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

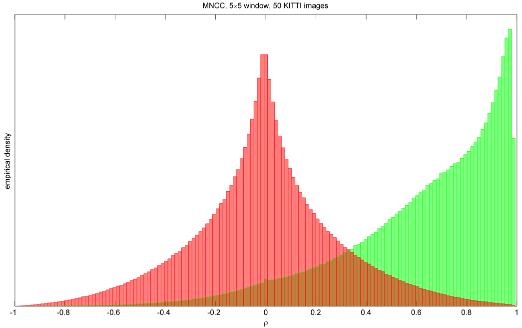
http://cmp.felk.cvut.cz
mailto:sara@cmp.felk.cvut.cz
phone ext. 7203

rev. December 13, 2022

Open Informatics Master's Course

►A Principled Approach to Similarity

Empirical Distribution of MNCC ρ for Matches (green) and Non-Matches (red)



- histograms of ρ computed from 5×5 correlation window

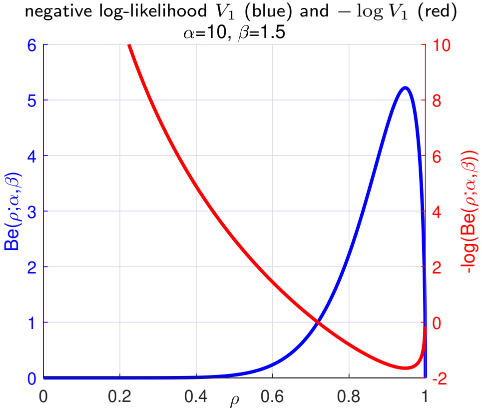
ρ: bigger is better

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

Match Likelihood



- ρ is just a normalized measurement
- 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 cost

$$V_1\big(\rho(l,r)\big) = -\log p_1\big(\rho(l,r)\big) \qquad \text{smaller is better} \tag{37}$$

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

►A Principled Approach to Matching: Formulating 'What We Want'

- given matching M in table T, what is the likelihood of observed data D?
- data – all cost pairs $(V_0,\, V_1)$ in the matching table T
- matches – pairs $p_i = (l_i, r_i) \in M \subset T, \quad i = 1, \ldots, n$
- 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) constant number of variables in T

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

$V_1(D \mid p)$ – negative log-probability of data D at matched pixel p (37)
$V_0(D \mid p)$ – ditto at unmatched pixel p →175 and →176

- matching problem

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

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

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

▶(cont'd) Log-Likelihood Ratio

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

$$
\begin{aligned}
V(D \mid M, T) &= \sum_{p \in M} V_1(D \mid p) + \sum_{p \in T \setminus M} V_0(D \mid p) \quad + \overbrace{\sum_{p \in M} V_0(D \mid p) - \sum_{p \in M} V_0(D \mid p)}^{0} \\
&= \sum_{p \in M} \underbrace{\Big(V_1(D \mid p) - V_0(D \mid p)\Big)}_{-L(D \mid p)} + \underbrace{\sum_{p \in T \setminus M} V_0(D \mid p) + \sum_{p \in M} V_0(D \mid p)}_{\sum_{p \in T} V_0(D \mid p) = \text{const}}
\end{aligned}
$$

- hence

$$
\arg \min_{M \in \mathcal{M}(T)} V(D \mid M) = \arg \max_{M \in \mathcal{M}(T)} \sum_{p \in M} L(D \mid p) \tag{38}
$$

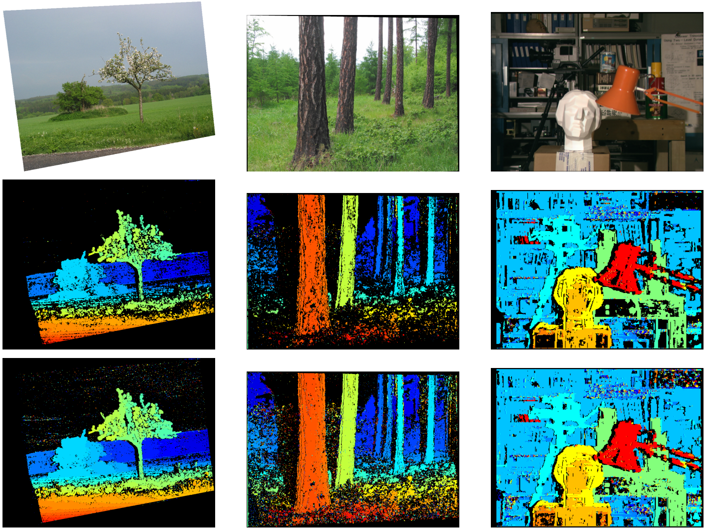
$L(D \mid p)$ – logarithm of matched-to-unmatched likelihood ratio (bigger is better)

why this way: we want to use maximum-likelihood on the entire T

- (38) is max-cost matching (maximum assignment) for the maximum-likelihood (ML) matching problem
  - use Hungarian (Munkres) algorithm and threshold the result with τ: $L(D \mid p) > \tau \geq 0$

or approximate the problem by sacrificing symmetry to speed and use dynamic programming

Some Results for the Maximum-Likelihood (ML) Matching

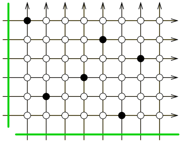


- unlike the WTA we can efficiently control the density/accuracy tradeoff with τ black = no match
- middle row: threshold τ for $L(D \mid p)$ set to achieve error rate of 3% (and 61% density results)
- bottom row: threshold τ set to achieve density of 76% (and 4.3% error rate results)

▶Basic Stereoscopic Matching Models

- notice many small isolated errors in the ML matching
- Q: how to reduce the noisiness? A: a stronger model

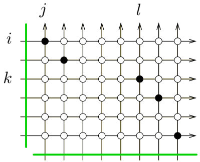
Potential models for M (from weaker to stronger)



incoherent

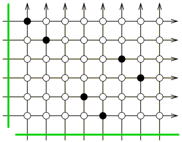
1. Uniqueness: Every image point matches at most once
   - excludes semi-transparent objects
   - used in the ML matching algorithm (but not in the WTA algorithm)
2. Monotonicity: Matched pixel ordering is preserved →181
   - for all $(i,j) \in M, (k,l) \in M, \quad k > i \Rightarrow l > j$

