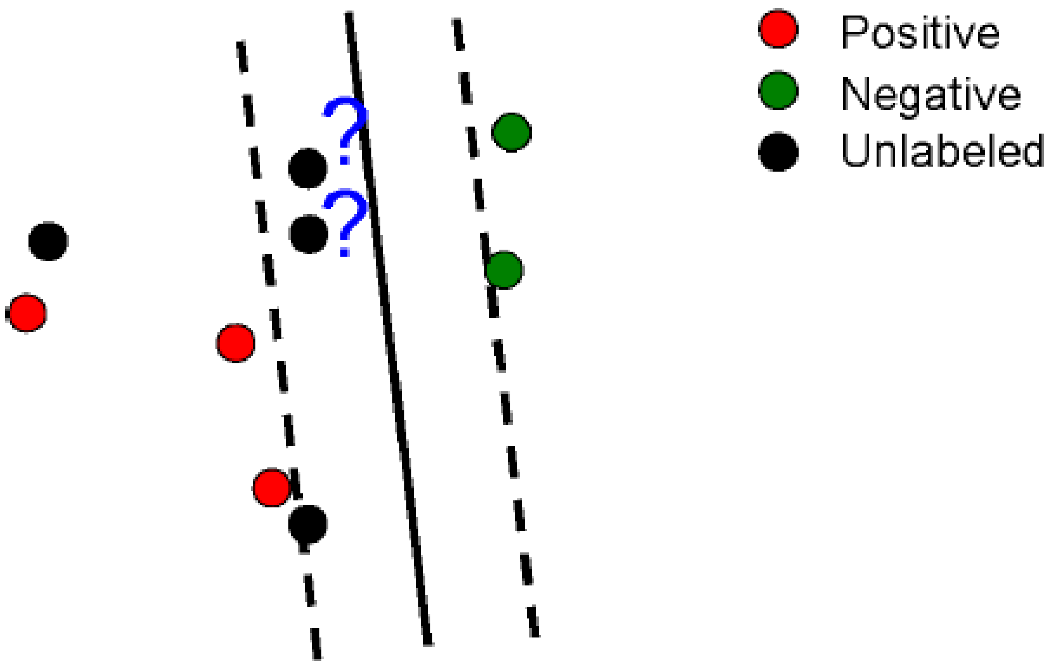
# What's It Going to Cost You?: Predicting Effort vs. Informativeness for Multi-Label Image Annotations

*Sudheendra Vijayanarasimhan and Kristen Grauman*

Deepti Ghadiyaram

# Active Learning

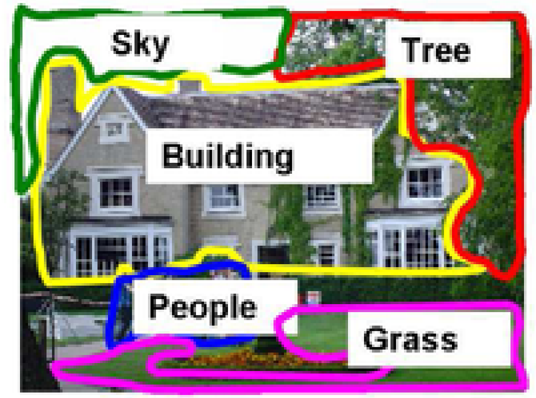



Source : http://vision.cs.utexas.edu/projects/active-mil/nips08.pdf

**Active selection is particularly complex for visual category learning.**

1) Real world images contain multiple objects


VS.




*Active learner must assess the value of an image containing some Unknown combination of categories.*

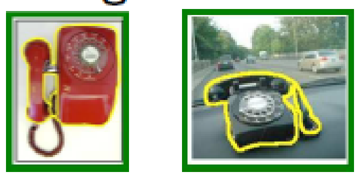
Source : http://vision.cs.utexas.edu/projects/active-prediction/

# 2) Different levels of information

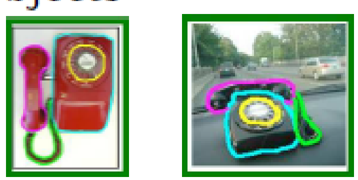
**Less expensive to obtain**

- **Weak labels:** informing about presence of an object


- **Strong labels:** outlines demarking the object



- **Stronger labels:** informing about labels of parts of objects

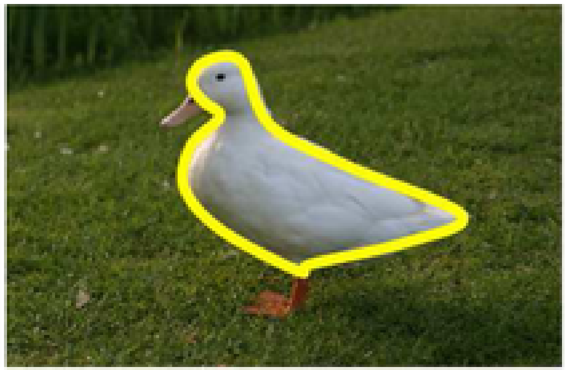


**More expensive to obtain**

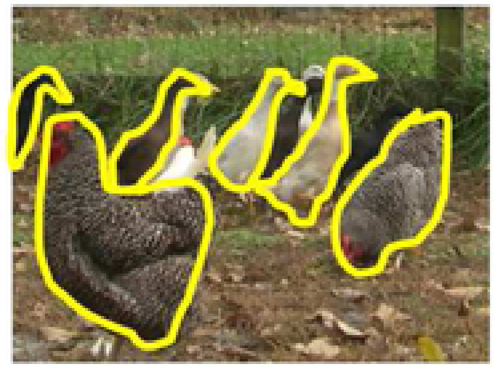
*Active learner must specify what type of annotation is currently most Helpful.*

Source : http://vision.cs.utexas.edu/projects/active-prediction/

3) Manual Effort dependent on annotation type and image content.



