Foreground Clustering for Joint Segmentation and Localization in Videos and Images

Abhishek Sharma

Problem Introduction

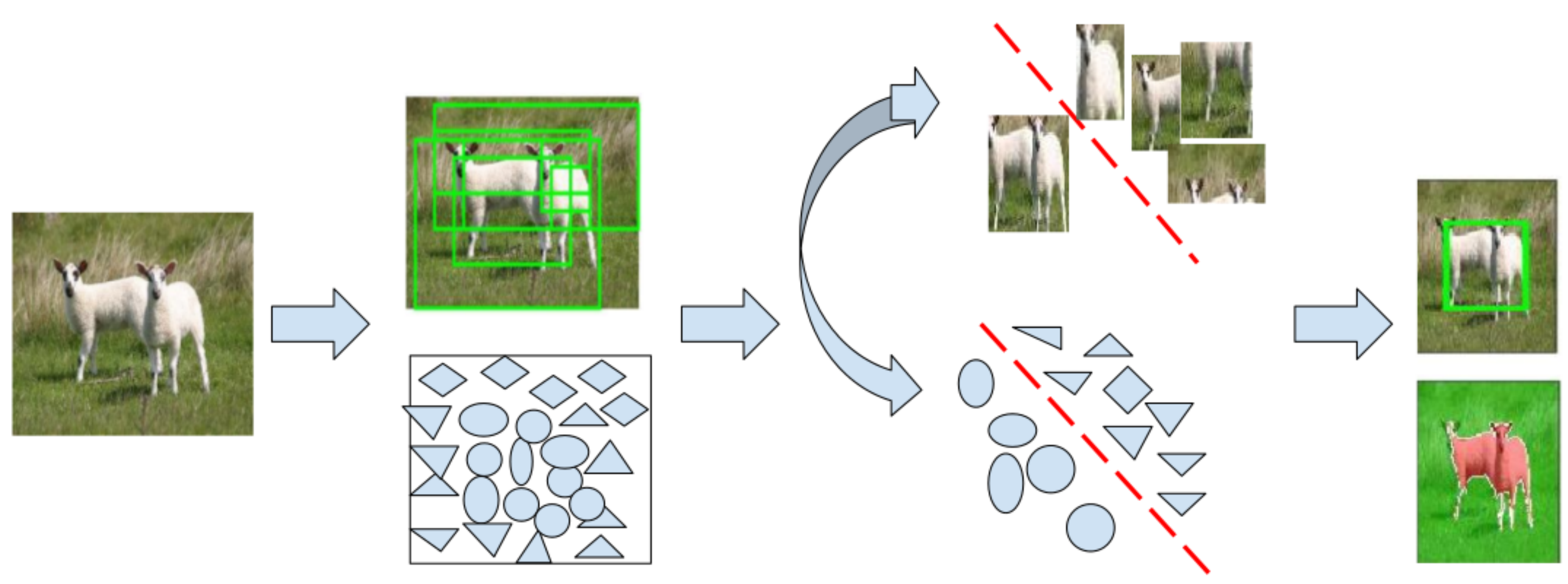
- Segmentation and Localization are similar tasks and yet modelled separately in images, videos and 3D data under **weak** supervision.
- How to exploit semantic cues of boxes to guide segmentation and leverage low level appearance cues at superpixel level to improve localization.
- Can we define a notion of similarity in a totally discriminative classifier to model video data where the background tends to be not discriminative.

Background: Discriminative Clustering

- Over all labelling, find one that gives max margin classifier(Xu *et al.* NIPS 05)
- For square loss, problem reduces to convex optimization and closed form solution exist for this problem [1]:

$$\min_{y \in \{0,1\}^n, w \in \mathbb{R}^d} ||y - \mathcal{X}w - b\mathbf{1}||^2 + \beta ||w||^2,$$

Contribution 1: Learning from Constraints



- Key Idea: If an object localization classifier considers some bounding box to be a background, this should enforce the segmentation classifier that superpixels in this bounding box are more likely to be background and vice-versa.