     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 in 3-Label Dynamic Programming (3LDP) [SP]

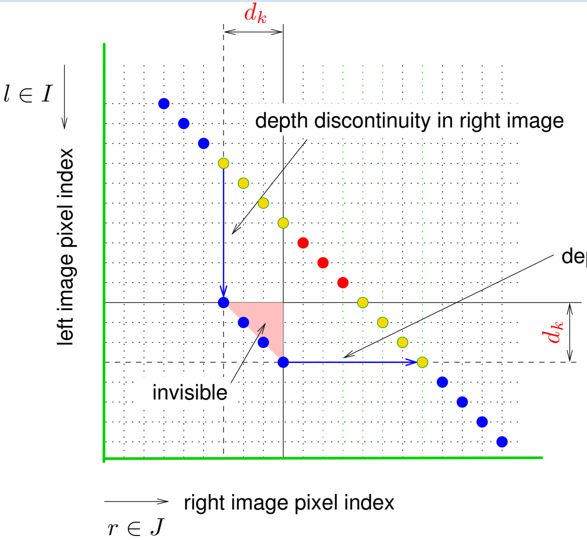


monotonic coherent

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 in Stable Segmented 3LDP [Aksoy et al. PRRS 2008]
4. (Piecewise) binocular continuity: The scene images continuously w/o self-occlusions
   - disparities do not differ much in neighboring pixels (except at object boundaries)
   - full binocular continuity too strong, except in some applications
   - piecewise binocular continuity is combined with monotonicity in 3LDP

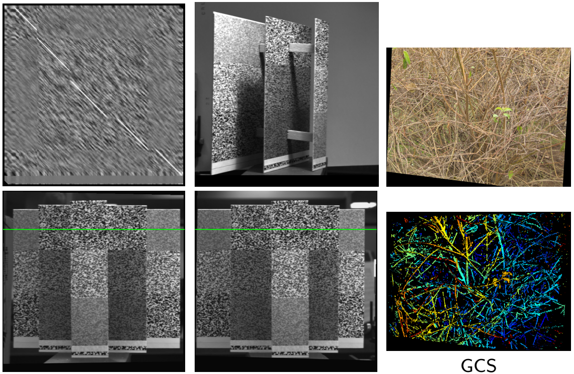


non-monotonic coherent



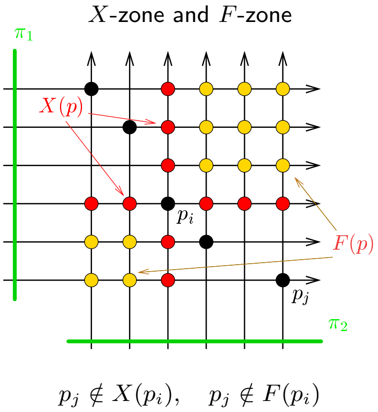
- binocularly visible foreground points
- binocularly visible background pts violating ordering
- monocularly visible points

d_k critical disparity



GCS

- this leads to the concept of 'forbidden zone'



- **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

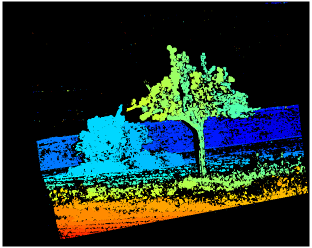
  ⊛ 2: how many are there maximal monotonic matchings? (e.g. 27 for $n = 4$; hard!)

- 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**

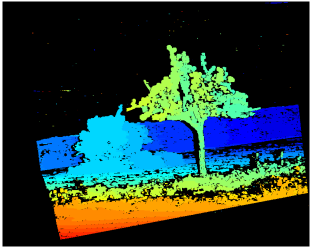
Some Results: AppleTree

left image

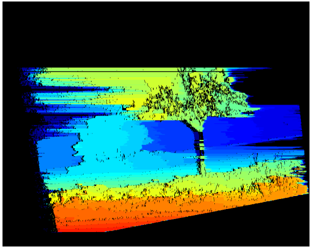
right image



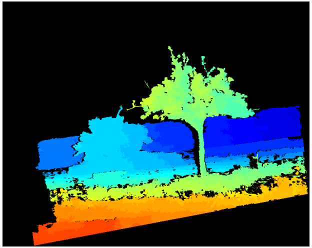
ML →178



3LDP w/ordering
[SP]



naïve DP
[Cox et al. 1992]



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

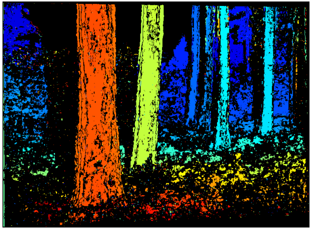
- 3LDP parameters α_i, V_e learned on Middlebury stereo data

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

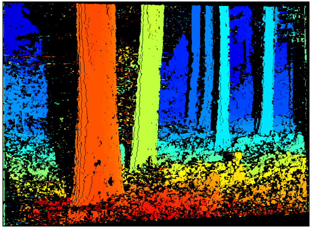
Some Results: Larch

left image

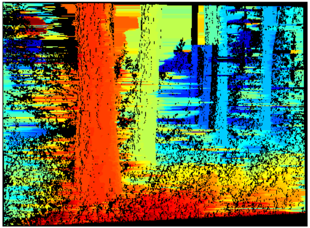
right image



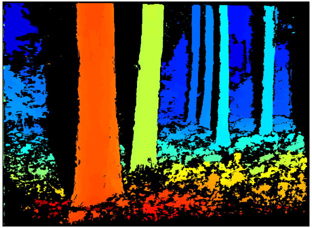
ML →178



3LDP w/ordering [SP]



naïve DP



Stable Segmented 3LDP

- naïve DP: no mutual occlusion model, ignores symmetry, has no similarity distribution model, ignores $T \setminus M$
- but even 3LDP has errors in mutually occluded region
- Stable Segmented 3LDP: few errors in mutually occluded region since it uses a coherence model

Algorithm Comparison

Marroquin's Winner-Take-All (WTA →172)

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

Maximum Likelihood Matching (ML →178)

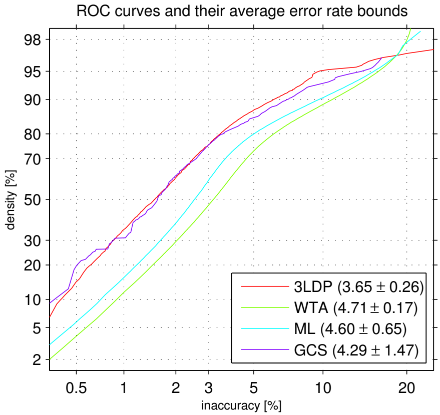
- semi-dense disparity map
- many small isolated errors
- models basic occlusion
- $O(N^3 \log(NV))$ algorithm — max-flow by cost scaling

MAP with Min-Cost Labeled Path (3LDP)

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

Stable Segmented 3LDP

- better than 3LDP — 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!)

