Inferring 3D Cues from a Single Image

Wei-Cheng Su

THE UNIVERSITY OF TEXAS AT AUSTIN

UT ECE

ELECTRICAL & COMPUTER ENGINEERING

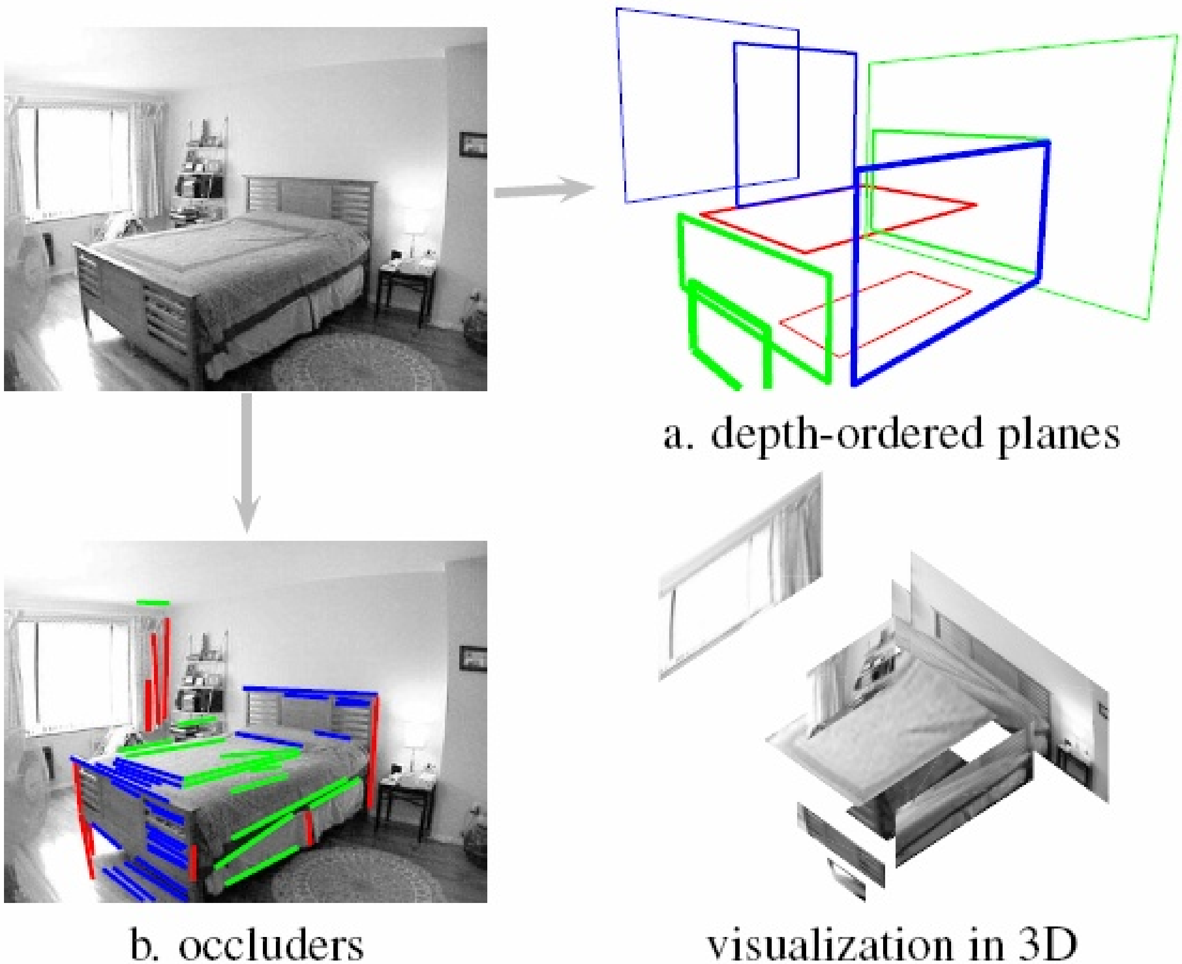
Motivation

- Human can estimate the 3D information from a single image easily. But how about computers?
- Possible cues: defocus, texture, shading, perspective, object size…

Outline

- Inferring Spatial Layout from A Single Image via Depth-Ordered Grouping, by Stella X. Yu, Hao Zhang, and Jitendra Malik, Workshop on Perceptual Organization in Computer Vision, 2008
- Depth Estimation using Monocular and Stereo Cues, by A. Saxena, J. Schulte, and A. Ng. IJCAI 2007
- Comparison

Inferring Spatial Layout from A Single Image via Depth-Ordered Grouping

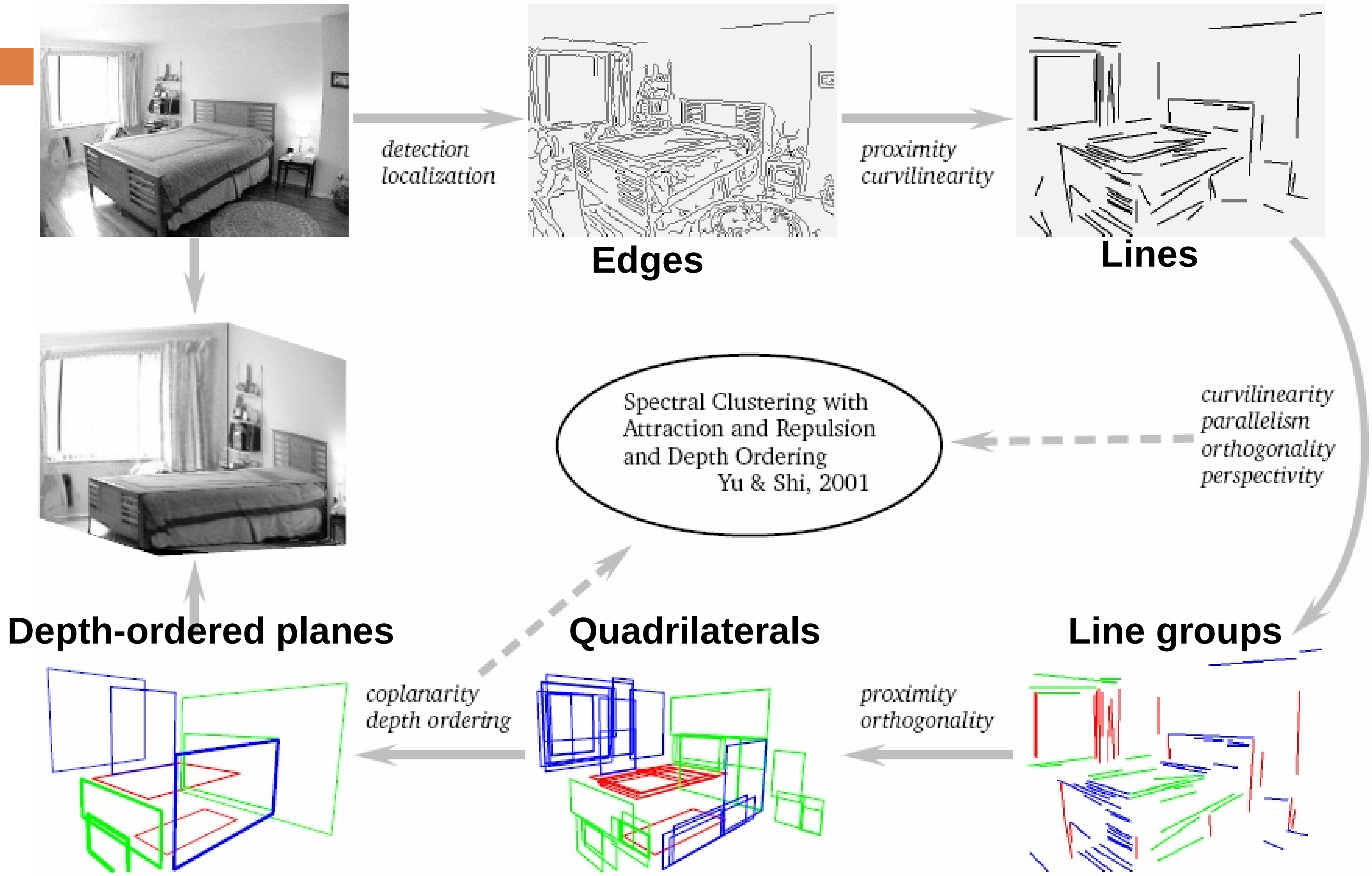


[Yu, Zhang, and Malik, Workshop on Perceptual Organization in Computer Vision 2008]

Goal

- Infer 3D spatial layout from a single 2D image
- Based on grouping
- Focus on indoor scenes

