Visual Parsing with Weak Supervision

Jia Xu

Department of Computer Sciences
University of Wisconsin-Madison

2015-07-30

THE UNIVERSITY of WISCONSIN MADISON

Research Goal

Teach Computer to See at/beyond Human Level

- Interpret/summarize/organize visual data on the Internet
- Help the disabled population (e.g., the blind)

Visual Parsing

Fundamental Task

- Semantically parse every pixel in images and videos

Visual Parsing

Fundamental Task

- Semantically parse every pixel in images and videos
- First step towards high level applications

Self-driving Car

Unmanned Aerial Vehicle

Wearable Glasses

Visual Parsing

Fundamental Task

Turning Visual Data Into Knowledge

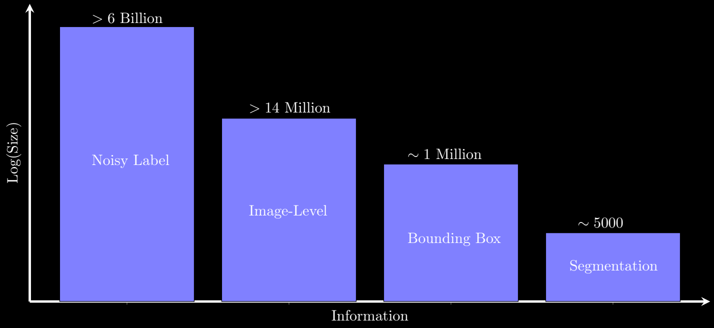
flickr facebook YouTube

Everyday > 3.5 million > 300 million $> 150,000$ hours

- Never Ending Language Learning (Mitchell et al., 2009)
- Never Ending Image Learner (Chen et al., 2013)

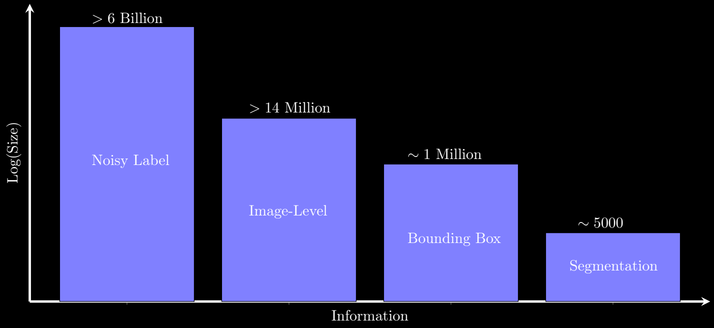
Challenges

Modern Image Dataset



Challenges

Modern Image Dataset



Much fewer segmentations are annotated for videos!

Motivation

Bottleneck of Fully Supervised Methods

- Full annotation is expensive to collect and limited at size

Motivation

Bottleneck of Fully Supervised Methods

- Full annotation is expensive to collect and limited at size

Why Weakly Supervised Learning

- Weak supervision is easier to obtain: e.g., gaze

Motivation

Bottleneck of Fully Supervised Methods

- Full annotation is expensive to collect and limited at size

Why Weakly Supervised Learning

- Weak supervision is easier to obtain: e.g., gaze
- Large datasets with side/weak annotations are readily available: metadata, tags, text

Motivation

Bottleneck of Fully Supervised Methods

- Full annotation is expensive to collect and limited at size

Why Weakly Supervised Learning

- Weak supervision is easier to obtain: e.g., gaze
- Large datasets with side/weak annotations are readily available: metadata, tags, text
- Visual data presents the physical world: shape, geometry, context

My Thesis Research

- How can we utilize weakly labeled data effectively for the visual parsing task?
- When human comes into the visual parsing loop, how can we minimize user effort while still achieving satisfactory parsing results?

Roadmap

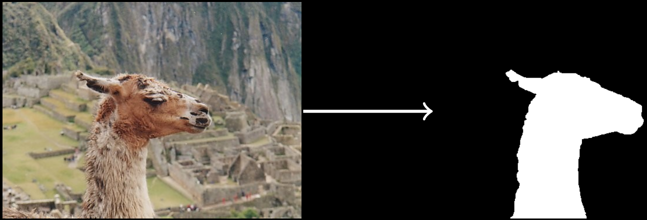
Chapter	Parsing Task	Weak Supervision	Publication
Ch. 2	Object Segmentation	User Indication	CVPR 2013
Ch. 3	Scene Parsing	Image-level Tags	CVPR 2014
Ch. 4	Scene Parsing	Image-level Tags Bounding Boxes Partial Labels	CVPR 2015a
Ch. 5	Video Segmentation	Side Knowledge	ICCV 2013
Ch. 6	Video Summarization	Human Gaze	CVPR 2015b

Roadmap

Chapter	Parsing Task	Weak Supervision	Publication
Ch. 2	Object Segmentation	User Indication	CVPR 2013
Ch. 3	Scene Parsing	Image-level Tags	CVPR 2014
Ch. 4	Scene Parsing	Image-level Tags Bounding Boxes Partial Labels	CVPR 2015a
Ch. 5	Video Segmentation	Side Knowledge	ICCV 2013
Ch. 6	Video Summarization	Human Gaze	CVPR 2015b

Object Segmentation

Object Segmentation



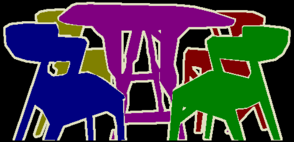
Main Challenges

1. Semantic gap: what is an object?

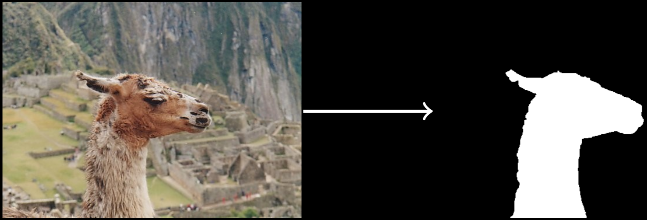
Object Segmentation

Main Challenges

1. Semantic gap: what is an object?
2. Ambiguity of user intention: which object do you want?



Interactive Object Segmentation



Main Challenges

1. Semantic gap: what is an object?
2. Ambiguity of user intention: which object do you want?

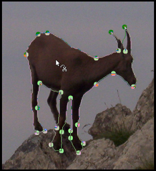
A few user scribbles can make segmentation much easier!