A Summary of This Course Highlights

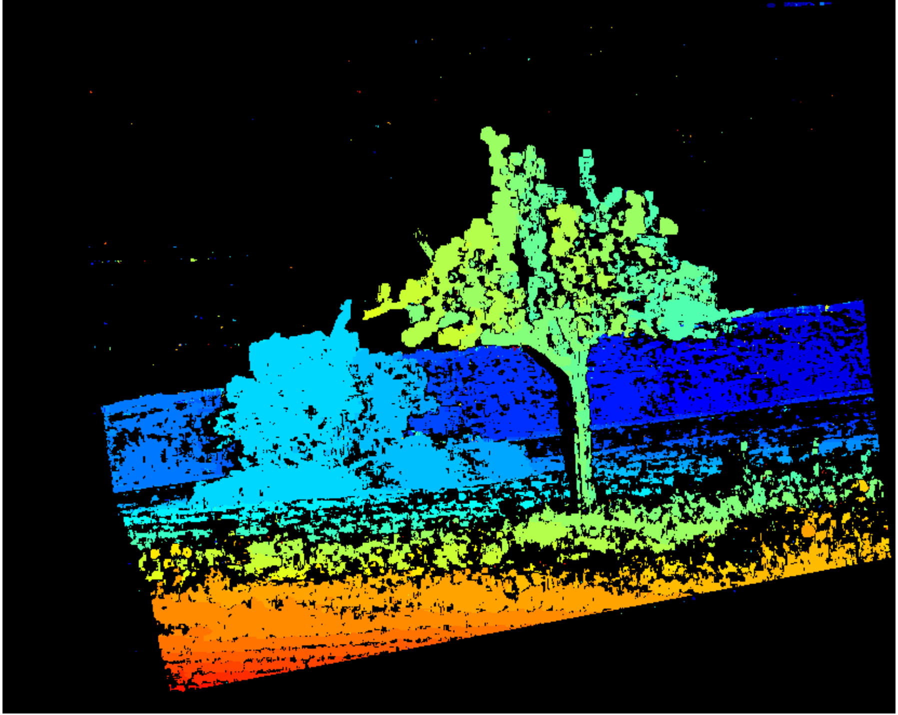
- homography as a two-image model
- epipolar geometry as a two-image model
- core algorithms for 3D vision:
  - simple intrinsic calibration methods
  - 6-pt alg for camera resection and 3-pt alg for exterior orientation (calibrated resection)
  - 7-pt alg for fundamental matrix, 5-pt alg for essential matrix
  - essential matrix decomposition to rotation and translation
  - efficient accurate triangulation
  - robust matching by RANSAC sampling
  - camera system reconstruction
  - efficient bundle adjustment
  - stereoscopic matching basics
- statistical robustness as a way to work with partially unknown information

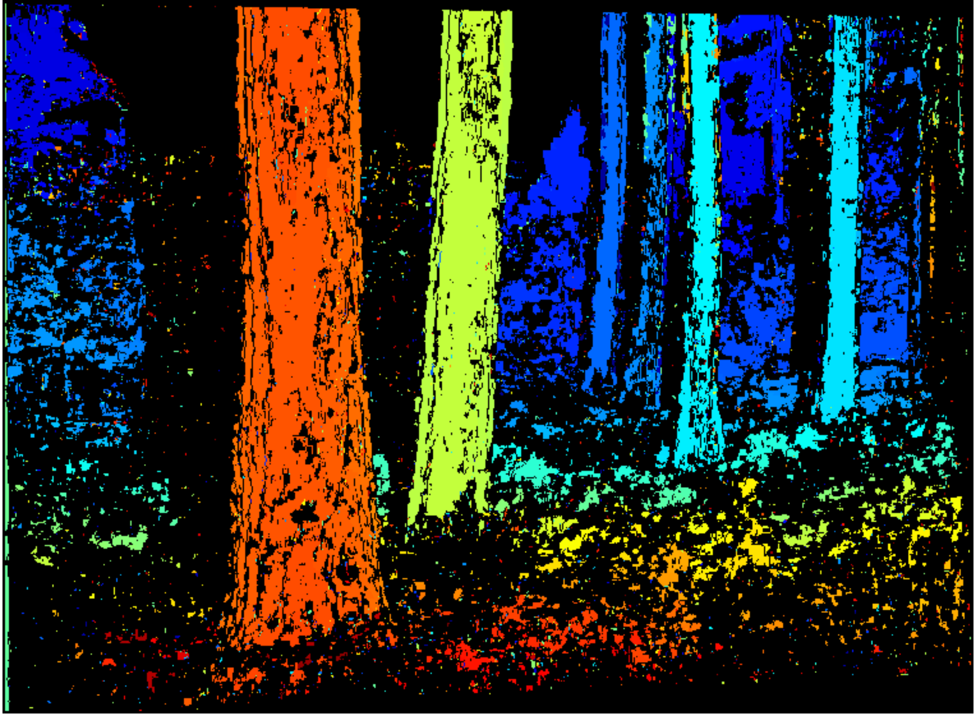
What can we do with these tools?

