ANNOTATOR RATIONALES FOR VISUAL RECOGNITION

Jeff Donahue and Kristen Grauman
Department of Computer Science – The University of Texas at Austin

Problem

Image labels alone are insufficient supervision for learning complex visual recognition tasks.

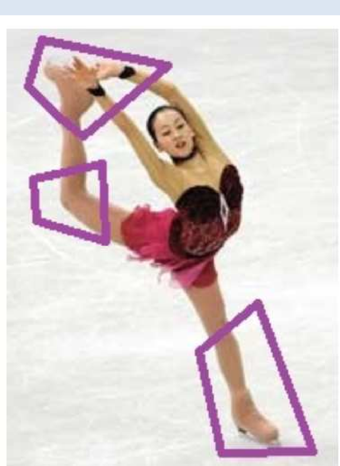
Is the coach's team winning?

Is the skater's form good?

Is she attractive?

Our Idea

- **Annotators should not only assign class labels (the "what"), but also give a *rationale* indicating their reasoning behind the label (the "why").**
- We propose two modes for visual rationales:
  Spatial: draw polygons around important image regions
  Attribute: name attributes most influential in label choice



Attribute	Rationale?
a_1: **pointed toes**	✓
a_2: **on ground**	
a_3: **balanced**	
a_4: **falling**	
a_5: **knee angled**	✓

Annotation task: *Is the skater's form good? How can you tell?*

SVM Training with Contrast Examples

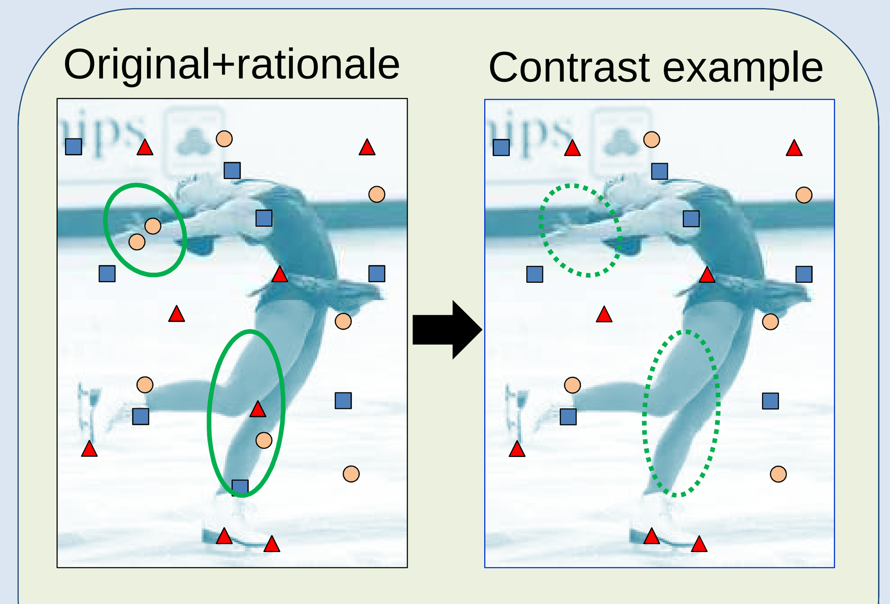
- Require classifier to treat ***contrast example*** that *lacks* the important features as "less positive" than the original.
- We adopt the SVM objective developed by Zaidan et al., [HLT 2007] for sentiment analysis in documents:

Minimize: $$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\sum_i \xi_i\right) + C_C\left(\sum_i \gamma_i\right)$$