Related work

- Region-based: Graphcut (Boykov and Jolly, 2001), Grabcut (Rother et al., 2004), Random Walks (Grady, 2006), Geodesic Shortest Path (Bai and Sapiro, 2009), Geodesic Star Convexity (Gulshan et al., 2010)
- Edge-based: Intelligent Scissors (Mortensen and Barrett, 1998), LabelMe (Russell et al., 2008)

GraphCut

GrabCut



Intelligent Scissors

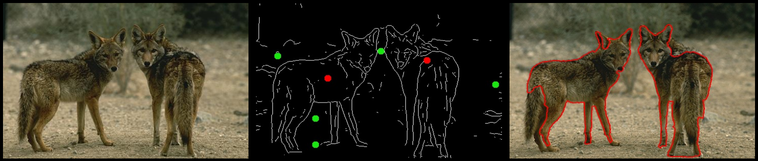
LabelMe

Our Ideas (EulerSeg)

Objective

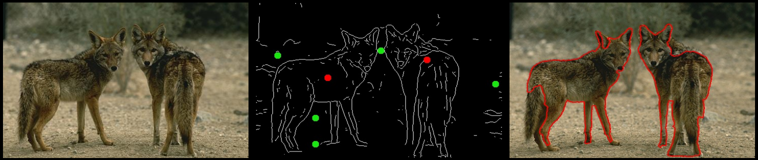
Modeling topological constraint while concurrently finding one or more minimum energy closed contours which satisfy:

- Foreground seeds must be “inside”
- Background seeds must be “outside”



[**X.**, Collins, Singh, CVPR 2013]

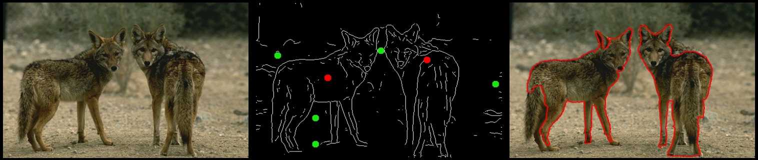
Our Ideas (EulerSeg)



Main Advantages

1. Basic primitives are edgelets
   (Little dependence on # of pixels)

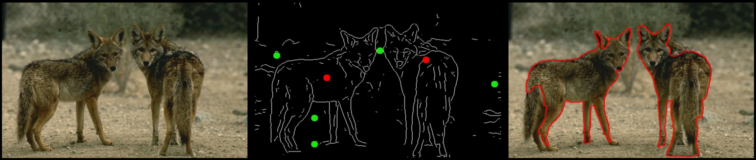
Our Ideas (EulerSeg)



Main Advantages

1. Basic primitives are edgelets
(Little dependence on # of pixels)
2. Dense strokes not needed to learn appearance model.
Results do *NOT* vary with seed location
(Interaction constraints are completely geometric in form)

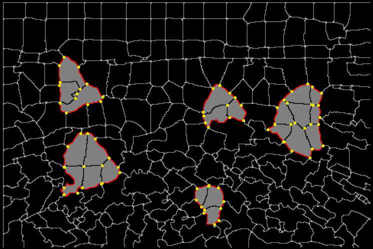
Our Ideas (EulerSeg)



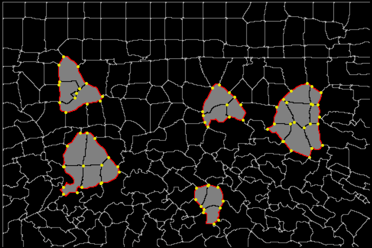
Main Advantages

1. Basic primitives are edgelets (Little dependence on # of pixels)
2. Dense strokes not needed to learn appearance model. Results do *NOT* vary with seed location (Interaction constraints are completely geometric in form)
3. Incorporating connectedness priors and specifying # of closures are easy (Euler characteristic)

Graph Representation

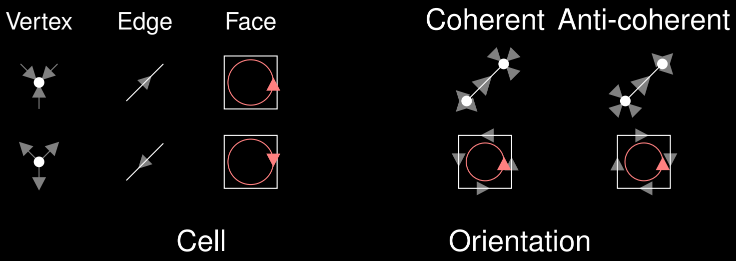


Graph Representation

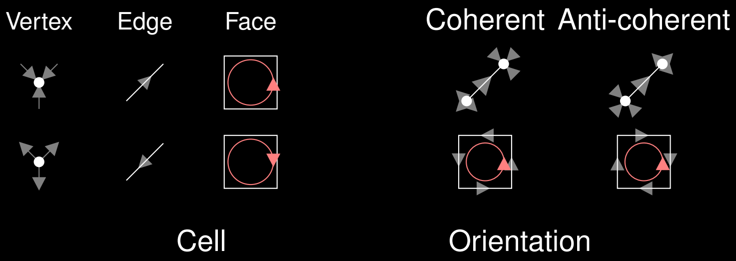


- $\mathbf{x}$: face indicator vector
- $\mathbf{y}$: edge indicator vector
- $\mathbf{z}$: vertex indicator vector
- $\mathbf{w}$: indicator vector for foreground boundary edges. Internal edges $\mathbf{y}_i \neq \mathbf{w}_i = 0$ are black, while boundary edges $\mathbf{y}_i = \mathbf{w}_i = 1$ are red

Discrete Calculus



Discrete Calculus

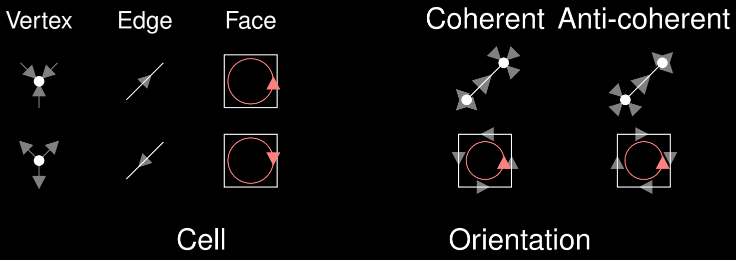


Vertex-edge Incidence Matrix: $A_1 = A, A_2 = A_1./D$

$$\mathbf{A}_{v_k, e_{ij}} = \begin{cases} 1 & k = i, j \\ 0 & \text{otherwise} \end{cases}$$

[Grady and Polimeni, 2010]

Discrete Calculus

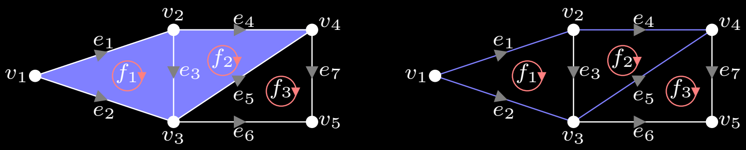


Edge-face Incidence Matrix: $C_1 = C, C_2 = |C|$

$$\mathbf{C}_{e,f} = \begin{cases} +1 & e \text{ is incident to } f \text{ and coherently oriented} \\ -1 & e \text{ is incident to } f \text{ and anti-coherently oriented} \\ 0 & \text{otherwise} \end{cases}$$

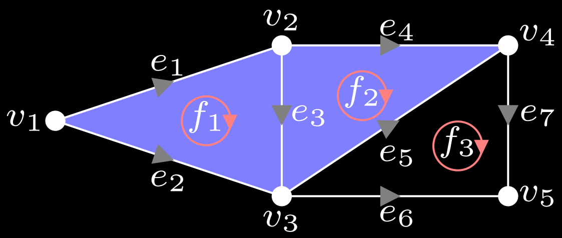
[Grady and Polimeni, 2010]