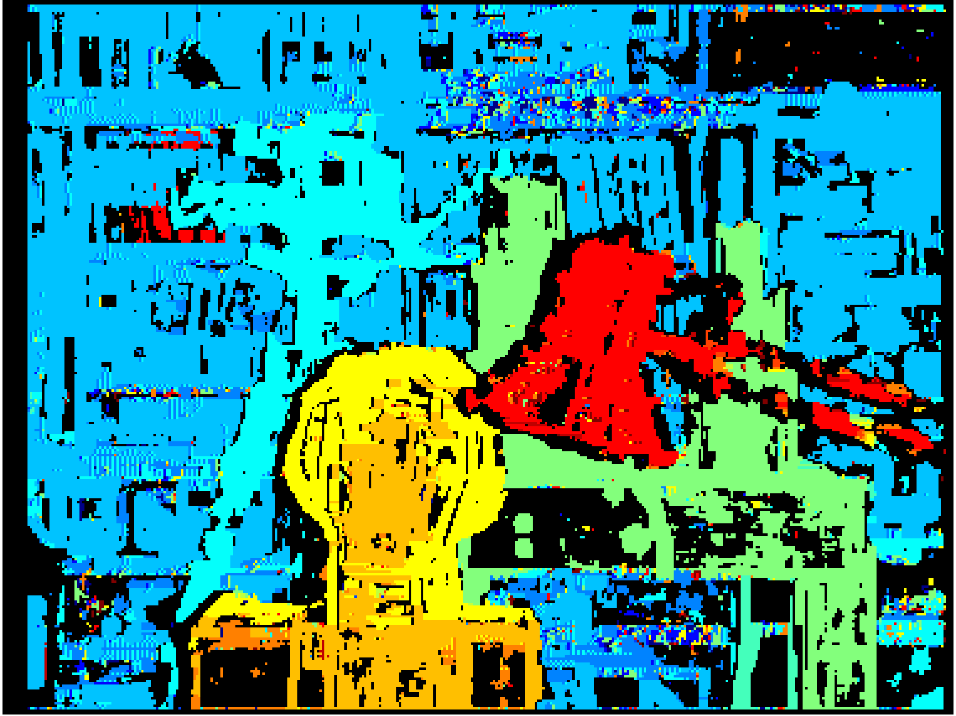
- perspective image rectification
- 3D scene reconstruction
- motion capture
- visual odometry
- robotic self-localization and mapping (SLAM) for navigation and motion planning

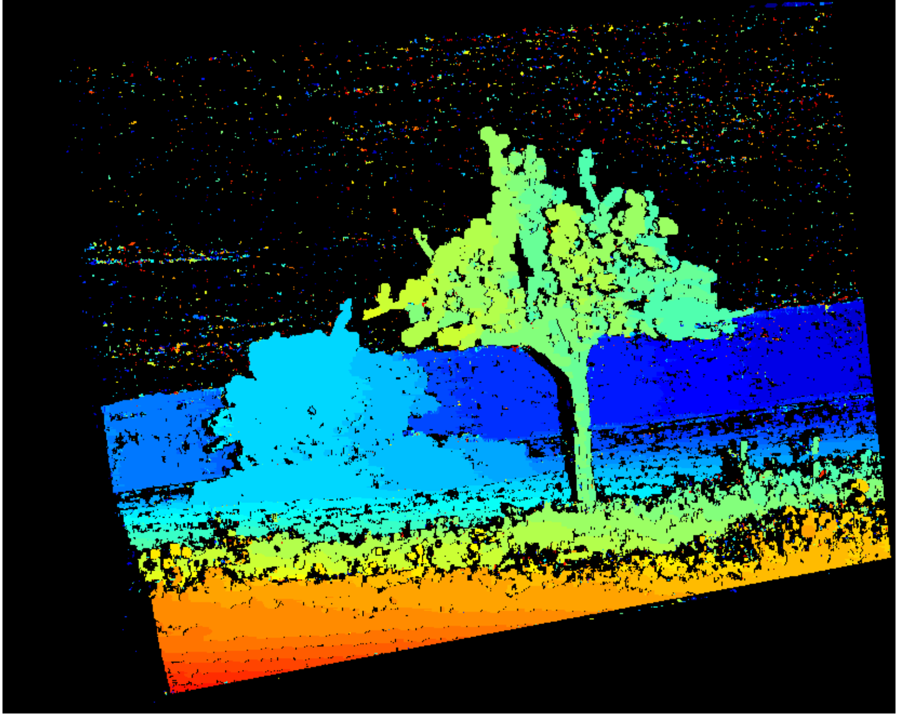
we did not cover 3D aggregation in scene maps

