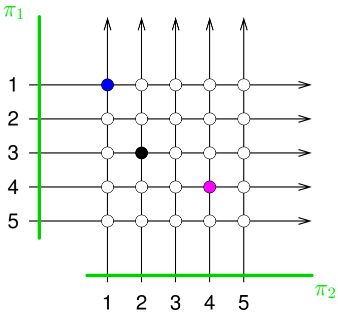
- in half-occlusion, every world point such as $X_1$ or $X_2$ is excluded by a binocularly visible surface point such as $Y_1$, $Y_2$, $Y_3$ $\Rightarrow$ decisions on correspondences are not independent
- in mutual occlusion this is no longer the case: any $X$ in the yellow zone is not excluded
  $\Rightarrow$ decisions in 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**

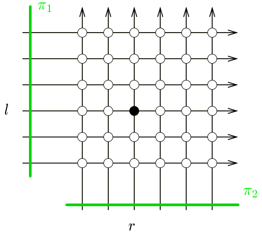
- rows and columns represent optical rays
- nodes: possible correspondence pairs
- full nodes: matches
- numerical values associated with nodes: descriptor similarities

see next

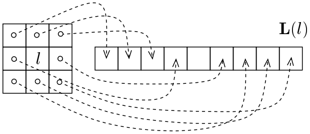
# ▶Constructing A Suitable Image Similarity Statistic

- 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:




- a simple block similarity is $\mathrm{SAD}(l, r) = \|\mathbf{L}(l) - \mathbf{R}(r)\|_1$ $\quad$ $L_1$ metric (sum of absolute differences)
- a scaled-descriptor similarity is $\quad \mathrm{sim}(l, r) = \dfrac{\|\mathbf{L}(l) - \mathbf{R}(r)\|^2}{\sigma_I^2(l, r)}$
- $\sigma_I^2$ – 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

$$\mathrm{sim}(l, r) = 1 - \underbrace{\frac{2\,\mathrm{cov}\big(\mathbf{L}(l), \mathbf{R}(r)\big)}{\mathrm{var}\big(\mathbf{L}(l)\big) + \mathrm{var}\big(\mathbf{R}(r)\big)}}_{\rho\left(\mathbf{L}(l), \mathbf{R}(r)\right)} \qquad (34)$$

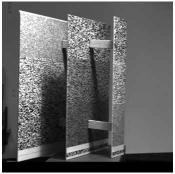
$\mathrm{var}(\cdot)$, $\mathrm{cov}(\cdot)$ is sample (co-)variance

- $\rho$ – MNCC – Moravec's Normalized Cross-Correlation statistic [Moravec 1977]

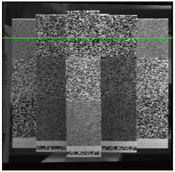
$$\rho^2 \in [0, 1], \qquad \mathrm{sign}\,\rho \sim \text{'phase'}$$

# How A Scene Looks in The Filled-In Matching Table

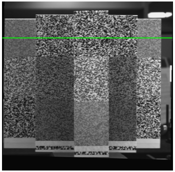
scene



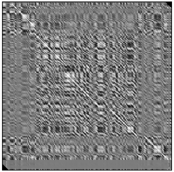
left image



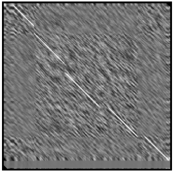
right image



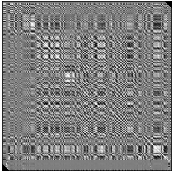
$5 \times 5$ window

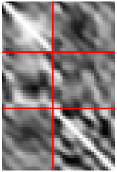


$11 \times 11$ window

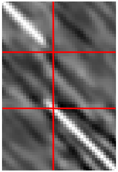


$3 \times 3$ window

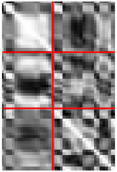




a good tradeoff



occlusion artefacts



undiscrimiable

- MNCC $\rho$ used ($\alpha = 1.5$, $\beta = 1$)
- high-correlation structures correspond to scene objects

**constant disparity**

- a diagonal in matching table
- zero disparity is the main diagonal

**depth discontinuity**

- horizontal or vertical jump in matching table

**large image window**

- better correlation
- worse occlusion localization

**repeated texture**

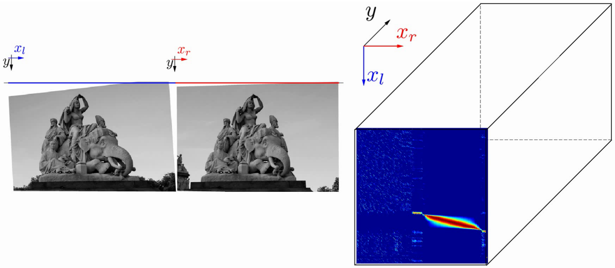
- horizontal and vertical block repetition

# Image Point Descriptors And Their Similarity

**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

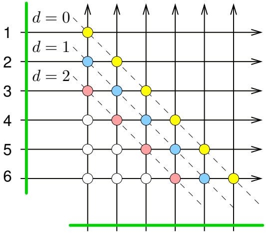


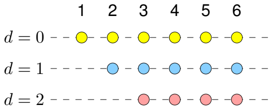

video

 R. Šára, CMP; rev. 8–Jan–2019

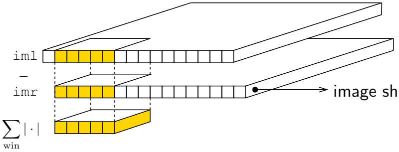
# ▶Marroquin's Winner Take All (WTA) Matching Algorithm

1. per left-image pixel: find the most similar right-image pixel using SAD →164
2. select disparity range — this is a critical weak point
3. represent the matching table diagonals in a compact form




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

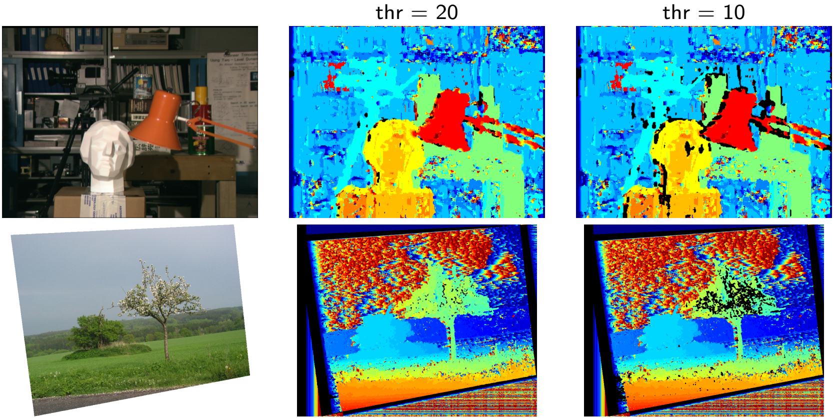



5. threshold results by maximal allowed dissimilarity

# 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

 % the size of each local patch; it is N=(2r+1)^2 except for boundary pixels
 N = boxing(ones(size(iml)), winsize);

