# Recognizing Actions Through Action-Specific Person Detection

Fahad Shahbaz Khan, Jiaolong Xu, Joost van de Weijer, Andrew D. Bagdanov, Rao Muhammad Anwer, and Antonio M. Lopez

***Abstract*— Action recognition in still images is a challenging problem in computer vision. To facilitate comparative evaluation independently of person detection, the standard evaluation protocol for action recognition uses an oracle person detector to obtain perfect bounding box information at both training and test time. The assumption is that, in practice, a general person detector will provide candidate bounding boxes for action recognition. In this paper, we argue that this paradigm is suboptimal and that action class labels should already be considered during the detection stage. Motivated by the observation that body pose is strongly conditioned on action class, we show that: 1) the existing state-of-the-art generic person detectors are not adequate for proposing candidate bounding boxes for action classification; 2) due to limited training examples, the direct training of action-specific person detectors is also inadequate; and 3) using only a small number of labeled action examples, the transfer learning is able to adapt an existing detector to propose higher quality bounding boxes for subsequent action classification. To the best of our knowledge, we are the first to investigate transfer learning for the task of action-specific person detection in still images. We perform extensive experiments on two benchmark data sets: 1) Stanford-40 and 2) PASCAL VOC 2012. For the action detection task (i.e., both person localization and classification of the action performed), our approach outperforms methods based on general person detection by 5.7% mean average precision (MAP) on Stanford-40 and 2.1% MAP on PASCAL VOC 2012. Our approach also significantly outperforms the state of the art with a MAP of 45.4% on Stanford-40 and 31.4% on PASCAL VOC 2012. We also evaluate our action detection approach for the task of action classification (i.e., recognizing actions without localizing them). For this task, our approach, without using any ground-truth person localization at test time, outperforms on both data sets state-of-the-art methods, which do use person locations.**

***Index Terms*— Action recognition, transfer learning, deep features.**

## I. Introduction

ACTION detection in still images is the task of localizing and classifying the actions of persons based on a single image. It is an extremely challenging problem due to factors such as person pose variation (e.g climbing), person context (e.g. gardening), and object-person interactions (e.g. phoning). Most research on the topic has focused on the *action classification* task in which the bounding box of the person performing the action is given at both training and testing time (i.e. localization is not part of the task). The rationale behind this has been that the harder task of *action detection* can be decomposed into a pre-processing step in which candidate persons are *proposed*, after which the action each detected person is performing can then be *classified*.

One of the main reasons for this two-step action detection methodology is that person detectors have become a mature technology over the last decade [9]. One of the most successful approaches, the Deformable Part Model (DPM), models the person as a structured constellation of parts [12]. Based on thousands of examples of labeled data, this method is able to learn a very accurate human model. However, human actions strongly condition the poses in which humans are expected to be observed: a person riding a horse is generally in a sitting position; somebody who is gardening is generally kneeling or slightly bent; and a person climbing might be in a contorted and infrequently observed pose (see Fig. 1). Based on this observation, we question the traditional two-step action detection methodology. Rather, we propose the use of action-specific person detection instead of general-purpose person detectors. Knowing what action class you are detecting provides valuable information on the poses in which the human is expected. Exploiting this information should lead to more accurate person localization, and consequently to better action classification results.

We consider two approaches to action-specific person detection. First, we investigate direct application of DPMs to each action class, by which we mean learning a complete DPM model from scratch for each action category.

Manuscript received March 19, 2015; revised June 26, 2015; accepted July 16, 2015. Date of publication August 5, 2015; date of current version August 18, 2015. This work was supported in part by the Svalbard Science Forum through the Collaborative Unmanned Aircraft Systems Project, in part by the VR through the ETT Project and through the Strategic Area for ICT Research ELLIIT, CADICS, in part by the Academy of Finland under Grant 255745 and Grant 251170, in part by the Data to Intelligence DIGILE SHOK Project under Project TIN2013-41751 and Project TIN2014-52072-P, in part by the Spanish Morocco Economic Competitiveness Project under Grant TRA2014-57088-C2-1-R, in part by the Spanish Ministry of Science through the Spanish DGT Project under Grant SPIP2014-01352, and in part by the Generalitat de Catalunya Project under Grant 2014-SGR-1506, and under Grant 2014-SGR-221. The work of A. D Bagdanov was supported by MICINN through Ramon y Cajal Fellowship. The work of J. Xu was supported by the Chinese Scholarship Council under Grant 2011611023. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dimitrios Tzovaras.

F. S. Khan is with the Computer Vision Laboratory, Department of Electrical Engineering, Linköping University, Linköping 581 83, Sweden (e-mail: fahad@cvc.uab.es).

J. Xu, J. van de Weijer, A. D. Bagdanov, and A. M. Lopez are with the Computer Vision Centre Barcelona, Barcelona 08193, Spain (e-mail: jiaolong@cvc.uab.es; joost@cvc.uab.es; bagdanov@cvc.uab.es; antonio@cvc.uab.es).

R. M. Anwer is with the Department of Information and Computer Science, Aalto University School of Science, Aalto FI-00076, Finland (e-mail: muhammad@cvc.uab.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2465147

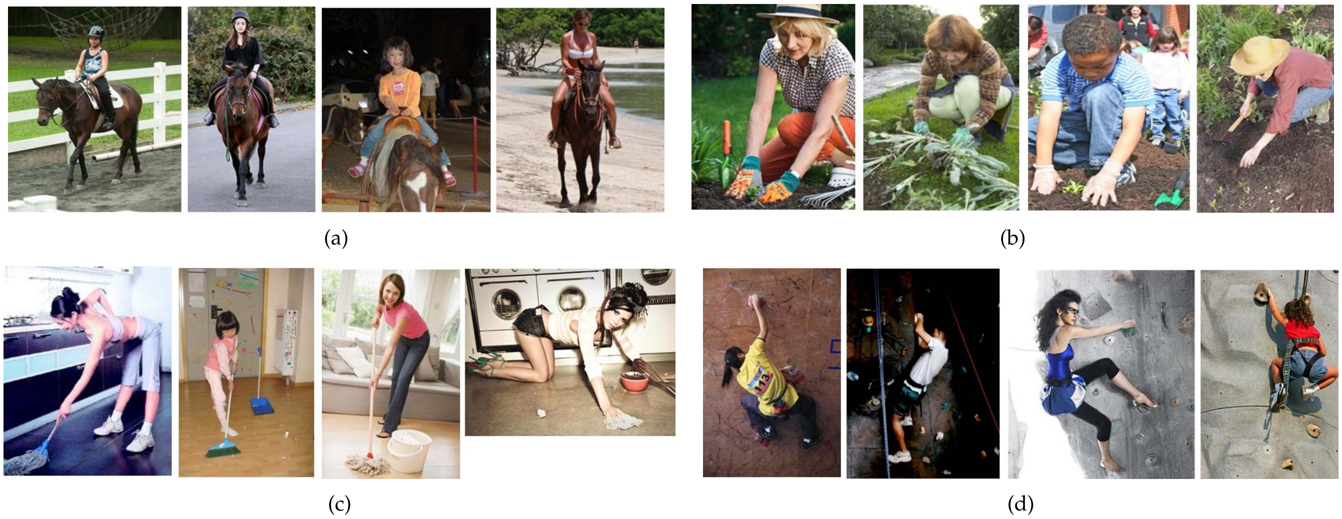


Fig. 1. Four example action categories: ‘riding horse’, ‘gardening’, ‘cleaning’ and ‘climbing’ from the Stanford-40 action dataset. These classes illustrate human actions strongly conditioned on pose. Action specific labels provide additional information on the poses in which the human is expected. (a) Action category: riding horse; Pose: sitting. (b) Action Category: gardening; Pose: kneeling. (c) Action Category: cleaning; Pose: bending. (d) Action Category: climbing; Pose: contorted.

