Keywords to Visual Categories: Multiple-Instance Learning for Weakly Supervised Object Categorization

Sudheendra Vijayanarasimhan and Kristen Grauman
Department of Computer Sciences, University of Texas at Austin

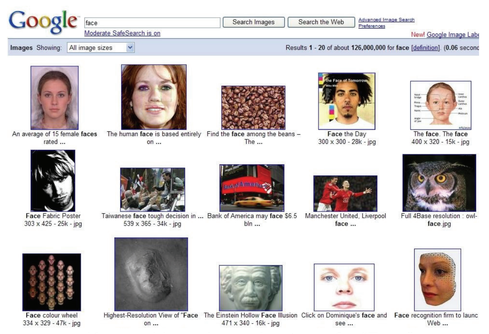
Problem

Carefully prepared labeled datasets are useful for learning visual category models, but they are expensive to obtain.

Images on the Web indexed using keywords are available in large quantities, but do not necessarily portray the query term and are much more variable than typical benchmark training sets.

Caltech101 "Face" training examples

vs.



Keyword-based search returns for "Face"

How can we learn reliable category models from images returned by keyword search?

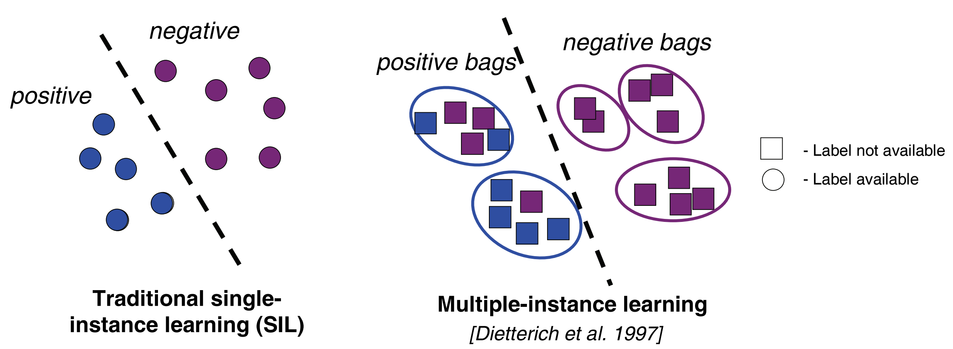
Our approach

We consider a *multiple-instance visual category learning* scenario to directly obtain discriminative models for specified categories.

- Large-margin solution with constraints to accommodate expected sparsity of good examples
- Iteratively improve multiple-instance classifier by automatically refining the representation

Previous approaches cluster to find visual themes [Sivic et al. 2005, Fergus et al. 2005, Li et al. 2007], but it is impossible to guarantee any clusters will correspond to desired topic. Others apply models known to work well with correctly labeled data [Fergus et al. 2004, Schroff et al. 2007], but extremely noisy data is problematic.

Multiple-Instance Learning (MIL)

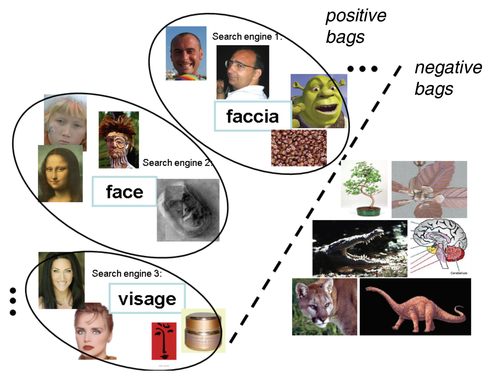


Traditional single-instance learning (SIL)

Multiple-instance learning *[Dietterich et al. 1997]*

In MIL, training examples are *bags* of instances, and we only know that each positive bag has at least one true positive instance. The classifier must predict labels for novel *instances*.

Algorithm Overview



Collect set of positive bags $\mathcal{X}_p$ from multiple search engines using the category name translated into multiple languages.

Collect set of negative bags $\mathcal{X}_n$ from images with known labels, or unrelated searches.

Let X denote a bag, x denote an instance, and let $\tilde{\mathcal{X}}_n = \{x | x \in X \in \mathcal{X}_n\}$ be all negative instances.

