Object Detection and Segmentation from Joint Embedding of Parts and Pixels

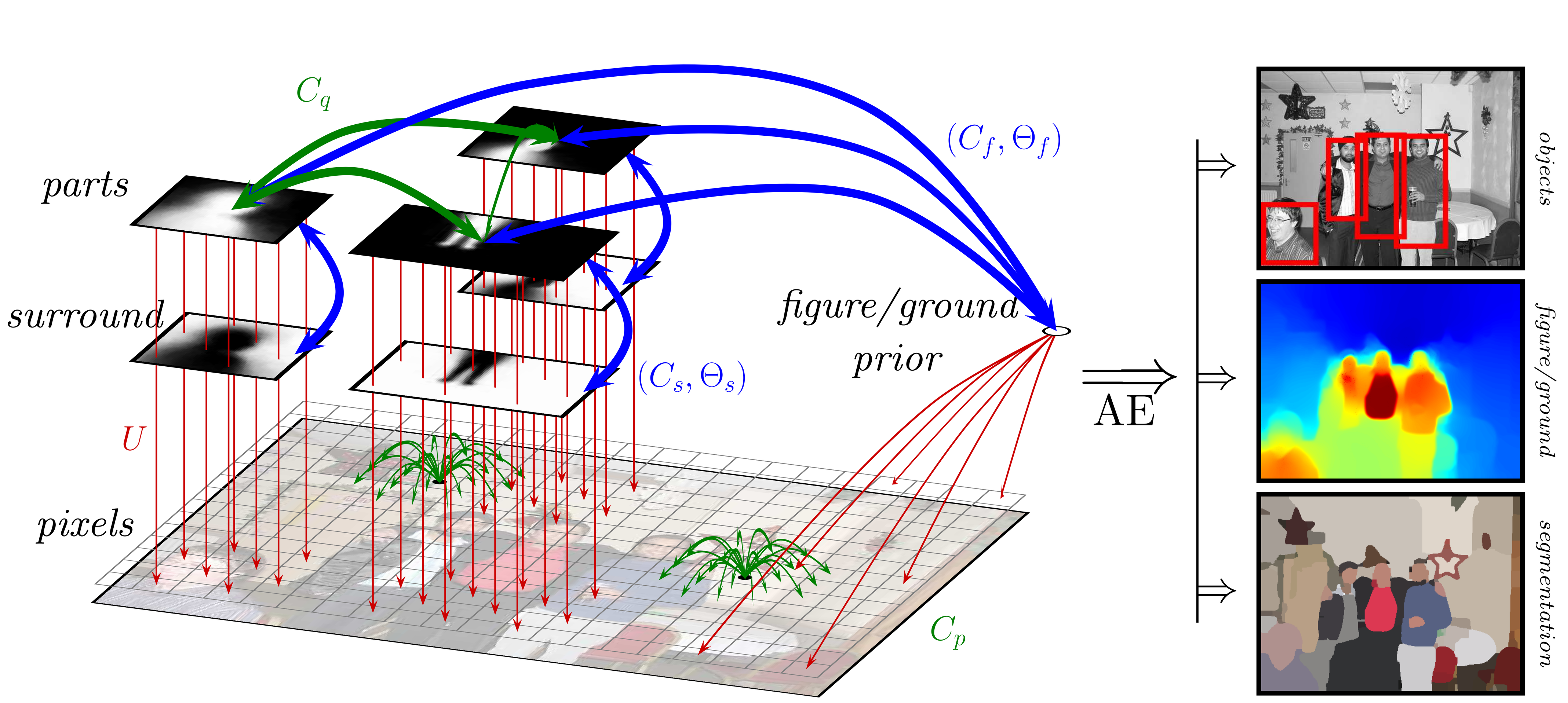
Michael Maire[1], Stella X. Yu[2], and Pietro Perona[1]

[1]*California Institute of Technology - Pasadena, CA 91125* [2]*Boston College - Chestnut Hill, MA 02467*

Abstract

We present a new framework in which image segmentation, figure/ground organization, and object detection all appear as the result of solving a single grouping problem. This framework serves as a perceptual organization stage that integrates information from low-level image cues with that of high-level part detectors. Pixels and parts each appear as nodes in a graph whose edges encode both affinity and ordering relationships. We derive a generalized eigenproblem from this graph and read off an interpretation of the image from the solution eigenvectors. Combining an off-the-shelf top-down part-based person detector with our low-level cues and grouping formulation, we demonstrate improvements to object detection and segmentation.

