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

Vincent Delaitre, Ivan Laptev and Josef Sivic

July 5, 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

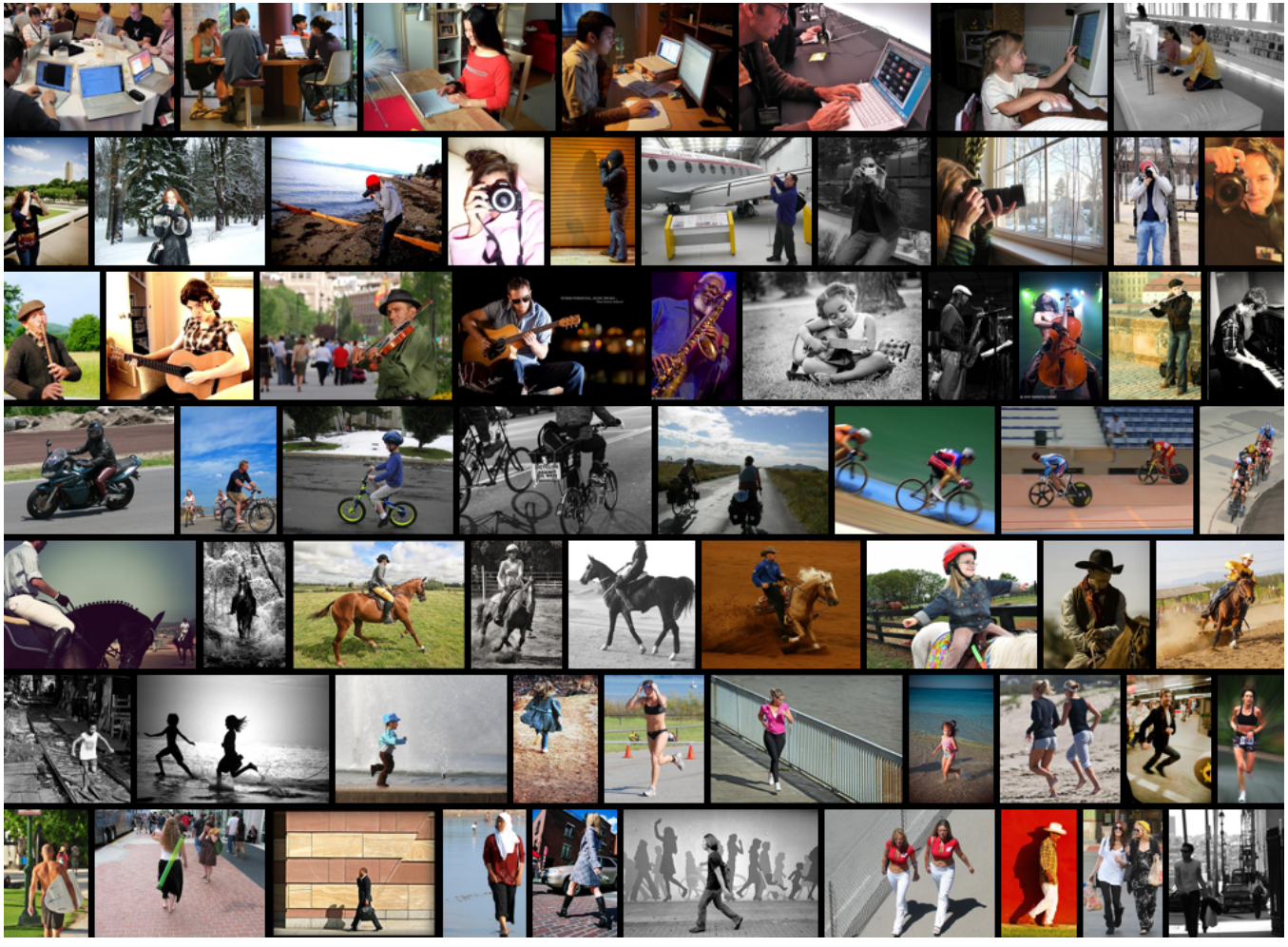
Playing music

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.

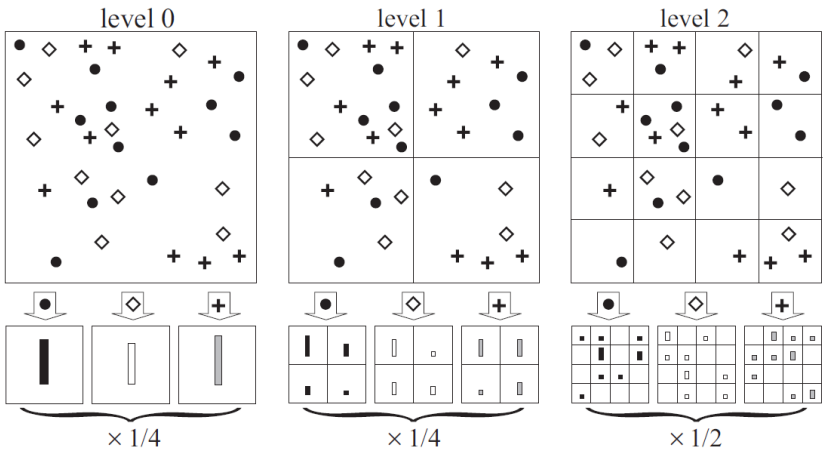
Bag-of-features classifier

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”**

B. **“Image”**

C1. **“Person+Background”**

C2. **“Person+Image”**