However, given that existing action recognition datasets only contain a few hundred examples of each action class, it might be hard to learn accurate DPMs. As a second approach, we aim to exploit state-of-the-art person detectors by adapting the DPM model to action-specific detection. We do so by posing action-specific person detection as a transfer learning problem [32]. Transfer learning allows us to transfer knowledge from a previously-learned task (i.e. generic person detection by DPM) to new tasks (i.e. action-specific person detection). We expect this approach to be less sensitive to few training examples than directly training action-specific detectors. Motivated by the recent success of convolutional neural networks [16], [26], [31], we use deep features to represent our person proposals for action classification. In our experiments, we show that an action-specific person detector based on transfer learning leads to significantly better results than a general person detector, even when action recognition for both is based on the same deep features.

To validate our proposed approach, we perform extensive experiments on two action recognition datasets: Stanford-40 and PASCAL VOC 2012. We show that the prevailing paradigm for action recognition is sub-optimal since an action class is strongly conditioned on expected human poses. We further show that action-specific person detection based on transfer learning yields superior performance compared to the de facto methodology based on general person detection. On the PASCAL VOC 2012 dataset, our approach yields a significant gain of 6.8% in mean average precision (MAP) compared to the state-of-the-art method [17] based on deep features and bottom-up person proposals for action detection. Additionally, we also evaluate our action detection method for the problem action classification achieving state-of-the-art performance without exploiting exact ground-truth information at test time.

In the next section we review the literature related to action recognition in still images. We describe our approach to transfer learning of action-specific person proposals in section 3. In section 4 we report on experiments conducted on the Stanford-40 and PASCAL 2012 action recognition datasets, and we conclude with a discussion of our contribution in section 5.

## II. Related Work

Recognizing human actions in still images is a challenging task due to large variations in human appearance and pose. In this section we briefly review the state-of-the-art in action recognition, detection and transfer learning related to our proposed approach.

*1) Person Detection:* Most state-of-the-art methods for person detection are based on the learning-from-examples paradigm [6], [12], [25]. Among the many approaches, the deformable, part-based method (DPM) by Felzenszwalb et al. [12] yields excellent performance for person detection. The conventional DPM model employs HOG features [6] for image representation, and Khan et al. [25] augmented the DPM detector with color information. The use of generic object proposals [1], [5], [30], [40] for object detection has gained significant attention in recent years. Girshick et al. [15] demonstrated significant improvement in object detection performance with an approach combining selective search [40] and CNN-based representations. Though selective search has been shown superior for generic object proposal, in the seminal work on selective search it was shown that proposals based on Deformable Part Models [12] outperform selective search for the person category of the PASCAL dataset [40].

*2) Action Classification in Still Images:* The conventional approach to action recognition assumes that person bounding boxes are provided both at training and test time. The task is to associate an action label with each given person instance (i.e. to classify the action). Most approaches [7], [24], [28], [35] employ the bag-of-words framework for action classification. Others tackle the problem by finding human-object interactions [33], [43], or take a human-centric approach that localizes persons and then finds an object and its relationship to the person instance [33]. Yao et al. [43] proposed a method that uses attributes and parts by learning a set of sparse attribute and part bases for action recognition.

*3) Action Detection in Images and Video:* The standard action classification pipeline which exploits exact knowledge of person location is unrealistic for real-world applications. Recently, several approaches have investigated the problem of action detection both in images [8], [17], [24] and videos [13], [30], [39]. In the *action detection* problem ground-truth person locations are not known at test time: the task is to simultaneously localize and classify the action category of each person instance. Khan et al. [24] proposed a color extension of the Deformable Part Model [12] for action detection. The work of Gkioxari et al. [17] proposed a convolutional neural network approach for the the task of action detection in still images. The method is built on the R-CNN framework [15] and employs region proposals generated using a multiscale combinatorial grouping method [2].

*4) Transfer Learning for Object Detection:* At the core of an object detector is a corresponding object classifier which, given an object proposal, determines if it actually corresponds to an object of interest. Underlying this approach is the assumption that the probability distributions of the training and testing data are the same. However, in practice this is not always the case, which results in a loss of accuracy in the learned classifier and consequently of associated detector. In some cases, the discrepancy between the test and train data probability distributions is due to changes in acquisition sensor or the application environment. *Domain adaptation* techniques are designed to tackle these situations [20], [22], [34], [41], [42]. In other cases, the discrepancy is due to the task in the application scenario differing from the task for which the classifiers were trained. Approaches addressing this type of problem are referred to as *transfer learning* [3], [4], [32].

Domain adaptation for person detection has recently been shown to be effective even for adapting models learned with virtual data to operate in the real world (holistic model [41], DPM [22]). In these approaches, only a few examples are used from the target domain. Adaptation consists of a retraining step, either by mixing source and target training data [41] or by just modifying the source model with the new target domain training data [22]. The latter case has the advantage of not having to revisit source domain training data, allowing faster training runs. In [22], DPMs are learned using Structural SVM (SSVM) and are domain-adapted by extension of the Adaptive SVM (A-SVM) [42] to operate with the SSVM (A-SSVM). In this paper, we use A-SSVM not to perform domain adaptation but for transfer learning. In particular, we will bias a generic person DPM towards action-specific person DPMs by using just a few action-specific training examples.

*5) Our Approach and the State-of-the-Art:* Based on the observation that actions strongly condition the poses in which humans are expected to appear, and the impressive results achieved by CNN features [15], in this paper we propose a technique to generate candidate person proposals for action detection in still images. We will show that generic person detectors [12] are not up to the task of generating robust, high quality action candidates, and that naive action-specific training of person detectors is also insufficient to generate high enough recall. Instead, we will use A-SSVM to transfer knowledge from generic person detectors, specializing them to action-specific ones that, combined with deep CNN features for classification, yield a significant improvement over the state-of-the-art on action recognition in still images. Figure 2 compares general and action-specific person detectors for action detection. The standard approach (in the middle row) assumes that the candidate person proposals are generated using a general person detector in an initial phase. On the top row, the direct training of a complete DPM model from scratch for each action class is shown. On the bottom row, we show our transfer learning pipeline that adapts general person DPM model to learn action-specific detectors.

## III. Person Proposals for Action Recognition

The standard approach to action detection consists of two steps: a general person detection step and an action-specific classification step. However, the strong relation between body pose and action suggests that this information could already be used in the detection phase. In this section we look into two approaches to action-specific person detection.

### A. Direct Learning of Action-Specific Person Detectors

A straightforward approach to obtaining action specific person proposals is to learn a detector for each action class. We call this approach *direct* learning of the action detection to distinguish it from our method based on transfer learning. We use the latest version (version 5.0) of the DPM detector [14] for all detectors learned directly or via transfer learning in this work.

Let $\mathcal{D} = (\mathbf{x}_1, y_1, \mathbf{h}_1), \ldots, (\mathbf{x}_N, y_N, \mathbf{h}_N) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{H}$ be the set of training examples, where $\mathcal{X}$ is the input space, $\mathcal{Y} = \{+1, -1\}$ is the label space, and $\mathcal{H}$ is the hypothesis or output space as usually defined for DPMs. Let $\Phi(\mathbf{x}, \mathbf{h})$ be a joint feature vector, where $\mathbf{h}$ is a latent variable not known during DPM training. Finally, let $\mathbf{w}$ be the usual vector of the DPM parameters (appearance and deformation parameters of the parts of all components).