Line-Based Depth-Ordered Grouping Model



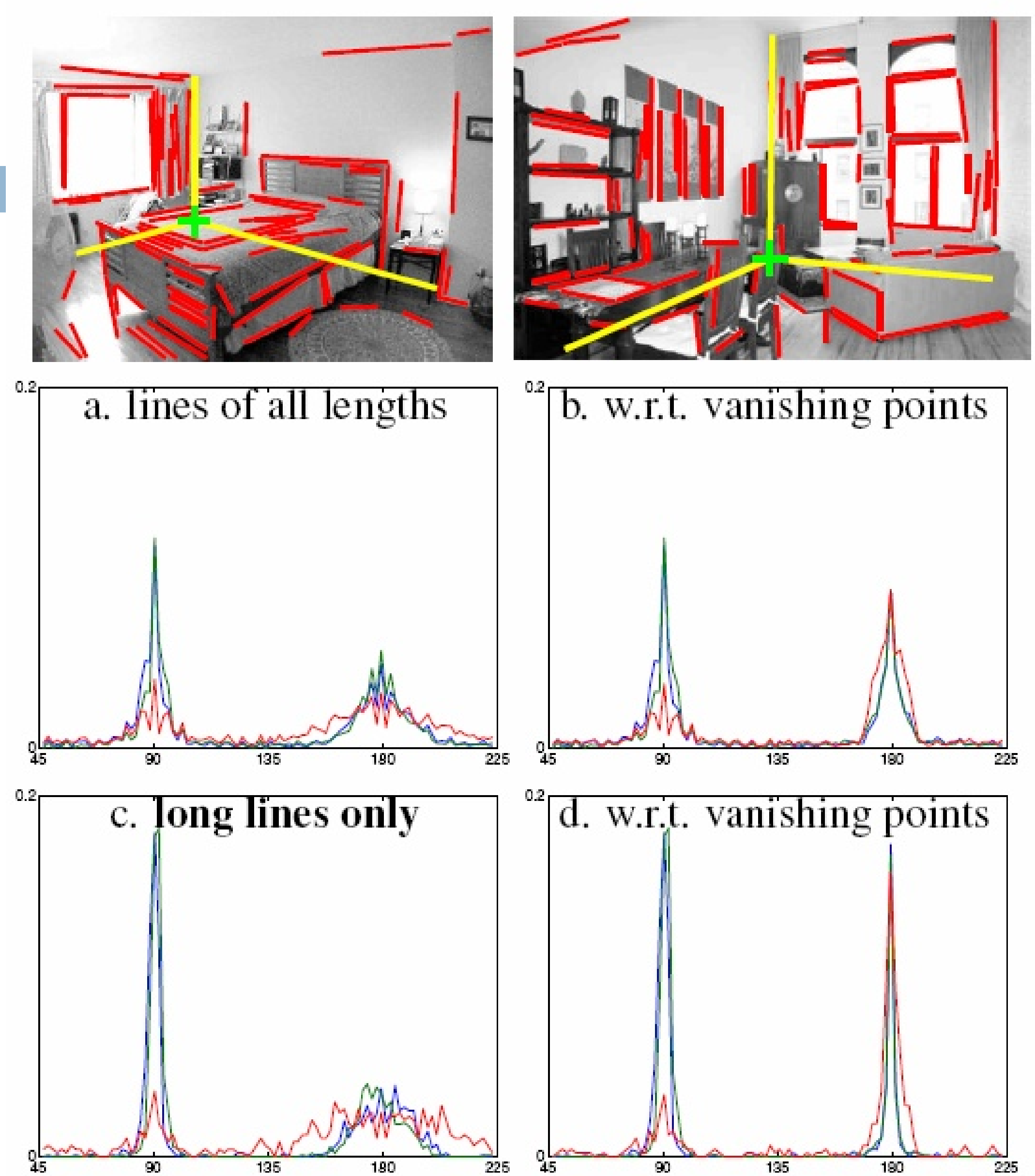
[Yu, Zhang, and Malik, Workshop on Perceptual Organization in Computer Vision 2008]

Edges

- The most time consuming operation
- Canny edge detection
- 5 seconds for a 400x400 image with a 2GHz CPU

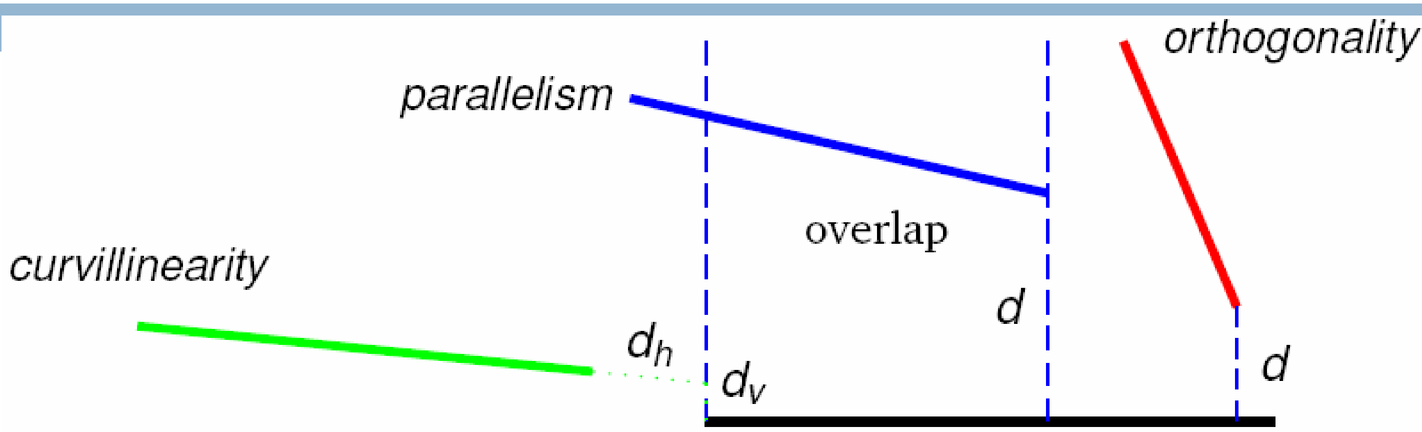
Lines

- Link edge pixels into line segments
- Short lines are ignored



[Yu, Zhang, and Malik, Workshop on Perceptual Organization in Computer Vision 2008]

Line Groups



$$A_{-} = \exp\left(-\frac{d_h^2}{2\sigma_{c1}^2} - \frac{d_v^2}{2\sigma_{c2}^2} - \frac{1-\cos^2\theta}{2\sigma_{c3}^2}\right)$$

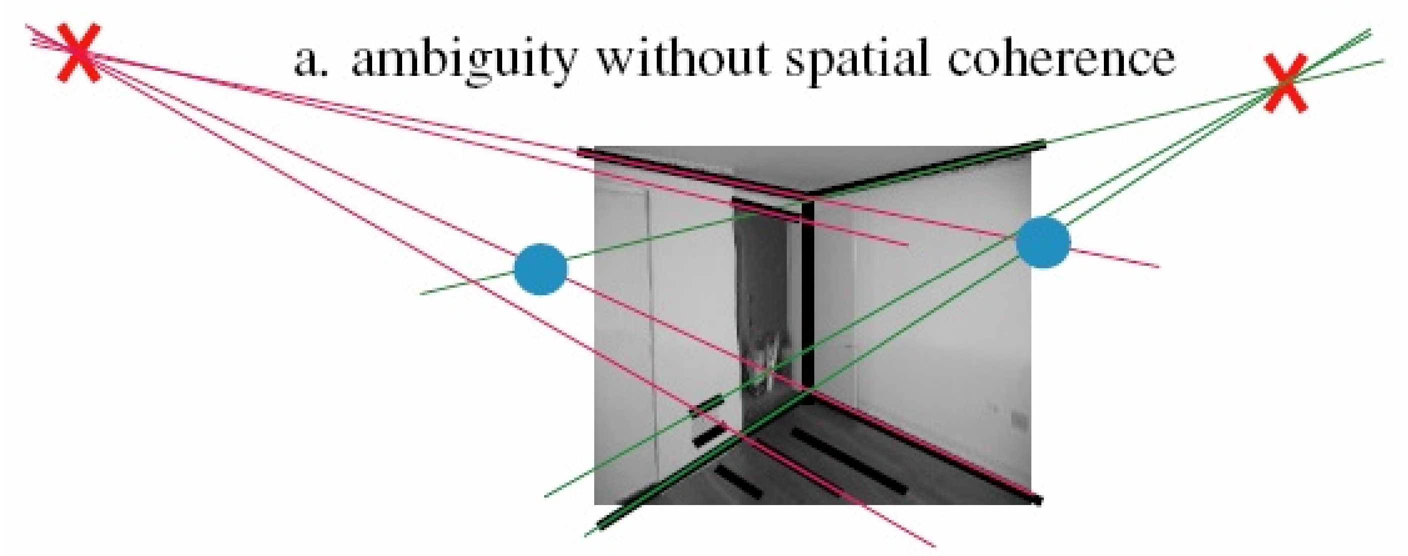
$$A_{\parallel} = \exp\left(-\frac{d^2}{2\sigma_{p1}^2} - \frac{(1-\mathrm{overlap})^2}{2\sigma_{p2}^2} - \frac{1-\cos^2\theta}{2\sigma_{p3}^2}\right)$$

$$R_{\perp} = \exp\left(-\frac{d^2}{2\sigma_{o1}^2} - \frac{\cos^2\theta}{2\sigma_{o2}^2}\right)$$

[Yu, Zhang, and Malik, Workshop on Perceptual Organization in Computer Vision 2008]

Line Groups

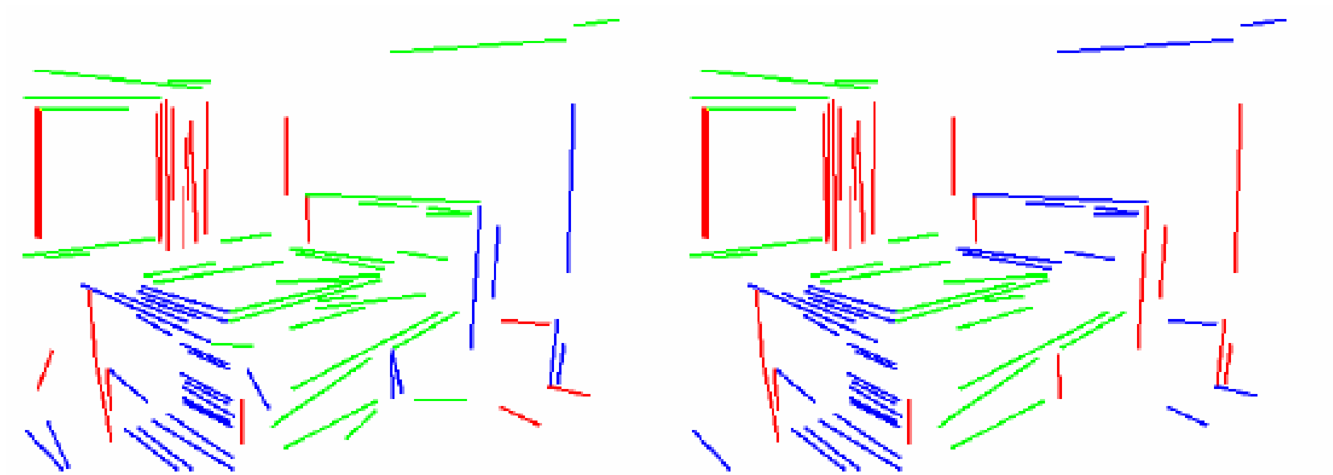
- Estimate vanish points (one for each of the three line clusters)