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

 % collect winners, threshold, and output disparity map
 [cmap,dmap] = min(V,[],3);
 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
```

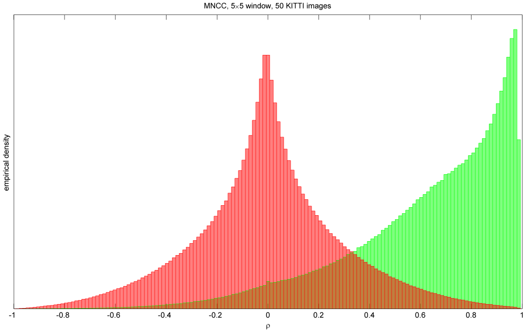
# WTA: Some Results




- 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

# ▶A Principled Approach to Similarity

Empirical Distribution of MNCC $\rho$ for Matches and Non-Matches




- histograms of $\rho$ computed from $5 \times 5$ correlation window
- KITTI dataset
  - $4.2 \cdot 10^6$ ground-truth (LiDAR) matches for $p_1(\rho)$ (green),
  - $4.2 \cdot 10^6$ random non-matches for $p_0(\rho)$ (red)

**Obs:**

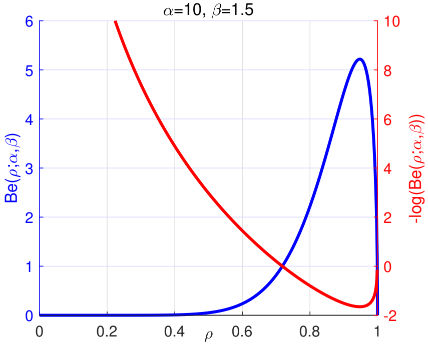
- non-matches (red) may have arbitrarily large $\rho$
- matches (green) may have arbitrarily low $\rho$
- $\rho = 1$ is improbable for matches

# Match Likelihood

- $\rho$ is just a statistic
- we need a probability distribution on $[0, 1]$, e.g. Beta distribution

$$p_1(\rho) = \frac{1}{B(\alpha, \beta)} \, |\rho|^{\alpha-1} (1 - |\rho|)^{\beta-1}$$

- note that uniform distribution is obtained for $\alpha = \beta = 1$
- when $\alpha = 2$ and $\beta = 1$ then $p_1(\cdot) = 2|\rho|$




- the mode is at $\sqrt{\frac{\alpha-1}{\alpha+\beta-2}} \approx 0.9733$ for $\alpha = 10,\ \beta = 1.5$