The model parameters $\mathbf{w}$ are learned by solving the following latent SSVM optimization:

$$\min_{\mathbf{w}} \quad \underbrace{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \max_{\widehat{y},\widehat{\mathbf{h}}}[\mathbf{w}'\Phi(\mathbf{x}_i, \widehat{\mathbf{h}}) + L(y_i, \widehat{y}, \widehat{\mathbf{h}})]}_{convex} \\ \underbrace{-C\sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w}'\Phi(\mathbf{x}_i, \mathbf{h})}_{concave}, \qquad (1)$$

where $C$ is the scalar penalty, $L(\cdot)$ is the loss function (we use 0-1 loss, i.e., returns 0 if $\widehat{y} = y_i$ and 1 otherwise). The latent SSVM optimization objective function (1) can be solved by the

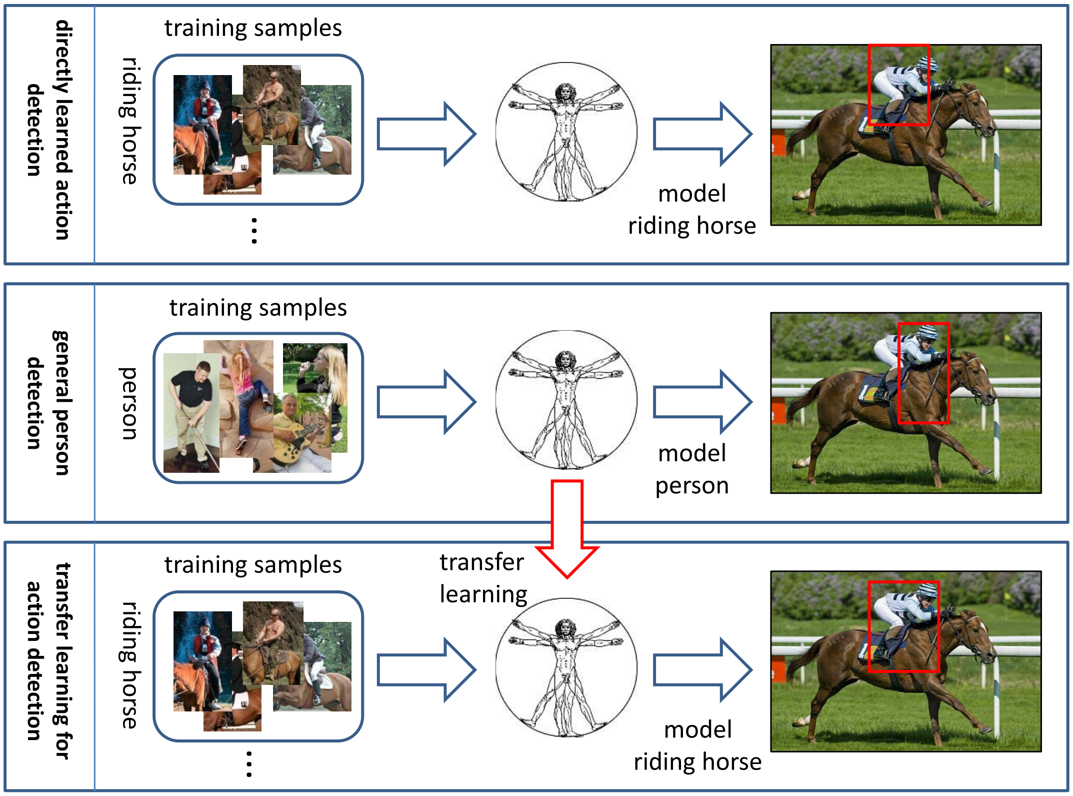


Fig. 2. General and action-specific person detectors. In the middle row we show the standard pipeline based on general person detection. On the top, we give the pipeline for action-specific person detection based on direct learning, while on the bottom we show our pipeline for action detection based on transfer learning in which the general person model is exploited to learn action-specific detectors.

Convex-Concave Procedure (CCCP) [14], which guarantees the convergence to a local minimum or a stationary point of the objective function.

In the direct approach to learning action-specific detectors, we learn a model $\mathbf{w}^a$ independently for ever action class $a$ using labeled person instances performing action $a$ as positive examples. Our intuition behind introducing action-specific detectors is that they should yield higher recall than the generic DPM for detections on persons performing the corresponding action.

In figure 3 we plot the detection recall of different types of person detectors on six action classes from the Stanford-40 dataset. We use the same number of detections (per image) for direct learning, general person and transfer learning based detectors. For general person we also significantly lower the standard threshold (25 times) to obtain extra detections (indicated by “low threshold” in figure 3). This yields extra detections ranging from 2 to 10 per image. No additional detections are generated by further lowering the threshold. We also compare with recall of the selective search based proposal approach [40], which yields around 2k proposals per image. The selective search proposal method has also been used in the RCNN framework [15], obtaining state-of-the-art results for the object detection task. Detection recall curves are commonly used to evaluate object proposals [5], [10]. We consider detections in all positive images (negative images are irrelevant for recall evaluation). Recall is shown as a function of the overlap percentage, where the standard PASCAL evaluation considers an overlap of 50% to be a correct detection [11].

We see in this figure that the selective search approach yields inferior recall compared to approaches based on the DPM framework. For some classes (e.g. ‘running’ and ‘looking through microscope’), direct learning yields higher recall than general person detection. However, for other classes (e.g. ‘playing violin’ and ‘smoking’), direct learning fails with significantly lower recall than the general person detector. The most likely explanation is that the lack of training data available for action detection (around 100 per class) prevents the action specific detectors to outperform the general person detector which is based on 10,000 training examples. In principle, direct learning is expected to improve its performance when given large amounts of labeled actions. Our approach based on transfer learning mitigates precisely this need for extensive labeling by using models from an already learned task.

### *B. Transfer Learning of Action-Specific Person Detectors*

From the analysis in the previous section we see that direct learning of action-specific person detectors does not outperform general person detectors in terms of generating good candidate detections for action recognition. In this section we propose a transfer learning approach to action-specific person detection that, instead of building new detectors from scratch, specializes the knowledge of a general person detector trained

