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

Vincent Delaitre[1]
vincent.delaitre@ens-lyon.org
Ivan Laptev[2]
ivan.laptev@inria.fr
Josef Sivic[2]
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)

Corrected version of the paper with updated results in Section 5

Abstract

Recognition of human actions is usually addressed in the scope of video interpretation. Meanwhile, common human actions such as "reading a book", "playing a guitar" or "writing notes" also provide a natural description for many still images. In addition, some actions in video such as "taking a photograph" are static by their nature and may require recognition methods based on static cues only. Motivated by the potential impact of recognizing actions in still images and the little attention this problem has received in computer vision so far, we address recognition of human actions in consumer photographs. We construct a new dataset available at [2] with seven classes of actions in 911 Flickr images representing natural variations of human actions in terms of camera view-point, human pose, clothing, occlusions and scene background. We study action recognition in still images using the state-of-the-art bag-of-features methods as well as their combination with the part-based Latent SVM approach of Felzenszwalb *et al.* [8]. In particular, we investigate the role of background scene context and demonstrate that improved action recognition performance can be achieved by (i) combining the statistical and part-based representations, and (ii) integrating person-centric description with the background scene context. We show results on our newly collected dataset of seven common actions as well as demonstrate improved performance over existing methods on the datasets of Gupta *et al.* [10] and Yao and Fei-Fei [22].

1 Introduction

Human actions represent essential content of many images. Recognizing human actions in still images will potentially provide useful meta-data to many applications such as indexing and search of large-scale image archives. Given the frequent interactions 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, see for example [3, 14, 17]. While the motion of people often provides discriminative cues for action classification,

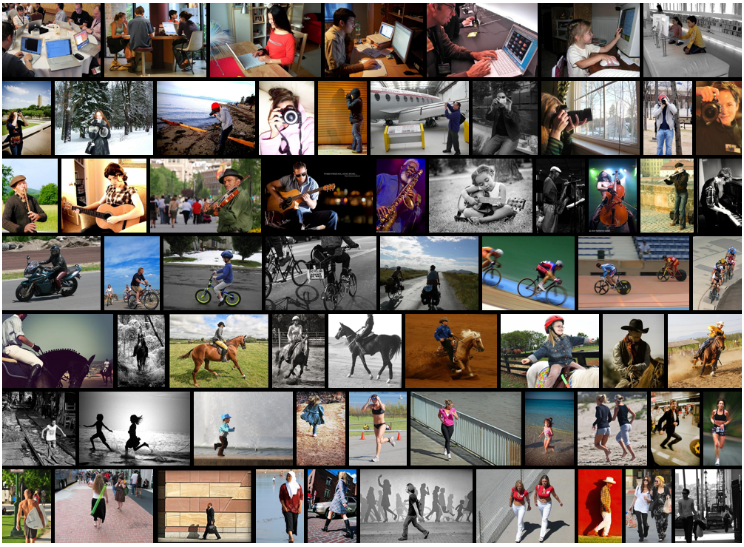


Figure 1: Example images from our newly collected dataset of seven action classes. Note the natural and challenging variations in the camera view-point, clothing of people, occlusions, object appearance and the scene layout present in the consumer photographs.

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 few related papers focused on specific domains, such as sports actions [10, 12, 16, 21] or, more recently, people playing musical instruments [22]. Learning from still images to recognize actions in video was investigated in [13].

The proposed methods [10, 12, 21] have mainly relied on the body pose as a cue for action recognition. While promising results have been demonstrated on sports actions [10, 12, 21], typical action images such as the ones illustrated in Figure 1 often contain heavy occlusions and significant changes in camera viewpoint and hence present a serious challenge for current body-pose estimation methods. At the same time, the presence of particular objects [10] and scene types [16] often characterizes the action and can be used for action recognition.

To deal with various types of actions in still images, we avoid explicit reasoning about body poses and investigate more general classification methods. We study action recognition in typical consumer photographs and construct a new dataset [1] with seven classes of actions in 911 images obtained from the Flickr photo-sharing web-site. Image samples in Figure 1 illustrate the natural and challenging variations of actions in our dataset with respect to the camera view-point, clothing of people, occlusions, object appearance and the scene layout.