An Example



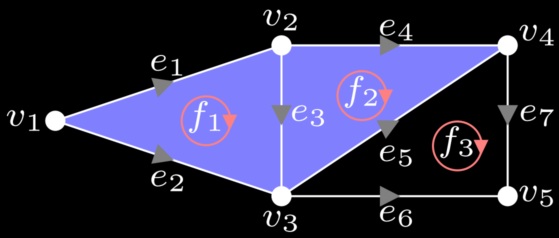
$$C = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \qquad \mathbf{b} = C\mathbf{x} = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}$$

Euler Characteristic



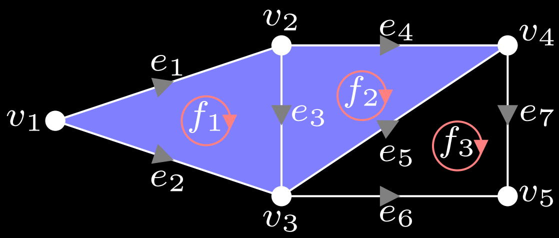
- Number of faces ($1^T \mathbf{x}$):

Euler Characteristic



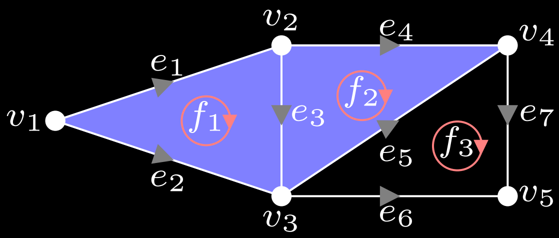
- Number of faces ($1^T\mathbf{x}$): 2
- Number of nodes ($1^T\mathbf{z}$):

Euler Characteristic



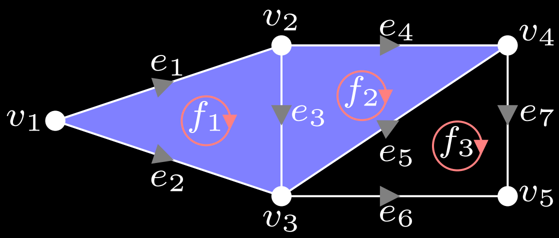
- Number of faces ($1^T\mathbf{x}$): 2
- Number of nodes ($1^T\mathbf{z}$): 4
- Number of edges ($1^T\mathbf{y}$):

Euler Characteristic



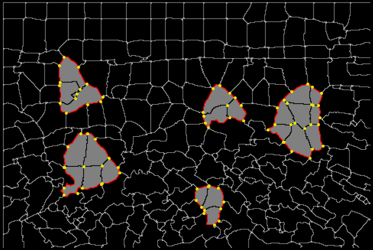
- Number of faces ($1^T\mathbf{x}$): 2
- Number of nodes ($1^T\mathbf{z}$): 4
- Number of edges ($1^T\mathbf{y}$): 5
- Number of connected components ($1^T\mathbf{x} + 1^T\mathbf{z} - 1^T\mathbf{y}$):

Euler Characteristic



- Number of faces ($1^T\mathbf{x}$): 2
- Number of nodes ($1^T\mathbf{z}$): 4
- Number of edges ($1^T\mathbf{y}$): 5
- Number of connected components ($1^T\mathbf{x} + 1^T\mathbf{z} - 1^T\mathbf{y}$): 1

Problem Formulation



Optimization Model

$$\min_{\mathbf{w},\mathbf{x},\mathbf{y},\mathbf{z}} \quad f(\mathbf{w})$$

$$\text{s.t.} \quad \mathbf{w} = |C_1\mathbf{x}|, \quad 2\mathbf{y} = \mathbf{w} + C_2\mathbf{x},$$

$$A_2\mathbf{y} \le \mathbf{z} \le A_1\mathbf{y}, 1^T\mathbf{x} + 1^T\mathbf{z} - 1^T\mathbf{y} = n,$$

$$\mathbf{x}_1 \le \mathbf{x} \le 1 - \mathbf{x}_0, \quad w_i, x_j, y_k, z_l \in \{0, 1\}.$$

Ratio Objective

Input	Solution 1	Solution 2	Solution 3
			
	$\mathbf{N}^T\mathbf{w} = 38.48$	$\mathbf{N}^T\mathbf{w} = 164.77$	$\mathbf{N}^T\mathbf{w} = 389.61$

Ratio Objective

Input

Solution 1

$$\mathbf{N}^T\mathbf{w} = 38.48$$
$$\mathbf{D}^T\mathbf{w} = 52$$
$$\frac{\mathbf{N}^T\mathbf{w}}{\mathbf{D}^T\mathbf{w}} = 0.5721$$

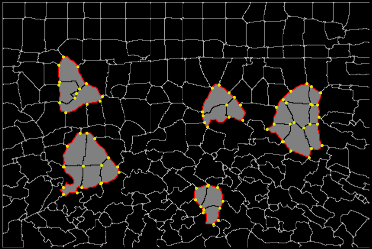
Solution 2

$$\mathbf{N}^T\mathbf{w} = 164.77$$
$$\mathbf{D}^T\mathbf{w} = 288$$
$$\frac{\mathbf{N}^T\mathbf{w}}{\mathbf{D}^T\mathbf{w}} = 0.7400$$

Solution 3

$$\mathbf{N}^T\mathbf{w} = 389.61$$
$$\mathbf{D}^T\mathbf{w} = 865$$
$$\frac{\mathbf{N}^T\mathbf{w}}{\mathbf{D}^T\mathbf{w}} = 0.4504$$

Problem Formulation



Optimization Model

$$\min_{\mathbf{w},\mathbf{x},\mathbf{y},\mathbf{z}} \quad \frac{\mathbf{N}^T\mathbf{w}}{\mathbf{D}^T\mathbf{w}}$$

$$\text{s.t.} \quad \mathbf{w} = |C_1\mathbf{x}|, \quad 2\mathbf{y} = \mathbf{w} + C_2\mathbf{x},$$

$$A_2\mathbf{y} \le \mathbf{z} \le A_1\mathbf{y}, 1^T\mathbf{x} + 1^T\mathbf{z} - 1^T\mathbf{y} = n,$$

$$\mathbf{x}_1 \le \mathbf{x} \le 1 - \mathbf{x}_0, \quad w_i, x_j, y_k, z_l \in \{0, 1\}.$$

Minimizing a Ratio Cost

Solved by minimizing

$$\psi(t, \mathbf{w}) = (\mathbf{N} - t\mathbf{D})^T \mathbf{w}$$

- Over feasible $\mathbf{w}$ for a sequence of chosen values of t
- With an initial finite bounding interval $[t_l, t_u]$