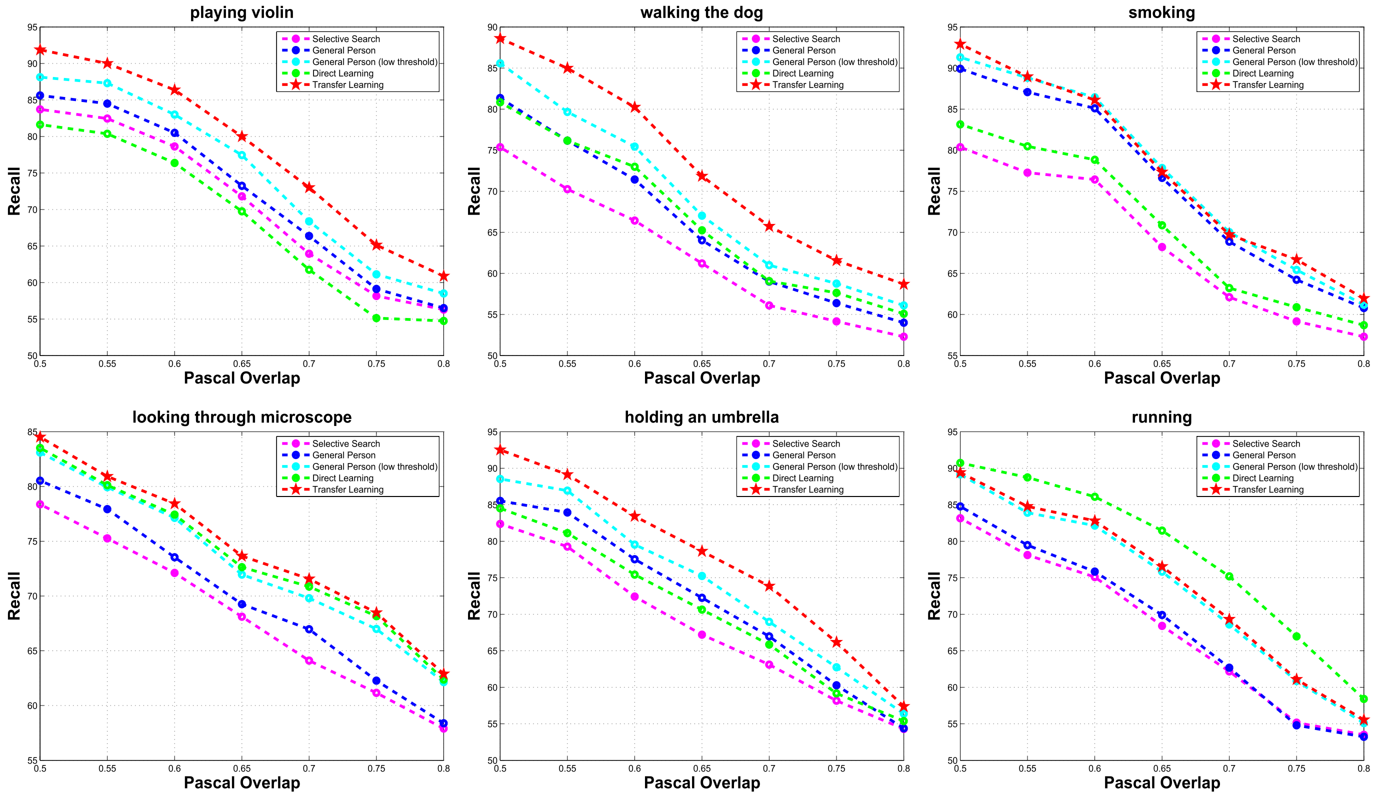


Fig. 3. Recall curves for person detection on six classes from the Stanford-40 dataset. In most cases, the transfer learning approach yields significantly improved person localizations compared to both the general and action-specific detectors.

on thousands of labeled bounding boxes to specific action classes using a small number of training examples from each. For transfer learning we apply the Adaptive SSVM (A-SSVM) approach [22], [42].

Let $\mathcal{D}_a$ be the labeled training examples for action $a$. Let $\mathbf{w}^G$ be the DPM parameter vector of the generic person detector learned according to equation (1) by using all the available training examples of a set $\mathcal{D}_G$ (i.e. irrespective of the action that the positive examples are performing). $\mathcal{D}_a$ can be a small subset of $\mathcal{D}_G$ or a completely different set (zero intersection). Now, by using $\mathcal{D}_a$ and $\mathbf{w}^G$ as input for an A-SSVM learning procedure we obtain the action specific person detector $\mathbf{w}^a$ for action $a$ by solving the following optimization problem:

$$\begin{aligned}&\min_{\mathbf{w}^a,\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}^a - \mathbf{w}^G\|^2 + C\sum_{i=1}^{N}\xi_i\\ &\text{s.t.} \quad \forall i, y, \mathbf{h}, \quad \xi_i \geq 0, \quad \forall(\mathbf{x}_i, y_i) \in \mathcal{D}_a\\ &\mathbf{w}^{a'}\Phi(\mathbf{x}_i, \mathbf{h}_i) - \mathbf{w}^{a'}\Phi(\mathbf{x}_i, \mathbf{h}) \geq L(y_i, y, \mathbf{h}) - \xi_i, \qquad (2)\end{aligned}$$

where $y_i$ and $\mathbf{h}_i$ are the ground truth label and object hypothesis, $y$ and $\mathbf{h}$ represent all the alternative output label and object hypotheses, and $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]'$ are slack variables. The regularization term $\|\mathbf{w}^a - \mathbf{w}^G\|^2$ in equation (2) adapts the original model $\mathbf{w}^G$ towards a new, action specific model $\mathbf{w}^a$ by regularizing the distance between them.

In the recall graphs of figure 3 we also plot the recall of action-specific person detectors trained using the transfer learning approach in equation (2). From these plots, we see that in all cases except the running action (where variation in pose is limited), our approach generates better candidates than both the general and action-specific detectors.

## C. Action Classification With Deep Features

Our action detection pipeline is as follows. For an action class we run the corresponding action-specific detector (either trained with direct learning or trained with transfer learning). On the output of the specific detector we then run the classifier trained for the same action class. Here we describe the classification approach we use.

Recently, features learned using Convolutional Neural Networks (CNNs) have shown significant performance gains in many computer vision application such as object recognition [31], object detection [15], video recognition [23] and face recognition [38]. These features (also sometimes called deep features) are extracted from hidden convolutional layers of CNNs. They are generic and result from training on a large amount of training data (e.g. ImageNet). We use the Very Deep Features pre-trained on ImageNet dataset from [37]. The network takes as input a fixed size image of size 224 × 224. Unlike conventional networks employing large receptive fields in the first layer, the very deep network uses small 3 × 3 receptive fields throughout. The receptive fields are convolved at each pixel with a stride of 1 pixel. The network has 5 max-pooling layers that perform spatial pooling over 2 × 2 pixel windows at a stride of 2 pixels. For more details, we refer to [37].[1]

[1] The deep network models available at: http://www.robots.ox.ac.uk/~vgg/research/very_deep/

For classification we use a feature representation obtained by concatenating the outputs of layers 16 and 19 from this network, yielding a 8192D feature vector representing each candidate action region. Classification is then performed using SVMs with linear kernels.

## IV. Experimental Results

In this section, we report on a series of action recognition experiments. We first describe the datasets and experimental protocols used, and then provide quantitative and qualitative results for action detection and action classification.

### *A. Datasets and Evaluation*

We evaluate the performance of our approach on the Stanford-40 and PASCAL VOC 2012 action recognition datasets. Stanford-40 consists of 9532 images of 40 different action classes such as gardening, fishing, applauding, cooking, brushing teeth, cutting vegetables, and drinking.[2] The dataset is divided into 4000 training images and 5532 test images. The PASCAL VOC 2012 dataset comprises 10 different action categories: phoning, playing instrument, reading, riding bike, riding horse, running, taking photo, using computer, walking and jumping.[3] Since the PASCAL 2012 test set is withheld by the organizers, for action detection we use the validation set for testing. The training set consists of 2296 images with 3134 instances (person bounding boxes and action labels). To augment the available data for training, we use the extra person bounding boxes and action label annotations provided by the authors of [17] on the PASCAL 2012 validation dataset. These additional annotations increase the instances in the validation set from 3144 to 5891.

To evaluate the performance of action detection and action classification, we follow the standard PASCAL detection protocol by computing average precision (AP) under the precision-recall curve. Given a set of proposals in an image, a detection is considered correct if both the overlap is more than 50% with the ground-truth and the action label is correctly classified for an instance [17], [24]. Our action-specific detectors propose candidate windows which are then classified into action categories.