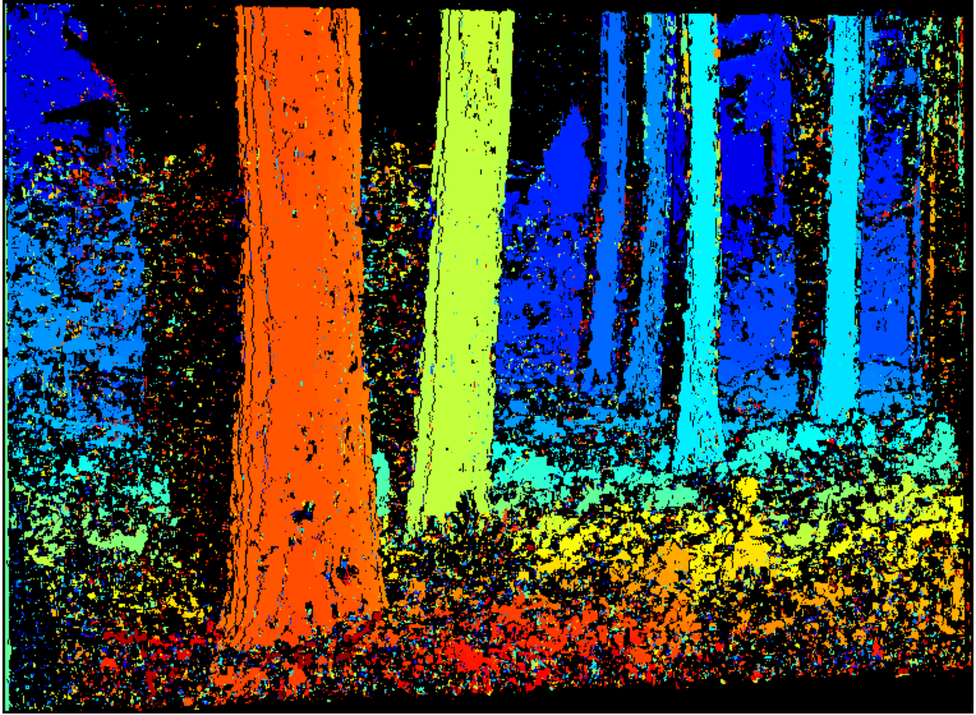
Thank You

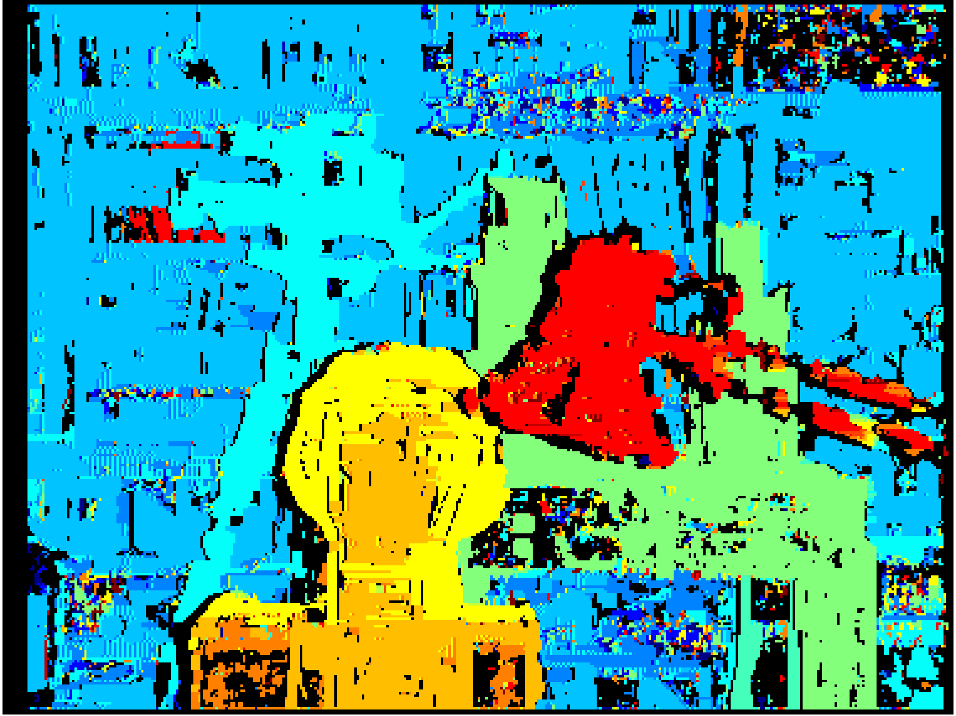


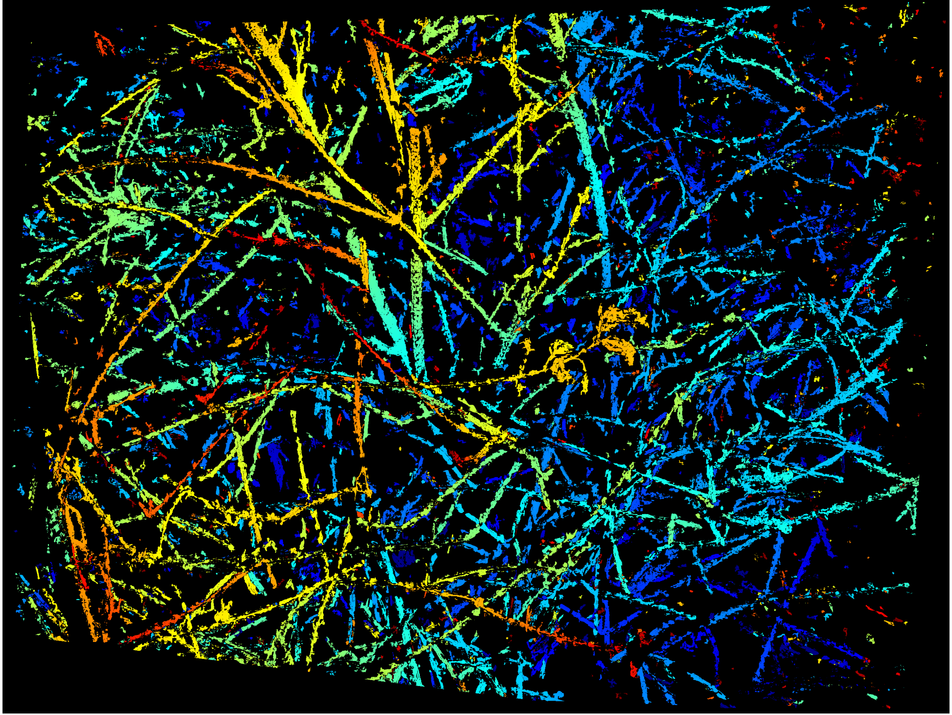


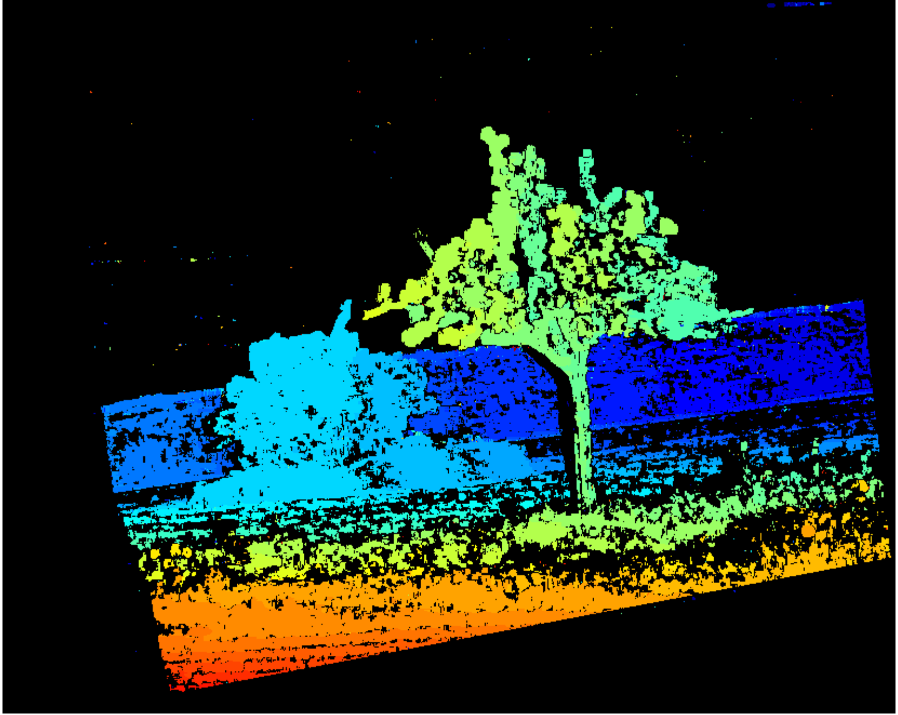