[Yu, Zhang, and Malik, Workshop on Perceptual Organization in Computer Vision 2008]

Line Groups

- A_ & $A_{||}$: measure how likely two lines belong to the same group – attraction
- $R_{\perp}$: measure how likely two lines belong to different groups – repulsion
- Pairwise attraction and repulsion in a graph cuts framework

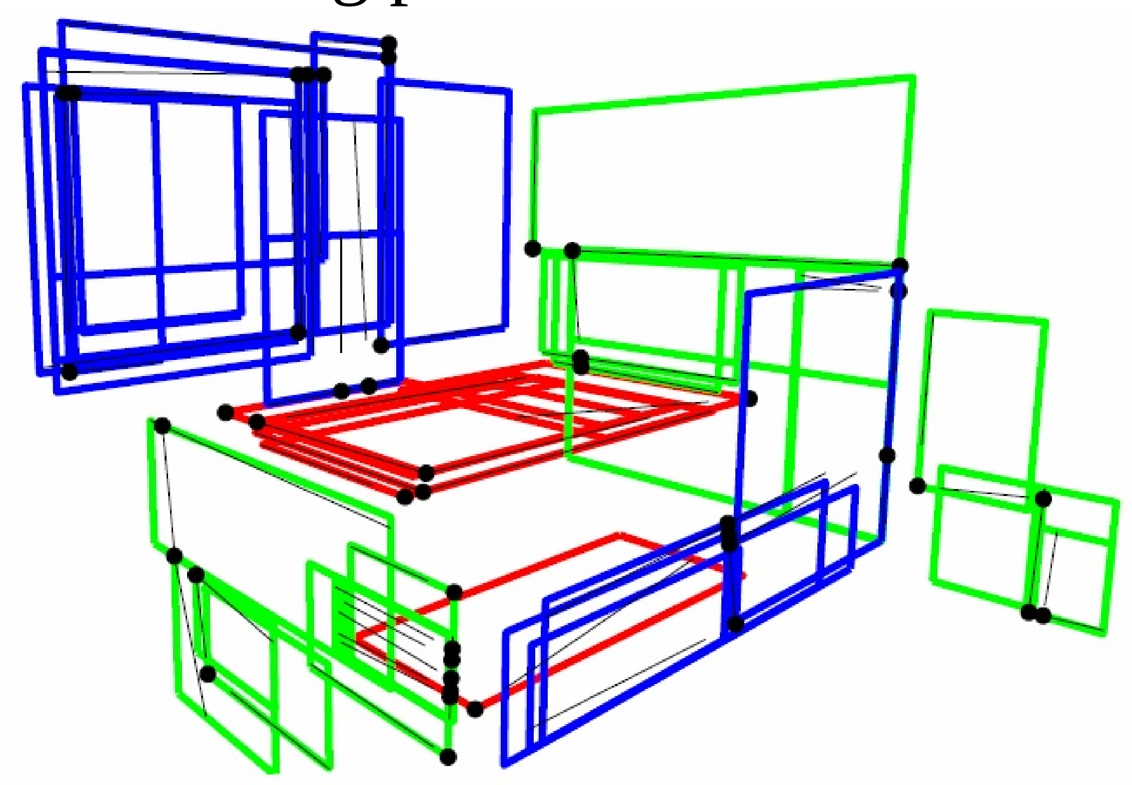


a: line grouping result   b: relabeled by vanishing points

[Yu, Zhang, and Malik, Workshop on Perceptual Organization in Computer Vision 2008]

Quadrilaterals

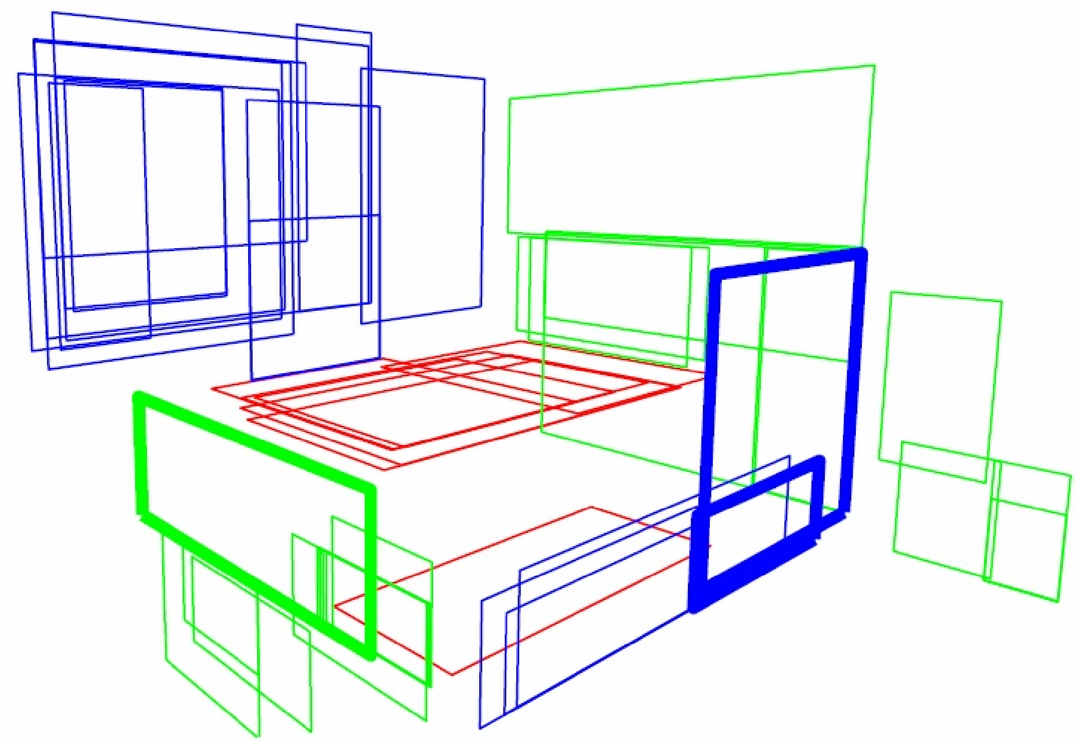
- Quadrilaterals are determined by adjacent lines and their vanishing points.



[Yu, Zhang, and Malik, Workshop on Perceptual Organization in Computer Vision 2008]

Depth Ordered Planes

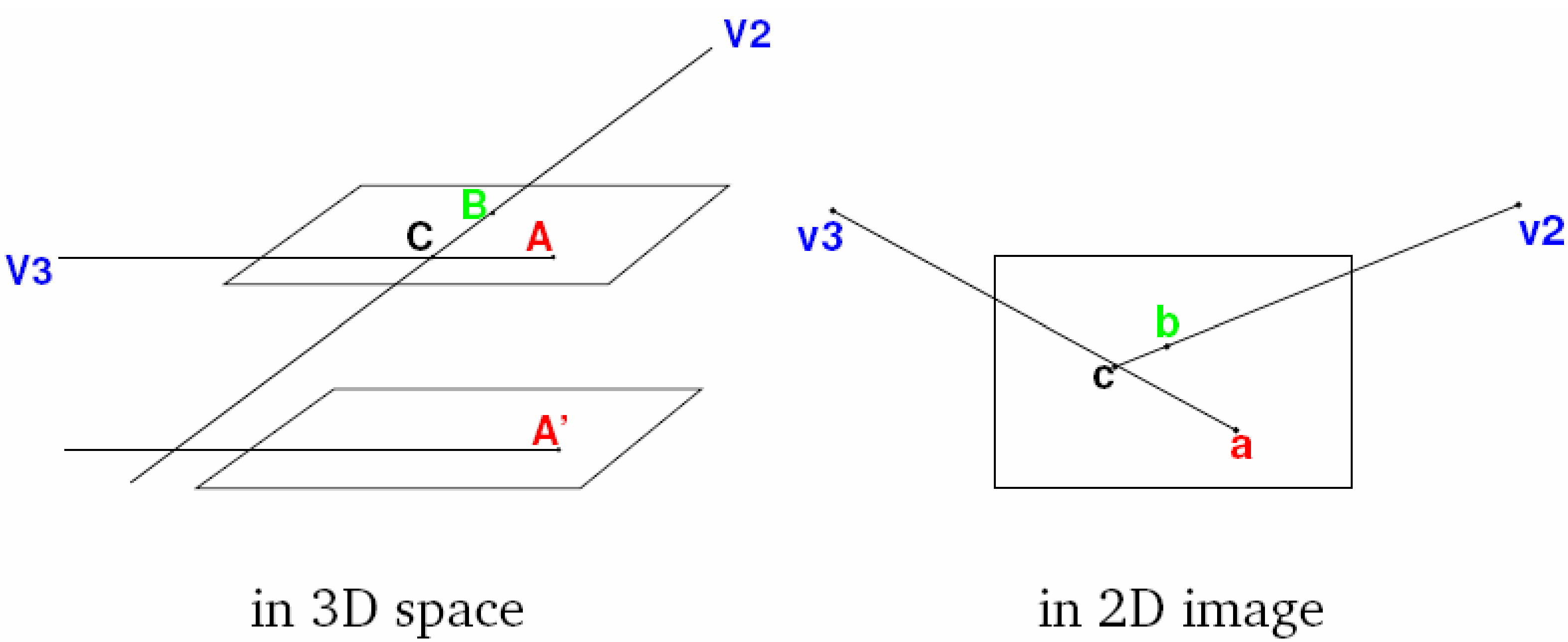
- Coplanarity: based on the degree of overlap, A□
- Rectify before measuring