### *B. Action Detection*

We begin by comparing baseline person proposal methods with the direct learning and transfer learning (TL) approaches to training action-specific person detectors. We then compare our approach with the state-of-the-art on action detection in still images and also give some qualitative evaluation.

*1) Person Proposals for Action Detection:* In action detection, the task is to simultaneously localize and classify the action associated with each person instance. We compare the general person proposal approach to both types of action-specific person detectors proposed in section 3.

TABLE I

Comparison of Our Transfer Learning (TL) Based Action Detection Approach With the Two Baseline Methods and the CN-HOG Detector. Performance Is Measured in MAP (%). We Use Same Set of Deep Features for All the Methods. On Stanford-40, Our Approach Improves Performance by 5.7% Compared to General-Person Detection. On PASCAL 2012, Our Approach Again Outperforms Both Direct-Specific and General-Person Methods

| Dataset | CN-HOG [24] | General-person | Direct-specific | TL |
|---|---|---|---|---|
| Stanford-40 | 27.5 | 39.7 | 37.6 | **45.4** |
| PASCAL 2012 | 25.4 | 29.3 | 28.6 | **31.4** |

For all three person proposal techniques we train a HOG-based generic person DPM model.[4] The three evaluated techniques are:

- *General-Person:* we train the DPM using the person instances from all action classes as positive training samples. To obtain negative samples, we use the images from the 19 non-person image classification classes of the PASCAL VOC 2012 set.
- *Direct-Specific:* for this first action specific approach we train a DPM model for each action category as described in section 3.1. These models are used to obtain action-specific person proposals on test images.
- *Transfer-Learning (TL):* we perform transfer learning to adapt the general-person source model to each specific action class using the approach discussed in Section 3.2.

After person proposal by one of these techniques, we use the deep feature classifier described in section 3.3 to recognize actions. To train action-specific classifiers, we use the positive bounding boxes of the respective class and the bounding boxes of other action classes as negative training samples. To obtain additional hard negatives, we apply a person detector on all training images and extract detections with an overlap threshold of 0.3 with any action class.

Table 1 shows the results of all three person proposal methods for action detection. On the Stanford-40 dataset, the general-person proposal approach achieves a mean AP of 39.7%, while direct-specific proposals are 2% lower at 37.6%. This is probably due to a lack of training data per action class. Further, the direct use of action-specific HOG detector outputs provides significantly inferior action detection performance (21.7%) compared to employing an action classification stage based on deep features (37.6%). Our transfer-learning method, which adapts the general person model to the specific action, significantly improves this performance with a mean AP of 45.4%. On this dataset, we also perform an experiment by training a DPM model using action class labels to initialize the model mixtures. This does improve performance over direct-specific DPM method but is still significantly inferior to our transfer learning based detection approach.

[2] The Stanford-40 dataset is available at http://vision.stanford.edu/Datasets/40actions.html

[3] PASCAL 2012 is available at: http://www.pascal-network.org/challenges/VOC/voc2012/

[4] The code for the DPM detector is available at: http://www.cs.berkeley.edu/~rbg/latent/index.html

TABLE II

COMPARISON WITH STATE-OF-THE-ART RESULTS ON PASCAL VOC 2012 AUGMENTED VALIDATION SET FOR ACTION DETECTION. OUR APPROACH YIELDS A SIGNIFICANT GAIN OF 6.0% IN MEAN AP OVER THE BEST REPORTED RESULTS IN THE LITERATURE

| | phoning | playingmusic | reading | ridingbike | ridinghorse | running | takingphoto | usingcomputer | walking | jumping | mean AP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Action RCNN [17] | 8.4 | 9.4 | 4.7 | 9.6 | 19.0 | 16.4 | 10.8 | 3.1 | 3.7 | 14.4 | 10.0 |
| Det RCNN [17] | 9.1 | 11.9 | 7.5 | 9.9 | 20.1 | 24.4 | 5.3 | 1.8 | 9.3 | 19.5 | 11.9 |
| Action-Det RCNN [17] | 22.4 | 28.4 | 16.8 | 26.2 | 35.2 | 28.1 | 22.7 | 15.8 | **20.6** | **29.6** | 24.6 |
| HOG [12] | 19.7 | 18.3 | 13.4 | 32.6 | 45.7 | 46.5 | 10.4 | 18.1 | 18.3 | 18.4 | 24.1 |
| CN-HOG [24] | 21.5 | 19.3 | 14.8 | **32.9** | **48.2** | **47.7** | 7.9 | 22.2 | 18.5 | 21.1 | 25.4 |
| Our approach | **30.1** | **43.6** | **24.9** | 32.0 | 38.6 | 39.1 | **30.6** | **35.6** | 11.4 | 27.6 | **31.4** |

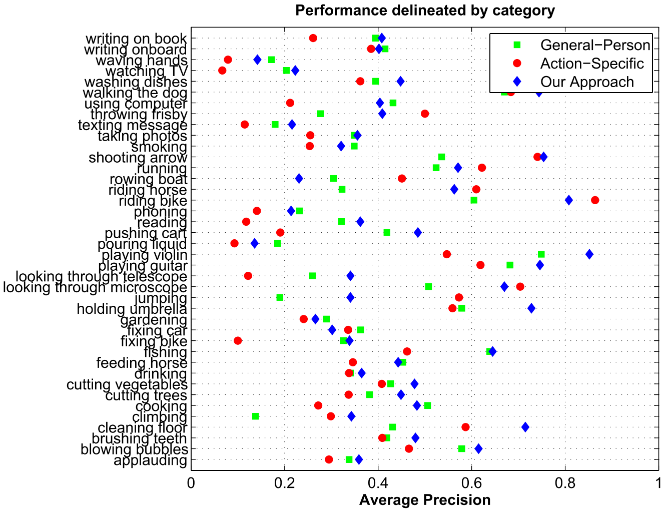


Fig. 4. Per-category performance comparison of our action detection approach with the two baseline methods on the Stanford-40 dataset. Our approach outperforms the two conventional methods on 24 out of 40 action classes.

Similarly, on the augmented PASCAL VOC 2012 validation set the general-person approach slightly outperforms direct-action proposals, while transfer learning yields a gain of 2.1% compared to the general-person method. In conclusion, these results clearly show that the two step paradigm of a general person detector and an action classifier is suboptimal, and that it is important to already take the action class into account during detection.

In figure 4 we give a per-category performance comparison of our approach with the two baseline methods on the Stanford-40 dataset. Our approach improves the detection results on 24 out of 40 action classes. Especially, remarkable accuracy gains are obtained on the holding-an-umbrella (+14.9%), cleaning the floor (+12.8%), playing violin (+10.3%), and looking through a telescope (+8.1%) action classes. These actions all have very distinctive poses related to them (such as kneeling when cleaning the floor, and bending when looking through telescope) which are not well-described by the general person model.

*2) Comparison With the State-of-the-Art:* We compare our action detection approach with state-of-the-art methods from the literature. The HOG-based DPM detector [12] and Opponent-HOG (OPP-HOG) were evaluated on the action detection task in [24]. The HOG obtains a mean AP of 21.7%, and OPP-HOG and CN-HOG further improve the detection performance with a mean AP of 25.7% and 27.5%, respectively. Our approach based on transfer learning (TL) improves the state-of-the-art (CN-HOG) by 17.9% in mean AP on this dataset.