Low effort



High effort

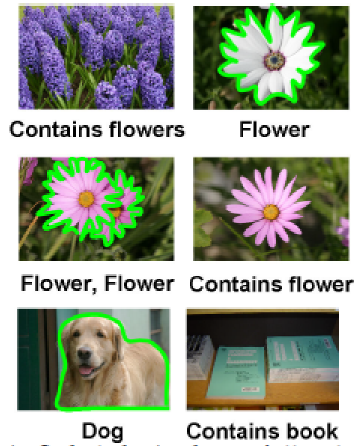
*Active learner should take into account the actual manual effort required to label the images.*

Source : http://vision.cs.utexas.edu/projects/active-prediction/

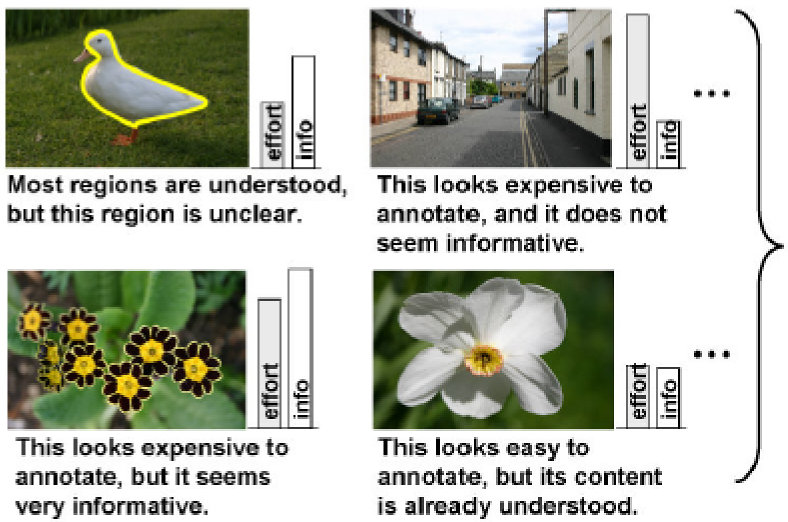
# PROBLEM STATEMENT

*How do we effectively learn from a mixture of strong and weak labels and select the most promising **{image + annotation type}** by balancing the **value of a new annotation** against the **time taken to receive it**.*

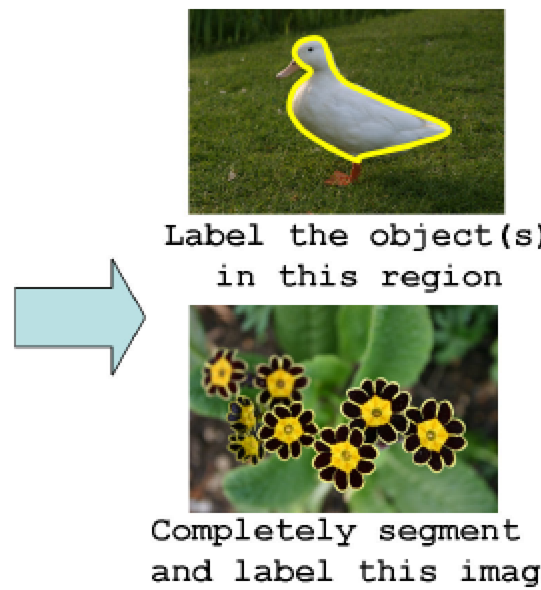
# PROPOSED APPROACH



(a) Labeled (and partially labeled) examples to build models



(b) Unlabeled and partially labeled examples to survey



(c) Actively chosen queries sent to annotators

# 3 types of annotations


Name an object

What class is this region?