We study performance of statistical bag-of-features representations combined with SVM classification [25]. In particular, we investigate person-centric representations and study the influence of background/context information on action recognition. We investigate a large set of parameters on the validation set and show a consistent generalization of results to the test set. In addition to statistical methods, we investigate the structural part-based

LSVM model of Felzenszwalb *et al.* [8] and demonstrate improved performance with the combination of both models. Based on the comparative evaluation on the datasets of [10] and [22] we demonstrate that previous methods relying on explicit body-pose estimation can be significantly outperformed by more generic recognition methods investigated in this paper.

The rest of the paper is organized as follows. In Section 2 we describe our new dataset for action recognition in still images and detail performance measures used in our evaluation. Sections 3 and 4 present the two recognition methods investigated in this paper and their combination. Section 5 provides extensive experimental evaluation of different methods and parameter settings on the three still-image action datasets.

2 Datasets and performance measures

We consider three datasets in this work: the datasets of Gupta *et al.* [10] and Yao and Fei-Fei [22], focused on sports and people playing musical instruments, respectively, as well as our newly collected dataset of actions in consumer photographs available at [2]. To avoid the focus on a specific domain and also investigate the effect of background (images in [10, 22] are cropped to eliminate background) we collect a new dataset of full (non-cropped) consumer photographs depicting seven common human actions: “Interacting with computers", “Photographing", “Playing a musical instrument", “Riding bike", “Riding horse", “Running" and “Walking". Images for the "Riding bike" action were taken from the Pascal 2007 VOC Challenge and the remaining images were collected from Flickr by querying on keywords such as “running people" or “playing piano". Images clearly not depicting the action of interest were manually removed. This way we have collected a total of 911 photos. Each image was manually annotated with bounding boxes indicating the locations of people. For these annotations we followed the Pascal VOC guidelines. In particular, we labeled each person with a bounding box which is the smallest rectangle containing its visible pixels. The bounding boxes are labelled as 'Truncated' if more than 15%-20% of the person is occluded or lies outside the bounding box. We also added a field “action" to each bounding box to list all actions being executed.We collected at least 109 persons for each class, split into 70 persons per class for training and the remaining ones for test. Example images for each of the seven classes are shown in figure 1.

Performance measures: We use two performance measures throughout the paper: (i) the *classification accuracy* and (ii) the *mean average precision (mAP)*. The classification accuracy is obtained as the average of the diagonal of the confusion table between different classes, and is a typical performance measure for multi-way classification tasks. To obtain mAP we first compute the area under the precision-recall curve (average precision) for each of the seven binary 1-vs-all action classifiers. mAP is then obtained as the mean of average precisions across the seven actions.

3 Bag-of-features classifier

Here we describe the spatial pyramid bag-of-features representation [15] with the Support Vector Machine (SVM) classifier [18] and the implementation choices investigated in this work. In particular we detail the image representation, the different kernels of the SVM classifier, and different methods for incorporating the person bounding box and the scene background information into the classifier.

Image representation: Images (or image regions given by a rectangular bounding box) are represented using SIFT descriptors sampled on 10 regular grids with increasing scales with spacing $s_i = \lfloor 12 \cdot 1.2^i \rfloor$ pixels for $i = 0, \cdots, 9$. The scale of features extracted from each grid is set to $w_i = 0.2 \cdot s_i$. Visual vocabularies are built from training descriptors using k-means clustering. We consider vocabularies of sizes $K \in \{256, 512, 1024, 2048, 4096\}$ visual words. Descriptors from both training and test sets are then assigned to one of the visual words and aggregated into a K-dimensional histogram, denoted further as the bag-of-features representation. Following the spatial pyramid representation of Lazebnik *et al.* [15] we further divide the image into 1×1 (Level 0), 2×2 (Level 1) and 4×4 (Level 2) spatial grids of cells. Local histograms within each cell are then concatenated with weights 0.25, 0.25 and 0.5 for levels 0, 1, and 2, respectively. This results in a $(1+4+16)K = 21K$ dimensional representation, where K is the vocabulary size. The weights of the different histogram levels are kept fixed throughout the experiments, but could be potentially learnt as shown in [4]. This representation captures a coarse spatial layout of the image (or an image region) and has been shown beneficial for scene classification in still images [15] and action classification in videos [14].

Support vector machine 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 $\sum_i x_i y_i$.

$\mathbf{x}$ and $\mathbf{y}$ denote visual word histograms, and γ and β are kernel parameters. For the χ^2 and intersection kernels, histograms are normalized to have unit L1 norm. For the RBF and linear kernels, histograms are normalized to have unit L2 norm [20]. Parameters γ and β of the χ^2 and RBF kernels, respectively, together with the regularization parameter of the SVM are set for each experiment by a 5-fold cross validation on the training set.

