Recognizing human actions in still images: a study of bag-of-features and part-based representations

Vincent Delaitre, Ivan Laptev and Josef Sivic

July 6, 2010

Introducing a new dataset

We collected a new challenging dataset for real-life human actions. It is composed of 968 images collected from Flickr representing natural variations in terms of camera view-point, human pose, clothing, occlusions and scene background.
Pictures are distributed among 7 different classes:

- Interacting with a computer
- Taking a photograph
- Playing music
- Riding bike
- Riding horse
- Running
- Walking

Interacting with a computer

Photographing

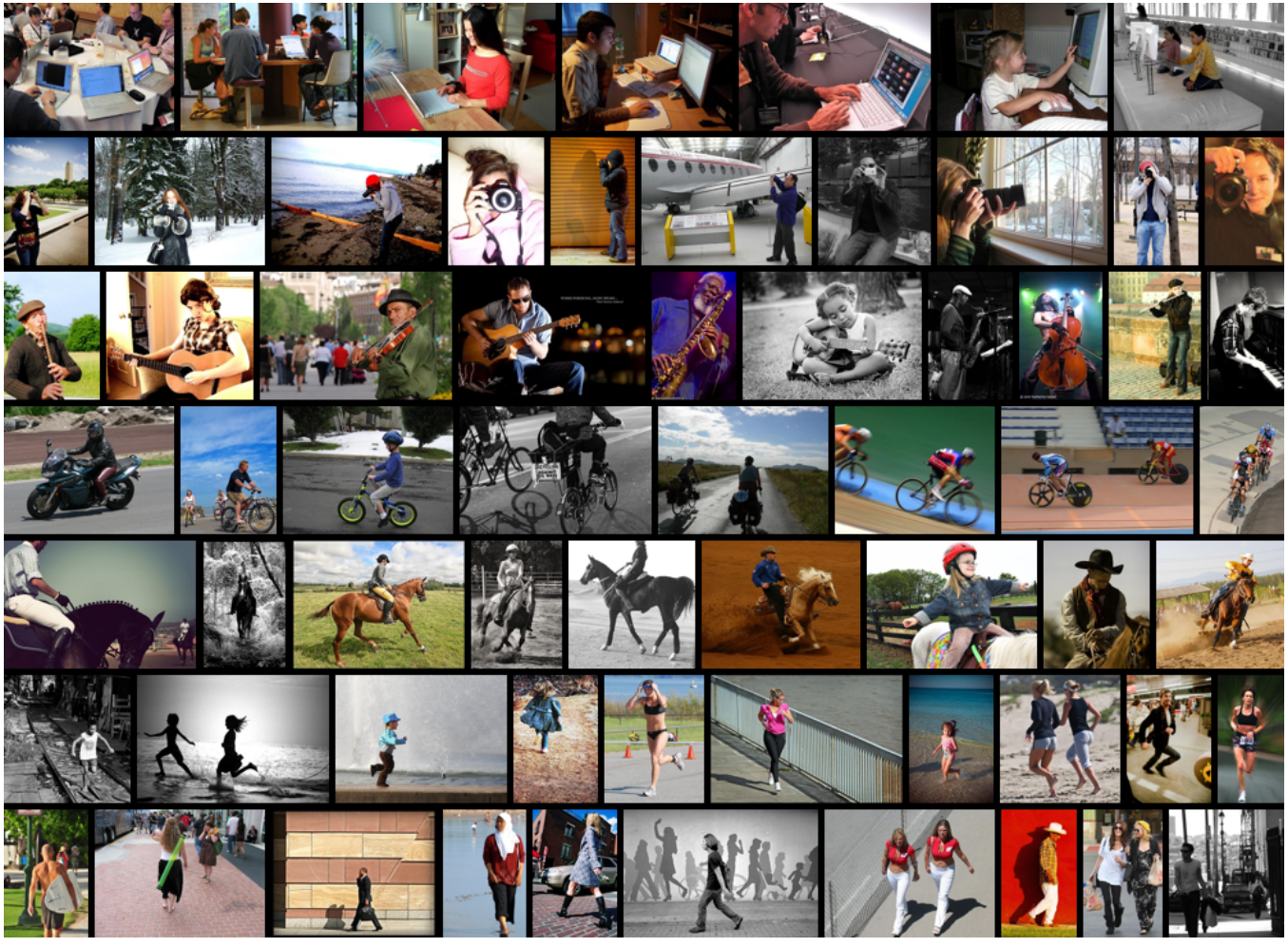
Playing music

Riding bike

Riding horse

Running

Walking



Classification task

Each person is annoted with a bounding box (smallest rectangle containing its visible pixels) and the action being executed.