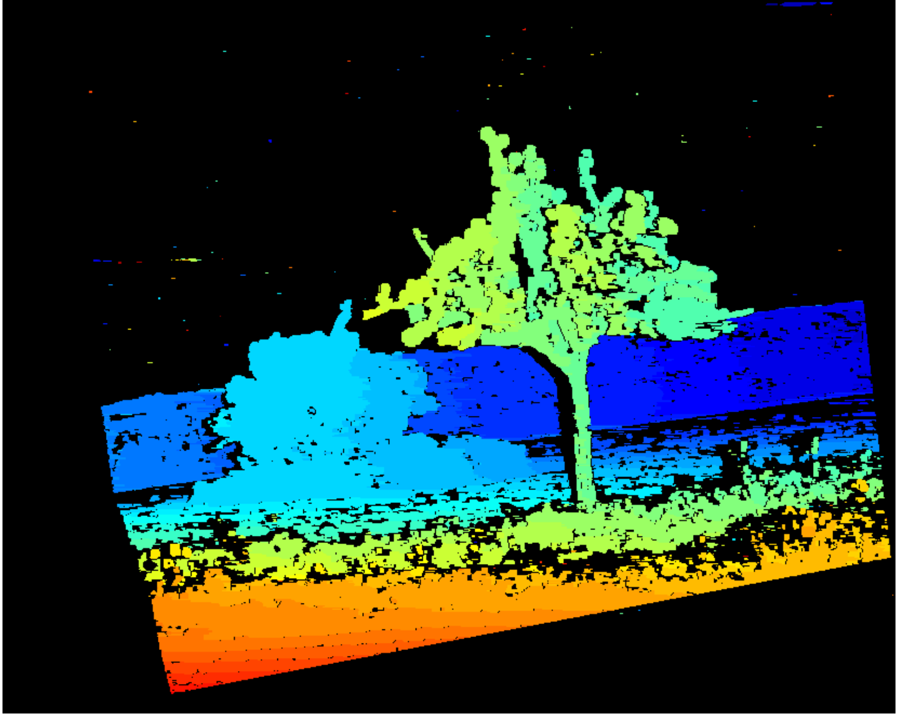


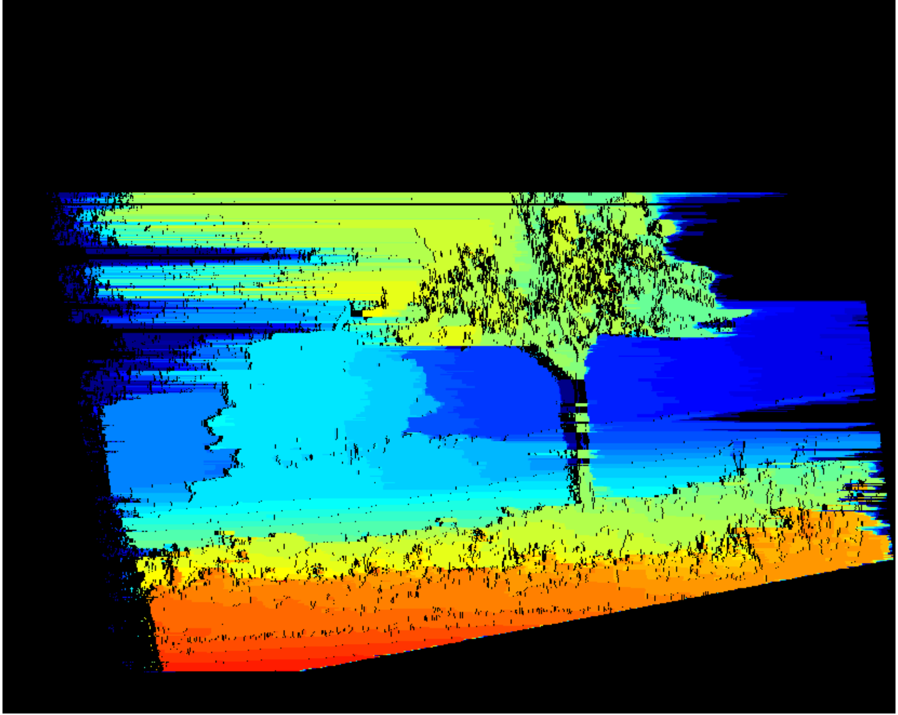


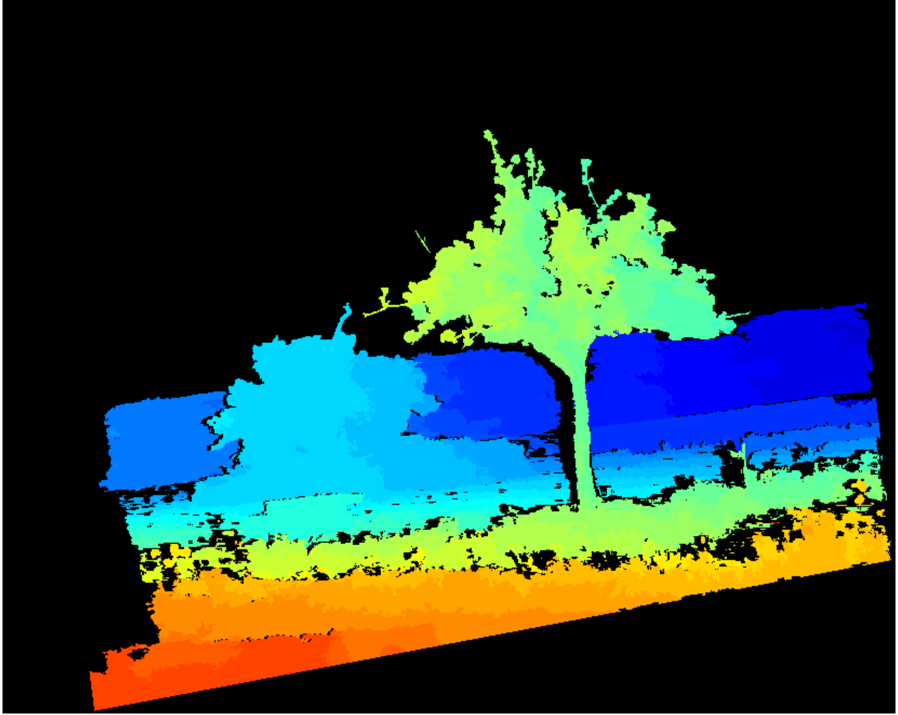


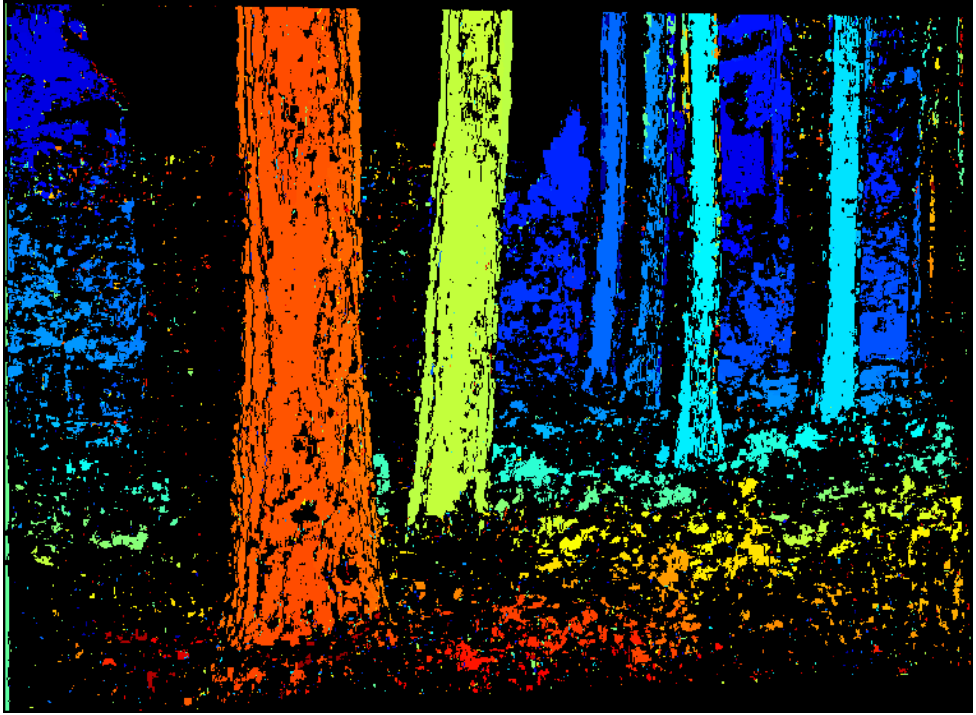