[Yu, Zhang, and Malik, Workshop on Perceptual Organization in Computer Vision 2008]

Depth Ordered Planes

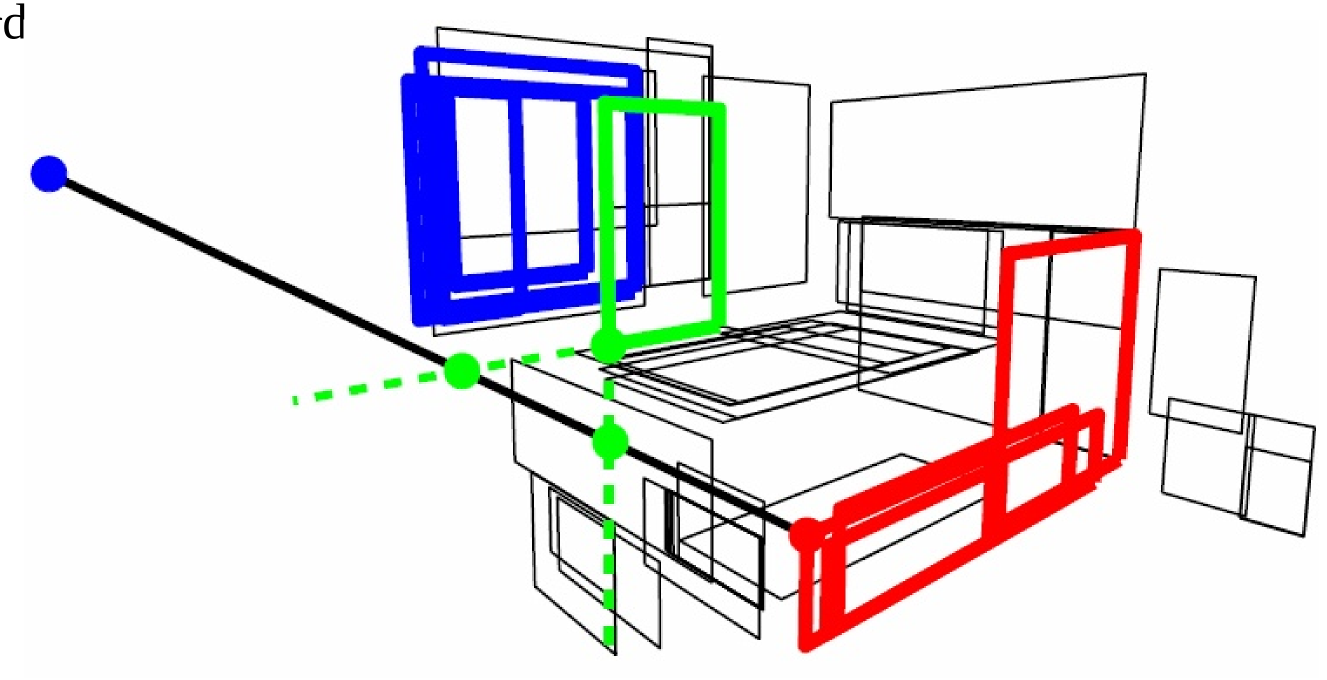
- Relative Depth



[Yu, Zhang, and Malik, Workshop on Perceptual Organization in Computer Vision 2008]

Depth Ordered Planes

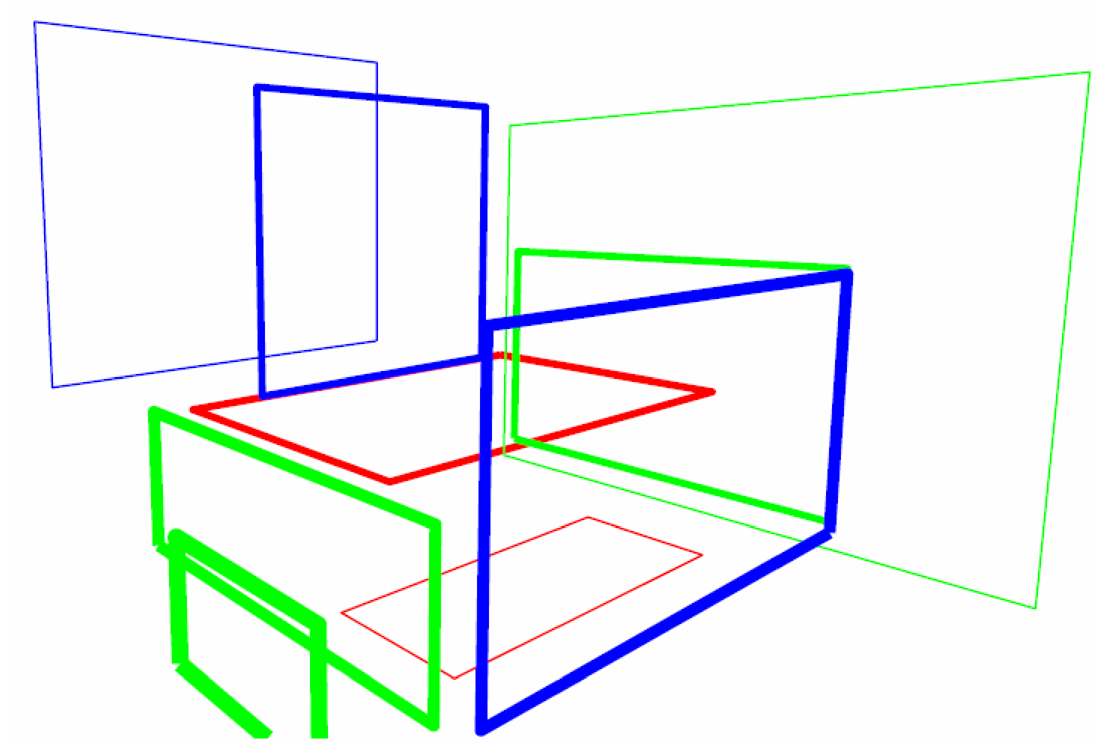
- The relative depth between two quadrilaterals is determined by the relative depth of their endpoints, R_d



[Yu, Zhang, and Malik, Workshop on Perceptual Organization in Computer Vision 2008]

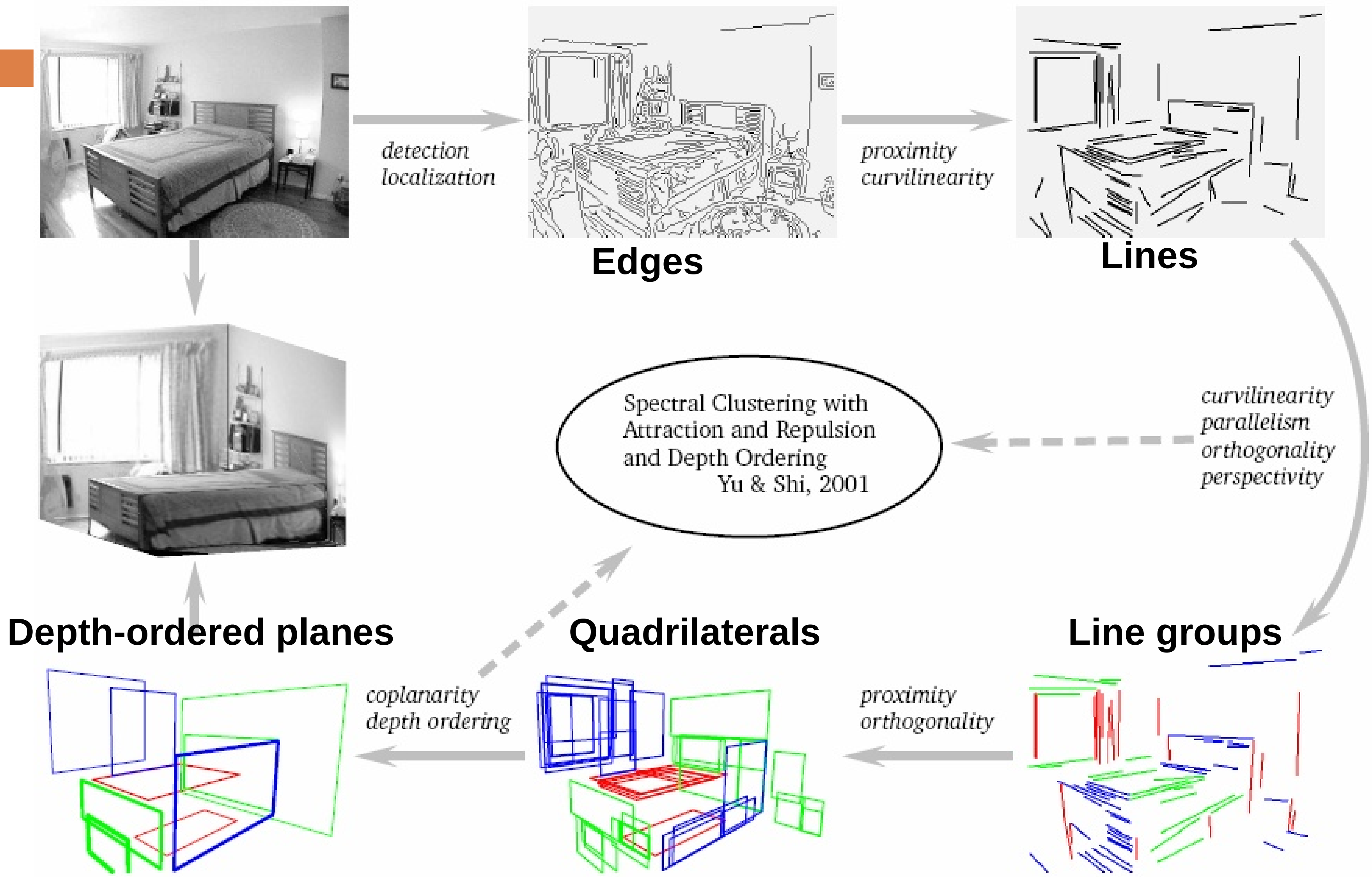
Depth Ordered Planes

- Pairwise attraction and directional repulsion in a graph cuts framework
  - Attraction: $A_{\square}$
  - Replusion: R_d



