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

Vincent Delaitre¹ vincent.delaitre@ens-lyon.org Ivan Laptev² ivan.laptev@inria.fr Josef Sivic² josef.sivic@ens.fr

¹ École Normale Supérieure de Lyon

² INRIA - Willow Project Laboratoire d'Informatique École Normale Supérieure CNRS/ENS/INRIA (UMR 8548)

Human actions represent essential content of many images. Recognizing human actions in still images will potentially provide useful metadata to many applications such as indexing and search of large-scale image archives. Given the frequent interaction of people with objects (e.g. "answer phone") and scenes (e.g. "walking around the corner"), human action recognition is also expected to help solving other related problems for still images such as object recognition or scene layout estimation.

Recognition of human actions has mostly been explored in video. While the motion of people often provides discriminative cues for action classification, many actions such as the ones illustrated in Figure 1 can be identified from single images. Moreover, several types of actions such as "taking a photograph" and "reading a book" are of static nature and may require recognition methods based on static cues only even if the video is available.

The goal of this work is to study recognition of common human actions represented in typical still images such as consumer photographs. This problem has received little attention in the past with the exception of a few related papers focused on actions in specific domains such as sports [2, 5] and people playing musical instruments [4]. The proposed methods have mainly relied on the body pose as a cue for action recognition. While promising results have been demonstrated, typical action images such as the ones illustrated in Figure 1 still present serious challenges for current body-pose estimation methods due to heavy occlusions and significant changes in camera viewpoint.

To deal with various types of actions in still images, we avoid explicit reasoning about body poses and investigate more general appearance-based classification methods. We study action recognition in typical consumer photographs and construct a new dataset with seven classes of actions in 968 images obtained from Flickr (Figure 1).

We investigate the performance of statistical bag-of-features (BOF) and spatial pyramid representations [3] combined with SVM classification. We examine a large set of parameters on the validation data and demonstrate consistent generalization of results to the test data. In particular, we investigate person-centric representations and study the influence of background/context information on action recognition. In addition to statistical methods, we also consider the structural part-based LSVM model of Felzenszwalb *et al.* [1]. Figure 2 illustrates results for both models individually as well for their combination. Figure 3 illustrates the complementarity of BOF and LSVM models on a few sample images from our dataset.

Based on the comparative evaluation on the datasets of [2] and [4] (c.f. Table 1), we finally demonstrate that the previous methods relying on explicit body-pose estimation can be significantly outperformed by more generic recognition methods investigated in this paper.

- [1] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 2009.
- [2] A. Gupta, A. Kembhavi, and L.S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE PAMI*, 31(10):1775–1789, October 2009.
- [3] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages II: 2169–2178, 2006.
- [4] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010.
- [5] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.



Figure 1: Example images from our dataset with seven action classes collected from Flickr. Note the natural and challenging variations in the camera view-point, clothing of people, occlusions, object appearance and scene layout present in consumer photographs.

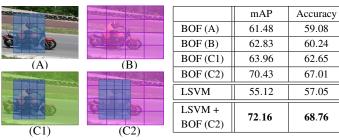


Figure 2: Different ways of integrating the background into the classifier: (A) Spatial pyramid matching (SPM) on person only - limited use of background, (B) SPM on the full image, (C1) 2 channels: SPM on person + bag-of-features (BOF) over the rest of the image, (C2) 2 channels: SPM on person + SPM over the full image. Table on the right shows the performance of (A), (B), (C1) and (D) as well as the performance of LSVM and its combination with C2 (combination is obtained by adding classification scores).

LSVM+BOF(d): Running	RidingBike	Photographing	PlayingMusic
BOF (d): RidingHorse	RidingBike	Inter. w/ comp.	PlayingMusic
LSVM: Running	RidingHorse	PlayingMusic	PlayingMusic

Figure 3: Examples of challenging images and corresponding classifications. LSVM typically outperforms BOF on images with confusing (blurred, textureless or unusual) background, but with clearly visible pose of people. Similarly, the combined method improves LSVM results mainly in cases where the camera viewpoint or the pose of the person are unusual.

Dataset	Gupta et al.		Yao and Fei-Fei [4]			
	[2]		Task 1		Task 2	
Method	mAP	Acc.	mAP	Acc.	mAP	Acc.
Gupta et al.	_	78.7	_	_	_	_
Yao and Fei-Fei	_	83.3	_	65.7	_	80.9
BOF (b)	91.3	85.0	76.9	71.7	87.7	83.7
LSVM	77.2	73.3	53.6	67.6	82.2	82.9
LSVM + BOF (b)	91.6	85.0	77.8	75.1	90.5	84.9

Table 1: Comparison with the methods of Gupta *et al.* [2] and of Yao and Fei-Fei [4, 5] on their datasets. 'Task 1' is the 7-class classification problem and 'Task 2' is the PPMI+ vs PPMI- (person playing music) problem (see [4] for details).