Minimizing a Ratio Cost

Solved by minimizing

$$\psi(t, \mathbf{w}) = (\mathbf{N} - t\mathbf{D})^T \mathbf{w}$$

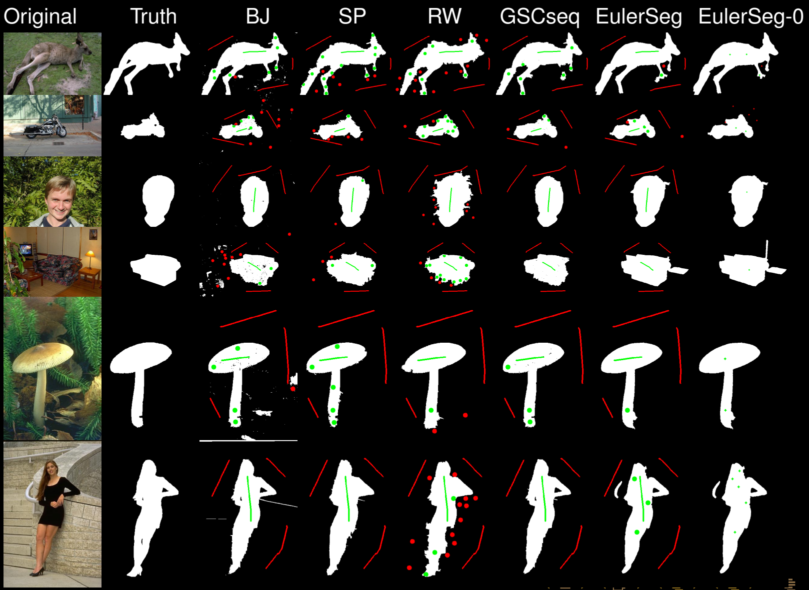
- Over feasible $\mathbf{w}$ for a sequence of chosen values of t
- With an initial finite bounding interval $[t_l, t_u]$

Pick $t_0 = \frac{t_l + t_u}{2}$, and let

$$\bar{\mathbf{w}} = \arg\min_{\mathbf{w}} \psi(t_0, \mathbf{w})$$

- $\psi(t_0, \bar{\mathbf{w}}) = 0$: $\mathbf{N}^T \bar{\mathbf{w}} / \mathbf{D}^T \bar{\mathbf{w}} = t_0$, terminate with solution t_0
- $\psi(t_0, \bar{\mathbf{w}}) < 0$: $\mathbf{N}^T \bar{\mathbf{w}} / \mathbf{D}^T \bar{\mathbf{w}} < t_0$, $t_u \leftarrow \mathbf{N}^T \bar{\mathbf{w}} / \mathbf{D}^T \bar{\mathbf{w}}$
- $\psi(t_0, \bar{\mathbf{w}}) > 0$: $\mathbf{N}^T \bar{\mathbf{w}} / \mathbf{D}^T \bar{\mathbf{w}} > t_0$, $t_l \leftarrow t_0$

Qualitative Results



Quantitative Evaluation

F-Measure

$$P = \frac{|A \cap T|}{|A|}, \quad R = \frac{|A \cap T|}{|T|}, \quad F = \frac{2PR}{P+R}$$

How much effort to reach $F = 0.95$ (using a robot user)?

Method	BJ	RW	SP	GSCseq	EulerSeg
User Scribbles	5.51	6.48	4.54	2.30	**2.06**

Seeds tell MORE than link/cannot link

[Gulshan et al., 2010]

Roadmap

Chapter	Parsing Task	Weak Supervision	Publication
Ch. 2	Object Segmentation	User Indication	CVPR 2013
Ch. 3	Scene Parsing	Image-level Tags	CVPR 2014
Ch. 4	Scene Parsing	Image-level Tags Bounding Boxes Partial Labels	CVPR 2015a
Ch. 5	Video Segmentation	Side Knowledge	ICCV 2013
Ch. 6	Video Summarization	Human Gaze	CVPR 2015b

Semantic Segmentation

Building

Tree

Boat

Person

Semantic Segmentation

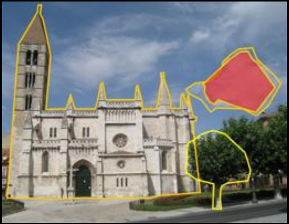
Building

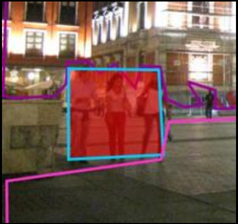
Tree

Boat

Person

Bad Object Labels

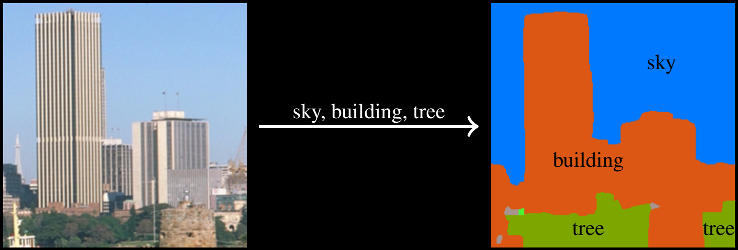




Weakly Supervised Semantic Segmentation

Motivation

- Annotation: presence of image classes
- Tags readily available in online photo collections
- Easier to obtain than segmentations



[**X.**, Schwing, Urtasun, CVPR 2014]

Cosegmentation

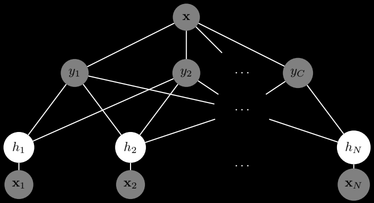
Concurrently segment common foreground objects from a set of images

[Collins, **X.**, Grady, Singh, CVPR 2012]
[Mukherjee, Singh, **X.**, Collins, ECCV 2012]
[Collins, Liu, **X.**, Mukherjee, Singh, ECCV 2014]

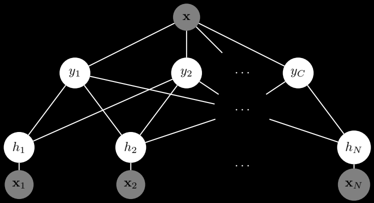
Latent Structured Prediction

Graphical Model

- Presence/absence of a class: $y_i \in \{0, 1\}$
- Semantic superpixel label: $h_j \in \{1, \ldots, C\}$
- Image evidence: x



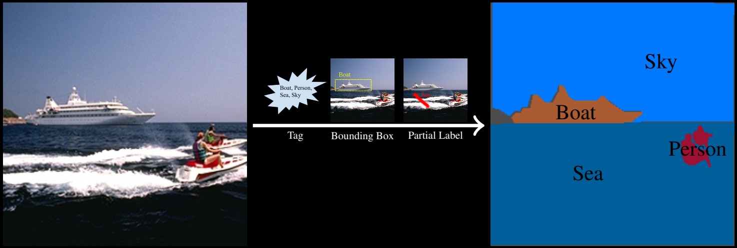
Learning/Inference with Tags



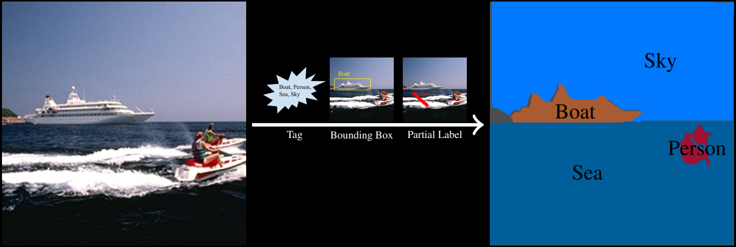
Inference without tags

[**X.**, Schwing, Urtasun, CVPR 2014]