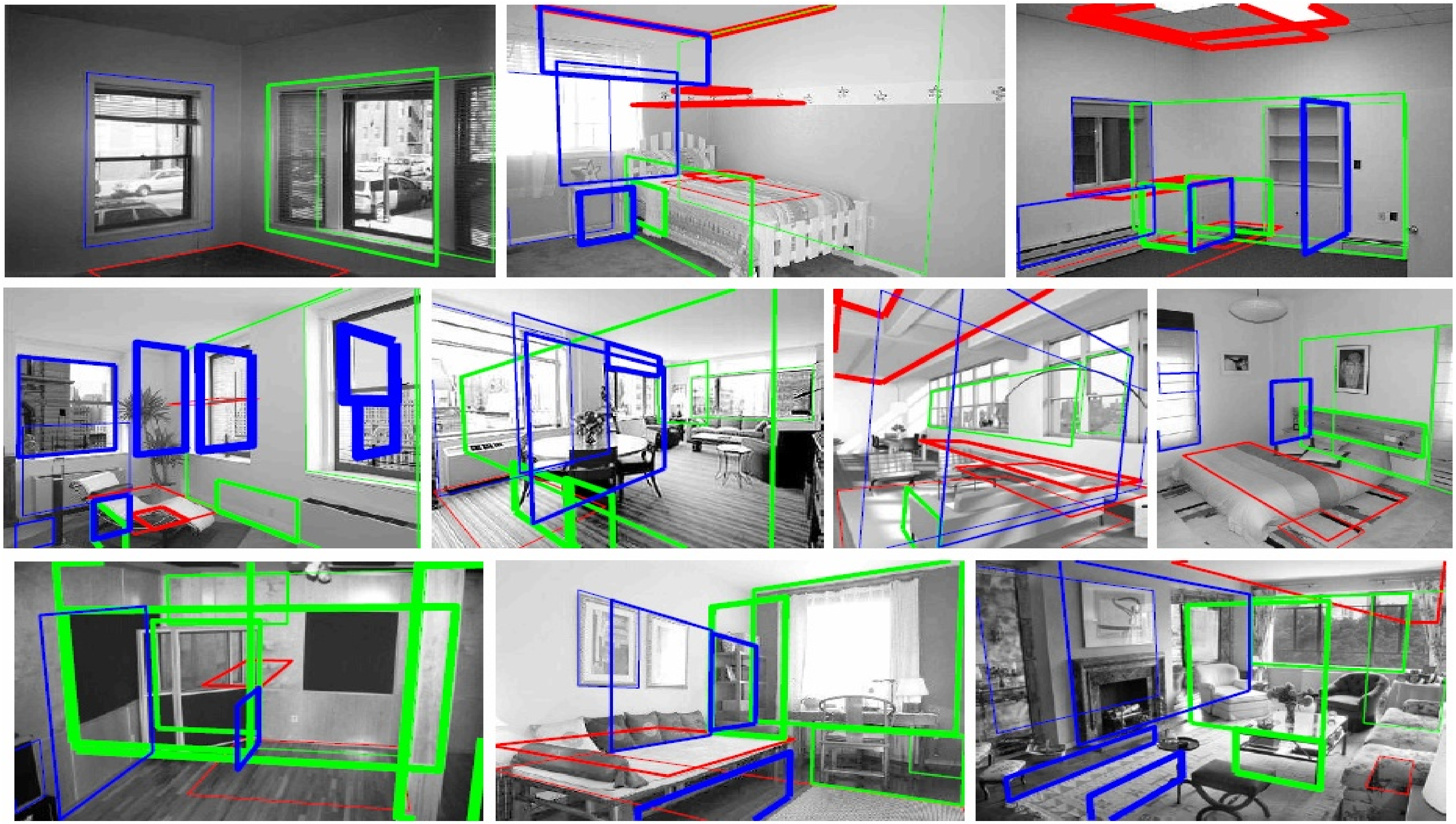
[Yu, Zhang, and Malik, Workshop on Perceptual Organization in Computer Vision 2008]

Line-Based Depth-Ordered Grouping Model



[Yu, Zhang, and Malik, Workshop on Perceptual Organization in Computer Vision 2008]

Results



[Yu, Zhang, and Malik, Workshop on Perceptual Organization in Computer Vision 2008]

Outline

- Inferring Spatial Layout from A Single Image via Depth-Ordered Grouping, by Stella X. Yu, Hao Zhang, and Jitendra Malik, Workshop on Perceptual Organization in Computer Vision, 2008
- Depth Estimation using Monocular and Stereo Cues, by A. Saxena, J. Schulte, and A. Ng. IJCAI 2007
- Comparison

Depth Estimation using Monocular and Stereo Cues

- Shortcomings of stereo vision
  - Fail for texture-less regions.
  - Inaccurate when the distance is large
- Monocular cues
  - Texture variations and gradients
  - Defocus
  - Haze
- Stereo and monocular cues are complementary
  - Stereo: image difference
  - Monocular: image content, prior knowledge about the environment and global structure are required.

Goal

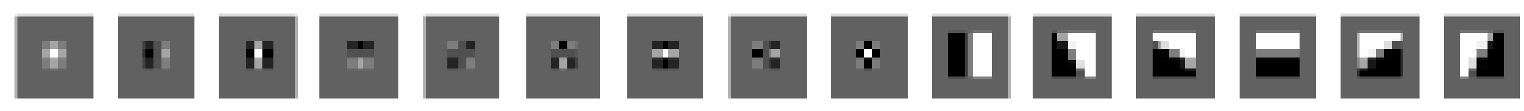
- 3-D scanner to collect training data
  - Stereo pairs
  - Ground truth depthmaps
- Estimate posterior distribution of the depths given the monocular image features and the stereo disparities
  - P(depths| monocular features, stereo disparities)

Visual Cues for Depth Estimation

- Monocular Cues
- Stereo Cues

Monocular Features

- 17 filters are used. 9 Laws' masks, 6 oriented edge filters, 2 color filters
  - Texture variation
  - Texture gradients
  - Color



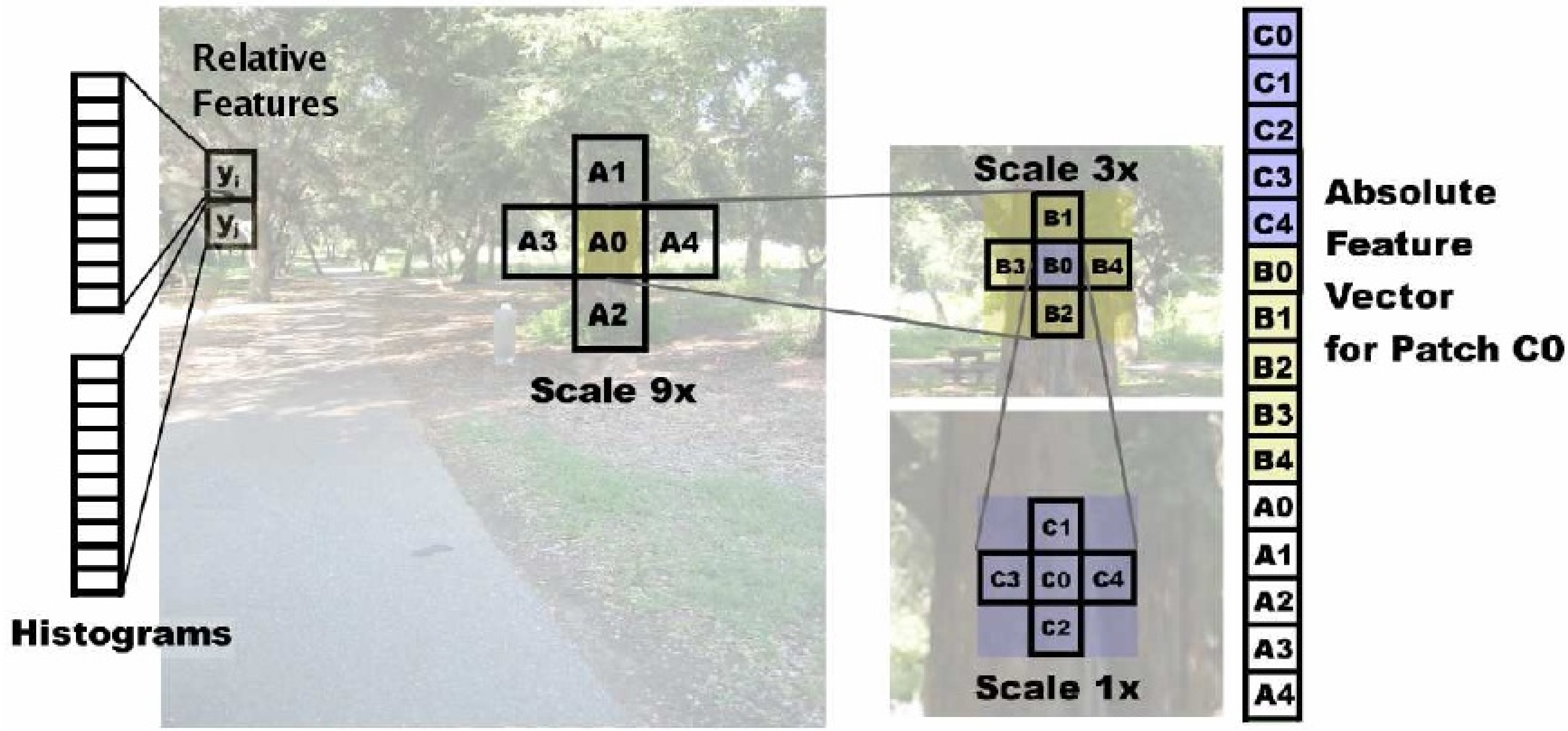
[Saxena, Schulte, and Ng, IJCAI 2007]