Incorporating the person bounding box into the classifier: Previous work on object classification [25] demonstrated that background is often correlated with objects in the image (e.g. cars often appear on streets) and can provide useful signal for the classifier. The goal here is to investigate different ways of incorporating the background information into the classifier for actions in still images. We consider the following four approaches:

A. **"Person":** Images are centred on the person performing the action, cropped to contain $1.5\times$ the size of the bounding box and re-sized such that the larger dimension is 300 pixels. This setup is similar to that of Gupta *et al.* [10], i.e. the person occupies the majority of the image and the background is largely suppressed.

B. **"Image":** The original images are resized to have the larger dimension at most 500 pixels. No cropping is performed. The person bounding box is not used in any stage of training or testing apart from evaluating the performance. Here the visual word histograms represent a mix of the action and the background.

C1. **"Person+Background'**: The original images are resized so that the maximum dimension of the 1.5× rescaled person bounding box is 300 pixels, but no cropping is performed. The 1.5× rescaled person bounding box is then used in both training and test to localize the person in the image and provides a coarse segmentation of the image into foreground (inside the rescaled person bounding box) and background (the rest of the image). The foreground and background regions are treated separately. The final kernel value between two images X and Y represented using foreground histograms $\mathbf{x}_f$ and $\mathbf{y}_f$, and background histograms $\mathbf{x}_b$ and $\mathbf{y}_b$, respectively, is given as the sum of the two kernels, $K(\mathbf{x},\mathbf{y}) = K_f(\mathbf{x}_f,\mathbf{y}_f) + K_b(\mathbf{x}_b,\mathbf{y}_b)$. The foreground region is represented using a 2-level spatial pyramid whereas the background is represented using a BOF histogram with no spatial binning.

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.

Note that approaches A, C1 and C2 use the manually provided person bounding boxes at both the training and test time to localize the person performing the action. This simulates the case of a perfectly working person detector [5, 8].

4 Discriminatively trained part-based model

We also investigate the performance of the discriminatively trained part-based model of Felzenszwalb *et al.* [8] (LSVM), which, in contrast to the bag-of-features approach, provides a deformable part-based representation of each action. The approach combines the strengths of efficient pictorial structure models [7, 9] with recent advances in discriminative learning of SVMs with latent variables [8, 24]. The approach has shown excellent human and object detection performance in the PASCAL visual recognition challenge [8]. In in this work we apply the model for classification (rather than detection with spatial localization) and focus on recognition of human actions rather than objects. Actions are modeled as multi-scale HOG templates with flexible parts. Similarly to the spatial pyramid bag-of-features representation described in section 3, we train one model for each action class in a 1-vs-all fashion. Positive training data is given by the 1.5× rescaled person bounding boxes for the particular action and negative training data is formed from all images of the other action classes. At test time, we take the detection with the maximum score, which overlaps the manually specified person bounding box in the test image more than 50%. The overlap is measured using the standard ratio of areas of the intersection over the union. The 50% overlap allows for some amount of scale variation between the model and the manual person bounding box. In cases when the person bounding box is not available the detection with the maximum score over the entire image is taken. We use the recently released version 4 of the training and detection code available at [1], which supports models with multiple mixture components for each part allowing for a wider range of appearances of each action. We train models with 8 parts and 3 mixture components.

Combining the part-based model with the bag-of-features classifier: The part-based model (LSVM) represents mostly the person and its immediate surroundings and largely ignores the background information. Hence, we also investigate combining the model with bag-of-feature classifiers described in section 3. We demonstrate in section 5 that such combination can significantly improve the classification performance of the LSVM approach. The two approaches are combined by simply adding together their classification scores with equal weighting. However, the weights could be potentially learnt. In a similar fashion,