In the following, we are interessted in the 7-class classification problem. The training set consists in 70 images of each type of action, so that at least 48 images per class remain for test.

We mesure the performances using:

i *the classification accuracy*: average of the diagonal of the confusion table

ii *the mean average precision (mAP)*: mean area under the precision-recall curve of each 1-vs-all classifiers.

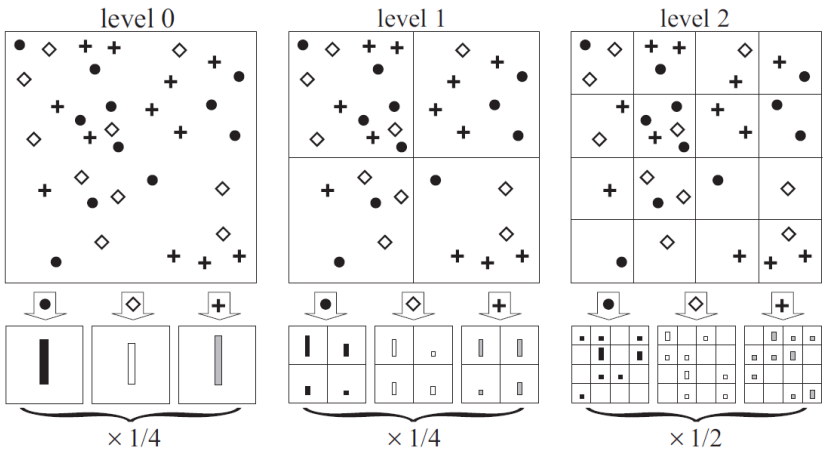
Bag-of-features classifier

Here we investigate the influence of various type of parameters in the classifier performances:

- **Image representation**: Images are represented using the spatial pyramid representation from Lazebnik *et al*. .
- **SVM Classification**: We use 1-vs-all classification scheme. We investigate the efficiency of different kernels.
- **Using context information**: We analyse the impact of the context using information provided by the bounding box.

Bag-of-features classifier: Image representation

- Features are extracted from multi-scale dense sampled SIFT descriptors.
- Visual vocabulary is built from k-means clustering. Size of the dictionnary $K \in \{256, 512, 1024, 2048, 4096\}$.
- Following Lazebnik *et al.* , we use a 2 levels spatial pyramid: image is divided into 1×1, 2×2 and 4×4 grids of cells leading to a $(1 + 4 + 16)K = 21K$ dimensional reprensation of an image.



Bag-of-features classifier: SVM Classification

Classification is performed with the SVM classifier using the 1-vs-all scheme, which, in our experiments, resulted in a small but consistent improvement over the 1-vs-1 scheme.
We investigate four different kernels:

1. the histogram intersection kernel, given by $\sum_i \min(x_i, y_i)$;
2. the χ^2 kernel, given by $\exp\{\frac{1}{\gamma}\sum_i \frac{(x_i-y_i)^2}{x_i+y_i}\}$;
3. the Radial basis function (RBF) kernel, given by $\exp\{\frac{1}{\beta}\sum_i (x_i - y_i)^2\}$; and
4. the linear kernel given by $\sum_i x_i y_i$,

where $\vec{x}$ and $\vec{y}$ denote visual word histograms of images X and Y, and γ and β are kernel parameters.

Bag-of-features classifier: Using context information

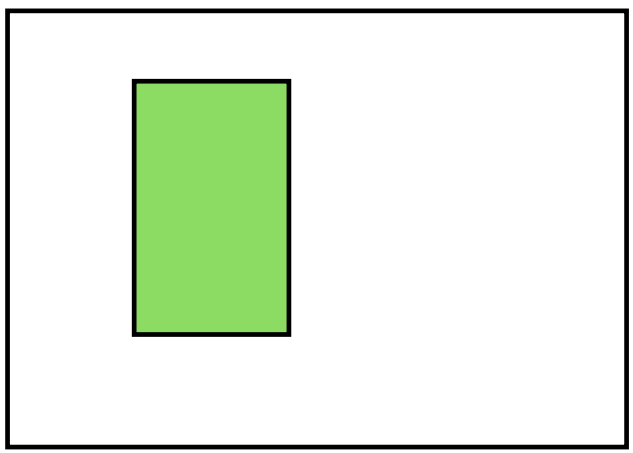
We consider the following four approaches:

A. **"Person"** Images cropped to 1.5× the size of the bounding box. Resize so that the larger dimension is 300 pixels.

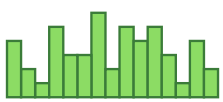
B. **"Image"**

C1. **"Person+Background"**

C2. **"Person+Image"**



Histogram:



Bag-of-features classifier: Using context information

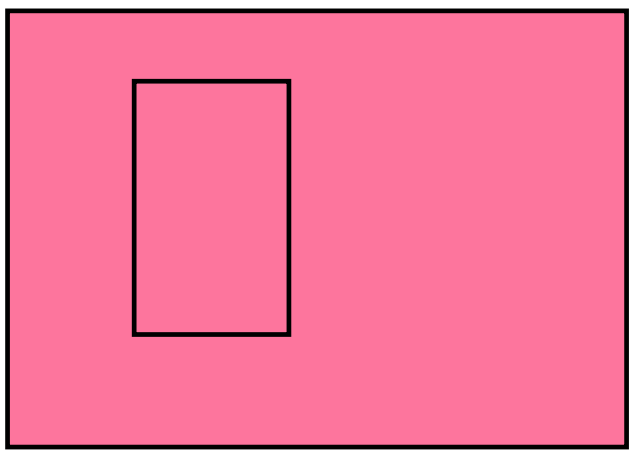
We consider the following four approaches:

A. **"Person"**

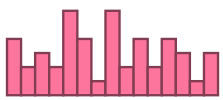
B. **"Image"** Histograms are computed on the full image. Image resized so that the larger dimension is 500 pixels.

C1. **"Person+Background"**

C2. **"Person+Image"**



Histogram:



Bag-of-features classifier: Using context information

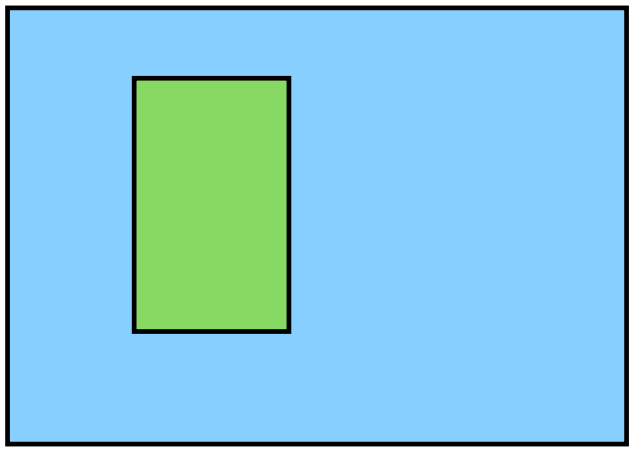
We consider the following four approaches:

A. **"Person"**

B. **"Image"**

C1. **"Person+Background"** Background is represented only with a BOF histogram. Kernel values between two images $\vec{x}$ and $\vec{y}$ is the sum of two kernels over the foreground and the background: $K(\vec{x}, \vec{y}) = K(\vec{x}_f, \vec{y}_f) + K(\vec{x}_b, \vec{y}_b)$

C2. **"Person+Image"**



Histogram:

Bag-of-features classifier: Using context information

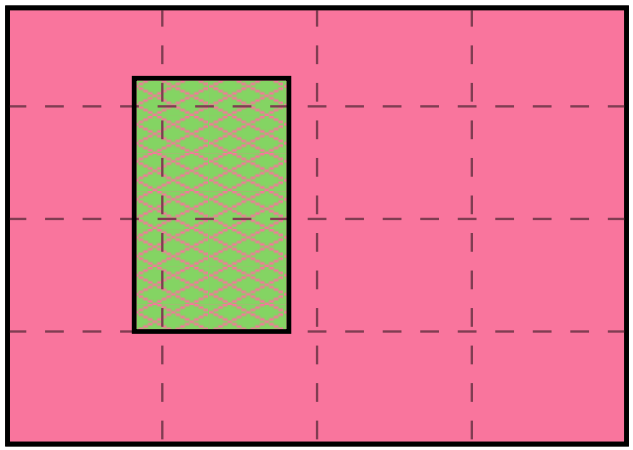
We consider the following four approaches:

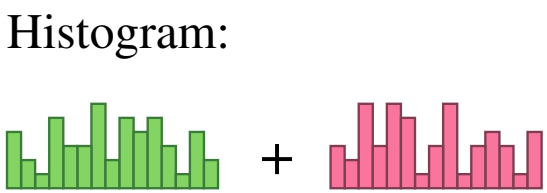
A. **"Person"**

B. **"Image"**

C1. **"Person+Background"**

C2. **"Person+Image"** This setup is similar to C1, however, instead of the background region, 2-level spatial pyramid representation of the entire image is used.