- if we chose $\beta = 1$ then the mode was at $\rho = 1$
- perfect similarity is 'suspicious' (depends on expected camera noise level)
- from now on we will work with negative log-likelihood

$$V_1\big(\rho(l, r)\big) = -\log p_1\big(\rho(l, r)\big) \tag{35}$$

smaller is better

- we may also define similarity (and negative log-likelihood $V_0(\rho(l, r))$) for non-matches

# ▶A Principled Approach to Matching

- given matching $M$ what is the likelihood of observed data $D$?
- data – all pairwise costs in matching table $T$
- matches – pairs $p_i = (l_i, r_i), \quad i = 1, \dots, n$
- matching: partitioning matching table $T$ to matched $M$ and excluded $E$ pairs

$$T = M \cup E, \quad M \cap E = \emptyset$$

- matching cost (negative log-likelihood, smaller is better)

$$V(D \mid M) = \sum_{p \in M} V_1(D \mid p) + \sum_{p \in E} V_0(D \mid p)$$

$V_1(D \mid p)$ – negative log-probability of data $D$ at matched pixel $p$ (35)
$V_0(D \mid p)$ – ditto at unmatched pixel $p$ →170 and →171

- matching problem

$$M^* = \arg \min_{M \in \mathcal{M}(T)} V(D \mid M)$$

$\mathcal{M}(T)$ – the set of all matchings in table $T$

- symmetric: formulated over pairs, invariant to left ↔ right image swap

# ▶(cont'd) Log-Likelihood Ratio

- we need to reduce matching to a standard polynomial-complexity problem