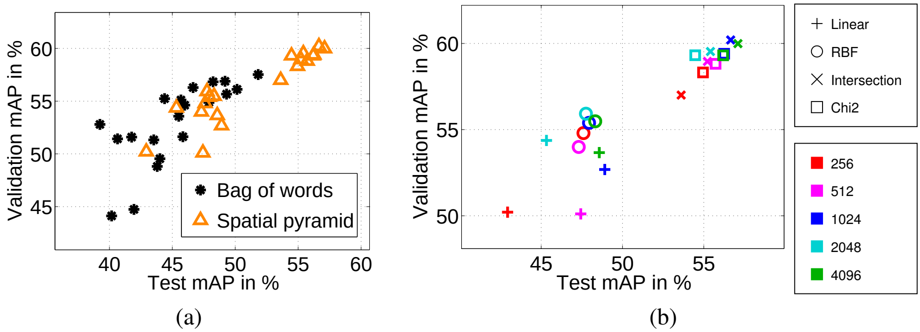


Figure 2: Classification performance (cross-validation mAP vs. test mAP) for different parameter settings for the BOF method "A. Person". The best results are at the top right portion of the graph. (a) Spatial pyramid vs. the bag-of-feature representation. (b) Classification performance for different combinations of kernels and vocabulary sizes using the spatial pyramid representation. Best viewed in color. The standard deviation (not shown in the plots) of the validation mAP is typically 2-3%.

combining scene-level classifiers with object detectors was shown to improve object detection results in the PASCAL 2009 object detection challenge [1].

5 Results

We first evaluate different parameter settings for the bag-of-features classifier. Equipped with a well tuned classifier we examine different ways of incorporating the foreground (person) and background (scene context) information. Next, we compare and combine the bag-of-features classifier with the structured part-based model. Finally, we show results on the datasets of Gupta *et al.* [10] and Yao and Fei-Fei [22].

Setting parameters for the bag-of-features method: We first evaluate in detail different parameter settings (kernel type, vocabulary size, spatial representation) for bag-of-features method A, where images are cropped to contain mostly the person performing the action and the background is suppressed. We have found that the pattern of results across different parameter settings for methods B and C is similar to A and hence their detailed discussion is omitted from the paper.

Figure 2 shows plots of the classification performance obtained from the 5-fold cross-validation on the training set against the classification performance on the test set. First, we note that both cross-validation and test performance are well correlated, which suggests that the cross-validation results can be used to select the appropriate parameter setting. It is clear from figure 2(a) that spatial pyramid representation outperforms the vanilla bag-of-features model with no spatial binning. Examining figure 2(b), all kernels show similar trend of improvement towards larger vocabulary sizes. However the χ^2 and intersection kernels convincingly outperform the linear and RBF kernels. The best results (in terms of the lowest cross-validation error) are obtained for the spatial pyramid representation, intersection kernel, and vocabulary size 1,024 and we use this parameter setting for the rest of the paper.

How to model background context? Here we examine the different approaches for incorporating the background information into the bag-of-features action classifier (methods A-C). The overall results are summarized using the classification accuracy and the mean average precision in table 1 (rows A-C2). Average precision across different action classes is shown in table 2 (columns A-C2).

	Method	mAP	Accuracy
A.	BOF Person	56.65	55.90
B.	BOF Image	54.00	53.97
C1.	BOF Person+Background	57.60	56.78
C2.	BOF Person+Image	59.64	58.86
	LSVM	50.21	55.07
	LSVM + C2	**62.88**	**62.15**

Table 1: The overall classification performance for the different methods.[1]

Action / Method	A	B	C1	C2	LSVM	LSVM+C2
(1) Inter. w/ Comp.	51.59	51.61	57.34	58.15	30.21	**58.50**
(2) Photographing	31.78	30.65	24.65	35.39	28.12	**37.41**
(3) Playing Music	63.38	69.05	68.43	**73.19**	56.34	73.11
(4) Riding Bike	76.46	78.15	82.31	82.43	68.70	**83.26**
(5) Riding Horse	66.18	56.33	67.43	69.60	60.12	**77.03**
(6) Running	51.34	39.00	45.06	44.53	51.99	**53.34**
(7) Walking	55.79	53.24	**57.95**	54.18	55.97	57.51
mAP	56.65	54.00	57.60	59.64	50.21	**62.88**

Table 2: Per-class average precision across different methods.[1]

Focusing on the person by cropping the image and removing the background (method A) results in a slightly better overall performance than method B where we use the entire image, including the background, with no knowledge about the location of the person. However, for some actions ("Playing Music" and "Riding Bike") using the background (method B) is beneficial and reduces their confusion with other classes.

The overall performance can be further improved by treating and matching the foreground and background separately using two separate kernels (methods C1 and C2). This holds for all classes except "Running" and "Walking" where using background (method C2) slightly increases the confusion with the other action classes comparing to method A (and specially with "Riding Bike" and "Riding Horse" which are also outdoor scenes). In addition, representing the background with a spatial pyramid (C2) performs better overall than the vanilla BOF histogram (C1) with no spatial information. The overall benefit of treating foreground and background regions separately is inline with the recent experimental evidence from object and image classification [19, 25].