System Diagram



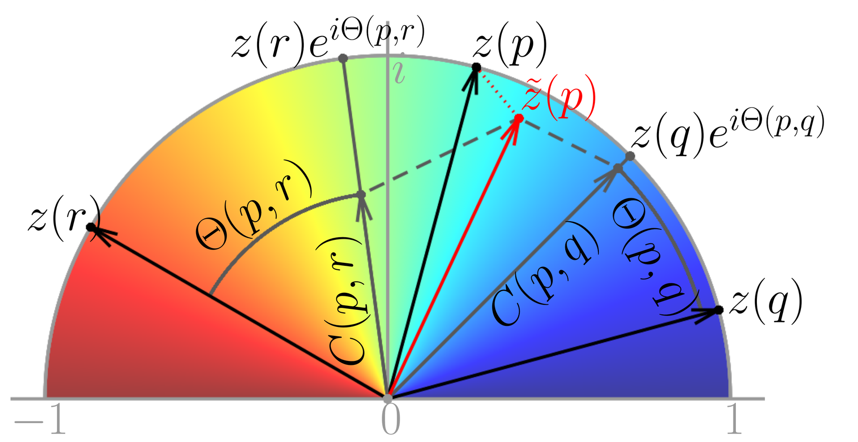
Integration: Angular Embedding

Given:

- Pairwise ordering relationships $\Theta(\cdot,\cdot)$
- Confidence on each relationship $C(\cdot,\cdot)$

Recover:

- Ordering $\theta(p)$ by embedding: $p \to z(p) = e^{i\theta(p)}$



Subject to:

- Linear constraints on solution in columns of U

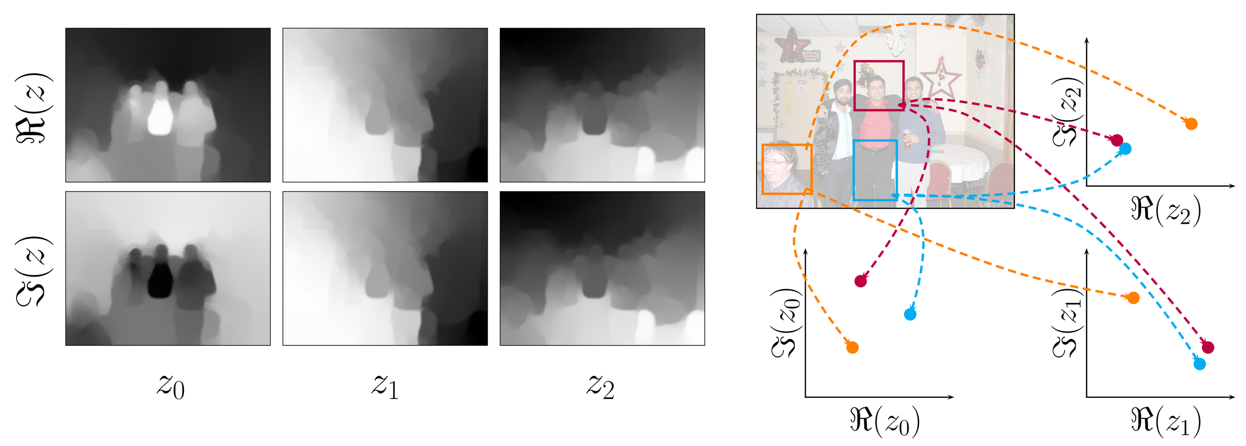
Minimize: $\varepsilon = \sum_p \frac{\sum_q C(p,q)}{\sum_{p,q} C(p,q)} \cdot |z(p) - \tilde{z}(p)|^2$ (subject to U)

Relax to generalized eigenproblem $QPQz = \lambda z$ where:

$$P = D^{-1}W$$
$$Q = I - D^{-1}U(U^T D^{-1} U)^{-1} U^T$$

with D and W defined as: $D = Diag(C1_n)$ and $W = C \bullet e^{i\Theta}$

Eigenvectors $\{z_0, z_1, ..., z_{m-1}\}$ embed pixels and parts into $\mathbb{C}^m$:



ONR MURI N00014-06-1-0734, ONR MURI 1015 G NA127, and ARL Cooperative Agreement W911NF-10-2-0016 supported this work.

Graph Setup: Pixel and Part Relations

Node types: pixels (p), parts (q), surround (s), figure/ground prior (f)

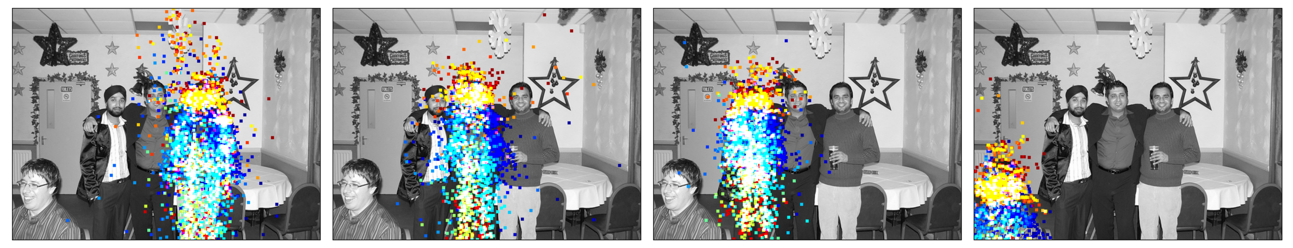
$$C = \begin{bmatrix} C_p & 0 & 0 & 0 \\ 0 & \alpha \cdot C_q & \beta \cdot C_s & \gamma \cdot C_f \\ 0 & \beta \cdot C_s^T & 0 & 0 \\ 0 & \gamma \cdot C_f^T & 0 & 0 \end{bmatrix} \quad \Theta = \Sigma^{-1} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \Theta_s & \Theta_f \\ 0 & -\Theta_s^T & 0 & 0 \\ 0 & -\Theta_f^T & 0 & 0 \end{bmatrix}$$

