

# Multimodal semi-supervised learning for image classification

Matthieu Guillaumin, Jakob Verbeek, Cordelia Schmid

LEAR team, INRIA Grenoble, France



# Motivation and goal

- Images often come with additional textual info.



WIKIPEDIA  
The Free Encyclopedia

[Main page](#)  
[Contents](#)  
[Featured content](#)  
[Current events](#)  
[Random article](#)

▼ Interaction  
[About Wikipedia](#)  
[Community portal](#)  
[Recent changes](#)  
[Contact Wikipedia](#)  
[Donate to Wikipedia](#)  
[Help](#)

New features 🌟 [Log in](#) / [create account](#)

Article [Discussion](#) Read [Edit](#) [View history](#)


## Golden Gate Bridge

From Wikipedia, the free encyclopedia

**The [Golden Gate Bridge](#)** is a [suspension bridge](#) spanning the [Golden Gate](#), the opening of the [San Francisco Bay](#) into the [Pacific Ocean](#). As part of both [U.S. Route 101](#) and [California State Route 1](#), it connects the city of [San Francisco](#) on the northern tip of the [San Francisco Peninsula](#) to [Marin County](#). The Golden Gate Bridge was the [longest suspension bridge span](#) in the world when it was completed during the year 1937, and has become one of the most internationally recognized symbols of [San Francisco](#), [California](#), and of the [United States](#). Since its completion, the span length has been surpassed by eight other bridges. It still has the second longest suspension bridge main span in the United States, after the [Verrazano-Narrows Bridge](#) in [New York City](#). In 1999, it was ranked fifth on the [List of America's Favorite Architecture](#) by the [American Institute of Architects](#).

**Coordinates:** 37°49′11″N 122°28′43″W﻿ / ﻿

### Golden Gate Bridge



**Carries** 6 lanes of  [101](#)  [US 101](#) / [SR 1](#) , pedestrians and bicycles

*Spans* [Golden Gate](#)

*Harnraet in Fletcher's creek likes nie*

- Videos with scripts and subtitles, ...

# Goal of this work

- Visual object category recognition,
- Leveraging user tags available on **flickr**:



## Tags

- wow
- San Fransisco
- Golden Gate Bridge
- SBP2005
- top-f50
- fog
- SF Chronicle 96 hours

# Overview of the talk

(A) Data sets and features

(B) Learning scenarios using images with tags

(1) Supervised multimodal classification

(2) Multimodal semi-supervised scenario

(3) Weakly supervised learning

# Data sets of images with tags

- PASCAL VOC 07,  $\approx 10000$  images, 804 Flickr tags, 20 classes.



*Flickr tags:* india  
*Class labels:* cow



aviation, airplane, airport  
aeroplane

- MIR Flickr, 25000 images, 457 Flickr tags, 38 classes.



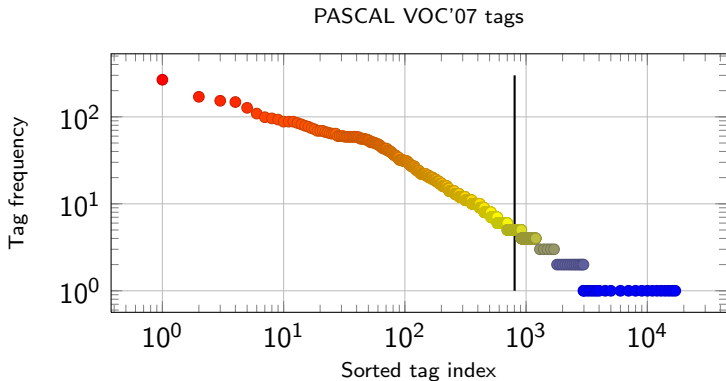
*Flickr tags:* desert, nature, landscape, sky  
*Class labels:* clouds, plant life, sky, tree



rose, pink  
flower, plant life

# Flickr tags as textual features

- Restrict to the most frequent tags.



- Binary vector of tag presence/absence.
- Linear kernel counts the number of shared tags.