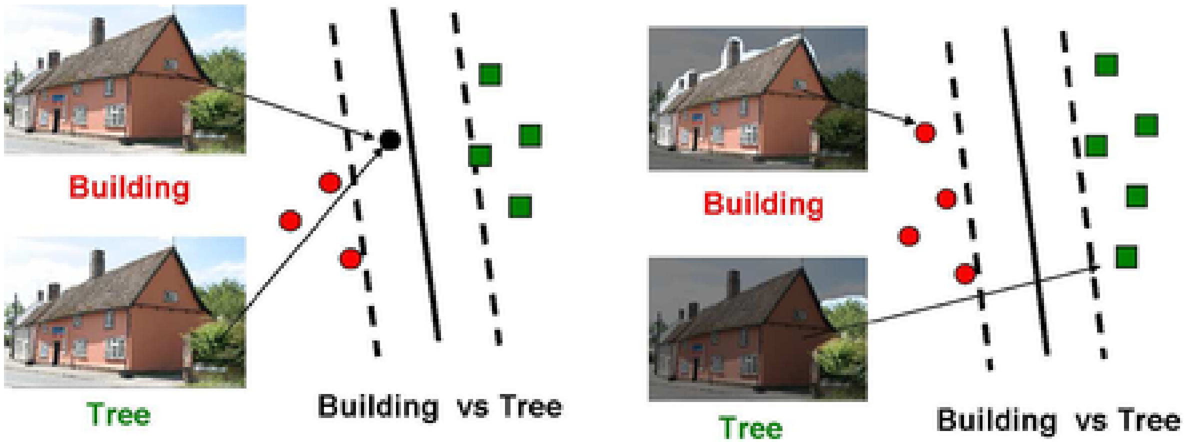
Segment this image

Source : http://vision.cs.utexas.edu/projects/active-prediction/

# Key Ideas of the approach

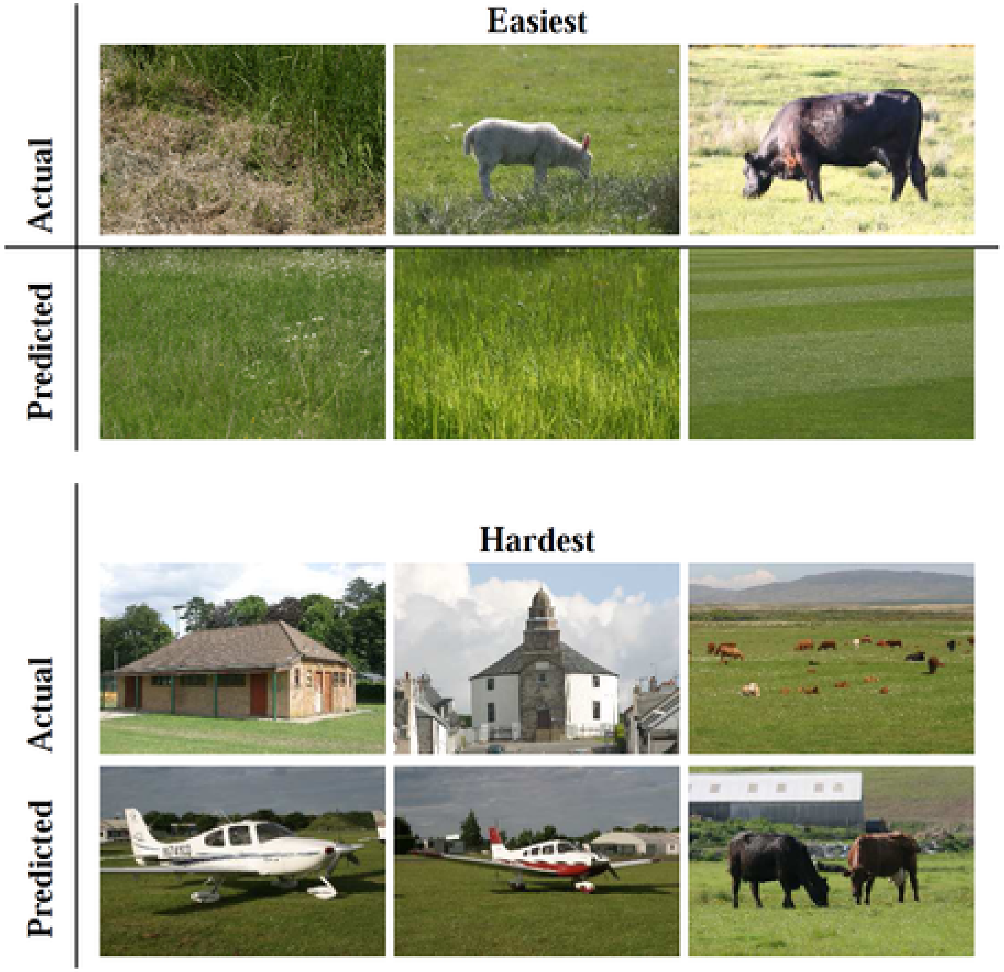
# 1. Multi-Label multiple-instance learning.

Multi-label Set Kernel based classifier




Source : http://vision.cs.utexas.edu/projects/active-prediction/

# 2. Predicting the cost of annotation based on image content.




Source : http://vision.cs.utexas.edu/projects/active-prediction/

# 3. Predicting the informativeness of an annotation *(z)*

Change in the *Total Misclassification Risk* resulted from z- Cost for obtaining that annotation

$$\begin{aligned} VOI(\mathbf{z}) &= T\left(\mathcal{X}_L, \mathcal{X}_U, \mathcal{X}_P\right) - T\left(\hat{\mathcal{X}}_L, \hat{\mathcal{X}}_U, \hat{\mathcal{X}}_P\right) \qquad (6) \\ &= \mathcal{R}(\mathcal{X}_L) + \mathcal{R}(\mathcal{X}_U) + \mathcal{R}(\mathcal{X}_P) \\ &\quad - \left(\mathcal{R}(\hat{\mathcal{X}}_L) + \mathcal{R}(\hat{\mathcal{X}}_U) + \mathcal{R}(\hat{\mathcal{X}}_P)\right) - \mathcal{C}(\mathbf{z}), \end{aligned}$$

# EXPERIMENTS

Label 3: Building, Sky

Label 4: Aeroplane, Grass, Sky

Label 5: Cow, Grass, Mountain

# 1. ACTIVE V/S RANDOM SELECTION

# Experimental Setup

**Active Learner**

| | |
|---|---|
| **Task** | **Actively** choose between **tags, regions and complete segmentation** |
| **Training & Validation Data** | Super pixel segments |
| **Test Data** | Ground Truth Segments |
| **Number of iterations(number of samples added to classifier)** | 30 |
| **Initial Training set size** | 3 ( 1 bag/image per class) |
| **Unlabeled Data** | 87 |

**Random Selection**

| | |
|---|---|
| **Task** | **Randomly** choose between **tags, regions and complete segmentation** |
| **Training & Validation Data** | Super pixel segments |
| **Test Data** | Ground Truth Segments |
| **Number of iterations(number of samples added to classifier)** | 30 |
| **Initial Training set size** | 3 ( 1 bag/image per class) |
| **Unlabeled Data** | 87 |

Risk /Cost prediction is not a part of random selection

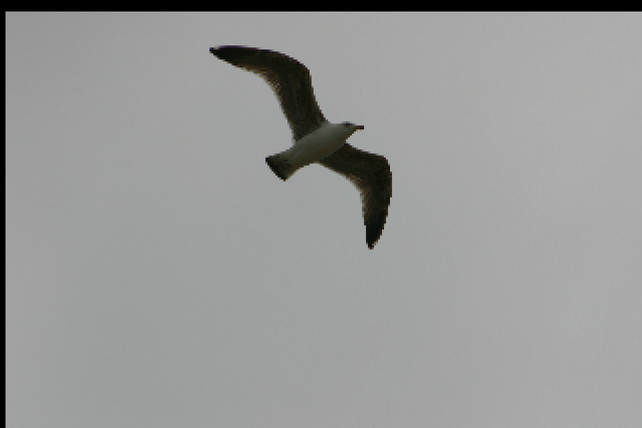
# ACTIVE LEARNER IN ACTION


Learner: What class is this region?
Oracle: Sky

Learner: What class is this region?
Oracle: Grass

Learner: Name an object
Oracle: Grass

Learner: What class is this region?
Oracle: Sky



Learner: What class is this region?
Oracle: Sky

# RANDOM SELECTION


Random: What class is this region?
Oracle: Sky

Random: Segment this image

Random: Name an object
Oracle: Sky

Random: What class is this region?
Oracle: Grass

Random: Segment this image


Random: What class is this region?
Oracle: Sky

Random: Segment this image

Random: Name an object
Oracle: Sky

Random: What class is this region?
Oracle: Grass

Random: Segment this image

(At the end of 30 iterations)

**Observation:** Requesting for complete segmentation of few images doesn't *necessarily* yield to better classification performance or reduction in risk.