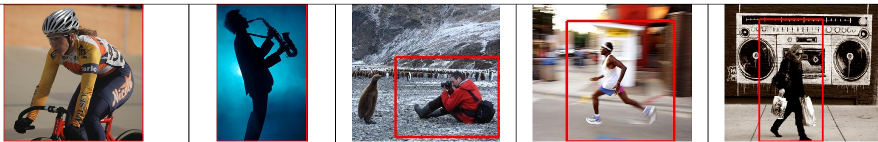
Part-based model vs. bag-of-features classifier: Here we compare the performance of the bag-of-features classification method (C2), the structured part-based model (LSVM) and their combination (LSVM+C2). The overall results are summarized using the classification accuracy and mean average precision in the last three rows of table 1. Average precision across different action classes is shown in the last three columns of table 2. The bag-of-features classifier (C2) and structured part-based model (LSVM) have comparable accuracy but the average precision of (C2) is better for all classes but "Running" and "Walking". It might be due to our choice to limit the part-based model to 3 mixture components: this could be insufficient for classes as "Interacting with computer", "Photographing" or "Playing music" where there is a large number of viewpoints (including close-ups) and music instruments. Overall, the combined (LSVM+C2) approach performs best and significantly

[1] Results have been updated from [6] to be consistent with the corrected version of the dataset available at [1].

Action	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Inter. w/ Comp.	**82.05**	0.00	15.39	2.56	0.00	0.00	0.00
(2) Photographing	12.99	**29.87**	15.58	3.90	6.49	12.99	18.18
(3) Playing Music	12.71	15.26	**66.10**	2.54	2.54	0.00	0.85
(4) Riding Bike	0.00	2.88	5.04	**74.10**	7.91	2.16	7.91
(5) Riding Horse	0.00	0.00	7.02	12.28	**73.69**	1.75	5.26
(6) Running	2.47	6.17	7.41	8.64	0.00	**51.85**	23.46
(7) Walking	4.92	8.20	0.00	0.00	11.47	18.03	**57.38**

Table 3: Confusion table for the best performing method (LSVM+C2). Accuracy (average of the diagonal): 62.15%.[1]

improves over (C2) on classes like "Running" and "Walking" where (LSVM) was better, but also interestingly on "Riding horse" which suggests that the bag-of-features classifier and the part-based model use complementary information for this class. These variations across classes are likely due to the varying levels of consistency of the human pose (captured well by structured part-based models), and the overall scene (captured well by the bag-of-features classifier). The full confusion table for the overall best performing method (LSVM+C2) is shown in table 3. While accuracy is over 65% on actions like "Interacting with computer", "Playing music", "Riding bike" or "Riding horse" other actions are more challenging, e.g. "Photographing" (accuracy 30%) is often confused with actions where people mostly stand as "Playing music", "Running", or "Walking". The last two classes have accuracy around 55% and are often confused with each other. Examples of images correctly classified by the combined LSVM+C2 method are shown in figures 3 and 4. Examples of challenging images misclassified by the LSVM+C2 method are shown in figure 5. We have found that the combined LSVM+C2 method often improves the output of the bag-of-features classifier (C2) on images with confusing (blurred, textureless or unusual) background, but where the pose of the person is very clear and the LSVM model provides a confident output. Similarly, the combined method appears to improve the vanilla LSVM results mainly in cases where camera viewpoint or the pose of the person are unusual.



LSVM+C2:	RidingBike	PlayingMusic	Photographing	Running	Walking
C2:	RidingHorse	Inter. w/ comp.	RidingBike	Inter. w/ comp.	Running

Figure 3: Example images correctly classified by the combined LSVM+C2 method (labels in the 2nd row), but misclassified by the C2 bag-of-features approach (labels in the 3rd row).

LSVM+C2:	Photographing	Inter. w/ comp.	RidingHorse	RidingBike	PlayingMusic
LSVM:	PlayingMusic	PlayingMusic	RidingBike	RidingHorse	Photographing

Figure 4: Example images correctly classified by the combined LSVM+C2 method (labels in the 2nd row), but misclassified by the part-based LSVM approach (labels in the 3rd row).

LSVM+C2:	Walking	Running	Photographing	Photographing	PlayingMusic
G.T.:	Running	Photographing	PlayingMusic	Walking	Inter. w/ comp.