- we convert the matching cost to an 'easier' sum

$$V(D \mid M) = \sum_{p \in M} V_1(D \mid p) + \sum_{p \in E} V_0(D \mid p) \quad + \overbrace{\sum_{p \in M} V_0(D \mid p) - \sum_{p \in M} V_0(D \mid p)}^{0}$$

$$= \underbrace{\sum_{p \in M} \Big( V_1(D \mid p) - V_0(D \mid p) \Big)}_{-L(D \mid p)} + \underbrace{\sum_{p \in E} V_0(D \mid p) + \sum_{p \in M} V_0(D \mid p)}_{\sum_{p \in T} V_0(D \mid p) = \text{const}}$$

- hence

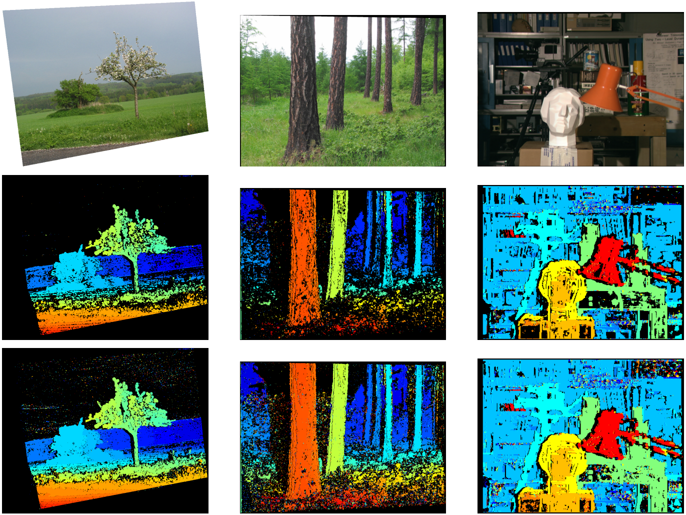
$$\arg \min_{M \in \mathcal{M}(T)} V(D \mid M) = \arg \max_{M \in \mathcal{M}(T)} \sum_{p \in M} L(D \mid p) \tag{36}$$

$L(D \mid p)$ – logarithm of matched-to-unmatched likelihood ratio (bigger is better)

why this way: we want to use maximum-likelihood but our measurement is all data $D$

- (36) is max-cost matching (maximum assignment) for the maximum-likelihood (ML) matching problem
  - it must contain no pairs $p$ with $L(D \mid p) < 0$
  - use Hungarian (Munkres) algorithm and threshold the result based on $L(D \mid p) > T$
  - or step back: sacrifice symmetry to speed and use dynamic programming

# Some Results for the Maximum-Likelihood (ML) Matching



- unlike the WTA we can efficiently control the density/accuracy tradeoff — black = no match