- An image is divided into rectangular patches, a single depth value is estimated for each patch

Monocular Features

- Absolute features
  - Sum-squared energy of each filter outputs over each patch
  - To capture global information, 4 neighboring patches at 3 spatial scales are concatenated.
  - Feature vector: (1+4)*3*17 = 255 dimensions
- Relative features
  - 10-bin histogram formed by the filter outputs of pixels in one patch. 10*17 = 170 dimensions

Monocular Features



[Saxena, Schulte, and Ng, IJCAI 2007]

Stereo Cues

- Use the sum-of-absolute-differences correlation as the metric score to find correspondences
- Find disparity
- Calculate the depth

Probabilistic Model

- Markov Random Field model
- P(d|X), X: monocular features of the patch, stereo disparity, and depths of other parts of the image

$$P_G(d|X;\theta,\sigma) = \frac{1}{Z_G}\exp\left(-\frac{1}{2}\sum_{i=1}^{M}\left(\frac{(d_i(1)-d_{i,\mathrm{stereo}})^2}{\sigma_{i,\mathrm{stereo}}^2} + \frac{(d_i(1)-x_i^T\theta_r)^2}{\sigma_{1r}^2} + \sum_{s=1}^{3}\sum_{j\in N_s(i)}\frac{(d_i(s)-d_j(s))^2}{\sigma_{2rs}^2}\right)\right)$$

the depth and stereo disparity

the depth and the features of patch i

Smoothness constraint

Learning

- θ_r : maximizing $p(d|X; \theta_r)$ of the training data. Assume all σ's are constant.
- Model σ^2_{2rs} as a linear function of the patches i and j's relative depth features y_{ijs}.
  - $\sigma^2_{2rs} = u_{rs}^T |y_{ijs}|$
- Model σ^2_{1r} as a linear function of x_i
  - $\sigma^2_{1r} = v_r^T x_i$

Laplacian Model

$$P_L(d|X;\theta,\lambda) = \frac{1}{Z_L}\exp\left(-\sum_{i=1}^{M}\left(\frac{|d_i(1)-d_{i,\text{stereo}}|}{\lambda_{i,\text{stereo}}} + \frac{|d_i(1)-x_i^T\theta_r|}{\lambda_{1r}} + \sum_{s=1}^{3}\sum_{j\in N_s(i)}\frac{|d_i(s)-d_j(s)|}{\lambda_{2rs}}\right)\right)$$

- The histogram of (d_i – d_j) is close to a Laplacian distribution empirically
- Laplacian is more robust to outliers
- Gaussian is not able to give depthmaps with sharp edges

Experiments

- Laser scanner on a panning motor
  - 67x54
- Stereo cameras
  - 1024x768
- 257 stereo pairs+depthmaps are obtained
  - 75% used for training, 25% used for testing
- Scenes
  - Natural environments
  - Man-made environments
  - Indoor environments

[Saxena, Schulte, and Ng, IJCAI 2007]

Experiments

- Baseline
- Stereo
- Stereo(smooth, Lap)
- Mono(Gaussian)
- Mono(Lap)
- Stereo+Mono(Lap)

$$P_G(d|X;\theta,\sigma) = \frac{1}{Z_G}\exp\left(-\frac{1}{2}\sum_{i=1}^{M}\left(\frac{(d_i(1)-d_{i,\text{stereo}})^2}{\sigma_{i,\text{stereo}}^2} + \frac{(d_i(1)-x_i^T\theta_r)^2}{\sigma_{1r}^2} + \sum_{s=1}^{3}\sum_{j\in N_s(i)}\frac{(d_i(s)-d_j(s))^2}{\sigma_{2rs}^2}\right)\right)$$

$$P_L(d|X;\theta,\lambda) = \frac{1}{Z_L}\exp\left(-\sum_{i=1}^{M}\left(\frac{|d_i(1)-d_{i,\text{stereo}}|}{\lambda_{i,\text{stereo}}} + \frac{|d_i(1)-x_i^T\theta_r|}{\lambda_{1r}} + \sum_{s=1}^{3}\sum_{j\in N_s(i)}\frac{|d_i(s)-d_j(s)|}{\lambda_{2rs}}\right)\right)$$

Results

Table 1: The average errors (RMS errors gave similar results) for various cues and models, on a log scale (base 10).

ALGORITHM	ALL	CAMPUS	FOREST	INDOOR
BASELINE	.341	.351	.344	.307
STEREO	.138	.143	.113	.182
STEREO (SMOOTH)	.088	.091	.080	.099
MONO (GAUSSIAN)	.093	.095	.085	.108
MONO (LAP)	.090	.091	.082	.105
STEREO+MONO (LAP)	**.074**	**.077**	**.069**	**.079**

[Saxena, Schulte, and Ng, IJCAI 2007]

Results

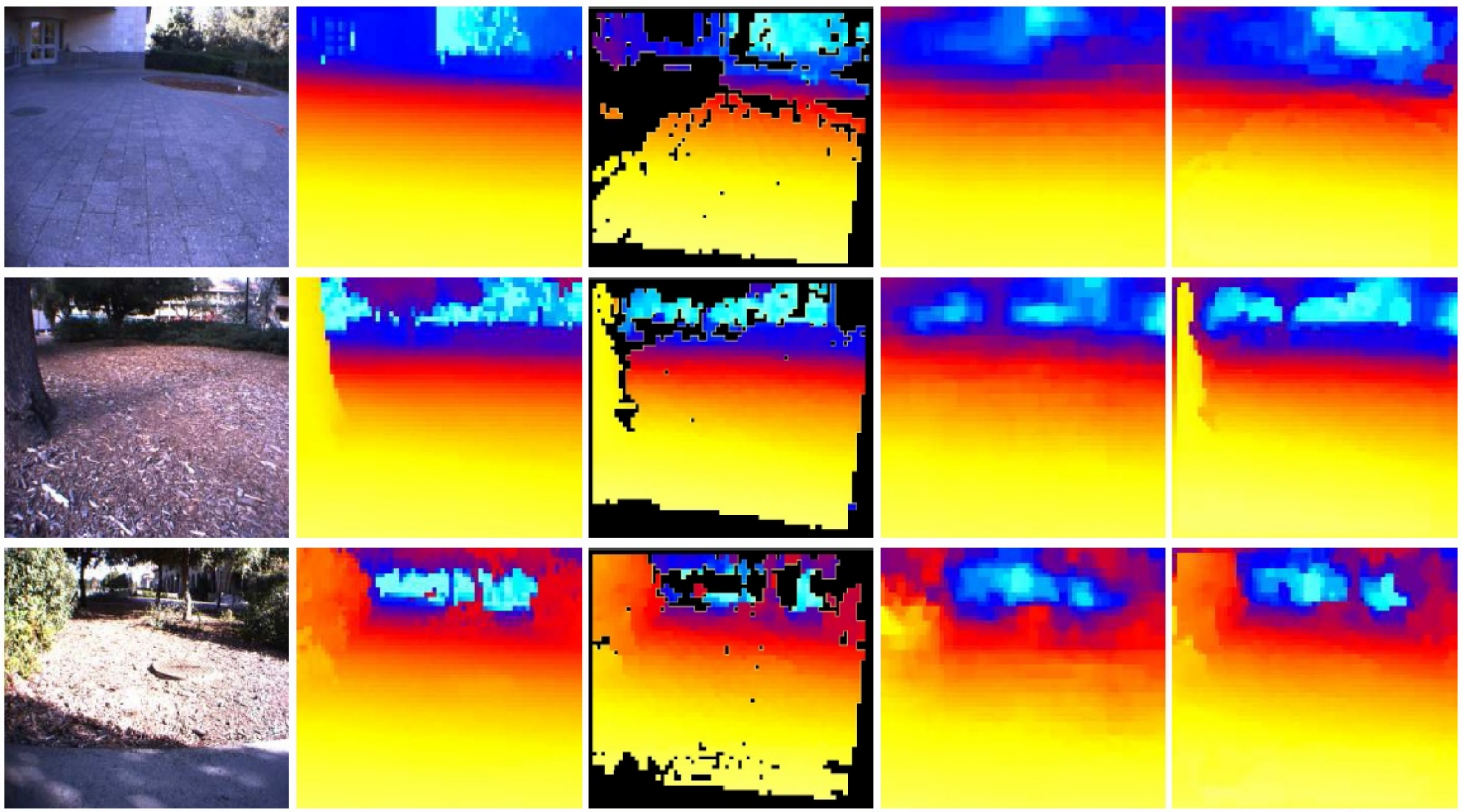
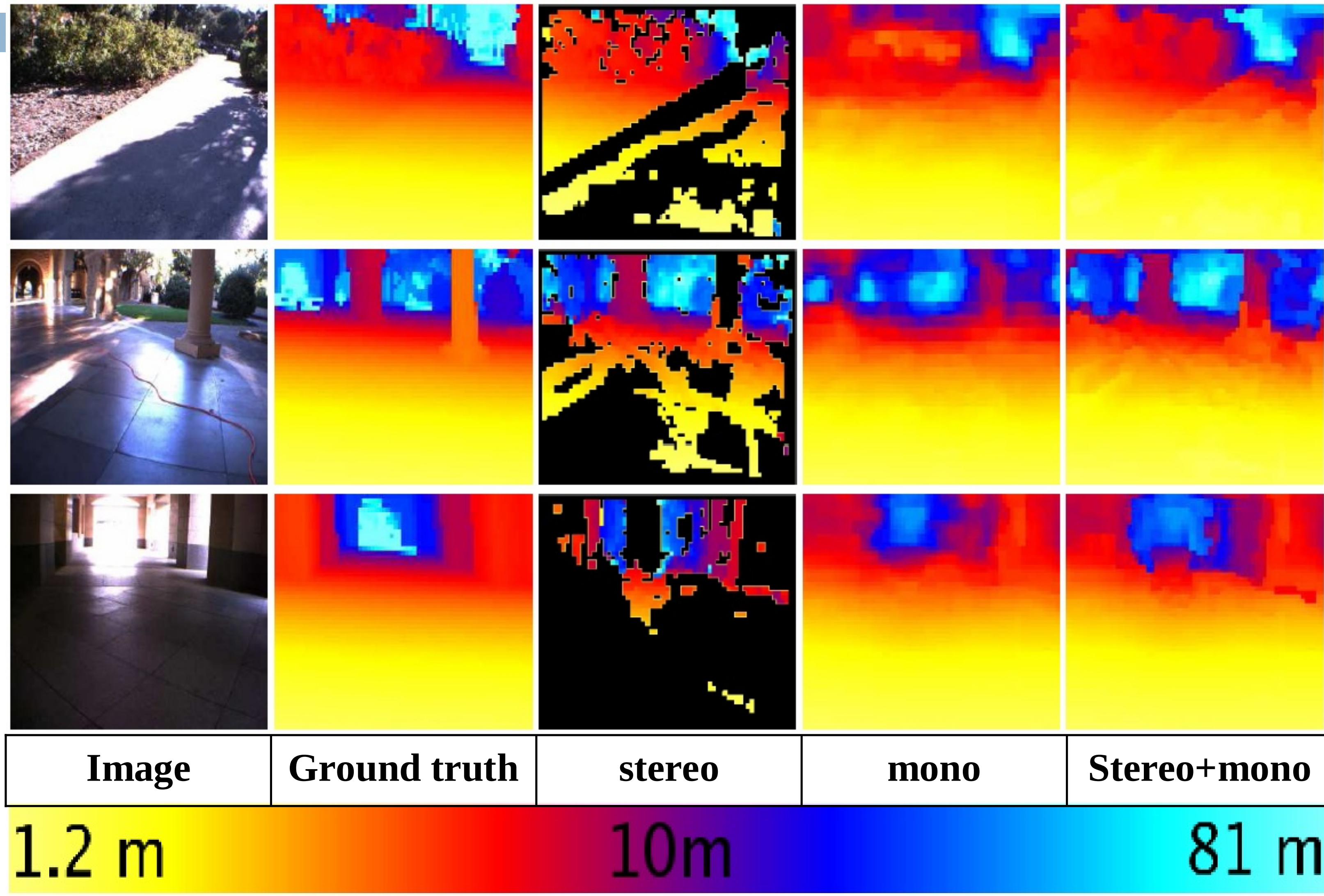