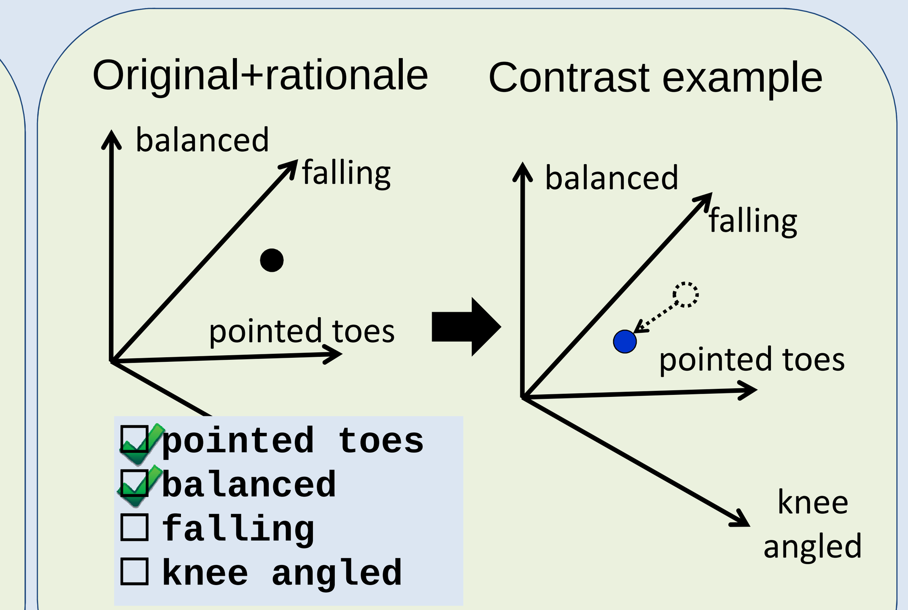
Subject to: $$\forall i \quad y_i \boldsymbol{w}^T \boldsymbol{x}_i \geq 1 - \xi_i$$
$$\forall i \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i - \boldsymbol{w}^T\boldsymbol{v}_i) \geq \mu(1-\gamma_i) \qquad \xi_i, \gamma_i \geq 0$$

where $\boldsymbol{x}_i$ is the i-th training example, $\boldsymbol{v}_i$ is its corresponding contrast example, and y_i is the class label {1, -1}.

Visual Rationales → Contrast Examples



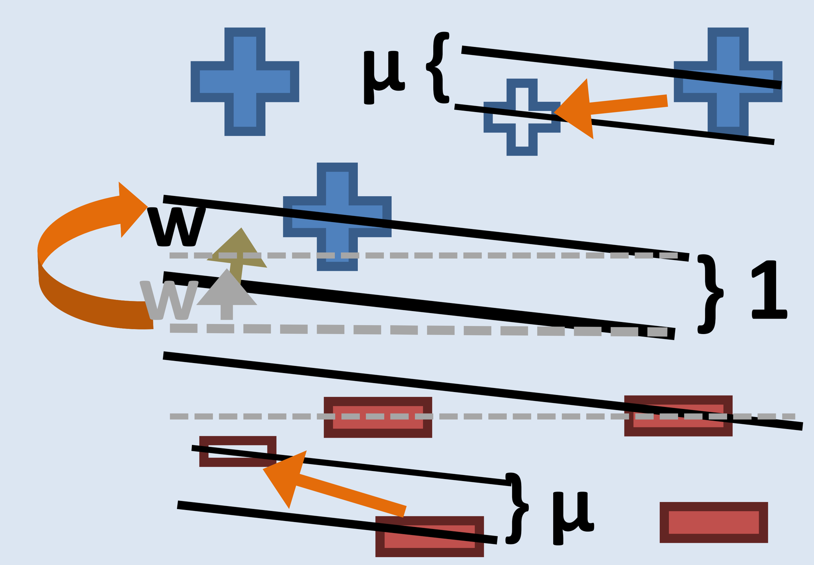
Spatial rationale



Attribute rationale

Impact on Classifier

Contrast examples refine the resulting hyperplane



Results: Scene Categorization

- Test our *spatial rationales* on 15 Scene Categories dataset with annotations from 545 unique MTurk workers
- **Task**: Name the scene type