- middle row: threshold $T$ for $L(D \mid p)$ set to achieve error rate of 3% (and 61% density results)
- bottom row: threshold $T$ set to achieve density of 76% (and 4.3% error rate results)

# ▶Basic Stereoscopic Matching Models

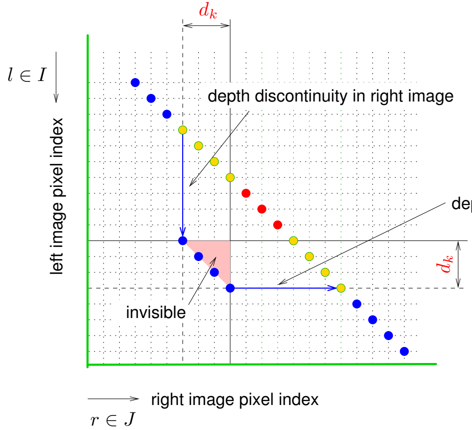
- notice many small isolated errors in the ML matching
- we need a stronger model

**Potential models for $M$ (from weaker to stronger)**

1. Uniqueness: Every image point matches at most once
   - excludes semi-transparent objects
   - used by the ML matching algorithm (but not by the WTA algorithm)
2. Monotonicity: Matched pixel ordering is preserved
   - For all $(i,j) \in M, (k,l) \in M, \quad k > i \Rightarrow l > j$

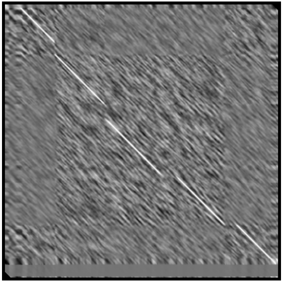
     Notation: $(i,j) \in M$ or $j = M(i)$ – left-image pixel $i$ matches right-image pixel $j$
   - excludes thin objects close to the cameras
   - used by 3LDP [SP]
3. Coherence: Objects occupy well-defined 3D volumes
   - concept by [Prazdny 85]
   - algorithms are based on image/disparity map segmentation
   - a popular model (segment-based, bilateral filtering and their successors)
   - used by Stable Segmented 3LDP [Aksoy et al. PRRS 2008]
4. Continuity: There are no occlusions or self-occlusions
   - too strong, except in some applications

# Understanding Occlusion Structure in Matching Table



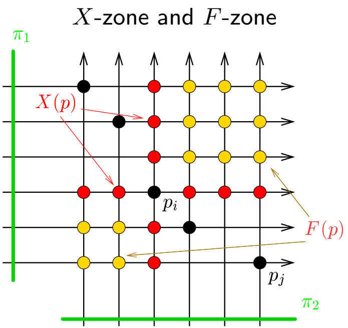