- Pixel-Pixel Affinity C_p determined by *intervening contour*
- Part-Part Affinity C_q depends on:
  - Part detection scores (using *poselets* of [2])
  - Pairwise part pose compatibility
- Part-Surround Repulsion (C_s, Θ_s):
  - Repulsion increases with part detector score
  - f acts as global surround node in (C_f, Θ_f)
- Constraints U:
  - Require part/surround nodes to agree with pixels they cover
  - Part embedding must equal mean embedding of its member pixels

Output: Decoding Eigenvectors

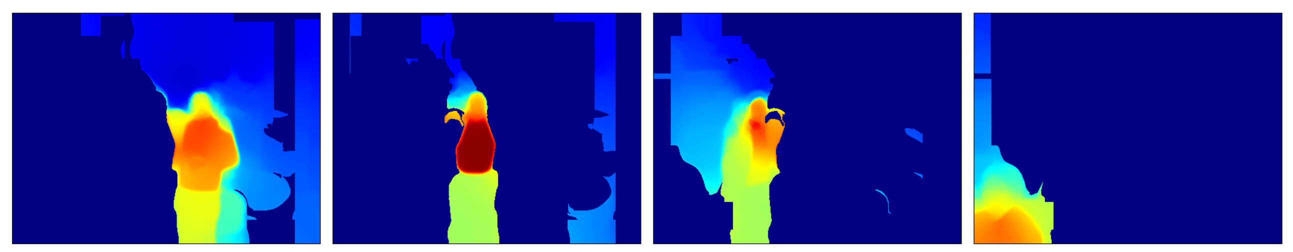
By design, the locations of the nodes in $\mathbb{C}^m$ are meaningful, in terms of both ordering (given by z_0) and clustering (given by $z_1, ..., z_{m-1}$). We "decode" the following from the eigenvectors:

- Figure/Ground Organization
  - $\angle z_0$ defines a global ordering, separating figure from ground [3]
- Image Segmentation
  - Embedding maps similar pixels to similar locations in $\mathbb{C}^m$
  - Using eigenvector gradients, ∇z_k, and image morphology, agglomeratively cluster pixels into a region hierarchy [1]
- Detected Object Instances
  - Agglomeratively cluster part nodes in $\mathbb{C}^m$ into object instances Q_i



  - Predict bounding boxes from clustered parts
- Segmentation of Each Object
  - Assign pixels to object instances:

$$p_k \to \underset{Q_i}{\operatorname{argmin}} \left\{ \min_{q_j \in Q_i} \{Dist(p_k, q_j)\} \right\}$$

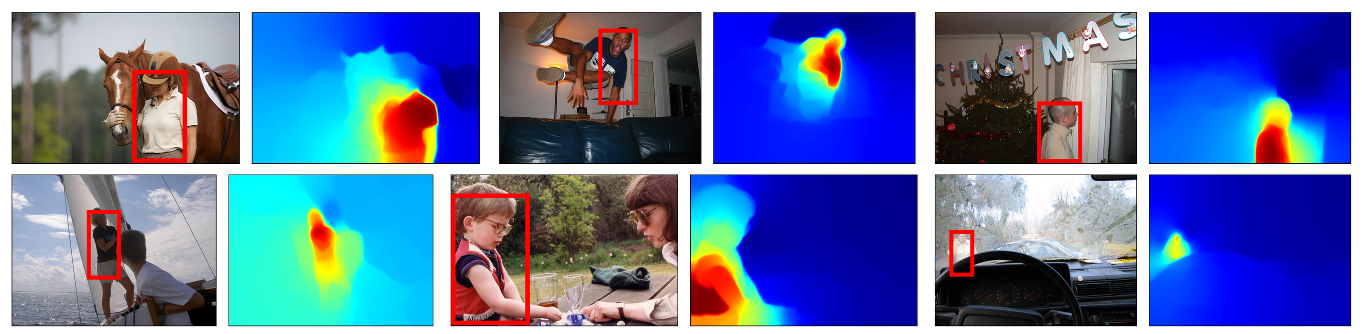


Results: Better Segmentation

Detections **Poselet Mask** **F/G Mask** **Segmentation**

Our final result (right column) scores 41.1 on the PASCAL VOC 2010 person segmentation task, compared to 35.5 for the poselet baseline.

Results: Improved Detection



Even on the segmentation benchmark, most of the gain (worth a score of 39.5) results from detection of otherwise missed people (examples above). Thus, integrating low-level cues boosts detection performance.

References

[1] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. Contour Detection and Hierarchical Image Segmentation. *PAMI*, 2011.

[2] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting People Using Mutually Consistent Poselet Activations. *ECCV*, 2010.

[3] M. Maire. Simultaneous Segmentation and Figure/Ground Organization using Angular Embedding. *ECCV*, 2010.

[4] S. X. Yu. Angular Embedding: A Robust Quadratic Criterion. *PAMI*, 2011.