Contribution 2: Foreground Model

- Bring a notion of similarity in a purely discriminative model by including a histogram matching term that minimizes the discrepancy between the segmented foreground.
- A histogram can be written as a **vector** $h = \mathcal{H}y$ where $H_{ij} = 1$ if the feature associated with pixel j falls in bin number i of the histogram, and $H_{ij} = 0$. y is a binary indicator variable for pixels.
- Norm of vector difference is convex by definition.

Optimization Problem for one Image

Given a set of m bounding boxes per image, with a binary vector z in $\{0,1\}^m$, n superpixels with a binary vector y in $\{0,1\}^n$, and for each bounding box, a set S_i of its superpixels and the corresponding binary vector x_i in $\{0,1\}^{|S_i|}$:

$$\min_{y,z} \quad E(y,z) \quad \text{under the constraints:}$$

$$\gamma |S_i| z_i \leq \sum_{j \in S_i} x_{ij} \leq \eta |S_i| z_i \qquad \text{for} \quad i = 1, \ldots, m, \tag{1}$$

$$\sum_{i: j \in S_i} x_{ij} \leq \sum_{i: j \in S_i} z_i, \qquad \text{for} \quad j = 1, \ldots, n, \tag{2}$$

$$\mathcal{P}_i y = x_i, \qquad \text{for} \quad i = 1, \ldots, m. \tag{3}$$

$$\sum_{i=1}^{m} z_i = 1 \tag{4}$$

Experimental Evaluation

Baselines

- Sal. : only minimizes the saliency term and picks the most salient one.
- Loc. : optimizes the localization problem alone.
- Seg. : optimizes the segmentation problem alone.
- (Seg. + Loc.): optimizes the combined problem of segmentation and localization.
- Ours(full): optimizes (Seg. + Loc.) + Foreground model.

Result on Youtube Video Dataset [5]

Table: Video Colocalization on Youtube Objects dataset.

Metric	Sal.	[3]	Loc.	(Loc.+Seg)	Ours(full)	[4]
CorLoc.	28	31	35	49	54	56

Table: Video segmentation on Youtube Objects dataset.

Metric	Sal.	Seg.	(Seg. +Loc.)	Ours(full)	FST [5]
IoU.	43	49	56	61	53

Image Colocalization Results

Table: Image Colocalization Comparison on Object Discovery dataset.

Metric	Sal.	Loc.	TJLF14 [6]	Ours(full)	CSP15 [2]
CorLoc.	68	75	72	80	84

Table: Image Colocalization Comparison on Pascal VOC 2007.

Metric	Sal.	Loc.	TJLF14 [6]	Ours(full)	CSP15 [2]
CorLoc.	33	40	39	51	68

Conclusion and Future Work

- We proposed a simple framework based on two different level of visual representations that uses linear constraints as a means to transfer intrinsic information in an unsupervised manner.
- The key idea of transferring knowledge between tasks via spatial relation is very general and will encourage frameworks such as constrained CNN to model multiple tasks under weak supervision.
- Source Code: **https://github.com/Not-IITian/Foreground-Clustering-for-Joint-Segmentation-and-Localization**

References

- F. Bach and Z. Harchaoui.
  Diffrac: a discriminative and flexible framework for clustering.
  In *NIPS*, 2007.
- M. Cho, S. Kwak, C. Schmid, and J. Ponce.
  Unsupervised object discovery and localization in the wild.
  In *CVPR*, 2015.
- A. Joulin, K. Tang, and L. Fei-Fei.
  Efficient image and video co-localization with frank-wolfe algorithm.
  In *ECCV*, 2014.
- S. Kwak, M. Cho, I. Laptev, J. Ponce, and C. Schmid.
  Unsupervised object discovery and tracking in video collections.
  In *ICCV*, 2015.
- A. Papazoglou and V. Ferrari.
  Fast object segmentation in unconstrained video.
  In *CVPR*, 2013.
- K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei.
  Co-localization in real-world images.
  In *CVPR*, 2014.