# Active Learning v/s Random sampling

At the end of 30 iterations…

| | Active | Random ( average of 5 runs). |
|---|---|---|
| Execution time ( in secs) | 261 | **0** |
| Cost | **288** | 326.967 |
| Avg AUROI | **0.979867** | 0.966164 |
| Risk | **42.4003** | 44.07574 |

# Information as a function of iterations

Steep gain from the first few picks.

The most informative selections are made in the first few iterations.

Consistent gain in information in active selection.

For a given number of instances, active learner ensures the best possibile system.

2) WHAT TYPE OF ANNOTATION TO REQUEST?

1) Name an object

2) What class is this region?

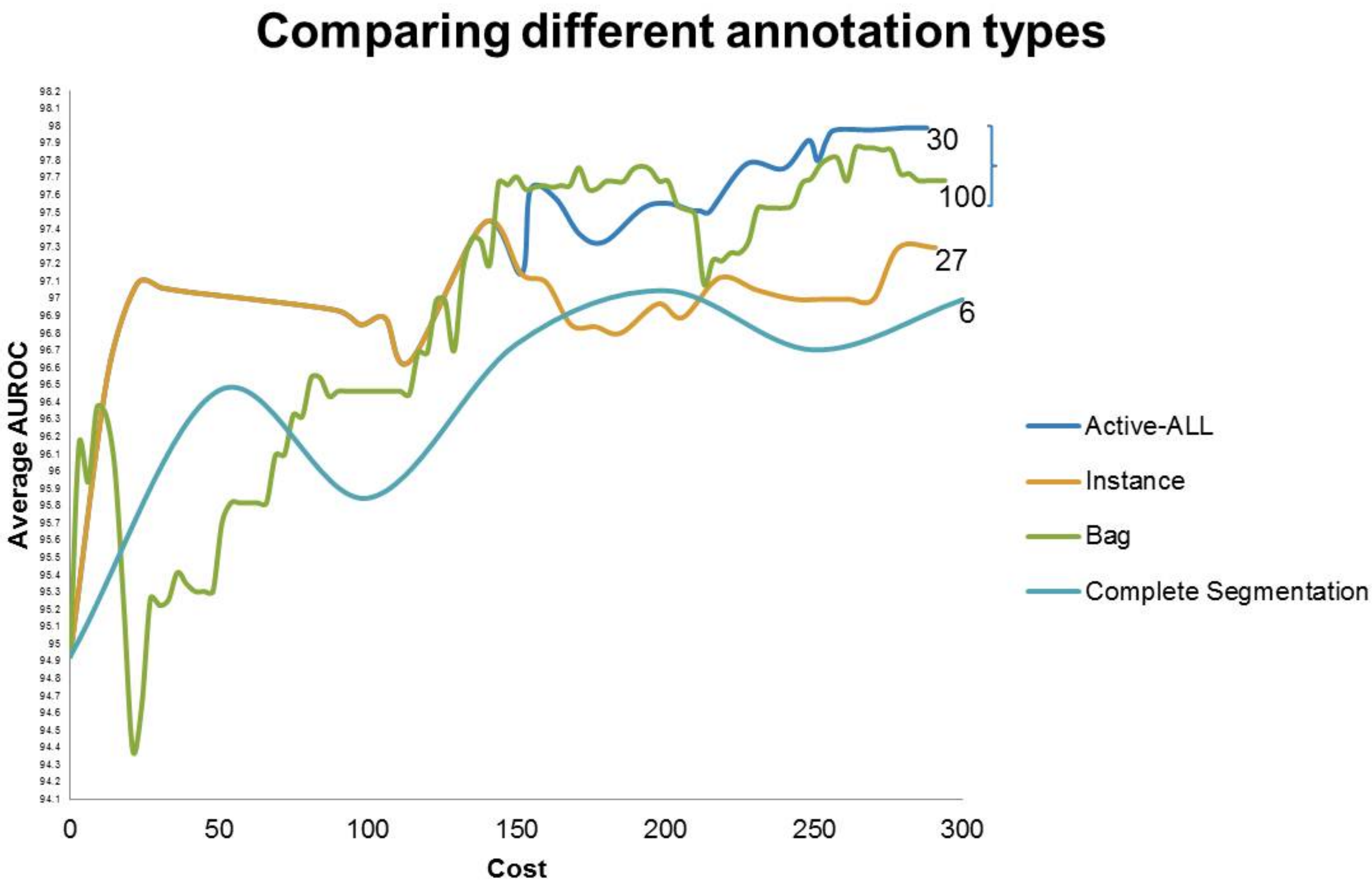

3) Segment this image

4) Any of the above

Source : http://vision.cs.utexas.edu/projects/active-prediction/

# Experimental Setup

| | |
|---|---|
| **Task** | Actively choose between<br>Case#1  Only tags<br>Case#2 : Only regions<br>Case#3   Only Complete segmentation<br>Case#4 Any of the above |
| **Training & Validation Data** | Super pixel segments |
| **Test Data** | Ground Truth Segments |
| **Number of iterations ( number of samples added to the training data)** | Variable ( Ranges between 6-100 for each of the above cases) |
| **Initial Training set size** | 3 ( 1 bag per class) |
| **Unlabeled Data** | 87 |




**Observation**

- A combination of annotation types is more beneficial than a fixed annotation type.

At the end of 30 iterations..

| | Active-ALL | Active-bag | Active-instance | Complete segmentation |
|---|---|---|---|---|
| Execution time ( in secs) | 261 | 121 | **23** | 49 |
| Cost | 288 | **84** | 331.667 | 789 |
| Global Mean Accuracy | **89.6774** | 88.3871 | 87.7419 | 87.0968 |
| Avg AUROI | **0.979867** | 0.965411 | 0.973077 | 0.9723 |
| Risk | **42.4003** | 39.7979 | 45.003 | 46.8725 |

Execution time is proportional to the number of instances/bags to be considered.

Better accuracy can be achieved by combining different types of annotations.