In table 2 we compare our approach with the state-of-the-art on the augmented PASCAL VOC 2012 validation dataset. From [17], the action R-CNN network obtains a mean AP of 10.0%, the detection R-CNN further improves this with mean AP of 11.9%, and the network trained jointly for detection and action recognition improves the performance with a mean AP of 24.6%. We also performed experiments by training action-specific HOG and CN-HOG based DPM detectors. The HOG and CN-HOG based DPM action detectors of achieve comparable results at 24.1% and 25.4% mean AP, respectively. Our approach significantly improves on the state-of-the-art with a mean AP of 31.4%.

*3) Error Analysis:* We analyze the types of errors made by our action detection approach using the protocol proposed by Hoiem et al. [21] for diagnosing errors in generic object detectors. Three types of false positive errors are distinguished: Loc (localization errors), BG (confusion with background) and Oth (other errors, which is correct localization of a person but misclassification of action label). We compare the transfer learning method with direct learning on the Stanford-40 dataset. Transfer learning significantly reduces errors due to localization and confusion with background compared to direct learning method. Background errors are reduced by 4% using our approach, and localization errors are similarly reduced by 4% compared to the direct learning method. This shows that our transfer learning based person proposals efficiently localize persons while rejecting the background. Figure 5 shows the percentage of errors of each type identified in the top scoring 25-3200 false positives using our approach and direct learning method on four action classes: playing violin, walking with dog, waving hands and feeding a horse. On most action classes, our approach significantly reduces localization errors and background confusion.

Figure 6 shows several example action detection results based on action-specific person proposals (in green) and our action specific proposals obtained through transfer learning (in red). These examples clearly suggest that our approach leads to generally better localization of person instances and also higher and more precise action classification scores.

### C. Action Classification

We also validate the performance of our action detection approach for the task of action classification. In the standard protocol for action classification, bounding box information

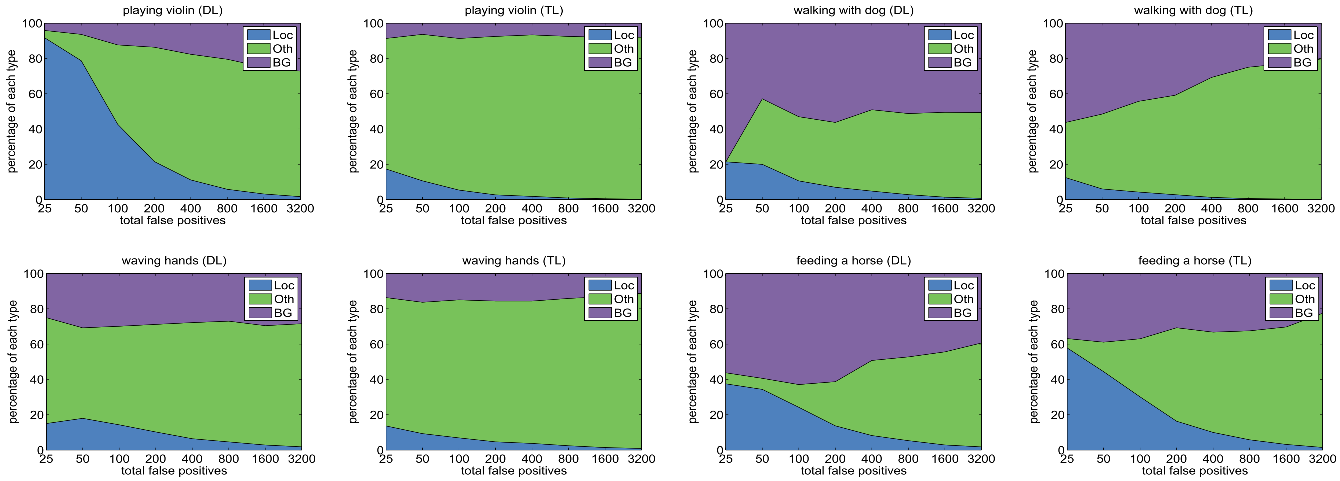


Fig. 5. Action detection error analysis using direct learning (DL) and transfer learning (TL) based methods. Top row: comparison on 'playing violin' and 'walking with dog' action classes. Bottom row: comparison on 'waving hands' and 'feeding a horse' classes. In most cases our TL method significantly reduces the percentage of false positive errors due to localization (Loc) and background confusion (BG).

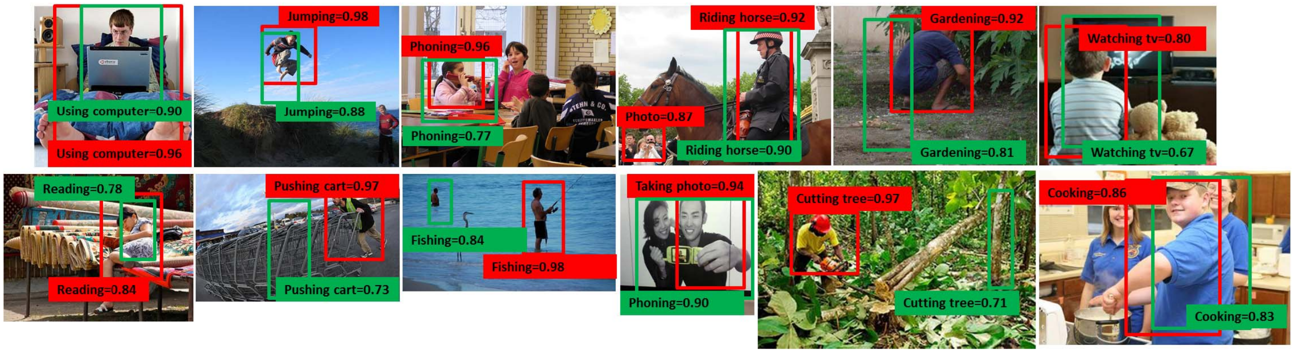


Fig. 6. Example action detection results based on action-specific person proposals (green) and proposals obtained through transfer learning (red). Transfer learning of action-specific proposals leads to generally better person localizations and higher classification scores for the correct class.

TABLE III

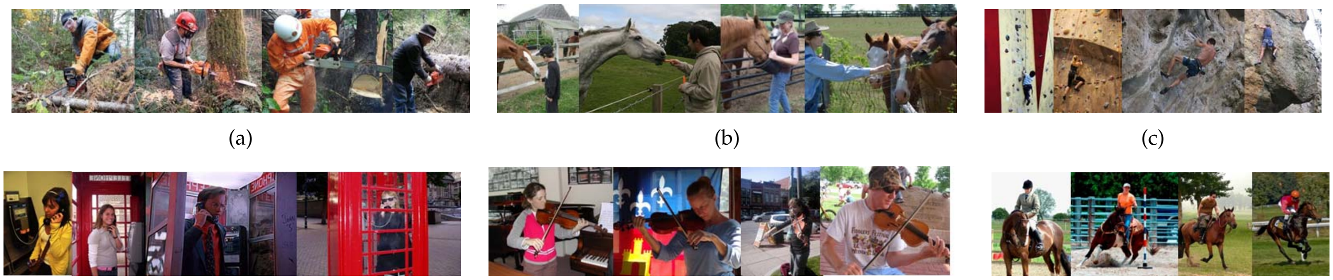
COMPARISON OF OUR ACTION CLASSIFICATION APPROACH WITH THE STATE-OF-THE-ART ON STANFORD-40. NOTE THAT OUR APPROACH DOES NOT USE ANY GROUND-TRUTH INFORMATION AT TEST TIME. OUR APPROACH YIELDS A SIGNIFICANT GAIN OF 22.4% IN MEAN AP OVER THE BEST REPORTED RESULTS IN THE LITERATURE

| Method | EPM [36] | SB [43] | CF [24] | SMP [27] | Ours |
|---|---|---|---|---|---|
| mAP | 45.2 | 45.7 | 51.9 | 53.0 | **75.4** |

of each person instance is provided both at train and test time. The task is then to classify the action category for each person instance. Typically, action classification approaches use both the bounding box of a person together with full image representation. In our experiments, we do not use any ground truth bounding box information at test time. Instead, we use the outputs from our action detector as an approximate location of a person. For the Stanford-40 dataset, since there is one person instance per image, we use the output from our action detector with maximum confidence for the action class respectively. We further combine it with a representation of the full image.