Image	Ground truth	stereo	mono	Stereo+mono

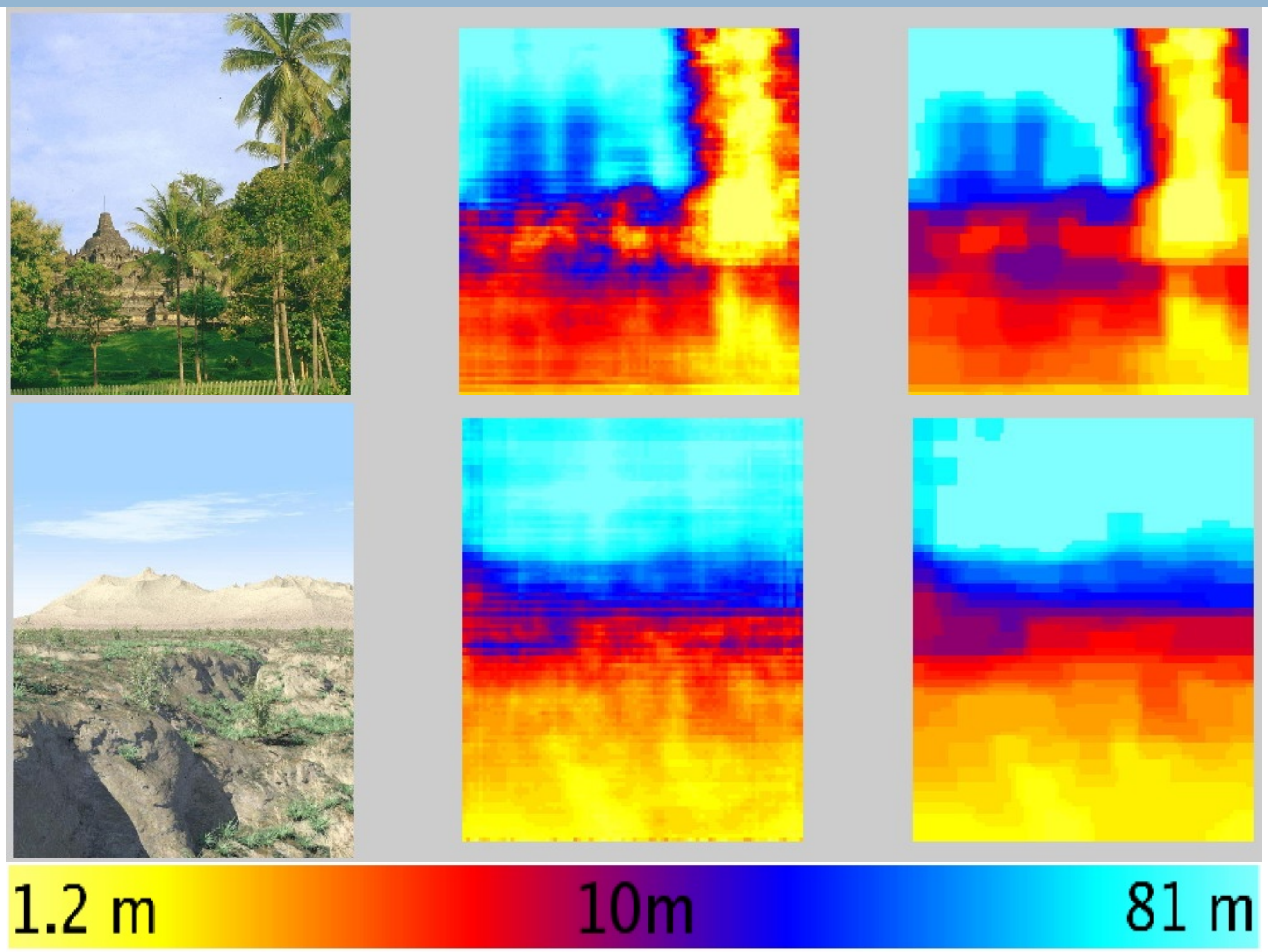
[Saxena, Schulte, and Ng, IJCAI 2007]

Results



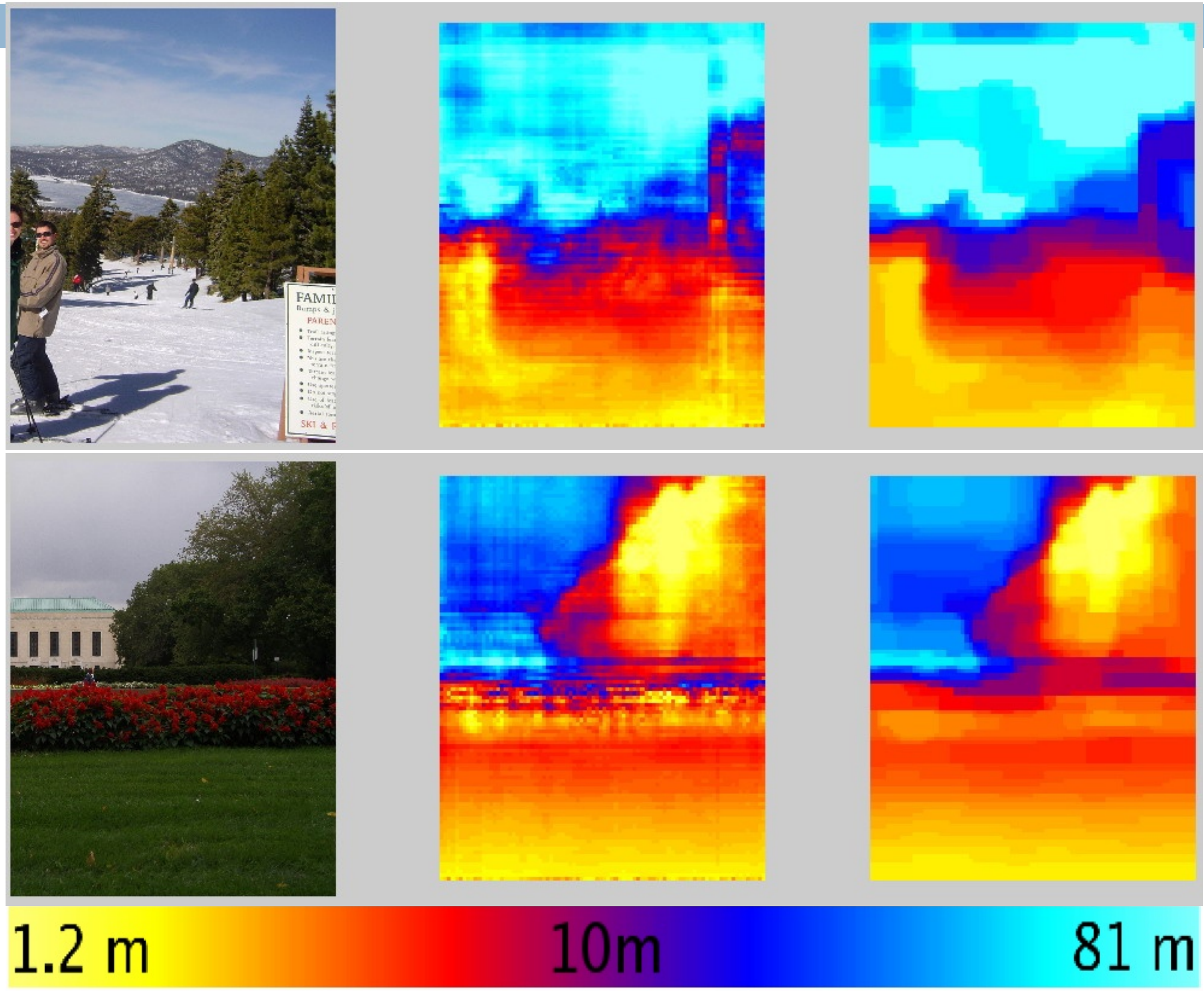
[Saxena, Schulte, and Ng, IJCAI 2007]