How About Other Forms of Weak Supervision



How About Other Forms of Weak Supervision



Unified Model

$$\min_{W,H} \quad \frac{1}{2}\mathrm{tr}(W^T W) + \lambda \sum_{p=1}^{n} \xi(W; \mathbf{x}_p, \mathbf{h}_p)$$

$$\text{s.t.} \quad H\mathbf{1}_C = \mathbf{1}_n, H \in \{0,1\}^{n \times C}$$

$$H \in \mathcal{S}$$

[**X.**, Schwing, Urtasun, CVPR, 2015]

Max-Margin Objective

Denote

- $X = [\mathbf{x}_1^T, \mathbf{x}_p^T, \cdots, \mathbf{x}_n^T] \in R^{n \times d}$: feature matrix
- $H = [\mathbf{h}_1^T, \mathbf{h}_p^T, \cdots, \mathbf{h}_n^T] \in \{0, 1\}^{n \times c}$: hidden label matrix
- $W \in R^{d \times c}$: feature weighting matrix

Max-Margin Objective

Denote

- $X = [\mathbf{x}_1^T, \mathbf{x}_p^T, \cdots, \mathbf{x}_n^T] \in R^{n \times d}$: feature matrix
- $H = [\mathbf{h}_1^T, \mathbf{h}_p^T, \cdots, \mathbf{h}_n^T] \in \{0,1\}^{n \times c}$: hidden label matrix
- $W \in R^{d \times c}$: feature weighting matrix

$$\min_{W,H} \quad \frac{1}{2}\mathrm{tr}(W^T W) + \lambda \sum_{p=1}^{n} \sum_{c=1}^{C} \xi(\mathbf{w}_c; \mathbf{x}_p, h_p^c)$$

where

- $$\xi(\mathbf{w}_c; \mathbf{x}_p, h_p^c) = \begin{cases} \max(0, 1 + (\mathbf{w}_c^T \mathbf{x}_p)), & h_p^c = 0 \\ \mu^c \max(0, 1 - (\mathbf{w}_c^T \mathbf{x}_p)), & h_p^c = 1 \end{cases}$$
- $$\mu^c = \frac{\sum_{p=1}^{n} 1(h_p^c == 0)}{\sum_{p=1}^{n} 1(h_p^c == 1)}$$

[Zhao et al., 2008, Zhao et al., 2009]

Supervision Space as Constraints

- Unlabeled/Cosegmentation/Transductive: $\mathcal{S} = \emptyset$
- Image level tags: $\mathcal{S} = \{H \leq BZ, B^T H \geq Z\}$
- Bounding boxes: $\mathcal{S} = \{H \leq \hat{B}\hat{Z}, \hat{B}^T H \geq \hat{Z}\}$
- Semi-supervision $\mathcal{S} = \{H_\Omega = \hat{H}_\Omega\}$

Supervision Space as Constraints

- Unlabeled/Cosegmentation/Transductive: $\mathcal{S} = \emptyset$
- Image level tags: $\mathcal{S} = \{H \leq BZ, B^T H \geq Z\}$
- Bounding boxes: $\mathcal{S} = \{H \leq \hat{B}\hat{Z}, \hat{B}^T H \geq \hat{Z}\}$
- Semi-supervision $\mathcal{S} = \{H_\Omega = \hat{H}_\Omega\}$

An Example (2 images, 5 superpixels (2+3), 3 classes)

$$B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad Z = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \quad H = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$H \leq BZ = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad B^T H = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 2 \end{bmatrix} \geq Z$$

Optimization Model

$$\min_{W,H} \quad \frac{1}{2}\mathrm{tr}(W^T W) + \lambda \sum_{p=1}^{n} \xi(W; \mathbf{x}_p, \mathbf{h}_p)$$

$$\text{s.t.} \quad H\mathbf{1}_C = \mathbf{1}_n, H \in \{0,1\}^{n \times C}$$

$$H \in \mathcal{S}$$

Optimization Model

$$\min_{W,H} \quad \frac{1}{2}\text{tr}(W^T W) + \lambda \sum_{p=1}^{n} \xi(W; \mathbf{x}_p, \mathbf{h}_p)$$

$$\text{s.t.} \quad H\mathbf{1}_C = \mathbf{1}_n, H \in \{0,1\}^{n \times C}$$

$$H \in \mathcal{S}$$

Observations

- Challenge: non-convex mixed integer programming

Optimization Model

$$\min_{W,H} \quad \frac{1}{2}\mathrm{tr}(W^T W) + \lambda \sum_{p=1}^{n} \xi(W; \mathbf{x}_p, \mathbf{h}_p)$$

$$\text{s.t.} \quad H\mathbf{1}_C = \mathbf{1}_n, H \in \{0,1\}^{n\times C}$$

$$H \in \mathcal{S}$$

Observations

- Challenge: non-convex mixed integer programming
- Optimization problem is bi-convex, i.e., it is convex w.r.t. W if H is fixed, and convex w.r.t. H if W is fixed
- Constraints are linear and they only involve the super-pixel assignment matrix H

Learning Algorithm

$$\min_{W,H} \quad \frac{1}{2}\mathrm{tr}(W^T W) + \lambda \sum_{p=1}^{n} \xi(W; \mathbf{x}_p, \mathbf{h}_p)$$

$$\text{s.t.} \quad H\mathbf{1}_C = \mathbf{1}_n, H \in \{0,1\}^{n \times C}$$

$$H \in \mathcal{S}$$

Alternating Between

- Fix H solve for W independent of classes (1-vs-all linear SVM)
- Fix W infer super-pixel labels H in parallel w.r.t images (small LP instances)

Learning Algorithm

Alternating Between

- Fix H solve for W independent of classes (1-vs-all linear SVM)
- Fix W infer super-pixel labels H in parallel w.r.t images (small LP instances)

Inference

$$\begin{aligned} \max_{H} \quad & \mathrm{tr}((XW)^T H) \\ \text{s.t.} \quad & H\mathbf{1}_C = \mathbf{1}_n, H \in \{0,1\}^{n\times C}, \\ & H \in \mathcal{S} \end{aligned}$$

Learning Algorithm

Alternating Between

- Fix H solve for W independent of classes (1-vs-all linear SVM)
- Fix W infer super-pixel labels H in parallel w.r.t images (small LP instances)

Inference

$$\begin{aligned} \max_{H} \quad & \mathrm{tr}((XW)^T H) \\ \text{s.t.} \quad & H\mathbf{1}_C = \mathbf{1}_n, H \in \{0,1\}^{n \times C}, \\ & H \in \mathcal{S} \end{aligned}$$

Proposition

Fixing W solving for H using a linear program gives the integral optimal solution.

Theoretical Guarantee

Proposition

Fixing W solving for H using a linear program gives the integral optimal solution.

Proof.