# 3. Ground Truth Segments v/s Super pixel Segments

**Key Idea:** Use ground truth segments instead of super pixel segments while training and testing MIML classifier.

**Aim:** To understand the upper bound of the active learner ( limitations from using super pixel segments)

# How noisy are the super pixel segments?

| **For 90 images..** | | Number of instances. |
| --- | --- | --- |
| | Ground truth Segments | 152 |
| | Super pixel segments | 272 |

| Segment Type | Image | Instance Labels | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Super pixel** | **I1** | 5 | 5 | 3 | 5 | 5 | 5 |
| **True** | **I1** | 3 | 5 | | | | |
| | | | | | | | |
| **Super pixel** | **I2** | 3 | 5 | 5 | 5 | 4 | 5 |
| **True** | **I2** | 3 | 4 | 5 | | | |

True Learner v/s Noisy Learner : Case1

## True Learner

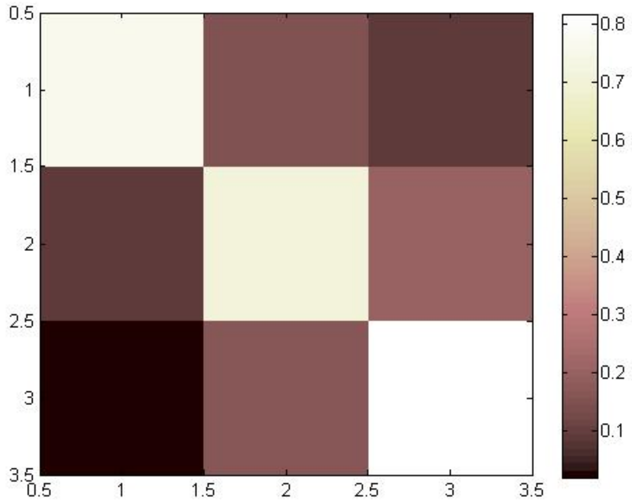
| | |
|---|---|
| **Task** | Actively choose between **tags, regions and complete segmentation** |
| **Training & Validation Data** | **Ground Truth segments** |
| **Test Data** | **Super pixel Segments** |
| **Number of iterations (number of samples added to classifier)** | 30 |
| **Initial Training set size** | 3 ( 1 bag per class) |
| **Unlabeled Data** | 87 |

## Noisy Learner

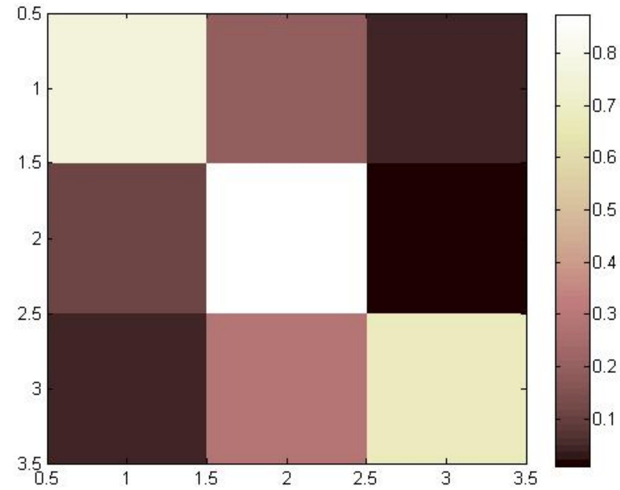
| | |
|---|---|
| **Task** | Actively choose between **tags, regions and complete segmentation** |
| **Training & Validation Data** | **Super pixel segments** |
| **Test Data** | **Super pixel Segments** |
| **Number of iterations (number of samples added to classifier)** | 30 |
| **Initial Training set size** | 3 ( 1 bag per class) |
| **Unlabeled Data** | 87 |

Confusion matrix

Noisy Learner




True Learner




Per class mean accuracy: 76.2958

Per class mean accuracy: 77.0973

(At the end of 30 iterations)

# True Learner v/s Noisy Learner : Case1

At the end of 30 iterations…

| | True Learner | Noisy Learner |
|---|---|---|
| Total Execution time ( in secs) | **53** | 195 |
| Cost | **128.000000** | 194.166667 |
| Global Mean Accuracy | **78.8235** | 75.2941 |
| Avg AUROI | **0.960086** | 0.936850 |
| Risk | **84.6304** | 98.0278 |

- Lesser number of instances to process during training and validation for the true learner

True Learner v/s Noisy Learner : Case2

### True Learner

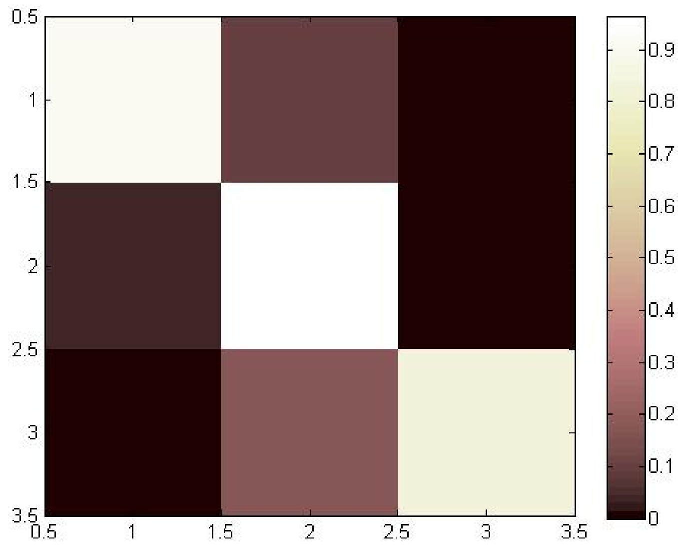
| | |
|---|---|
| **Task** | Actively choose between **tags, regions and complete segmentation** |
| **Training & Validation Data** | **Ground Truth segments** |
| **Test Data** | **Ground Truth segments** |
| **Number of iterations (number of samples added to classifier)** | 30 |
| **Initial Training set size** | 3 ( 1 bag per class) |
| **Unlabeled Data** | 87 |

### Noisy Learner

| | |
|---|---|
| **Task** | Actively choose between **tags, regions and complete segmentation** |
| **Training & Validation Data** | **Super pixel segments** |
| **Test Data** | **Ground Truth segments** |
| **Number of iterations (number of samples added to classifier)** | 30 |
| **Initial Training set size** | 3 ( 1 bag per class) |
| **Unlabeled Data** | 87 |