Bag-of-features classifier: Results

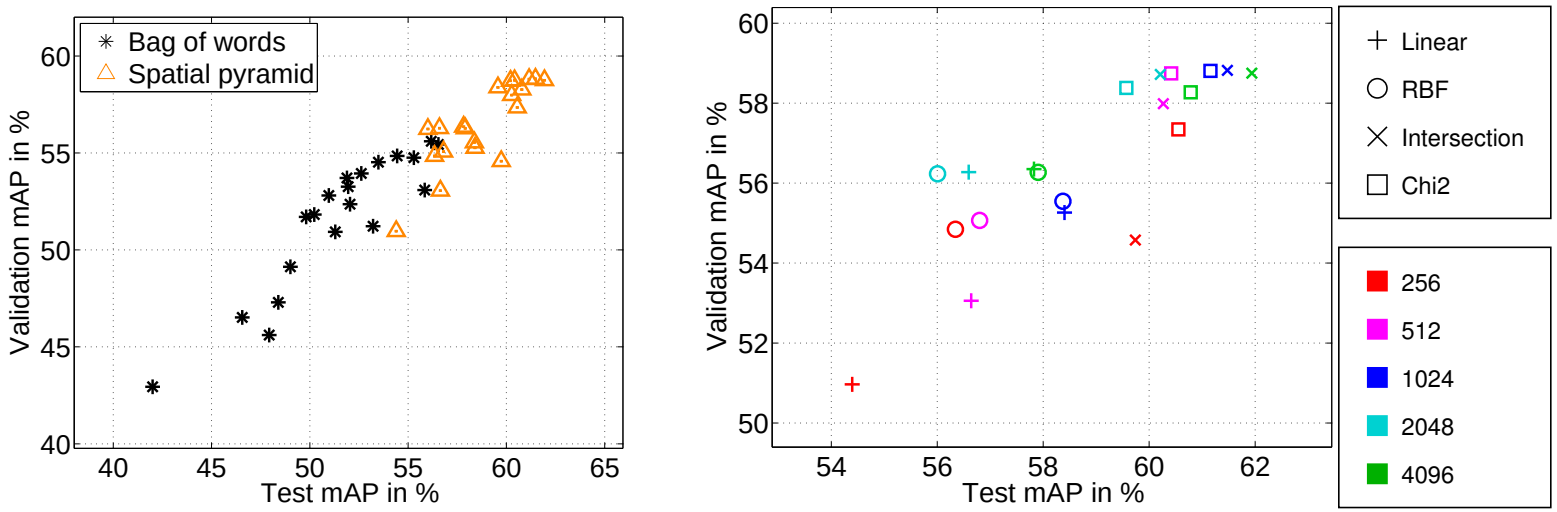


Figure: Classification performance (cross-validation mAP vs. test mAP) for different parameter settings for the method "A. Person". (a) 2-levels spatial pyramid vs. the bag-of-feature representation. (b) Classification performance for different combinations of kernels and vocabulary sizes using the 2-levels spatial pyramid representation. The standard deviation of the validation mAP is typically 2-3%.

Bag-of-features classifier: Results

In the following we use the best parameters for validation time of method "A. Person": visual dictionnary size K = 1024 and intersection kernel. We analyse the influence of context:

Method	mAP	Accuracy
A. Person	61.48	59.08
B. Image	62.83	60.24
C1. Person+Background	63.96	62.65
C2. BOF Person+Image	70.43	67.01

Table: The overall classification performance for the different methods.

The latent SVM

Combining part-based model and bag-of-features

Results

Action / Method	BOF C2	LSVM	LSVM+ BOF C2
(1) Inter. w/ Comp.	**84.21**	42.11	**84.21**
(2) Photographing	**35.53**	21.05	30.26
(3) Playing Music	62.39	**80.34**	70.94
(4) Riding Bike	80.85	63.83	**84.40**
(5) Riding Horse	**71.43**	67.86	**71.43**
(6) Running	55.00	51.25	**61.25**
(7) Walking	**79.66**	72.88	78.81
Average (mAP)	67.01	57.05	**68.76**

Table: Per-class accuracy across different methods.