Test Images from Internet



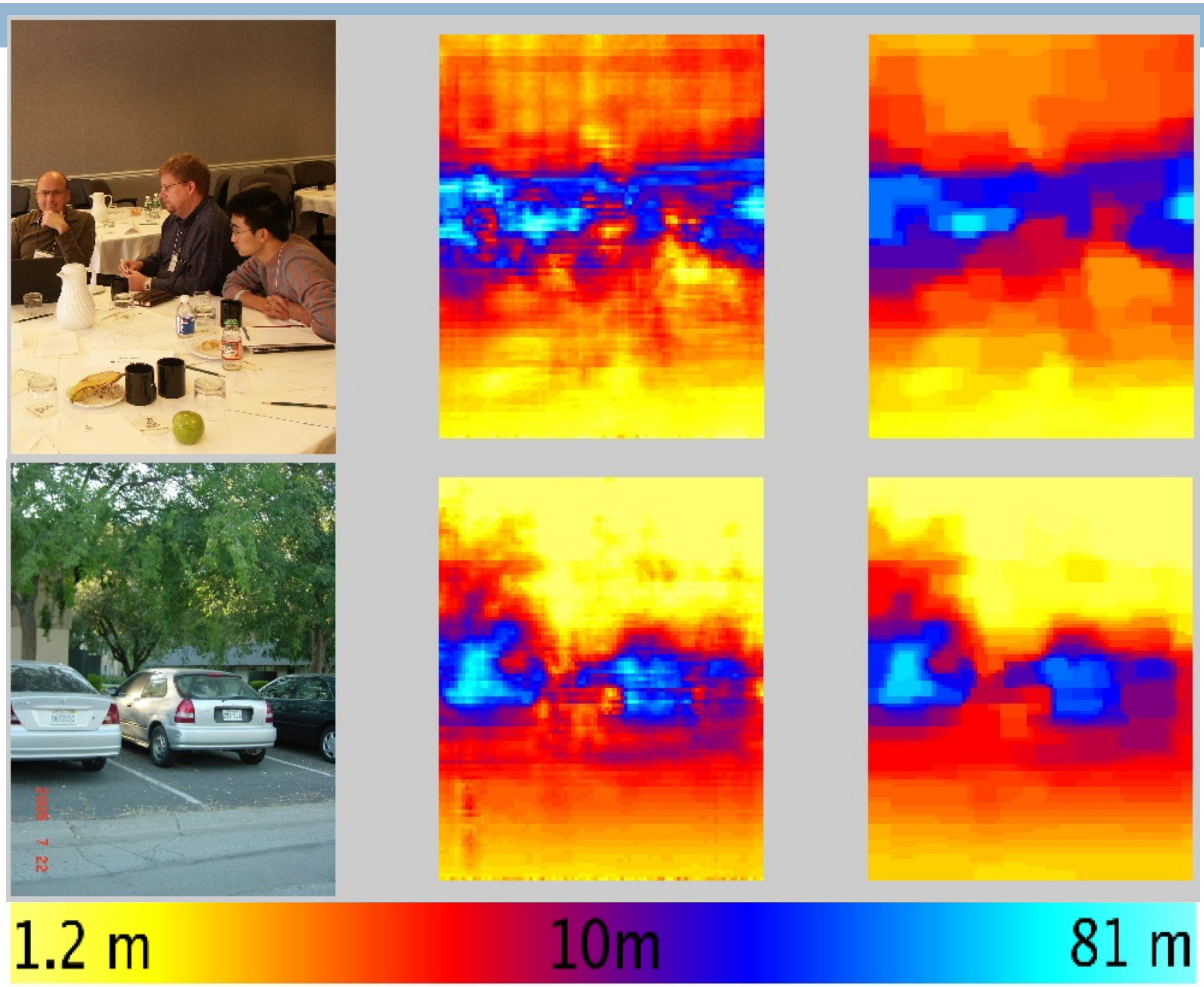
[http://ai.stanford.edu/~asaxena/learningdepth/others.html]

Test Images from Internet



[http://ai.stanford.edu/~asaxena/learningdepth/others.html]

Test Images from Internet



[http://ai.stanford.edu/~asaxena/learningdepth/others.html]

Results

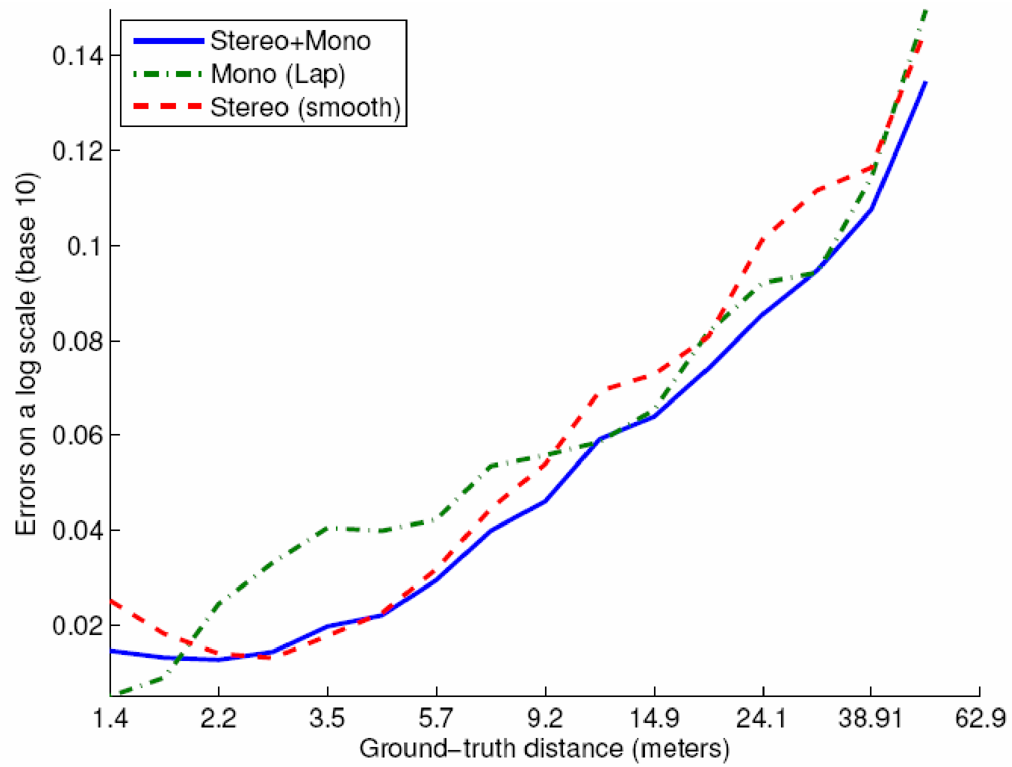


Figure 7: The average errors (on a log scale, base 10) as a function of the distance from the camera.

[Saxena, Schulte, and Ng, IJCAI 2007]

Outline

- Inferring Spatial Layout from A Single Image via Depth-Ordered Grouping, by Stella X. Yu, Hao Zhang, and Jitendra Malik, Workshop on Perceptual Organization in Computer Vision, 2008
- Depth Estimation using Monocular and Stereo Cues, by A. Saxena, J. Schulte, and A. Ng. IJCAI 2007
- Comparison

Comparison

- Depth order grouping [Zhang]
  - Geometrical
  - Learning is not required
  - Can be used only for indoor scenes
  - Estimate the relative depth between planes
  - Objects should be rectangular or quadrilaterals
- Depth estimation [Saxena]
  - Statistical
  - Learning is required.
  - May not generalize well on images very different from training samples
  - Can be used for both indoor and unstructured outdoor environments.
  - Estimate the absolute depth

Thank you

THE UNIVERSITY OF TEXAS AT AUSTIN
UT ECE
ELECTRICAL & COMPUTER ENGINEERING