Figure 5: Examples of challenging images misclassified by the combined LSVM+C2 method (labels in the 2nd row). The ground truth labels are shown in the 3rd row. Note the variation in viewpoint and scale as well as partial occlusion.

Dataset	Gupta et al.		Yao and Fei-Fei [22]			
			Task 1		Task 2	
Method	mAP	Acc.	mAP	Acc.	mAP	Acc.
Gupta *et al.* [10]	–	78.7	–	–	–	–
Yao and Fei-Fei [22, 23]	–	83.3	–	65.7	–	80.9
BOF Image (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 Image (B)	**91.6**	**85.0**	**77.8**	**75.1**	**90.5**	**84.9**

Table 4: Comparison with the method of Gupta *et al.* [10] and of Yao and Fei-Fei [22, 23] on their datasets. 'Task 1' is the 7-class classification problem and 'Task 2' is the PPMI+ vs PPMI- problem (see [22]).

Comparison with the state of the art[10, 22, 23]: For both datasets, no person bounding box information is used in training or test (method B). However, as the images in the original dataset are already cropped and centred to contain mostly the person of interest the approach is comparable with method A on our dataset. For the sport dataset of Gupta *et al.* [10] we have cross-validated again the parameters of the bag-of-features classifier and found that bigger vocabularies ($K = 4096$) perform better on this dataset. Other parameters (the intersection kernel and spatial pyramid binning) remain the same. For the Person Playing Musical Instrument dataset of Yao and Fei-Fei [22] we adopted a denser sampling of the SIFT features with initial spacing of 6 pixels to adapt to the smaller size of the images. Moreover we used a 3 level spatial pyramid and a LSVM with 9 parts to have a denser spatial coverage. Other parameters (the intersection kernel and $K = 1024$) remain the same. As shown in table 4, both the BOF and LSVM+BOF methods outperform the approach of Gupta *et al.* and Yao and Fei-Fei by 1.7% to 9.4%.

6 Conclusions

We have studied the performance of the bag-of-features classifier and the latent SVM model [8] on the task of action recognition in still images. We have collected a new challenging dataset of more than 900 consumer photographs depicting seven everyday human actions. We have demonstrated on this data, as well as two existing datasets of person-object interactions [10, 22], that (i) combining statistical and structured part-based representations and (ii) incorporating scene background context can lead to significant improvements in action recognition performance in still images. Currently, almost all tested methods (except the image-level classifier B) use the manually provided person bounding boxes. Next, we plan to investigate incorporating real person detections [5, 8] into the classifier.

Acknowledgements: We are grateful for financial support from the MSR-INRIA laboratory and the Quaero Programme, funded by OSEO.

References

[1] http://people.cs.uchicago.edu/∼pff/latent/.

[2] Willow actions database. http://www.di.ens.fr/willow/research/stillactions/.

[3] A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *IEEE PAMI*, 23(3):257–276, 2001.

[4] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *CIVR*, 2007.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages I:886–893, 2005.

[6] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *Proc. BMVC.*, 2010. updated version, available at http://www.di.ens.fr/willow/research/stillactions/.

[7] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 2009.

[9] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computer*, 22(1):67–92, January 1973.

[10] 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, 2009.

[11] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *ICCV*, 2009.

[12] N. Ikizler, R. G. Cinbis, S. Pehlivan, and P. Duygulu. Recognizing actions from still images. In *Proc. ICPR*, 2008.

[13] N. Ikizler-Cinbis, R. G. Cinbis, and S. Sclaroff. Learning actions from the Web. In *ICCV*, 2009.

[14] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[15] 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.

[16] L. J. Li and L. Fei-Fei. What, where and who? Classifying events by scene and object recognition. In *ICCV*, 2007.

[17] T. B. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 103(2-3):90–126, 2006.

[18] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond.* MIT Press, Cambridge, MA, 2002.

[19] J. Uijlings, A. Smeulders, and R. Scha. What is the spatial extent of an object? In *CVPR*, pages 770–777, 2009.

[20] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV*, 2009.

[21] Y. Wang, H. Jiang, M. S. Drew, Z. N. Li, and G. Mori. Unsupervised discovery of action classes. In *CVPR*, pages II: 1654–1661, 2006.

[22] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010.

[23] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.

[24] C. Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, 2009.

[25] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *IJCV*, 73(2): 213–238, 2007.