(Sketch) The main idea of our proof is to show our coefficient matrix is totally unimodular. By Grady 2010: If A is totally unimodular and b is integral, then linear programs of forms like $\{\min \mathbf{c}^T\mathbf{x} \mid A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\}$ have integral optima, for any $\mathbf{c}$. Hence, the LP relaxation gives the optimal integral solution. $\square$

Computation Efficiency

Model Nature

- Decomposable
- Parallelizable
- Theoretical guarantee of relaxation quality

Computation Efficiency

Model Nature

- Decomposable
- Parallelizable
- Theoretical guarantee of relaxation quality

Running time

- orders of magnitude faster than the state-of-the-art (20 min v.s. 24 hours)
- 10 ms to test one image

Experimental Evaluation

Datasets

- SIFT-Flow (a.k.a, LabelMe): 2688 images, 33 classes
- MSRC: 591 images, 21 classes

Accuracy Metric

- Per-pixel: the fraction of the number of pixels classified rightly over the number of pixels to be classified in total
- Per-class: the average of accuracy of all the classes

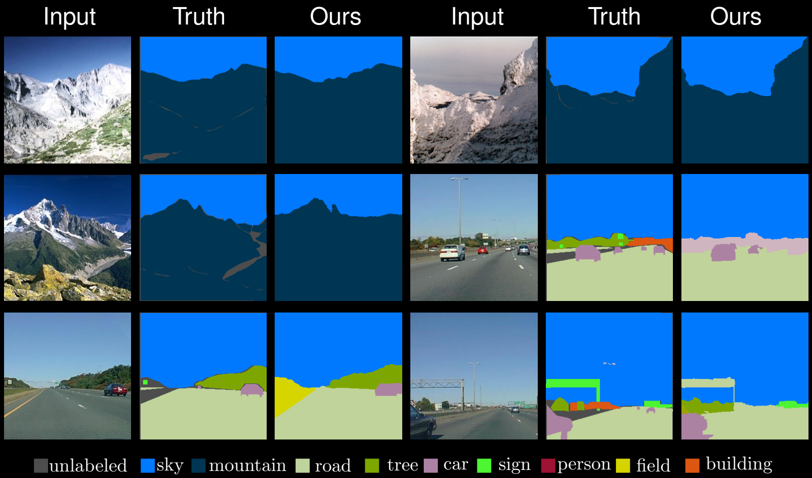
Comparison to State-of-the-art on Sift-Flow

Method	Supervision	Per-class	Per-pixel
Liu et al., 2011 (PAMI)	full	24	76.7
Farabet et al., 2012 (ICML)	full	29.5	78.5
Farabet et al., 2012 (ICML) balanced	full	46.0	74.2
Eigen et al., 2012 (CVPR)	full	32.5	77.1
Singh et al., 2013 (CVPR)	full	33.8	79.2
Tighe et al., 2013 (IJCV)	full	30.1	77.0
Tighe et al., 2014 (CVPR)	full	39.3	78.6
Yang et al., 2014 (CVPR)	full	48.7	79.8
Vezhnevets et al., 2011 (ICCV)	weak (tags)	14	N/A
Vezhnevets et al., 2012 (CVPR)	weak (tags)	22	51
Xu et al., 2014 (CVPR)	weak (tags)	27.9	N/A
Ours (1-vs-all)	weak (tags)	**32.0**	**64.4**
Ours (ILT)	weak (tags)	**35.0**	**65.0**
Ours (1-vs-all + transductive)	weak (tags)	**40.0**	**59.0**
Ours (ILT + transductive)	weak (tags)	**41.4**	**62.7**

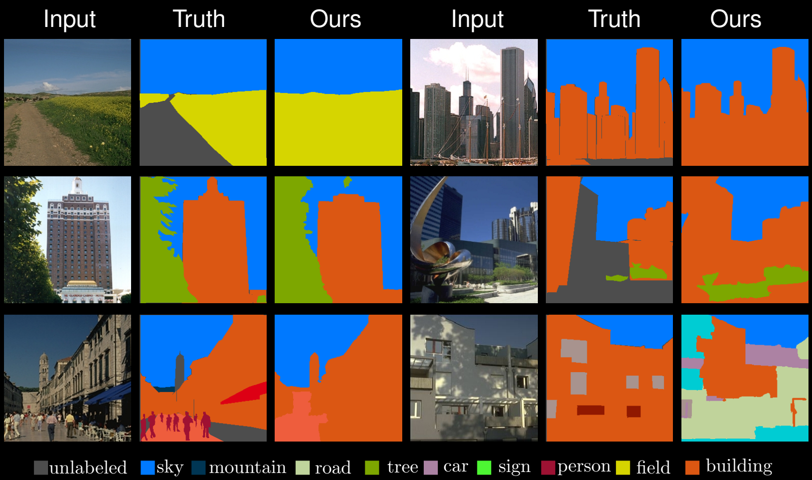
Comparison to State-of-the-art on MSRC

Method	Supervision	per-class	per-pixel
Shotton et al., 2008 (ECCV)	full	67	72
Yao et al., 2012 (CVPR)	full	79	86
Vezhnevets et al., 2011 (ICCV)	weak (tags)	67	67
Liu et al., 2012 (TMM)	weak (tags)	N/A	**71**
Ours	weak (tags)	**73**	70

Sample Results

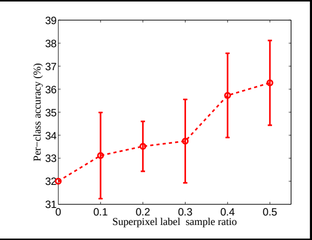


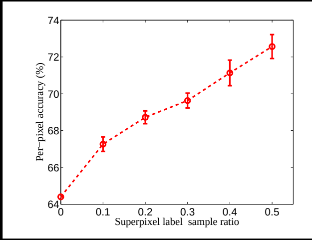
Sample Results (continued)



Other Forms of Weak Supervision

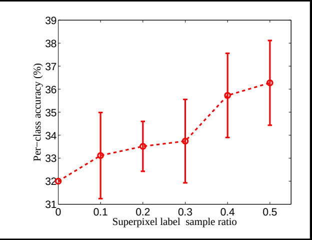
Semi-supervision

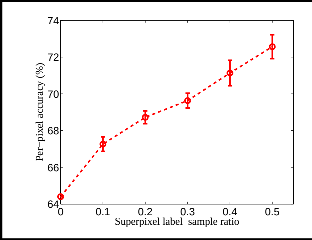




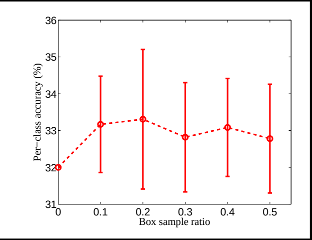
Other Forms of Weak Supervision

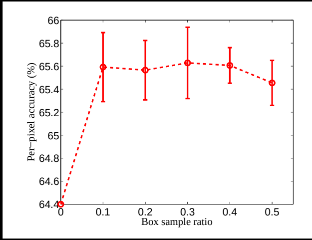
Semi-supervision





Bounding Box





Roadmap

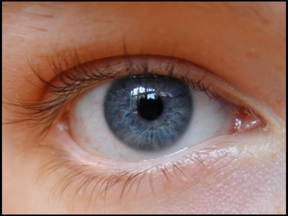
Chapter	Parsing Task	Weak Supervision	Publication
Ch. 2	Object Segmentation	User Indication	CVPR 2013
Ch. 3	Scene Parsing	Image-level Tags	CVPR 2014
Ch. 4	Scene Parsing	Image-level Tags Bounding Boxes Partial Labels	CVPR 2015a
Ch. 5	Video Segmentation	Side Knowledge	ICCV 2013
Ch. 6	Video Summarization	Human Gaze	CVPR 2015b