In table 3 we compare our approach with the state-of-the-art on Stanford-40 for the action classification task. The CF method [24] based on fusing multiple color descriptors obtains a MAP of 51.9%. The expanded part based method (EPM) [36] employs part based information and obtains a MAP of 45.2%. The semantic pyramid method (SMP) [24] based on constructing pyramids on different body parts obtains a MAP of 53.0%. Our approach, without using any bounding box information at test time, improves the state-of-the-art by 22.4% on this dataset. We also perform an experiment by using the bounding box information at test time in our pipeline obtaining a mean AP of 77.8%. The results clearly suggest that our approach performs favorably compared to using ground-truth information. Figure 7 shows top few predictions of six different action classes from the Stanford-40 dataset.

For the PASCAL VOC 2012 dataset, we evaluate our approach on the "Boxless Action Classification Taster" task (competition 11). Since there are multiple person instances in an image, a person in a test image is indicated only by a "single point" lying somewhere on their body. The objective of this taster competition is to evaluate action classifiers given no precise information (bounding box) of a person. We use the bounding boxes generated by our action-specific person detector. Again, for every person instance in the test set, we select the output from our action detector closest to the



(a) (b) (c)

(d) (e) (f)

Fig. 7. Example action classification results from the Stanford-40 dataset. Top correct predictions are shown of six action classes: ‘cutting tree’, ‘feeding a horse’, ‘climbing’, ‘phoning’, ‘playing violin’ and ‘riding a horse’. The proposed action classification approach combines the outputs from our action detector and holistic image classifier. (a) Action category: cutting tree. (b) Action category: feeding a horse. (c) Action category: climbing. (d) Action category: phoning. (e) Action category: playing violin. (f) Action category: riding a horse.

TABLE IV

Comparison With the State-of-the-Art on PASCAL VOC 2012 Test Set for the Action Classification Task. Our Approach, Without Using the Exact Bounding Box Information, Performs Favorably Compared to State-of-the-Art Methods. Our Method Provides Best Performance on 5 Out of 10 Action Categories

| | phoning | playingmusic | reading | ridingbike | ridinghorse | running | takingphoto | usingcomputer | walking | jumping | mean AP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MDF [31] | 46.0 | 75.6 | 45.3 | 93.5 | 95.0 | 86.5 | 49.3 | 66.7 | 69.5 | 78.4 | 70.2 |
| RMP [18] | 52.9 | 84.3 | 53.6 | **95.6** | **96.1** | **89.7** | 60.4 | 76.0 | 72.9 | 82.3 | 76.4 |
| WAB [19] | 49.5 | 67.5 | 39.1 | 94.3 | 96.0 | 89.2 | 44.5 | 69.0 | **75.9** | 79.6 | 70.5 |
| Action R-CNN [17] | 47.4 | 77.5 | 42.2 | 94.9 | 94.3 | 87.0 | 52.9 | 66.5 | 66.5 | 76.2 | 70.5 |
| Wholes and Parts [16] | 54.5 | 79.8 | 48.9 | 95.3 | 95.0 | 86.9 | **61.0** | 68.9 | 67.3 | 77.9 | 73.6 |
| Action Poselets [29] | 32.4 | 45.4 | 27.5 | 84.5 | 88.3 | 77.2 | 31.2 | 47.4 | 58.2 | 59.3 | 55.1 |
| Stanford | 44.8 | 66.6 | 44.4 | 93.2 | 94.2 | 87.6 | 38.4 | 70.6 | 75.6 | 75.7 | 69.1 |
| Oxford | 50.0 | 65.3 | 39.5 | 94.1 | 95.9 | 87.7 | 42.7 | 68.6 | 74.5 | 77.0 | 69.5 |
| Our approach | **62.4** | **91.3** | **61.1** | 93.3 | 95.1 | 84.1 | 59.8 | **84.5** | 53.0 | **84.9** | **77.0** |

provided reference point of a person. We further combine it with a representation of the full image.

In table 4 we compare our approach with the state-of-the-art on the PASCAL VOC 2012 test set for the action classification task. The mid-level deep image representation (MDF) method [31] obtains a MAP of 70.2%. The approach based on weak alignment of body part method (WAB) provides a MAP of 70.5% [19]. The method of [18] based on regularized max pooling (RMP) of feature vectors at multiple scale and windows obtains a MAP of 76.4%. The RMP approach employs both full image and bounding box information for feature extraction. Our approach, based on approximate localization using a single reference point, yields a mean AP of 77.0%. We also performed an experiment using the bounding box information at test time in our pipeline and obtained a mean AP of 81.3%. The notable difference is on the Taking Photo and Walking categories where the performance deteriorates when using our detector output. On the rest of the action categories, our approach provides similar performance compared to using exact bounding box information.

## V. Conclusion

In this paper we investigated the problem of action detection in still images. Most state-of-the-art approaches to action recognition exploit bounding box information at test time. In such a pipeline, it is assumed that the candidate persons are reliably proposed (e.g. using a general person detector) in an initial phase. Afterwards, action classification associates an action category label to each detected person instance. Since body pose is strongly conditioned on the action category, we argued that using a general person detector is sub-optimal for action recognition.

Instead, we proposed using action-specific person detectors to drive action detection in still images. We showed that a direct extension of the DPM framework to action-specific detection gives sub-optimal performance due to limited amounts of action training data. We then proposed a new approach by posing action detection as a transfer learning problem. Transfer learning is used to transfer knowledge from a previously-learned person detection task to the new task of action-specific person detection.

We evaluated our approach on the Stanford-40 and PASCAL 2012 action recognition datasets. For action detection, our approach based on transfer learning yields significant improvements compared to both general person and direct learning methods. On both datasets, our approach outperforms the state-of-the-art for action detection. This confirms our hypothesis that the action class label should be used when learning the detector. We also evaluated our approach on the action classification problem. Results show that our performance is superior to state-of-the-art methods that exploit ground truth bounding box information at test time.

It is worth mentioning that our transfer learning approach is non-parametric in that there is no parameter controlling the relative weight of source and target models. We expect that learning such a weighting will further improve performance. Another interesting direction is to investigate combining multiple transfer learning methods to obtain even better person proposals for action detection in still images.

## Acknowledgements

The authors also want to thank the NVIDIA Corporation for the support in the form of different GPU hardware units.

## References

[1] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.

[2] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 328–335.

[3] Y. Aytar and A. Zisserman, "Tabula rasa: Model transfer for object category detection," in *Proc. IEEE ICCV*, Nov. 2011, pp. 2252–2259.

[4] Y. Aytar and A. Zisserman, "Enhancing exemplar SVMs using part level transfer regularization," in *Proc. BMVC*, 2012, pp. 1–11.

[5] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300 fps," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 3286–3293.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2005, pp. 886–893.