- binocularly visible foreground points
- binocularly visible background pts violating ordering
- monocularly visible points

$d_k$ critical disparity



- this leads to the concept of 'forbidden zone'

# ▶Formally: Uniqueness and Ordering in Matching Table $T$




$$p_j \notin X(p_i), \quad p_j \notin F(p_i)$$

- **Uniqueness Constraint:**

  A set of pairs $M = \{p_i\}_{i=1}^{n}$, $p_i \in T$ is a matching iff
  $$\forall p_i, p_j \in M : \; p_j \notin X(p_i).$$

  $X$-zone, $p_i \notin X(p_i)$

- **Ordering Constraint:**

  Matching $M$ is monotonic iff
  $$\forall p_i, p_j \in M : \; p_j \notin F(p_i).$$

  $F$-zone, $p_i \notin F(p_i)$

- ordering constraint: matched points form a monotonic set in both images
- ordering is a powerful constraint: in $n \times n$ table we have monotonic matchings $O(4^n) \ll O(n!)$ all matchings

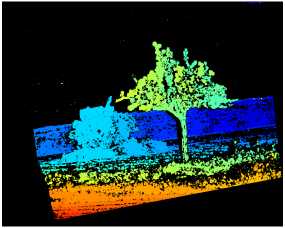
⊛ 2: how many are there maximal monotonic matchings? (e.g. 27 for $n = 4$; hard!)

- uniqueness constraint is a basic occlusion model
- ordering constraint is a weak continuity model — and partly also an occlusion model
- monotonic matching can be found by **dynamic programming**

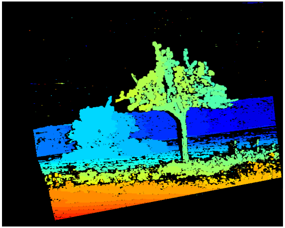
# Some Results: AppleTree

left image