Sparse MIL (sMIL)

To begin, we solve a large-margin decision problem with constraints as suggested in [Bunescu & Mooney, 2007] to reflect the fact that positive bags may contain as few as one true positive:

$$\text{minimize:} \quad \frac{1}{2}||w||^2 + \frac{C}{|\tilde{\mathcal{X}}_n|}\sum_{x\in\tilde{\mathcal{X}}_n}\xi_x + \frac{C}{|\mathcal{X}_p|}\sum_{X\in\mathcal{X}_p}\xi_X$$

Maximize the margin with minimal slack penalties.

$$\text{subject to:} \quad w\,\phi(x) + b \le -1 + \xi_x, \qquad \forall x \in \tilde{\mathcal{X}}_n$$

All negative instances should be labeled as negative.

$$w\frac{\phi(X)}{|X|} + b \ge \frac{2-|X|}{|X|} - \xi_X, \qquad \forall X \in \mathcal{X}_p$$

Sum constraints per positive bag: at least one instance in each bag is positive, but the rest can be negative.

$$\xi_x \ge 0, \xi_X \ge 0.$$

where $\phi(X) = \sum_{x\in X}\phi(x)$.

$$y(\hat{x}) = +1, \hat{x} \in X, \quad y(x) = -1, x \in X \setminus \{\hat{x}\}$$

$$\sum_{x\in X}\frac{\omega\phi(x)+b}{|X|} \ge \sum_{x\in X}\frac{y(x)}{|X|} - \xi_X \ge \frac{y(\hat{x})}{|X|} + \sum_{x\in X\setminus\{\hat{x}\}}\frac{y(x)}{|X|} - \xi_X \ge \frac{+1}{|X|} + \frac{1-|X|}{|X|} - \xi_X$$

The sparse MIL problem is convex, and reduces to single-instance learning when bags are of size 1.

Iterative refinement of positive bags

A limitation of the above objective is that the summed constraints mean each bag is mapped to the mean of its components; we would like instead for a bag to be represented mainly by the true positives within it.

We propose a new iterative refinement scheme for MIL that biases the bag representation towards instances that are likely to be positive:

$$\phi(X) = \frac{\sum_{i=1}^{|X|}\omega_i^{(t)}\phi(x_i)}{\sum_{i=1}^{|X|}\omega_i^{(t)}}$$

A positive bag is represented by a weighted sum of its instances.

$$\omega_i^{(0)} = \frac{1}{|X|}$$

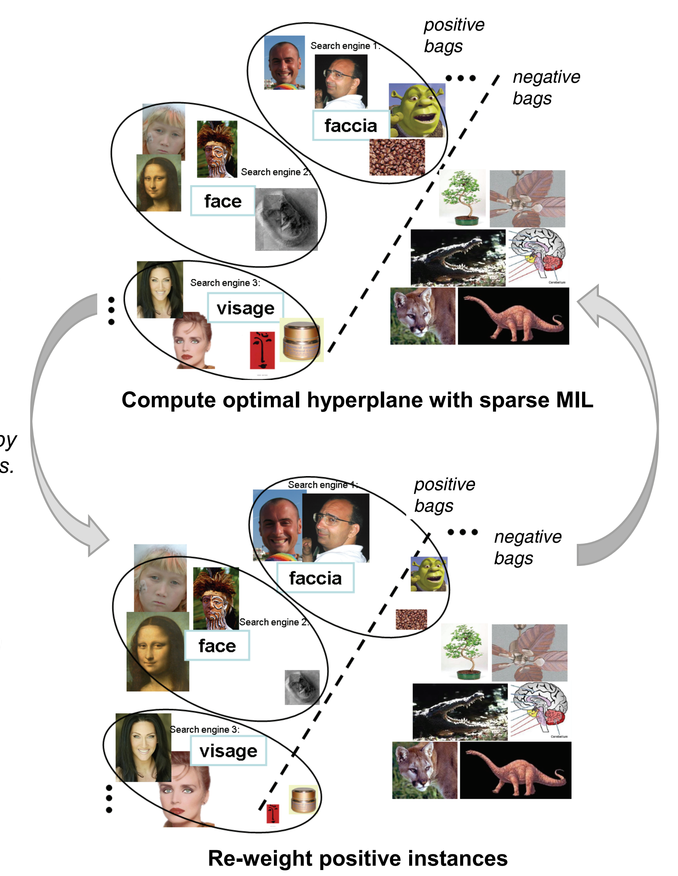
Initially all instances receive equal weight.

$$\omega_i^{(t)} = \omega_i^{(t-1)} e^{\frac{(y_i - y_m)}{\sigma^2}}$$

Weights are updated based on the distance of each instance from the current hyperplane.

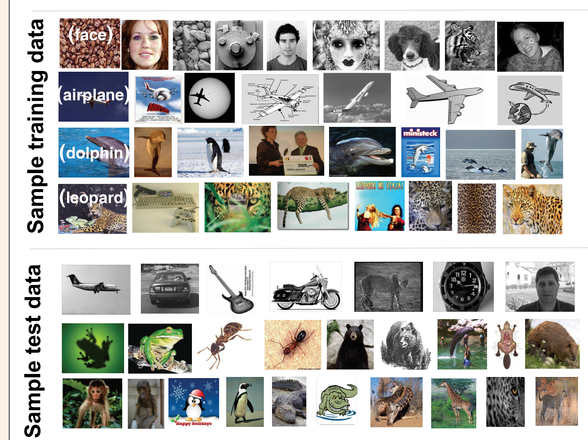
where $y_i = w\phi(x_i) + b$

$y_m = \text{argmax}_{x_i \in X}\, y_i$



Results

We evaluate our method for two kinds of tasks: categorizing novel examples, and re-ranking keyword search returns according to predicted relevance.