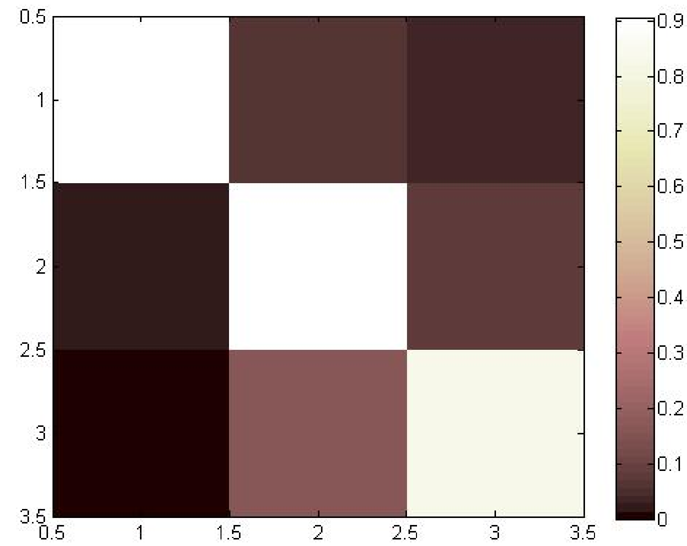
# True Learner v/s Noisy Learner : Case2

True Learner



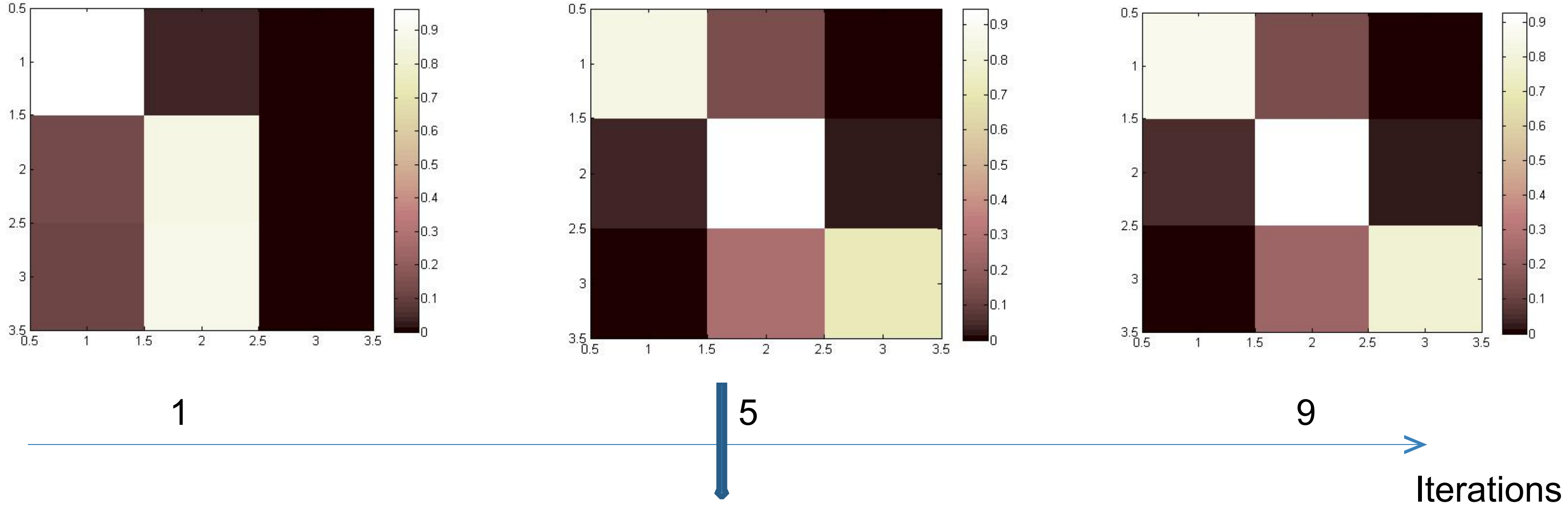

Per class mean accuracy: 90.012

Noisy Learner




Per class mean accuracy: 88.1252

# Initial misclassification with true segments





Learner: What class is this region?
Oracle: Grass

The first four iterations have always misclassified the images with class label=5, this is changed when the above image is added.

# True Learner v/s Noisy Learner : Case2

True Learner v/s Noisy Learner : Case2

# At the end of 30 iterations…

| | True Learner | Noisy Learner |
|---|---|---|
| Execution time ( secs) | 136 | 261 |
| Cost | 128. | 288 |
| Total Mean Accuracy | **91.6129** | 89.6774 |
| Avg AUROI | 0.987389 | 0.979867 |
| Risk | 38.1211 | 42.4003 |

True Learner v/s Noisy Learner : Case3

## True Learner

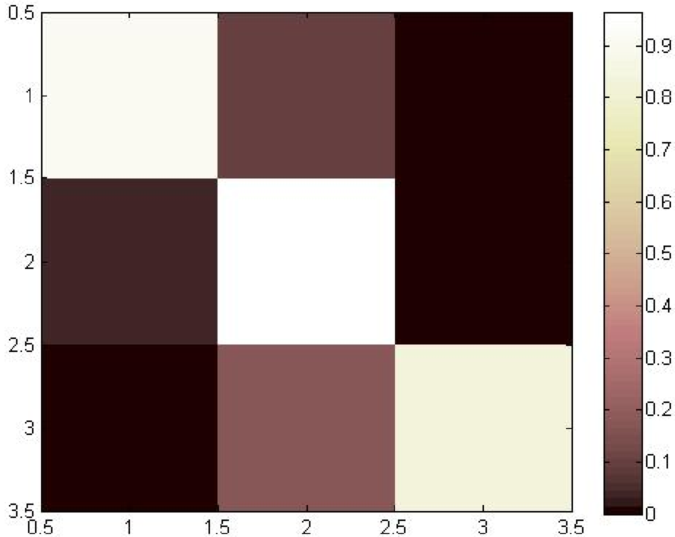
| | |
|---|---|
| **Task** | Actively choose between **tags, regions and complete segmentation** |
| **Training & Validation Data** | **Ground Truth segments** |
| **Test Data** | **Ground Truth segments** |
| **Number of iterations (number of samples added to classifier)** | 30 |
| **Initial Training set size** | 3 ( 1 bag per class) |
| **Unlabeled Data** | 87 |