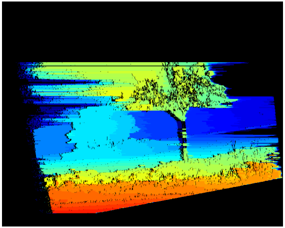
right image



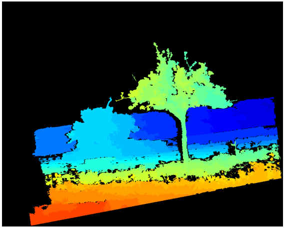
ML →173



3LDP w/ordering
[SP]



naïve DP
[Cox et al. 1992]



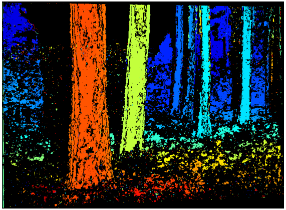
Stable Segmented 3LDP
[Aksoy et al. PRRS 2008]

- 3LDP parameters $\alpha_i$, $V_e$ learned on Middlebury stereo data http://vision.middlebury.edu/stereo/

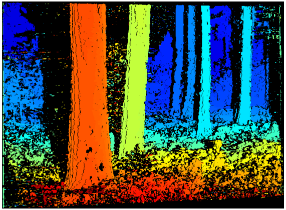
# Some Results: Larch

left image

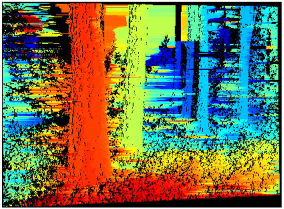
right image



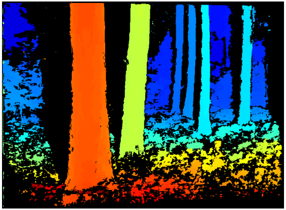
ML →173



3LDP w/ordering [SP]



naïve DP



Stable Segmented 3LDP

- naïve DP does not model mutual occlusion
- but even 3LDP has errors in mutually occluded region
- Stable Segmented 3LDP has few errors in mutually occluded region since it uses a coherence model

# Algorithm Comparison

## Marroquin's Winner-Take-All (WTA →167)

- the ur-algorithm — very weak model
- dense disparity map
- $O(N^3)$ algorithm, simple but it rarely works

## Maximum Likelihood Matching (ML →173)

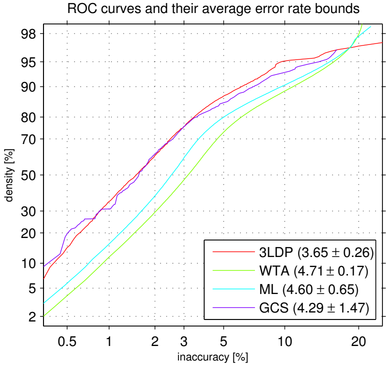
- semi-dense disparity map
- many small isolated errors
- models basic occlusion
- $O(N^3 \log(NV))$ algorithm — max-flow by cost scaling

## MAP with Min-Cost Labeled Path (3LDP)

- semi-dense disparity map
- models occlusion in flat, piecewise continuos scenes