[7] 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, pp. 97.1–97.11.

[8] C. Desai and D. Ramanan, "Detecting actions, poses, and objects with relational phraselets," in *Proc. 12th ECCV*, 2012, pp. 158–172.

[9] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[10] I. Endres and D. Hoiem, "Category-independent object proposals with diverse ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 222–234, Feb. 2014.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[12] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[13] A. Gaidon, Z. Harchaoui, and C. Schmid, "Actom sequence models for efficient action detection," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 3201–3208.

[14] R. Girshick, "From rigid templates to grammars: Object detection with structured models," Ph.D. dissertation, Dept. Comput. Sci., Univ. Chicago, Chicago, IL, USA, 2012.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 580–587.

[16] G. Gkioxari, R. Girshick, and J. Malik. (2014). "Actions and attributes from wholes and parts." [Online]. Available: http://arxiv.org/abs/1412.2604

[17] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. (2014). "R-CNNs for pose estimation and action detection." [Online]. Available: http://arxiv.org/abs/1406.5212

[18] M. Hoai, "Regularized max pooling for image categorization," in *Proc. BMVC*, 2014.

[19] M. Hoai, L. Ladicky, and A. Zisserman, "Action recognition from weak alignment of body parts," in *Proc. BMVC*, 2014.

[20] J. Hoffman, E. Rodner, J. Donahue, B. Kulis, and K. Saenko, "Asymmetric and category invariant feature transformations for domain adaptation," *Int. J. Comput. Vis.*, vol. 109, nos. 1–2, pp. 28–41, 2014.

[21] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing error in object detectors," in *Proc. 12th ECCV*, 2012, pp. 340–353.

[22] J. Xu, S. Ramos, D. Vazquez, and A. M. Lopez, "Domain adaptation of deformable part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2367–2380, Dec. 2014.

[23] A. Karpathy, S. Shetty, G. Toderici, R. Sukthankar, T. Leung, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 1725–1732.

[24] F. S. Khan, R. M. Anwer, J. van de Weijer, A. D. Bagdanov, A. M. Lopez, and M. Felsberg, "Coloring action recognition in still images," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 205–221, 2013.

[25] F. S. Khan, R. M. Anwer, J. van de Weijer, A. D. Bagdanov, M. Vanrell, and A. M. Lopez, "Color attributes for object detection," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 3306–3313.

[26] F. S. Khan, R. M. Anwer, J. van de Weijer, M. Felsberg, and J. Laaksonen, "Deep semantic pyramids for human attributes and action recognition," in *Proc. SCIA*, 2015, pp. 341–353.

[27] F. S. Khan, J. van de Weijer, R. M. Anwer, M. Felsberg, and C. Gatta, "Semantic pyramids for gender and action recognition," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3633–3645, Aug. 2014.

[28] F. S. Khan, J. van de Weijer, A. D. Bagdanov, and M. Felsberg, "Scale coding bag-of-words for action recognition," in *Proc. 22nd ICPR*, Aug. 2014, pp. 1514–1519.

[29] S. Maji, L. D. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 3177–3184.

[30] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid, "Spatio-temporal object detection proposals," in *Proc. ECCV*, 2014, pp. 737–752.

[31] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 1717–1724.

[32] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2009.

[33] A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 601–614, Mar. 2012.

[34] K. Saenko, B. Hulis, M. Fritz, and T. Darrel, "Adapting visual category models to new domains," in *Proc. 11th ECCV*, 2010, pp. 213–226.

[35] G. Sharma, F. Jurie, and C. Schmid, "Discriminative spatial saliency for image classification," in *Proc. CVPR*, 2012.

[36] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for human attribute and action recognition in still images," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 3506–3513.

[37] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: http://arxiv.org/abs/1409.1556

[38] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 1701–1708.

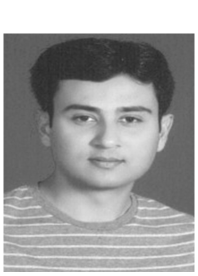
[39] D. Tran and J. Yuan, "Max-margin structured output regression for spatio-temporal action localization," in *Proc. NIPS*, 2012, pp. 350–358.

[40] J. R. R. Uijlings, K. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.

[41] D. Vázquez, A. M. López, J. Marín, D. Ponsa, and D. Gerónimo, "Virtual and real world adaptation for pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 797–809, Apr. 2014.

[42] J. Yang, R. Yan, and A. Hauptmann, "Cross-domain video concept detection using adaptive SVMs," in *Proc. ACM Multimedia*, 2007, pp. 188–197.

[43] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and F.-F. Li, "Human action recognition by learning bases of action attributes and parts," in *Proc. IEEE ICCV*, Nov. 2011, pp. 1331–1338.



**Fahad Shahbaz Khan** received the M.Sc. degree in intelligent systems design from the Chalmers University of Technology, Sweden, and the Ph.D. degree in computer vision from the Autonomous University of Barcelona, Spain. From 2012 to 2014, he was a Post Doctoral Fellow with the Computer Vision Laboratory, Linköping University, Sweden. He is currently a Research Fellow with the Computer Vision Laboratory, Linköping University. His research interests are in color for computer vision, object recognition, action recognition, and visual tracking. He has authored articles in high-impact computer vision journals and conferences in these areas.

**Jiaolong Xu** received the B.Sc. degree in information engineering and the M.Sc. degree in information and communication engineering from the National University of Defence Technology, China, in 2008 and 2010, respectively. He is currently pursuing the Ph.D. degree with the Advanced Driver Assistance Systems Group, Computer Vision Center, Universitat Autònoma de Barcelona. His research interests include pedestrian detection, virtual worlds, and machine learning.

**Joost van de Weijer** received the M.Sc. degree in applied physics from the Delft University of Technology, in 1998, and the Ph.D. degree from the University of Amsterdam, in 2005. From 2005 to 2007, he was a Marie Curie Intra European Fellow with the LEAR Team, INRIA Rhone-Alpes, France. From 2008 to 2012, he was a Ramon y Cajal Fellow with the Universidad Autonoma de Barcelona. He is currently a Senior Scientist with the Computer Vision Center Barcelona and a member of the LAMP Team. His main research is on the usage of color information in computer vision application.

**Andrew D. Bagdanov** received the Ph.D. degree in computer science from the University of Amsterdam. He is currently a Ramón y Cajal Fellow with the Computer Vision Center, Universitat Autònoma de Barcelona. His research spans a broad spectrum of computer vision, image processing, and machine learning.

**Rao Muhammad Anwer** received the master's degree in intelligent systems design from the Chalmers University of Technology, Sweden, and the Ph.D. degree in computer vision from the Autonomous University of Barcelona, Spain. He is a Post-Doctoral Research Fellow with the Department of Information and Computer Science, Aalto University School of Science, Finland. His research interests are in object detection, pedestrian detection, and action recognition.

**Antonio M. Lopez** received the B.Sc. degree in computer science from the Universitat Politcnica de Catalunya, in 1992, the M.Sc. degree in image processing and artificial intelligence from the Universitat Autnoma de Barcelona (UAB), in 1994, and the Ph.D. degree from UAB, in 2000. In 1996, he participated in the foundation of the Computer Vision Center at UAB, where he has held different institutional responsibilities, presently being the responsible for the research group on advanced driver assistance systems by computer vision. He has been responsible for public and private projects, and has co-authored over 100 papers all in the field of computer vision. Since 1992, he has been giving lectures with the Computer Science Department, UAB, where he is currently an Associate Professor.