Training data:

We download training examples for each category name using Google Image Search, Yahoo, and MSN, in five languages. On average, about 30% are good.

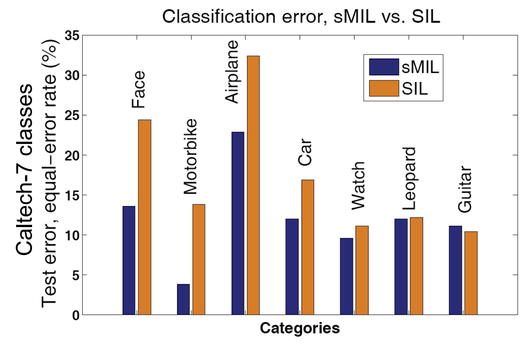
For comparison, we also obtain results when training with the same Google images downloaded in [Fergus et al. 2005].

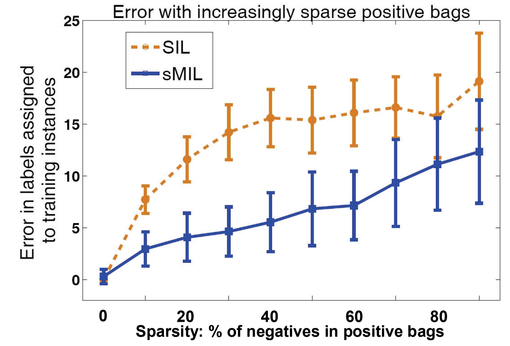
Test data:

Caltech 7 dataset: 2,148 images from seven classes

Animals dataset: 10,000 images from ten classes [Berg et al. 2006]

We represent images with bags of local SIFT features from four interest operators (Harris-Affine, DoG, edges, Kadir & Brady), and use an RBF kernel. Features were chosen to directly compare to previous work.

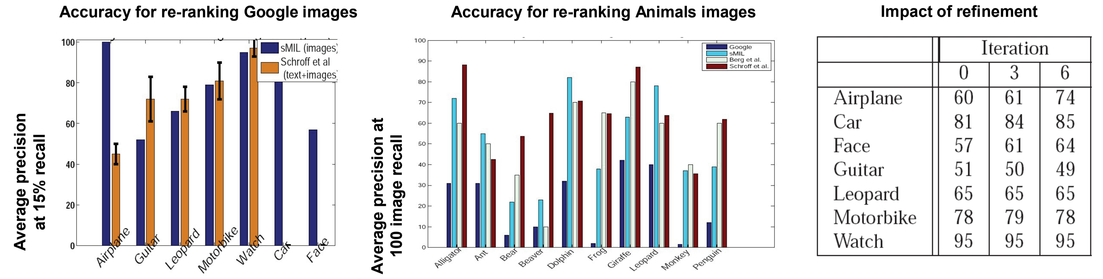




MIL is better suited to the sparse, noisy training data than the SIL baseline, and degrades much more gradually when given fewer true positives.

Ant. of manual supervision:	img labels	img labels	img labels +segment.	true img labels	none	none	none	none	none
Source of training data	Caltech	Caltech	Caltech	Google	Caltech	Noisy Caltech	Google	Google	multiple search engines
Method / Category	Fergus et al. 2003	Opelt et al. 2004	Leibe et al. 2004	SIL-SVM	Sivic et al. 2005	sMIL′	Fergus et al. 2005	sMIL	sMIL
Airplane	**7.0**	11.1	-	4.9	***3.4***	5.0	15.5	**10.7**	22.9
Car (rear)	9.7	8.9	**6.1**	10.7	21.4	***5.4***	16.0	**11.8**	12.0
Face	***3.6***	6.5	-	21.8	**5.3**	11.5	20.7	23.1	**13.6**
Leopard	***10.0***	-	-	11.1	-	-	13.0	12.4	**12.0**
Motorbike	6.7	7.8	**6.0**	4.0	15.4	***3.8***	6.2	***3.8***	***3.8***
Guitar	-	-	-	*6.9*	-	-	31.8	***8.2***	11.1
Wrist watch	-	-	-	*7.3*	-	-	19.9	**8.9**	9.6
Average error	-	-	-	9.5	-	-	17.59	11.27	12.14

Equal error rates on Caltech test data. Our results improve the state-of-the-art in unsupervised recognition, and are even competitive with fully supervised techniques.



	Iteration		
	0	3	6
Airplane	60	61	74
Car	81	84	85
Face	57	61	64
Guitar	51	50	49
Leopard	65	65	65
Motorbike	78	79	78
Watch	95	95	95

Using image content alone, our approach provides accurate re-ranking results, and for some classes improves precision more than methods employing both text and image features.