# Combination of several visual features

- RBF kernel on average distance between 15 image representations:
  - Bag-of-features histograms:
    - Harris interest points and dense grid,
    - SIFT [Lowe, 2004] and Hue [van de Weijer & Schmid, 2006],
    - K-means quantization.
  - Color histograms:
    - RGB, HSV and Lab colorspaces,
    - 16 bins per channel.
  - GIST [Oliva & Torralba, 2001],
  - 2 spatial layouts
    - Global,
    - 3 horizontal regions [Lazebnik *et al.*, 2006],
    - Only global for GIST.

# Learning scenarios using images with tags

- 1 Supervised multimodal classification
- 2 Multimodal semi-supervised scenario
- 3 Weakly supervised learning



# Supervised multimodal classification

- Flickr tags = additional features for classification.
- Tags also available at test time,
- MKL to combine visual and textual kernels.

DOG (+1)



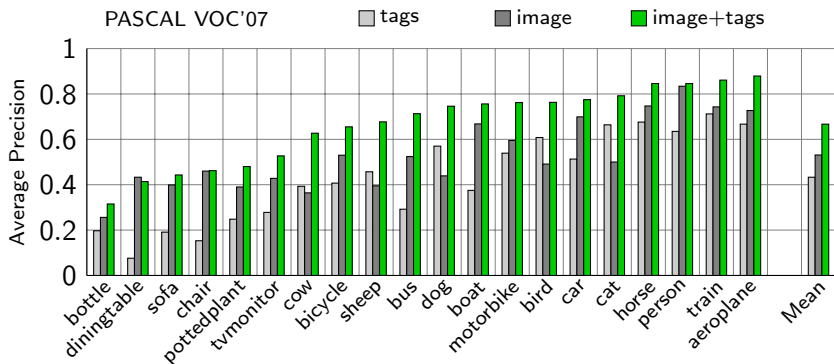
not DOG (-1)



DOG?



# Results of multimodal classification on PASCAL VOC 2007



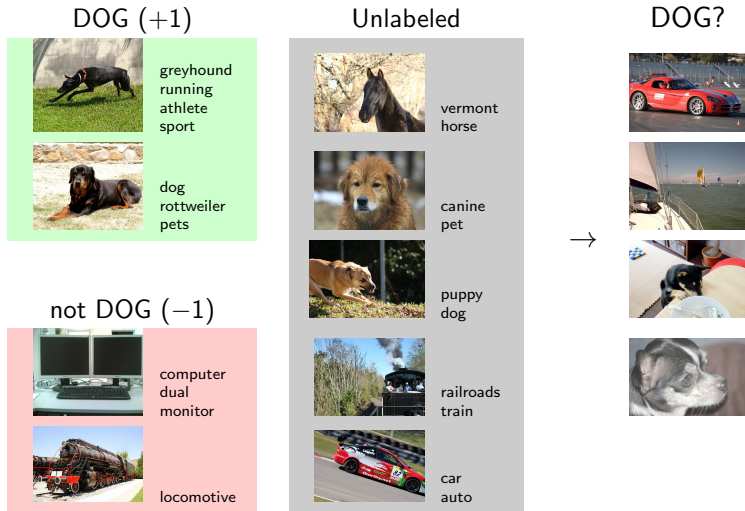
- Tags (0.43) < Image (0.53) < Image+tags (0.67)
- Winner of PASCAL VOC'07: 0.59.
- Similar observation for MIR Flickr.

# Learning scenarios using images with tags

- 1 Supervised multimodal classification
- 2 Multimodal semi-supervised scenario
- 3 Weakly supervised learning

# Multimodal semi-supervised scenario

- Large pool of additional unlabeled images with tags.
- Tags **NOT** available at test time: visual categorization.



# Three-step learning process

In a nutshell, predict labels for the unlabeled images:

- ① Train an MKL classifier on labeled images and tags.
- ② Score unlabeled data.
- ③ Train an image-only classifier. 2 options:
  - ① SVM:
    - Use unlabeled data with label from sign of MKL score,
    - Using only the sign, we dismiss the confidence of classification.
  - ② LSR:
    - Least-squares regression of MKL scores using the visual kernel,
    - Regularized using KPCA projection.