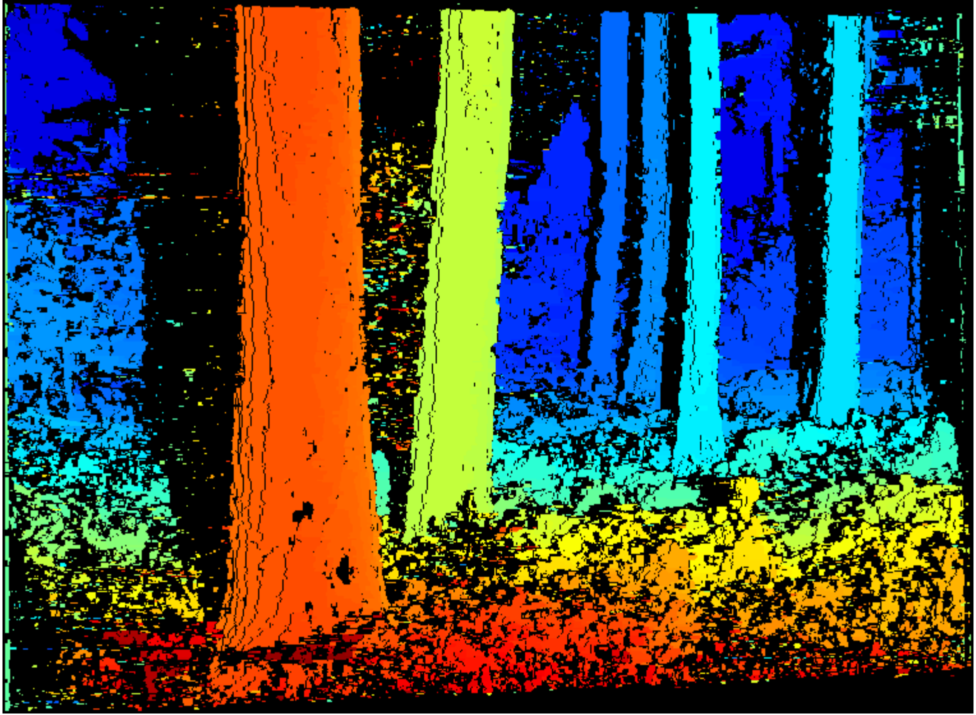


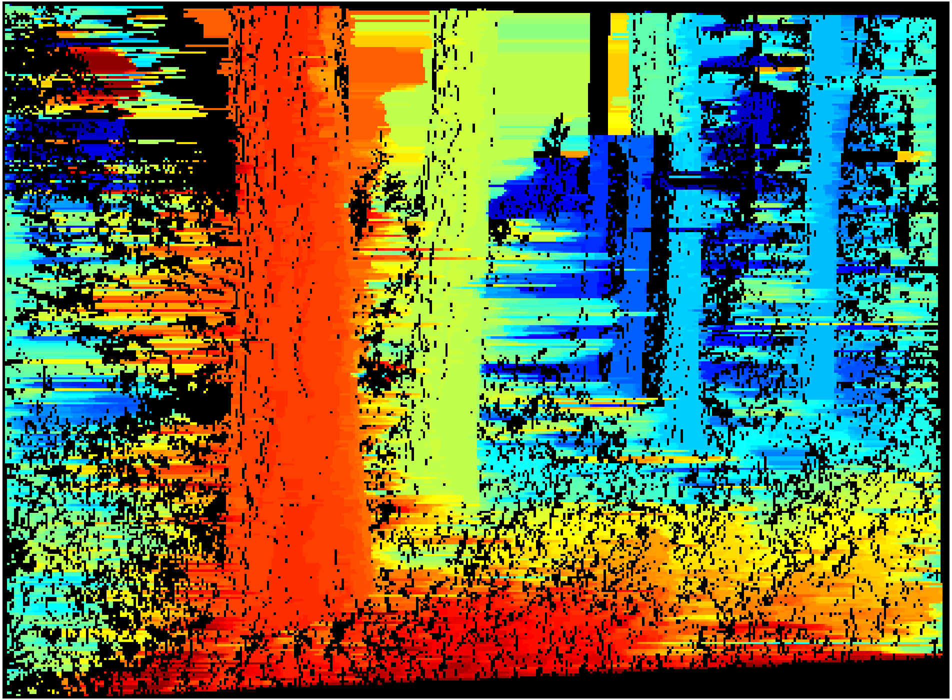


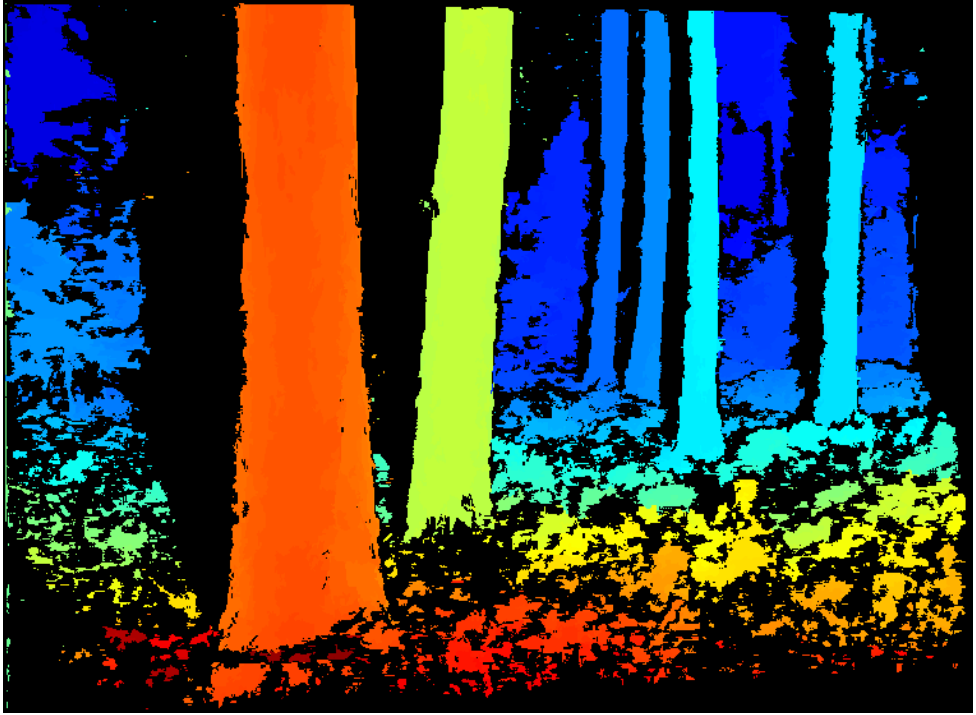


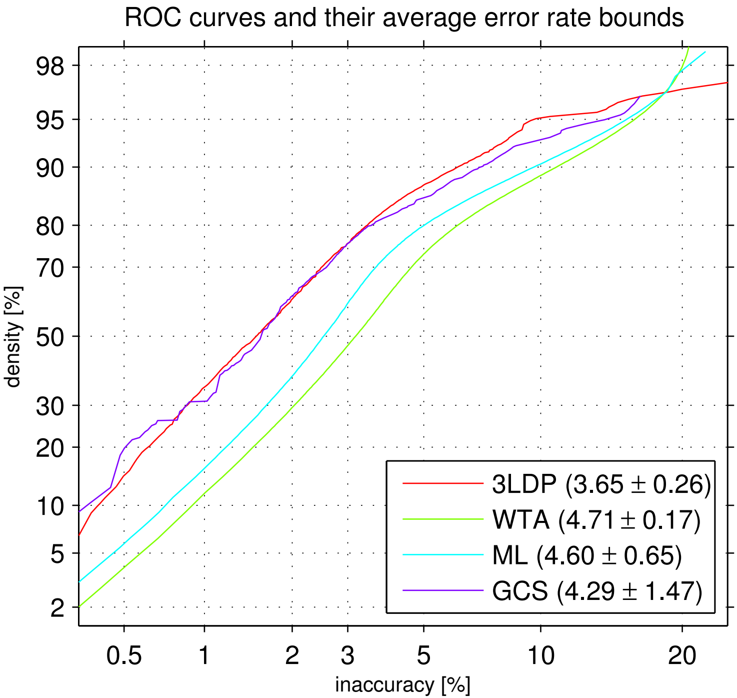











ROC curves and their average error rate bounds
98
95
90
80
70
50
30
20
10
5
2
density [%]
0.5
1
2
3
5
10
20
inaccuracy [%]
3LDP (3.65 ± 0.26)
WTA (4.71 ± 0.17)
ML (4.60 ± 0.65)
GCS (4.29 ± 1.47)