## Noisy Learner

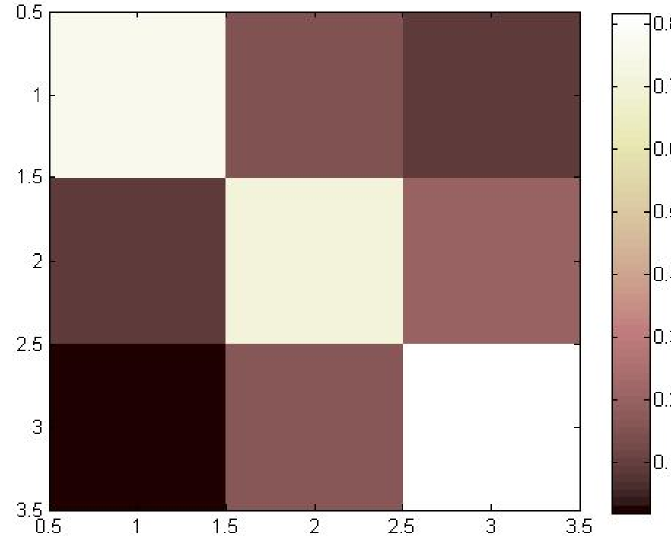
| | |
|---|---|
| **Task** | Actively choose between **tags, regions and complete segmentation** |
| **Training & Validation Data** | **Super pixel segments** |
| **Test Data** | **Super pixel segments** |
| **Number of iterations (number of samples added to classifier)** | 30 |
| **Initial Training set size** | 3 ( 1 bag per class) |
| **Unlabeled Data** | 87 |

- Noisy Learner  reaches the cost of that of a true learner  is in 30 iterations, in only 18 iterations of learning.
- True learner outperforms noisy learner by a very high margin.
- ***This test case demonstrates the upper bound of the active learner***

# Confusion Matrix




Per class mean accuracy: 90.012




Per class mean accuracy: 76.2958

(At the end of 30 iterations)

# Active Learning: True v/s Noisy Learner

At the end of 30 iterations…

| | True Learner | Noisy Learner |
|---|---|---|
| Execution time ( secs) | **53** | 195 |
| Cost | **128** | 194.167 |
| Total Mean Accuracy | **91.6129** | 75.2941 |
| Avg AUROI | **0.987389** | 0.936850 |
| Risk | **38.1211** | 98.0278 |

True learner is faster, more accurate and has lesser total risk.

Conclusion

| Test Cases | Risk |
|---|---|
| Train: Ground Truth Segments<br>Test: Ground Truth Segments | 38.1211 |
| Train: Super Pixel Segments<br>Test: Super Pixel Segments | 98.0278 |
| Train: Ground Truth Segments<br>Test: Super Pixel Segments | 84.6304 |
| Train: Super Pixel Segments<br>Test: Ground Truth Segments | 42.4003 |

→ Best case

→ Real System

- Super pixel segmentation imposes limitations on the performance of the active learner.
- The total risk of the system is lesser when ground truth segments are used.
- The computational cost of the system also varies with the correctness of the segments.

# 4) The contribution of each variable to the VOI for an annotation

$$
\begin{aligned}
VOI(\mathbf{z}) &= T\left(\mathcal{X}_L, \mathcal{X}_U, \mathcal{X}_P\right) - T\left(\hat{\mathcal{X}}_L, \hat{\mathcal{X}}_U, \hat{\mathcal{X}}_P\right) \qquad (6) \\
&= \mathcal{R}(\mathcal{X}_L) + \mathcal{R}(\mathcal{X}_U) + \mathcal{R}(\mathcal{X}_P) \\
&\quad - \left(\mathcal{R}(\hat{\mathcal{X}}_L) + \mathcal{R}(\hat{\mathcal{X}}_U) + \mathcal{R}(\hat{\mathcal{X}}_P)\right) - \mathcal{C}(\mathbf{z}),
\end{aligned}
$$

- Importance of cost prediction : **C(z)**
- Effect of the risk parameter : $r_L$
- VOI from Labeled data
- VOI from Partially labeled data.
- VOI from Unlabeled data.

Importance of cost prediction.

Effect of the risk parameter.

VOI from Labeled data.

VOI from Partially labeled data.

VOI from Unlabeled data.

$$
\begin{aligned}
VOI(\mathbf{z}) &= T\left(\mathcal{X}_L, \mathcal{X}_U, \mathcal{X}_P\right) - T\left(\hat{\mathcal{X}}_L, \hat{\mathcal{X}}_U, \hat{\mathcal{X}}_P\right) \qquad (6) \\
&= \mathcal{R}(\mathcal{X}_L) + \mathcal{R}(\mathcal{X}_U) + \mathcal{R}(\mathcal{X}_P) \\
&\quad - \left(\mathcal{R}(\hat{\mathcal{X}}_L) + \mathcal{R}(\hat{\mathcal{X}}_U) + \mathcal{R}(\hat{\mathcal{X}}_P)\right) - \boxed{\mathcal{C}(\mathbf{z})}
\end{aligned}
$$

# Experimental Setup

| | |
|---|---|
| **Task** | Actively choose between tags, regions and complete segmentations.<br><br>Case#1 – Without Annotation Cost.<br>Case#2 - With Annotation Cost. |
| **Training & Validation Data** | Super pixel segments |
| **Test Data** | Ground Truth Segments |
| **Number of iterations ( number of samples added to the training data)** | Variable ( Ranges between 6-100 for each of the above cases) |
| **Initial Training set size** | 3 ( 1 bag per class) |
| **Unlabeled Data** | 87 |

VOI only in terms of the estimate of misclassification

| Iterations=30 | Without Annotation Cost | **With Annotation Cost** |
|---|---|---|
| Total Cost | **756** | 288 |
| Global Mean Accuracy | 86.4516 | 89.6774 |

- Without C(z), VOI is measured only in terms of estimate risk of misclassification.
- Having the penalty on cost is useful in making better choices.

Importance of cost prediction.

Effect of the risk parameter.

VOI from Labeled data.