- has 'illusions' if ordering does not hold
- $O(N^3)$ algorithm

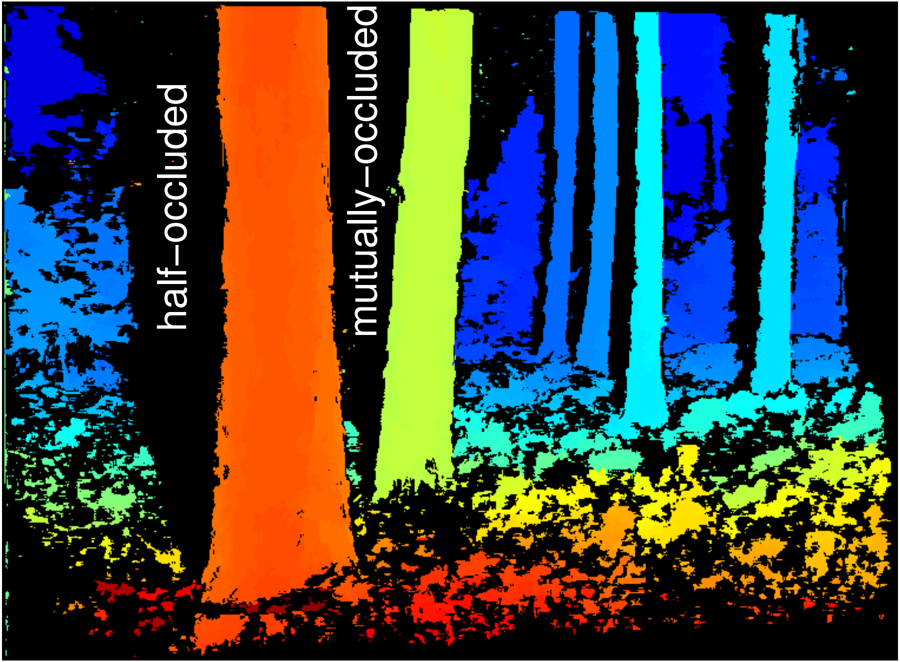
## Stable Segmented 3LDP

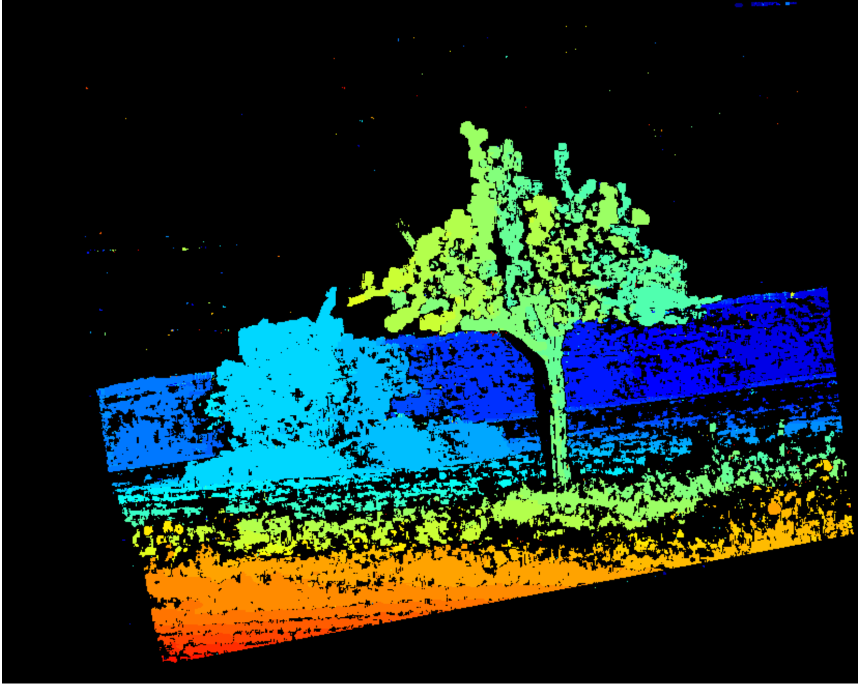
- better (fewer errors at any given density)
- $O(N^3 \log N)$ algorithm
- requires image segmentation — itself a difficult task

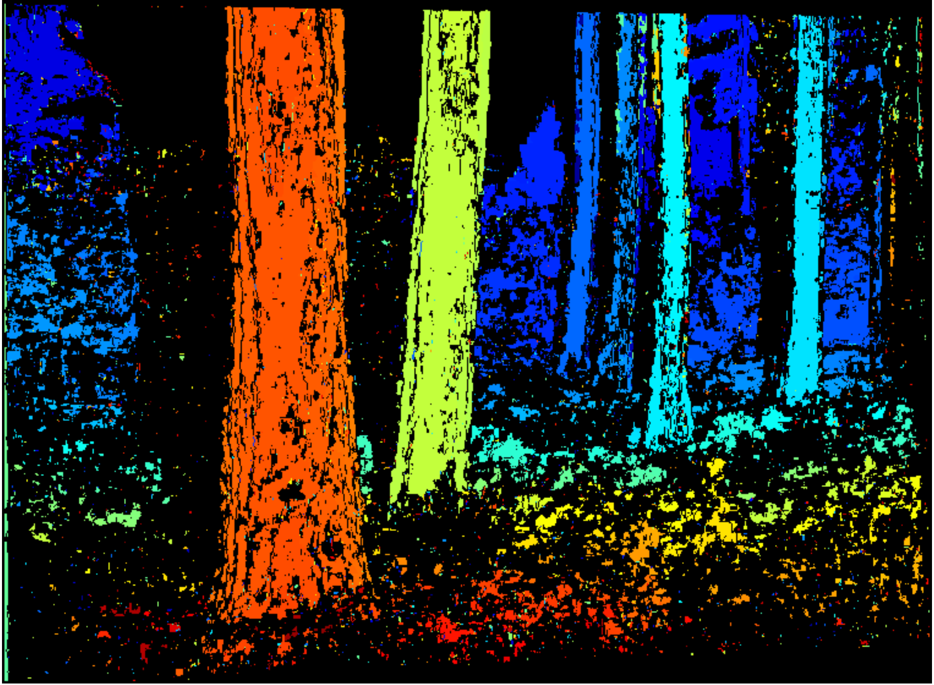


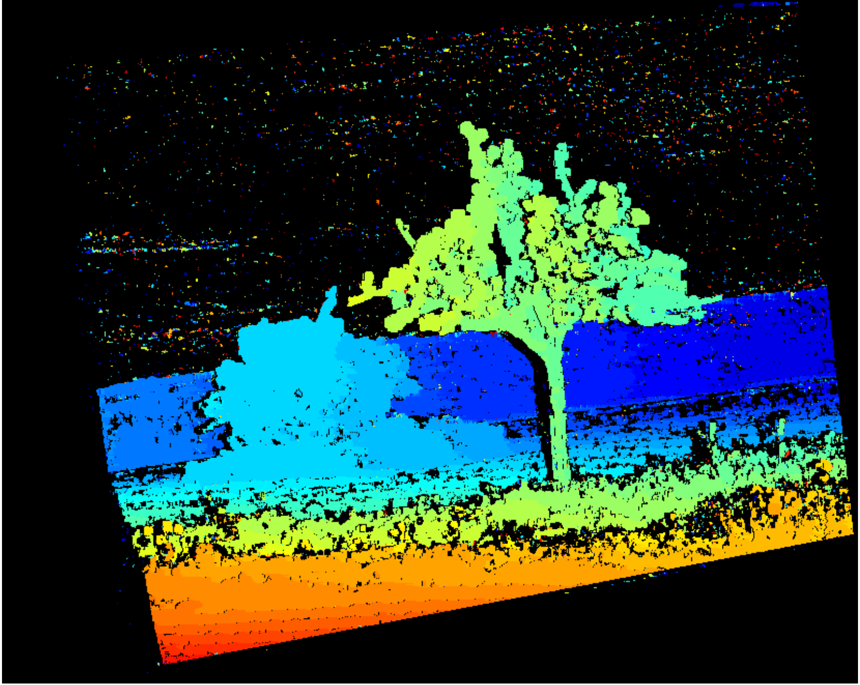

- ROC-like curve captures the density/accuracy tradeoff
- numbers: AUC (smaller is better)
- GCS is the one used in the exercises
- more algorithms at http://vision.middlebury.edu/stereo/ (good luck!)

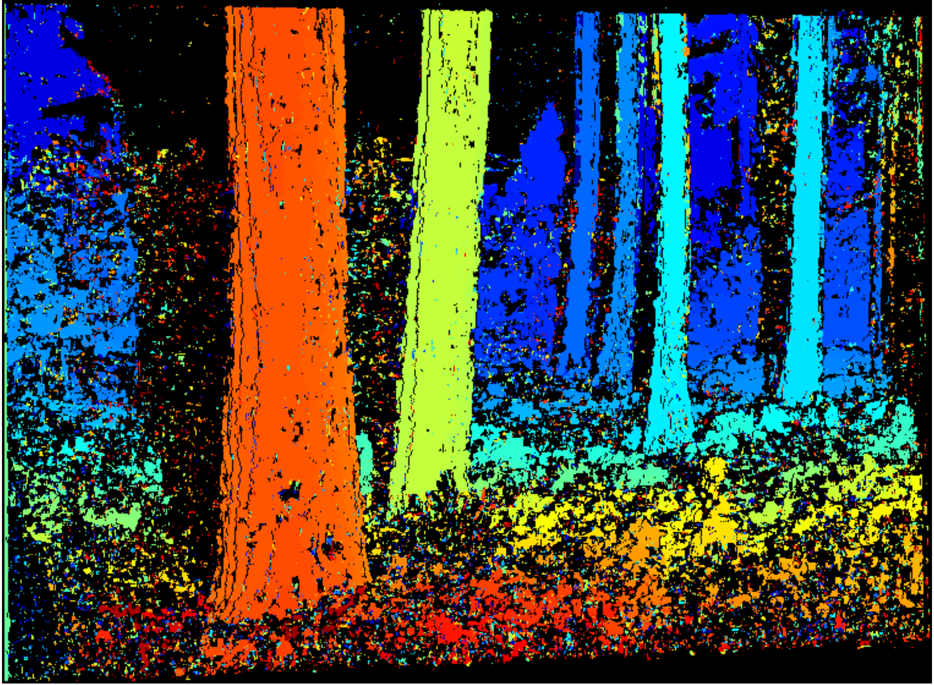
Thank You

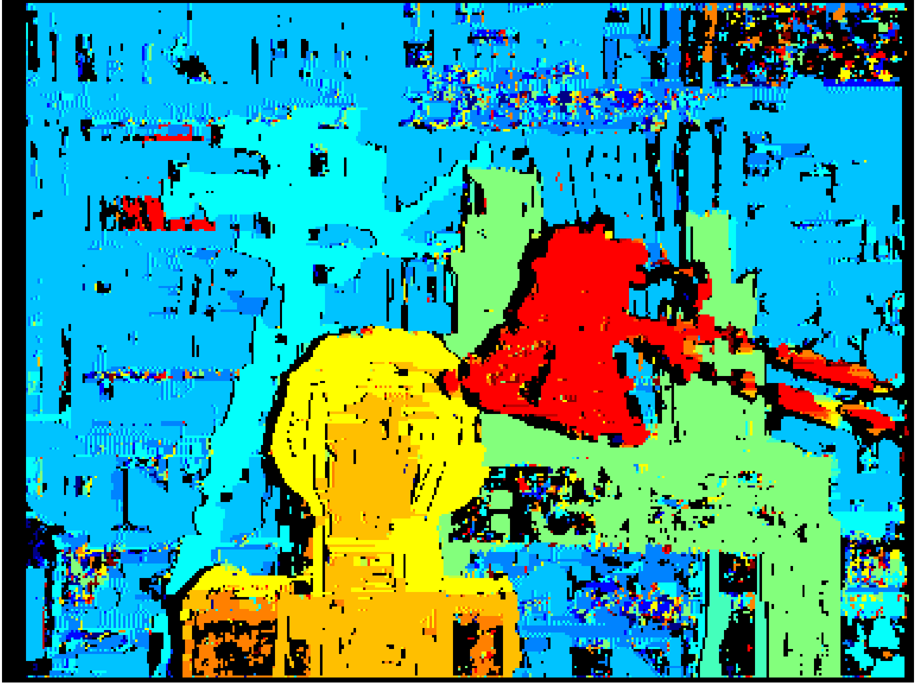

half–occluded
mutually–occluded

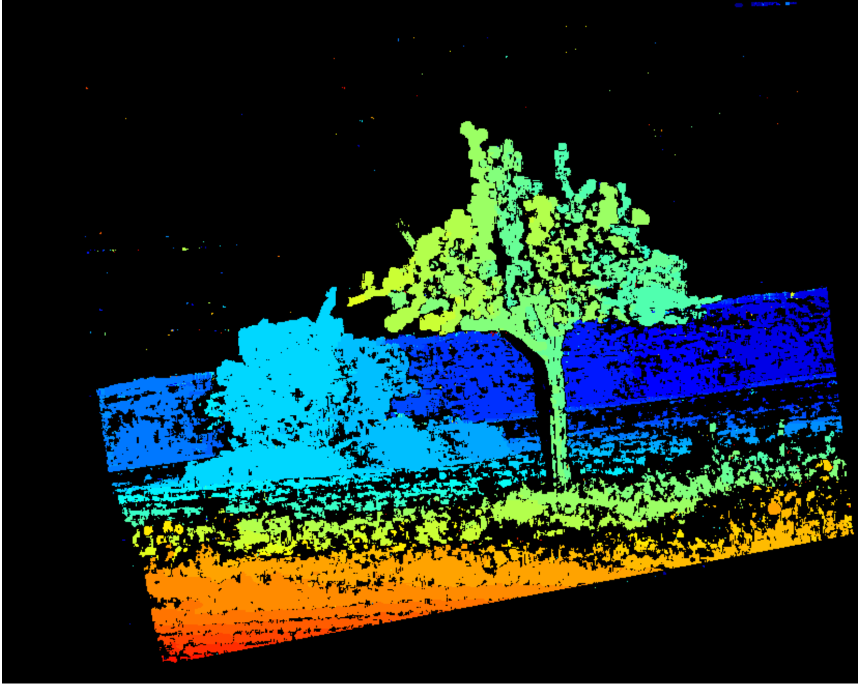


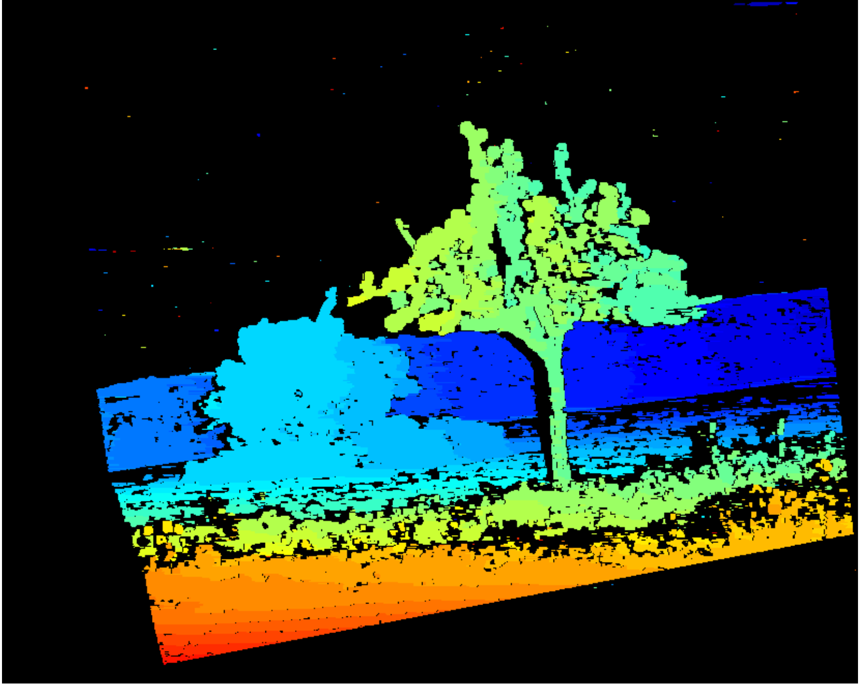


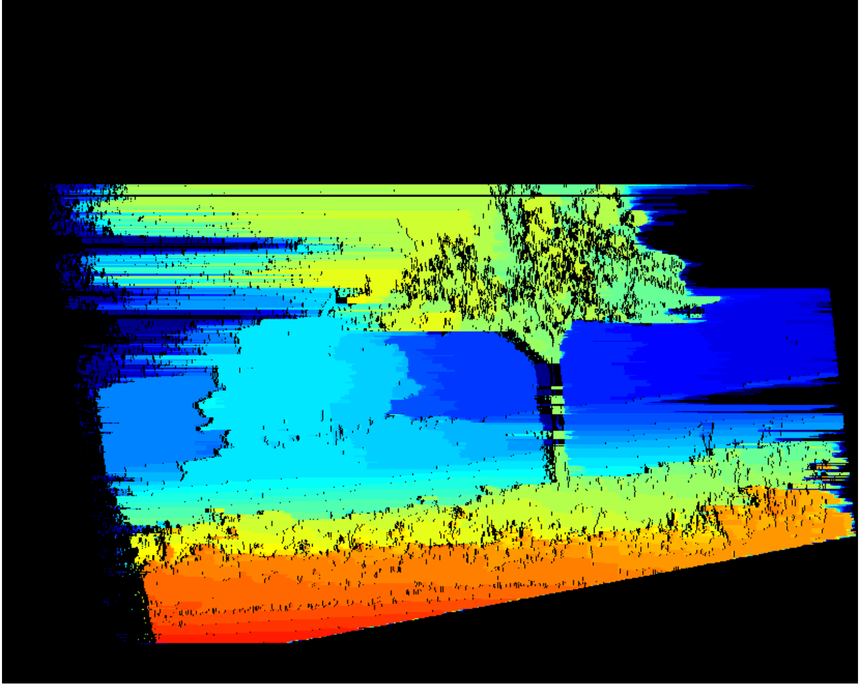


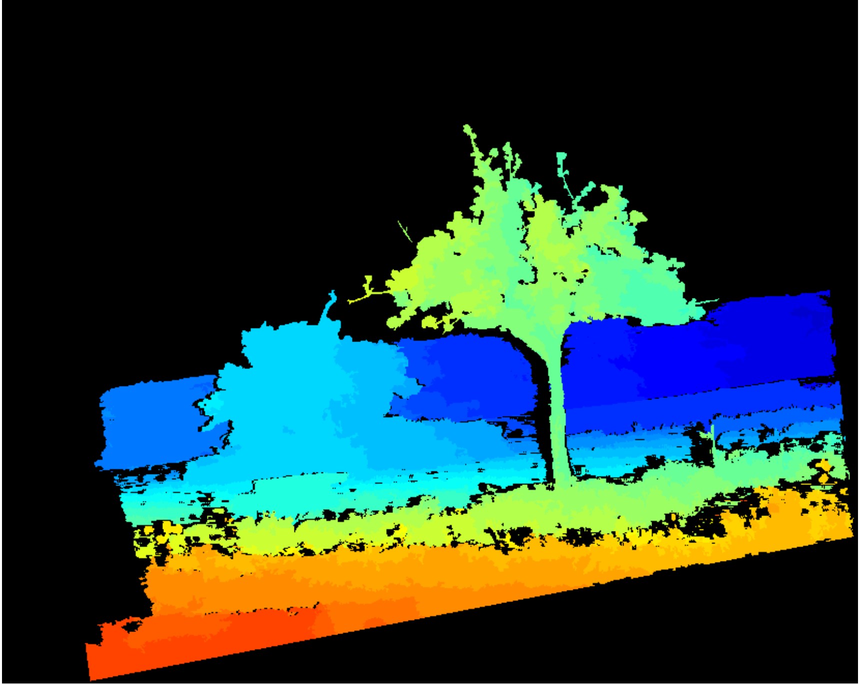


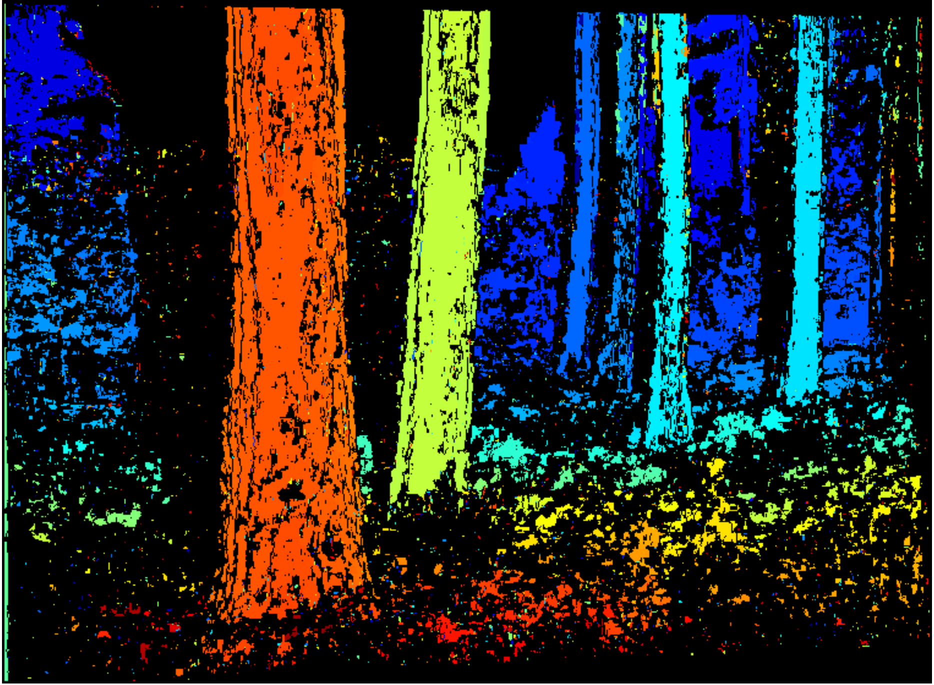


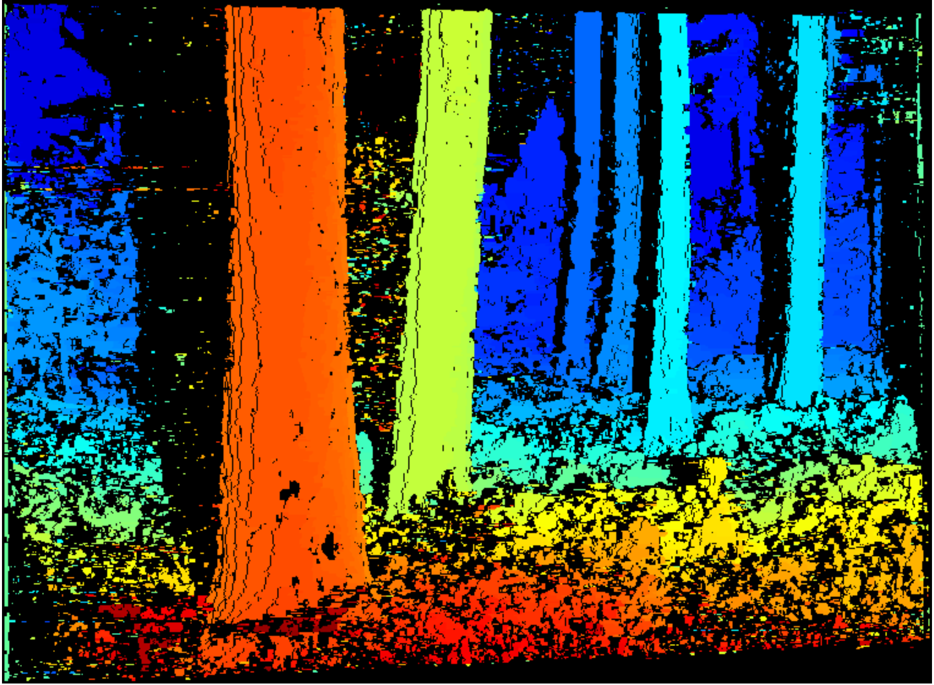


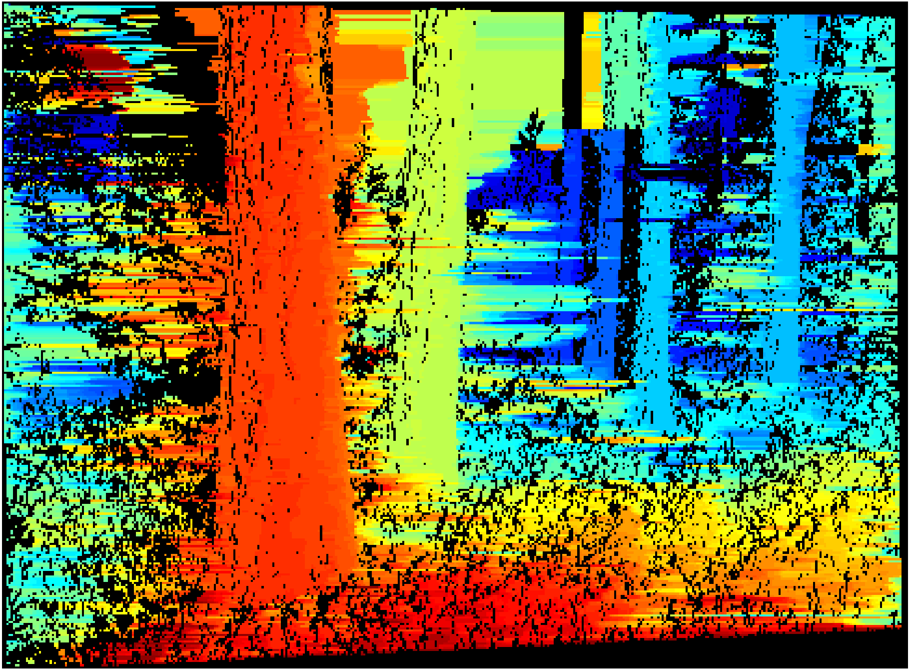


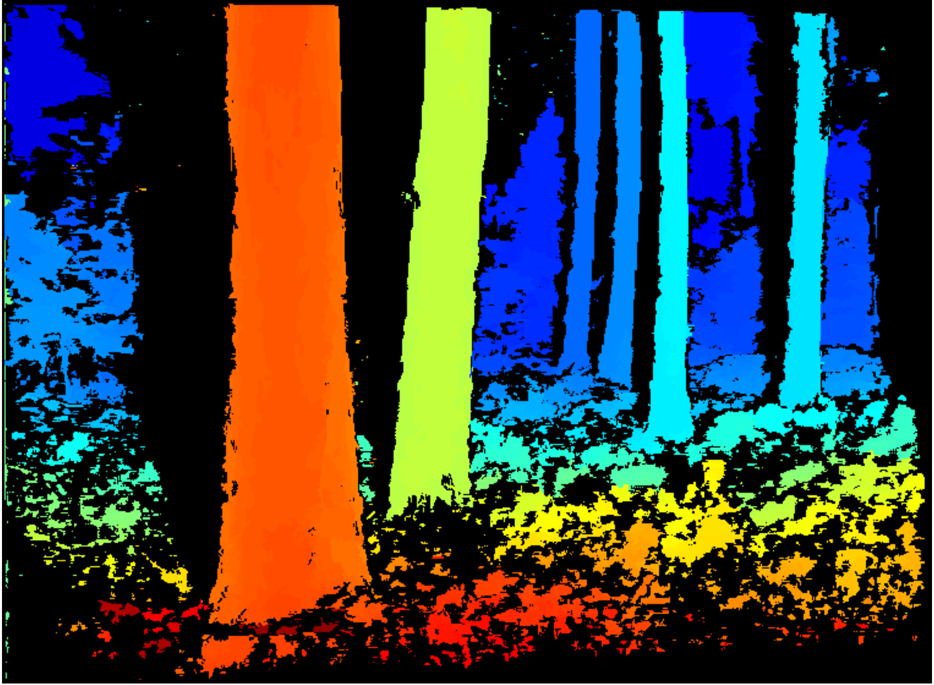


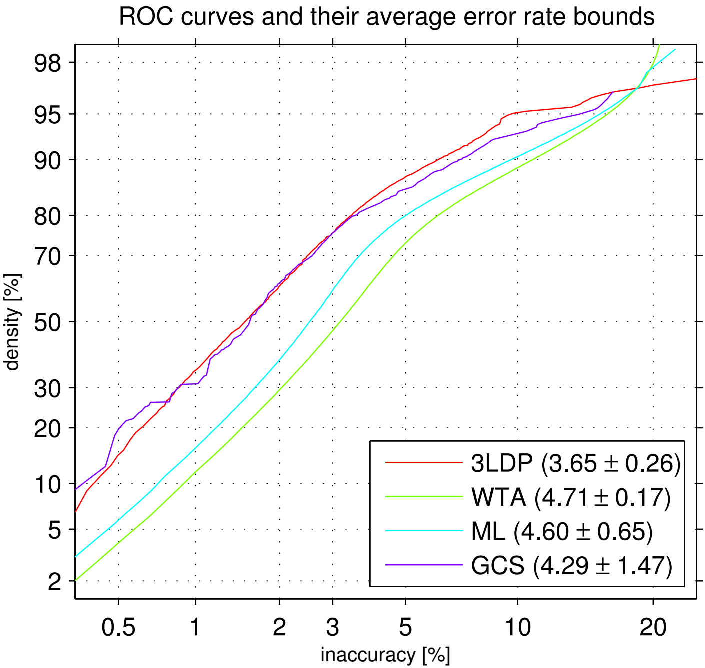











ROC curves and their average error rate bounds
98
95
90
80
70
50
30
20
10
5
2
density [%]
0.5
1
2
3
5
10
20
inaccuracy [%]
3LDP (3.65 ± 0.26)
WTA (4.71 ± 0.17)
ML (4.60 ± 0.65)
GCS (4.29 ± 1.47)