Results

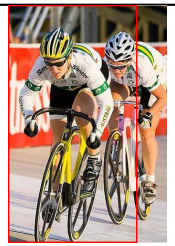
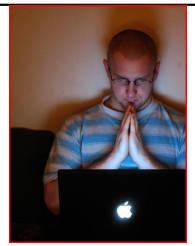
					
LSVM+C2:	PlayingMusic	Walking	RidingBike	Photograph.	PlayingMusic
C2:	Photograping	RidingHorse	???	???	???
LSVM:	???	???	PlayingMusic	RidingHorse	???

Figure: Example of images

The sports dataset

This is a sports dataset proposed by Gupta *et al.*with six classes:

The sports dataset

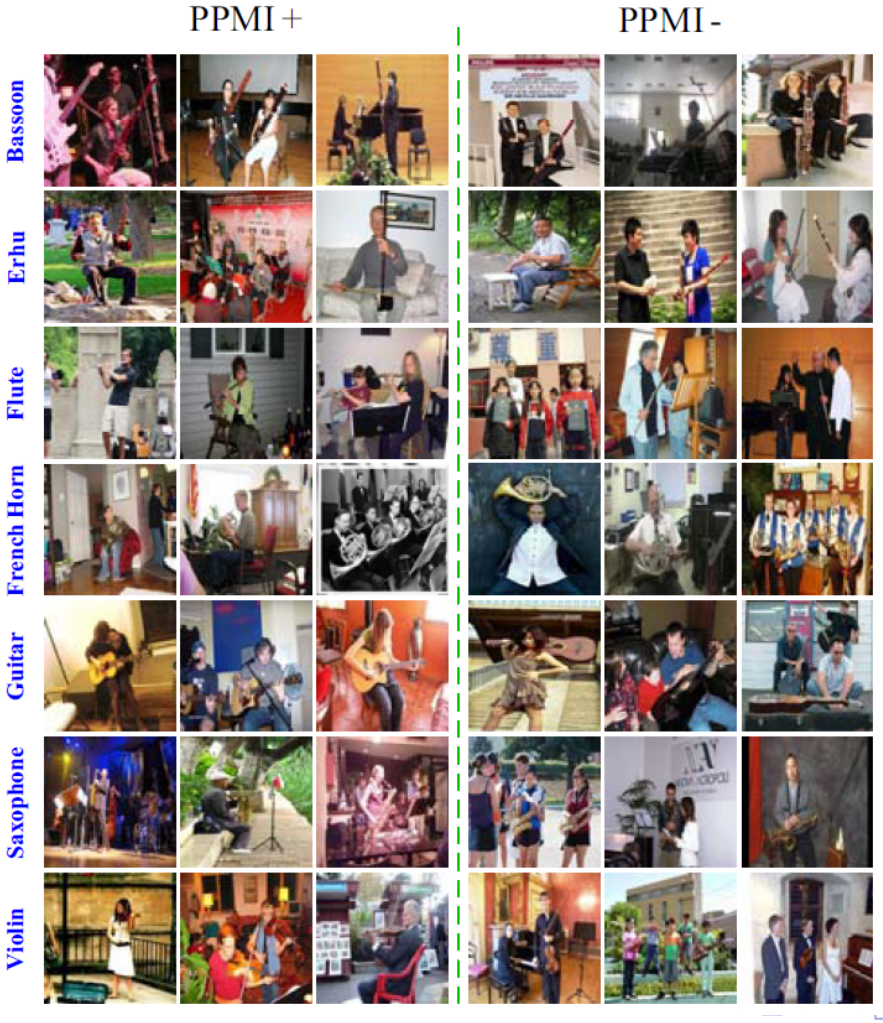
We kept the same parameters as the results reported previously:

Method	mAP	Accuracy
Gupta *et al.* [**?**]	–	78.67
BOF Image (B)	91.30	**85.00**
LSVM	77.19	73.33
LSVM + BOF Image (B)	**91.55**	**85.00**

Table: Comparison with the method of Gupta *et al.*on their dataset.

The People-playing-musical-instrument (PPMI) dataset

This is a dataset proposed by Fei-Fei *et al.*with people playing or only holding an instrument. There are seven different instruments.



The People-playing-musical-instrument (PPMI) dataset

Results...

Conclusion