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

Vincent DELAITRE, Ivan LAPTEV and Josef SIVIC
INRIA - Willow Project, Ecole Normale Supérieure, Paris, France (CNRS/ENS/INRIA UMR 8548)

## The Goal: We want to recognize human actions in still images:

*Running* *Skiing* *Writing* *Taking a photograph* *Sitting* *Reading a book*

## Motivations:

Most of the work is on video but:

- Some actions are static and video may not help.
- Human actions are a natural description of many images.

## Contributions:

- Creation of a new dataset for human actions.
- Quantitative evaluation of the statistical BOF model and the deformable part model.
- Combination of those two models.
- Investigation of the role of context.

## Related work:

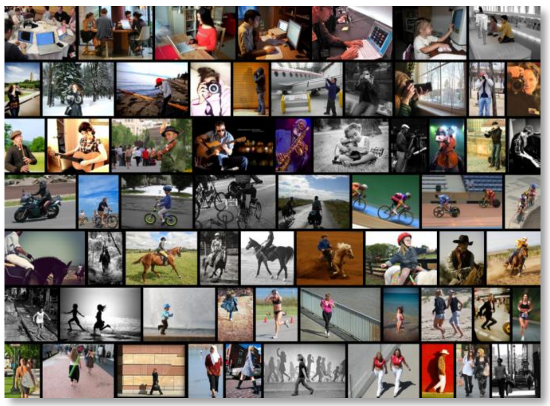
Related work on actions in still images has focused on specific domains such as sports or playing musical instruments. *(Gupta et al. PAMI09, Yao and Fei-Fei CVPR10).*

→ We want to study recognition of a more general set of human actions obtained from real images.

## A new database for action classification:

We collected 968 images from Flickr.
→ Large variations in terms of camera-viewpoints, human poses, clothing, occlusion, backgrounds, lighting, object appearance.

Interacting w. computer
Taking a photograph
Playing music
Riding bike
Riding horse
Running
Walking

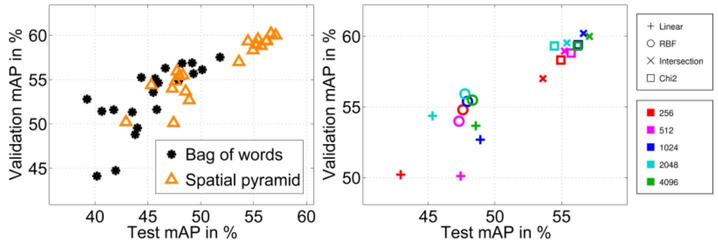


## The Spatial Pyramid Matching (SPM):

We use dense SIFT feature and investigated the role of different parameters:

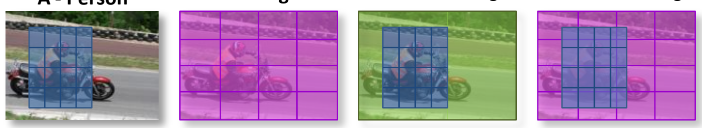
- **Visual vocabulary**: built from a k-means clustering with K centers. K ∈ {256, 512, 1024, 2048, 4096}.
- **Image signature**: spatial pyramid with L ∈ {0, 1, 2, 3} levels.
- **Different kernels**: Linear, RBF, Chi2, Intersection (the C and kernel parameters for the SVM are obtained by 5-fold cross-validation).

## Performance for different parameters:




## Using the context:

A - Person  B - Image  C1 - Person + Bckgnd  C2 - Person + Image



→ Taking the background into account improves performance:

| Context | mAP | Accuracy |
|---|---|---|
| A - Person | 56.7 | 55.9 |
| B - Image | 54.0 | 54.0 |
| C1 - Person + Background | 57.6 | 56.8 |
| C2 - Person + Image | <u>*59.6*</u> | <u>*58.9*</u> |

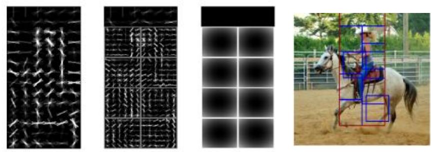
**mAP:** mean of the Average-Precision of each class.

**Accuracy:** mean of the diagonal of the confusion table.

## The latent – SVM (Felzenszwalb *et al.* PAMI 2008):

- State-of-the-art for object and person detection
- We used the default parameters (3 components, 8 parts)

Example of model for 'Riding Horse': filters and part placement.



## Performance of SPM and LSVM:

| Action / Method | LSVM | SPM (C2) | SPM (C2) + LSVM |
|---|---|---|---|
| Interacting w. computer | 30.2 | 58.2 | <u>*58.5*</u> |
| Photographing | 28.1 | 35.4 | <u>*37.4*</u> |
| Playing music | 56.3 | <u>*73.2*</u> | 73.1 |
| Riding bike | 68.7 | 82.4 | <u>*83.3*</u> |
| Riding horse | 60.1 | 69.6 | <u>*77.0*</u> |
| Running | 52.0 | 44.5 | <u>*53.3*</u> |
| Walking | 56.0 | 54.2 | <u>*57.5*</u> |
| Mean average precison | 50.2 | 59.6 | <u>*62.9*</u> |

- LSVM benefits from the strength of pictorial models.
- SPM takes advantages of the statistical representation of images.

→ Complementary models.

The combination 'SPM + LSVM' is achieved by adding the classification scores of both models.

## Combining both models:

| | | | | |
|---|---|---|---|---|
| SPM (C2) + LSVM | Photographing | Riding horse | Running | Photographing |
| SPM (C2) | Riding Bike | Riding horse | Int. w. comp. | Photographing |
| LSVM | Photographing | Riding bike | Walking | Photographing |

- SPM+LSVM often improves SPM on images with confusing (blurred, textureless, unusual) background but where the person is clearly visible.
- SPM+LSVM appears to improve the LSVM output where camera viewpoint or the pose of the person are unusual.

## Comparison with the state of the art:

- **The sports dataset [Gupta et al. PAMI09]**
  - Person-centered
  - Little background

| Method | mAP | Acc. |
|---|---|---|
| Gupta et al. | --- | 78.7 |
| Yao and Fei-Fei | --- | 83.3 |
| SPM (B) | 91.3 | <u>*85.0*</u> |
| LSVM | 77.2 | 73.3 |
| SPM (B) + LSVM | <u>*91.6*</u> | <u>*85.0*</u> |


- **The PPMI (Person Playing Musical Instrument) dataset [Yao and Fei-Fei CVPR10]**


**Task 1:** 7-class person playing musical instrument classification problem.

**Task 2:** Playing vs. non-playing musical instrument. 'Non-playing' means simply holding it.

| Method | Task 1 mAP | Task 1 Acc. | Task 2 mAP | Task 2 Acc. |
|---|---|---|---|---|
| Yao and Fei-Fei | --- | 80.9 | --- | 65.7 |
| SPM (B) | 87.7 | 83.7 | 76.9 | 71.7 |
| LSVM (9 parts) | 82.2 | 82.9 | 53.6 | 67.6 |
| SPM (B) + LSVM | <u>*90.5*</u> | <u>*84.9*</u> | <u>*77.8*</u> | <u>*75.1*</u> |