Typical *Tight* *"Artistic"*

- Scenes often **lack clear semantic boundaries** (e.g., city vs. street), making this a good task for rationales
- **Visual rationales outperform all three baselines for 13 of 15 classes**

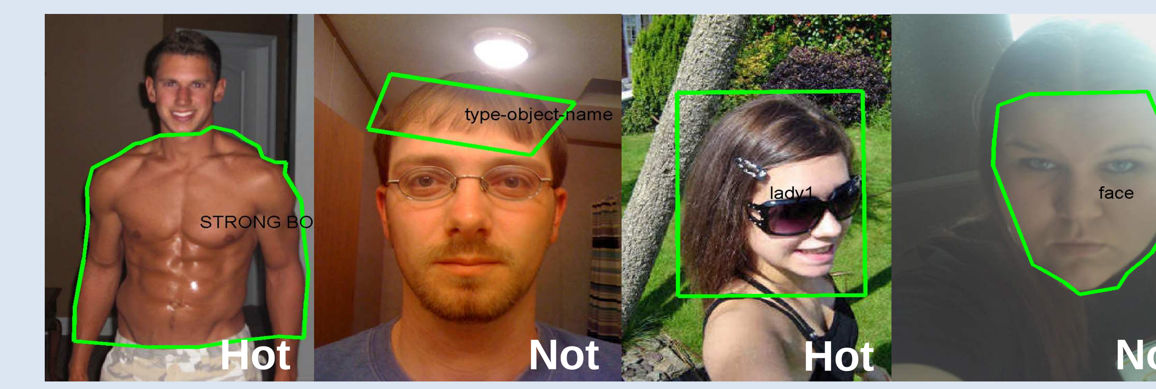
Classes w/ largest gains	Ours (mAP)	Originals Only	Rationales Only	Mutual Information
Kitchen	**0.1395**	0.1196	0.1277	0.1202
Living Rm	**0.1238**	0.1142	0.1131	0.1159
Inside City	**0.1487**	0.1299	0.1394	0.1245
Coast	**0.4513**	0.4243	0.4205	0.4129
Highway	**0.2379**	0.2240	0.2221	0.2112

Rationales != foreground segmentation

Rationales > discriminative feat. selection

Results: Hot or Not?

- Test our *spatial rationales* on hotornot.com using provided ratings +104 MTurk rationales
- **Task**: Classify male/female as "hot" (top 25%) or "not" (bottom 25%)



- **Visual rationales improve accuracy, especially for males**

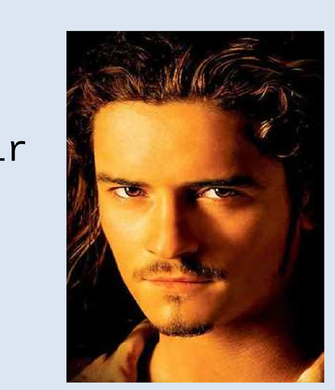
	Male		Female	
	N = 25	N = 100	N = 25	N = 100
Ours (Our Annotations)	**55.40%**	**60.01%**	53.13%	**57.07%**
Ours (MTurk Annotations)	53.73%	54.92%	53.83%	56.57%
Originals Only	52.64%	54.86%	**54.02%**	55.99%

Net savings in annotation effort, and better accuracy!

Results: Public Figure Attractiveness

- Test our *attribute rationales* on PubFig dataset
- **Task**: Classify public figure as attractive or not

Youth
Smiling
Straight Hair
Narrow Eyes



Youth
Black Hair
Goatee
Square Face
Shiny Skin
High Cheekbones

- **Large improvement, especially with "homogeneous rationales" for all classes**

	Homogeneous		Individual	
	Ours	Originals	Ours	Originals
Male	**68.14%**	64.60%	**62.35%**	59.02%
Female	**55.65%**	51.74%	51.86%	52.36%

Conclusions

- The "why" matters
- Positive results in multiple domains
- Rationales give deeper insight than a class label alone, especially useful in subjective tasks