Online Video Segmentation

- Background subspace is modeled on a Grassmannian manifold with online updating along the geodesic
- Spatially contiguous and structured foreground is modeled via group sparsity

[**X.**, Ithapu, Mukherjee, Rehg, Singh, ICCV 2013]

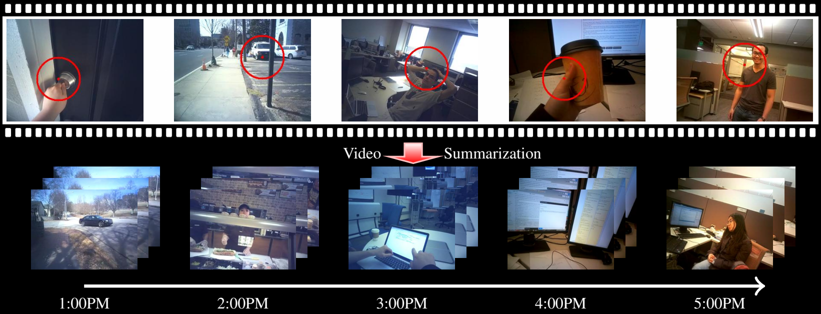
First Person Vision

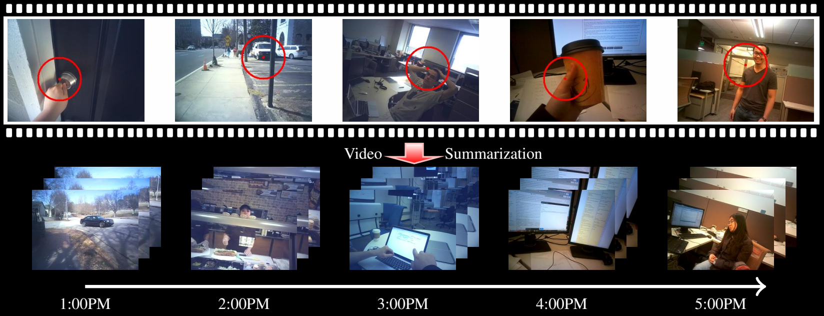

Motivation

- Life-logging with wearable cameras: SenseCam, GoPro, Google glass
- Memory aid
- Gaze provides a form of weak supervision: window of mind

Gaze-enabled Egocentric Video Summarization



Gaze-enabled Egocentric Video Summarization



What makes a good summary?

- Relevance
- Diversity
- Compactness
- Personalization

[**X**., Mukherjee, Li, Warnewr, Rehg, Singh, CVPR, 2015]

Relevance and Diversity Measurement

- Mutual Information

$$
\begin{aligned}
M(\mathcal{V}\backslash\mathcal{S};\mathcal{S}) &= H(\mathcal{V}\backslash\mathcal{S}) - H(\mathcal{V}\backslash\mathcal{S}|\mathcal{S}) \\
&= H(\mathcal{V}\backslash\mathcal{S}) + H(\mathcal{S}) - H(\mathcal{V})
\end{aligned}
$$

Relevance and Diversity Measurement

- Mutual Information

$$\begin{aligned} M(\mathcal{V}\backslash\mathcal{S};\mathcal{S}) &= H(\mathcal{V}\backslash\mathcal{S}) - H(\mathcal{V}\backslash\mathcal{S}|\mathcal{S}) \\ &= H(\mathcal{V}\backslash\mathcal{S}) + H(\mathcal{S}) - H(\mathcal{V}) \end{aligned}$$

- Entropy

$$H(\mathcal{S}) = \frac{1+\log(2\pi)}{2}|\mathcal{S}| + \frac{1}{2}\log(\det(L_{\mathcal{S}}))$$

Relevance and Diversity Measurement

- Mutual Information

$$
\begin{aligned}
M(\mathcal{V}\backslash\mathcal{S};\mathcal{S}) &= H(\mathcal{V}\backslash\mathcal{S}) - H(\mathcal{V}\backslash\mathcal{S}|\mathcal{S}) \\
&= H(\mathcal{V}\backslash\mathcal{S}) + H(\mathcal{S}) - H(\mathcal{V})
\end{aligned}
$$

- Entropy

$$
H(\mathcal{S}) = \frac{1+\log(2\pi)}{2}|\mathcal{S}| + \frac{1}{2}\log(\det(L_{\mathcal{S}}))
$$

- Maximizing

$$
M(\mathcal{S}) = \frac{1}{2}\log(\det(L_{\mathcal{V}\backslash\mathcal{S}})) + \frac{1}{2}\log(\det(L_{\mathcal{S}}))
$$

[Krause et al., 2008]

Relation to Determinantal Point Process

Positive semidefinite kernel matrix L indexed by elements of $\mathcal{V}$

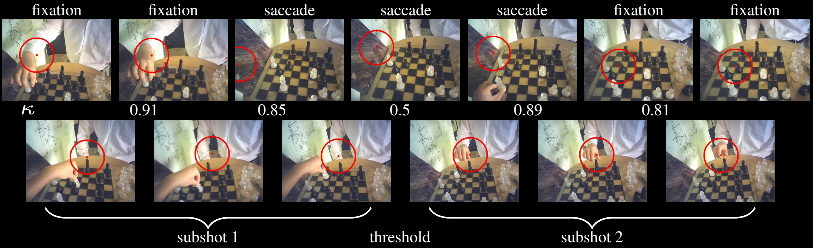
$$L_{ij} = \frac{\mathbf{v}_i^T}{\|\mathbf{v}_i\|} \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|}$$

For every $\mathcal{S} \in \mathcal{V}$, we define a diversity score

$$D(\mathcal{S}) = \log(\det(L_{\mathcal{S}}))$$

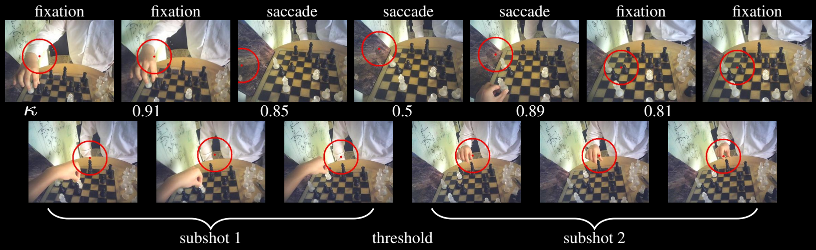
[Kulesza and Taskar, 2012]
(Acknowledgement to Jerry :)

Gaze in Video Summarization



- Better temporal segmentation: egocentric is continuous, but gaze is discrete

Gaze in Video Summarization



- Better temporal segmentation: egocentric is continuous, but gaze is discrete
- Personalization: attention measurement from gaze fixations

$$I(\mathcal{S}) = \sum_{i \in \mathcal{S}} c_i$$

Partition Matroid Constraint

Motivation

- Compactness: cardinality or knapsack constraint?
- High level supervision: timeline

Partition Matroid Constraint

Motivation

- Compactness: cardinality or knapsack constraint?
- High level supervision: timeline

Partition Matroid Construction

- Partition the video into b disjoint blocks $\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_b$
- Limit associated with each block
  $\mathcal{I} = \{\mathcal{A} : \quad |\mathcal{A} \cap \mathcal{P}_m| \leq f_m, m = 1, 2, \cdots, b\}$

[Bilmes, 2013]

Submodular Formulation

$$\max_{\mathcal{S}} \quad F(\mathcal{S}) = M(\mathcal{S}) + \lambda I(\mathcal{S})$$
$$\text{s.t.} \quad \mathcal{S} \in \mathcal{I}$$

Submodular Formulation

$$\max_{\mathcal{S}} \quad F(\mathcal{S}) = M(\mathcal{S}) + \lambda I(\mathcal{S})$$
$$\text{s.t.} \quad \mathcal{S} \in \mathcal{I}$$

Corollary

$F(\mathcal{S})$ is submodular.