# Experimental comparison

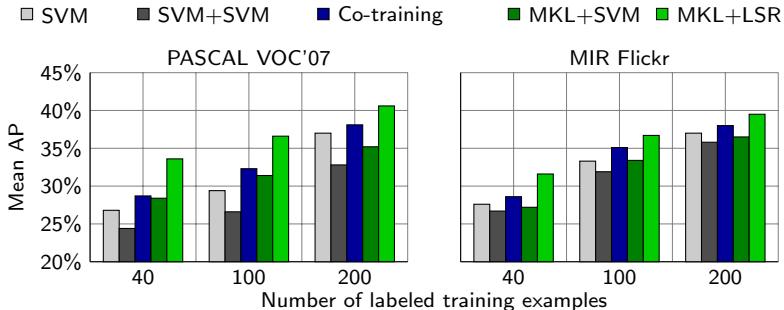
## Baselines:

- ① Supervised, image-only: **SVM**,
- ② Semi-supervised, image-only: **SVM+SVM**,
- ③ Semi-supervised, multimodal: **Co-training**, with SVM on images and SVM on tags. [Blum & Mitchell, 98]

## Our three-step learning approach (semi-supervised, multimodal):

- ① MKL learned on labeled images with tags, followed by visual-only SVM trained on labeled and unlabeled images: **MKL+SVM**,
- ② MKL, followed by LSR: **MKL+LSR**.

# Results of semi-supervised learning



- SVM+SVM worse than baseline.
- With little supervision, MKL+LSR is significantly better.
- With more supervision, differences shrink.

# Learning scenarios using images with tags

- 1 Supervised multimodal classification
- 2 Multimodal semi-supervised scenario
- 3 Weakly supervised learning

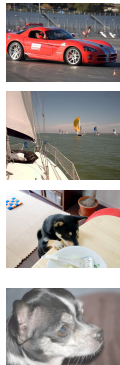


# Weakly supervised scenario

- For learning: no manual annotation, but Flickr tags,
- Other tags used as additional features.
- For evaluation: ground-truth labels.



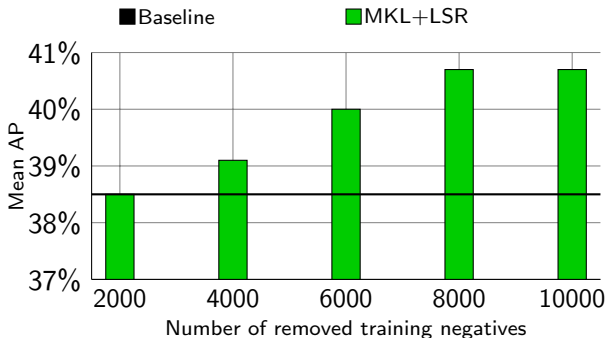
DOG?



# Weakly supervised setting

- Tags are noisy annotations:
  - Tag presence is relatively clean (82.0% precision)
  - Tag absence is relatively uninformative (17.8% recall)
- Our approach, modified:
  - 1 Learn a multimodal MKL with tag annotations,
  - 2 Rank training images and remove the images that yield highest MKL scores but do not have the tag,
  - 3 Fit LSR.
- Baseline: visual-only SVM learned on images with tag annotations.

# Results on 18 classes of MIR Flickr



- mAP on 18 MIR Flickr classes.
- On average, MKL+LSR outperforms SVM baseline:
  - SVM baseline better for 4 classes (up to +5.6%),
  - MKL+LSR better for 14 classes (up to +9.8%).

# Conclusion

- We considered using Flickr tags for 3 scenarios:
  - ① Supervised classification,
  - ② Semi-supervised learning of visual classifiers,
  - ③ Weakly supervised learning of visual classifiers.
- We proposed a three-step learning process:
  - ① Training of a multimodal classifier on labeled data,
  - ② Classification of the unlabeled data,
  - ③ Regression of the multimodal classifier.
- Our multimodal approach using Flickr tags improves over:
  - Visual-only SVM on all three scenarios,
  - Co-training for semi-supervised learning.

# Multimodal semi-supervised learning for image classification

Matthieu Guillaumin, Jakob Verbeek, Cordelia Schmid

LEAR team, INRIA Grenoble, France