VOI from Partially labeled data.

VOI from Unlabeled data.

# The effect of the risk parameter ($r_L$)

$$\mathcal{R}(\mathcal{X}_L) = \sum_{X_i \in \mathcal{X}_L} \sum_{l \in L_i} \boxed{r_l} (1 - p(l|X_i))$$

$$\mathcal{R}(\mathcal{X}_U) = \sum_{X_i \in \mathcal{X}_U} \sum_{l=1}^{C} \boxed{r_l} (1 - p(l|X_i)) \Pr(l|X_i),$$

$$\begin{aligned} \mathcal{R}(\mathcal{X}_P) &= \sum_{X_i \in \mathcal{X}_P} \sum_{l \in L_i} \boxed{r_l} (1 - p(l|X_i)) \\ &+ \sum_{l \in U_i} \boxed{r_l} (1 - p(l|X_i)) \, p(l|X_i), \end{aligned}$$

30

100

**Observations:**

- Without $r_L$, the effect of the risk estimations is negligible and choice of instances is dominated by C(z)
- We get better accuracy with lesser number of instances when risk estimation is also included.
- ***Thus, an equal contribution of both cost estimation and risk estimation leads to more informative learning.***

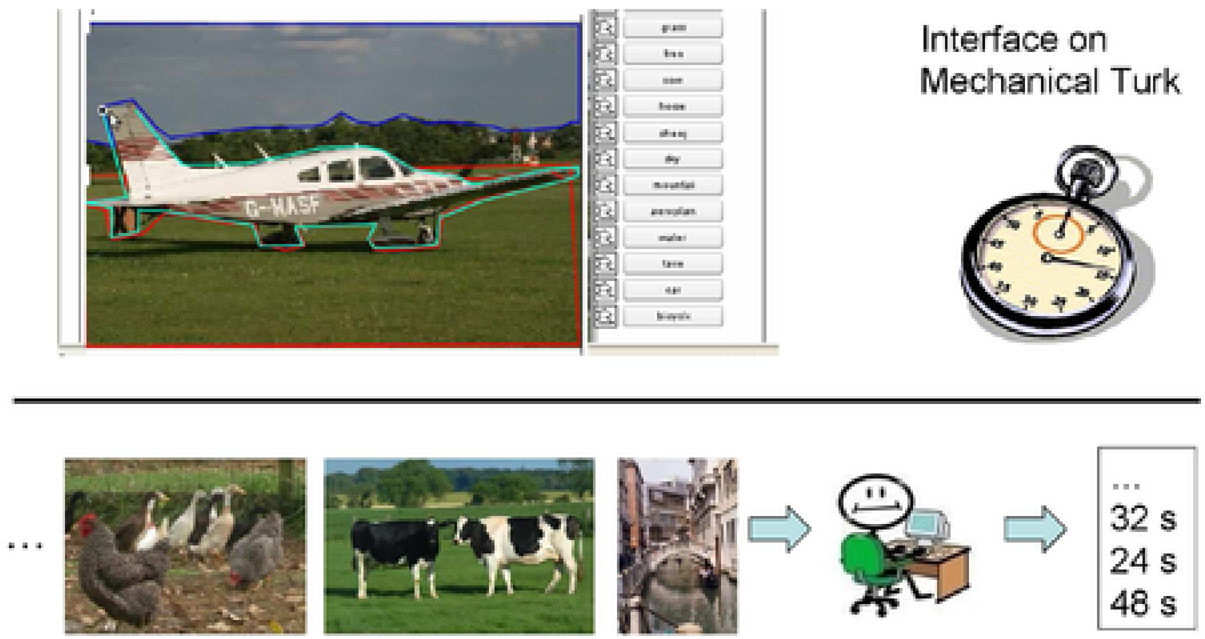
Importance of cost prediction.
Effect of the risk parameter.
VOI from Labeled data $R(X_L)$
VOI from Partially labeled data $R(X_P)$
VOI from Unlabeled data $R(X_U)$

$$\begin{aligned} VOI(\mathbf{z}) &= T(\mathcal{X}_L, \mathcal{X}_U, \mathcal{X}_P) - T\left(\hat{\mathcal{X}}_L, \hat{\mathcal{X}}_U, \hat{\mathcal{X}}_P\right) \quad (6) \\ &= \mathcal{R}(\mathcal{X}_L) + \mathcal{R}(\mathcal{X}_U) + \mathcal{R}(\mathcal{X}_P) \\ &\quad - \left(\mathcal{R}(\hat{\mathcal{X}}_L) + \mathcal{R}(\hat{\mathcal{X}}_U) + \mathcal{R}(\hat{\mathcal{X}}_P)\right) - \mathcal{C}(\mathbf{z}), \end{aligned}$$

| Iterations=30 | Without R(L) | Without R(P) | Without R(U) |
|---|---|---|---|
| Total Cost | **394** | 84 | 288 |

- Exclusion of R(L) leads to high risk and high total cost.
- This result shows the real contribution of each pool to the decision making.
- Since most changes are happening to the labeled pool of data, with every iteration, it has the highest contribution to the VOI.

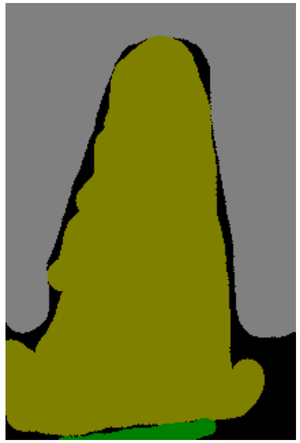
# 5) ANNOTATION DATA




| | |
|---|---|
| **Images** | 240 |
| **Users** | ~70 |

Source : http://vision.cs.utexas.edu/projects/active-prediction/

# Most picked v/s Least picked Images (Avg=22)

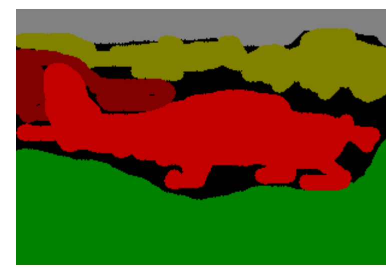



31


13

# **Least v/s most agreed upon Images (Avg= ~27)**


3.3545



149.32