Submodular Formulation

$$\max_{\mathcal{S}} \quad F(\mathcal{S}) = M(\mathcal{S}) + \lambda I(\mathcal{S})$$
$$\text{s.t.} \quad \mathcal{S} \in \mathcal{I}$$

Corollary

$F(\mathcal{S})$ is submodular.

Proposition

Greedy local search achieves a $\frac{1}{4}$-approximation factor for our constrained submodular maximization problem.

[Lee et al., 2010]
[Filmus and Ward, 2012]

Dataset Collection


- 5 subjects to record their daily lives
- 21 videos with gaze
- 15 hours in total

Annotation

Subjects group subshots into events.

Systematic Evaluation

Evaluation Metric

$$P = \frac{|A \cap T|}{|A|}, \quad R = \frac{|A \cap T|}{|T|}, \quad F = \frac{2PR}{P+R}$$

Systematic Evaluation

Evaluation Metric

$$P = \frac{|A \cap T|}{|A|}, \quad R = \frac{|A \cap T|}{|T|}, \quad F = \frac{2PR}{P+R}$$

F-measure on GTEA-GAZE+

Method	uniform	kmeans	uniform(gaze)	kmeans(gaze)	ours
F-measure	0.161	0.215 ± 0.016	0.526	0.475 ± 0.026	0.621

F-measure on Our New Dataset

Method	uniform	kmeans	uniform(gaze)	kmeans(gaze)	ours
F-measure	0.080	0.095 ± 0.030	0.476	0.509 ± 0.025	0.585

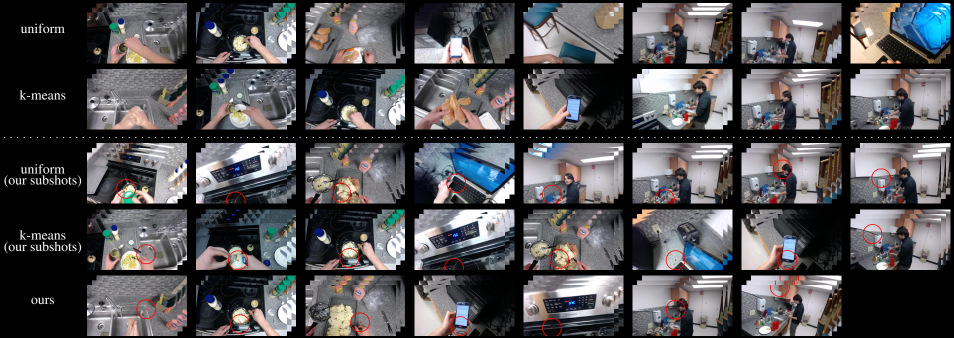
Qualitative Result

Results from GTEA-gaze+ pizza preparation video.

Qualitative Result

Results from our new dataset: our subject mixes a shake, drinks it, washes his cup, plays chess and texts a friend.

Qualitative Result



Results from our new dataset: our subject is cooking chicken and have a conversation with his roommate.

Summary

Thesis Contribution

- An efficient approach for interactive segmentation while minimizing human effort (Ch. 2)

Summary

Thesis Contribution

- An efficient approach for interactive segmentation while minimizing human effort (Ch. 2)
- A latent graphical model for semantic segmentation using only image level tags (Ch. 3)

Summary

Thesis Contribution

- An efficient approach for interactive segmentation while minimizing human effort (Ch. 2)
- A latent graphical model for semantic segmentation using only image level tags (Ch. 3)
- A unified model for semantic segmentation with various forms of weak supervision (Ch. 4)

Summary

Thesis Contribution

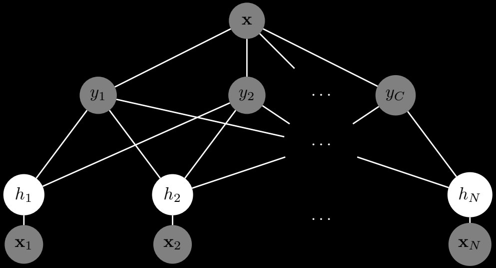
- An efficient approach for interactive segmentation while minimizing human effort (Ch. 2)
- A latent graphical model for semantic segmentation using only image level tags (Ch. 3)
- A unified model for semantic segmentation with various forms of weak supervision (Ch. 4)
- An online foreground/background video segmentation using Grassmannian subspace learning (Ch. 5)

Summary

Thesis Contribution

- An efficient approach for interactive segmentation while minimizing human effort (Ch. 2)
- A latent graphical model for semantic segmentation using only image level tags (Ch. 3)
- A unified model for semantic segmentation with various forms of weak supervision (Ch. 4)
- An online foreground/background video segmentation using Grassmannian subspace learning (Ch. 5)
- A submodular summarization framework for first person videos (Ch. 6)

Future: Joint Visual and Textual Parsing



- Enhance graphical model with richer prior knowledge: geometry (Hoeim et al., 2007), co-occurrence, etc.
- Other form of supervisions: Air Quality Index (AQI)
- Tackle noisy tags
- Extend to videos

Future: Egocentric/Robotic Vision

- Daily life logging / memory aid
- Predictive diagnosis for disease
- First-person vision for robotics
- Help the blind to sense the visual world

Acknowledgement

Thesis Committee

- Vikas Singh (advisor)
- Chuck Dyer
- Jerry Zhu
- Jude Shavlik
- Mark Craven

Funding

- UW-Epic RAship
- NSF RI 1116584
- NVIDIA Hardware Gift
- Adobe Gift

Collaborators

- Maxwell Collins (UW-Madison)
- Chuck Dyer (UW-Madison)
- Leo Grady (Heartflow)
- Vamsi Ithapu (UW-Madison)
- Hyunwoo Kim (UW-Madison)
- Yin Li (Georgia Tech)
- Zhe Lin (Adobe Research)
- Ji Liu (URochester)
- Lopa Mukherjee (UW-Whitewater)
- James M. Rehg (Georgia Tech)
- Alexander Schwing (UToronto)
- Xiaohui Shen (Adobe Research)
- Vikas Singh (UW-Madison)
- Raquel Urtasun (UToronto)
- Baba Vemuri (UFlorida)
- Jamieson Warner (UW-Madison)
- Jerry Zhu (UW-Madison)