Weakly supervised learning of interactions between humans and objects

Alessandro Prest, Cordelia Schmid, Vittorio Ferrari

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

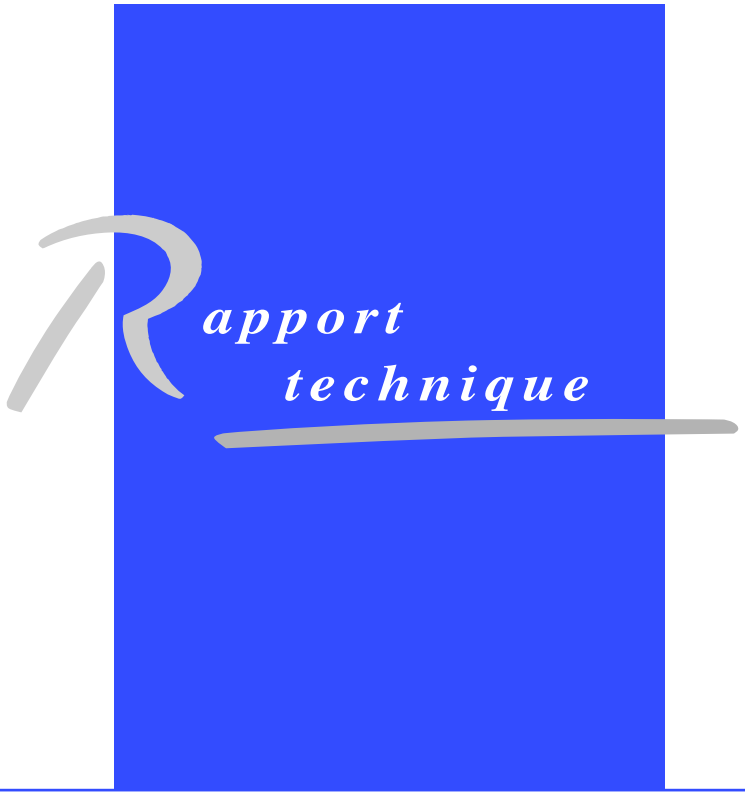
Weakly supervised learning of interactions between humans and objects

Alessandro Prest — Cordelia Schmid — Vittorio Ferrari

N° 0391

Septembre 2010

Vision, Perception and Multimedia Understanding



ISSN 0249-0803 ISRN INRIA/RT--0391--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE - RHÔNE-ALPES

Weakly supervised learning of interactions between humans and objects

Alessandro Prest* , Cordelia Schmid† , Vittorio Ferrari‡

Theme : Vision, Perception and Multimedia Understanding
Perception, Cognition, Interaction
Équipes-Projets LEAR

Rapport technique n° 0391 — Septembre 2010 — 23 pages

Abstract: We introduce a weakly supervised approach for learning human actions modeled as interactions between humans and objects. Our approach is human-centric: we first localize a human in the image and then determine the object relevant for the action and its spatial relation with the human. The model is learned automatically from a set of still images annotated *only* with the action label. Our approach relies on a human detector to initialize the model learning. For robustness to various degrees of visibility, we build a detector that learns to combine a set of existing part detectors. Starting from humans detected in a set of images depicting the action, our approach determines the action object and its spatial relation to the human. Its final output is a probabilistic model of the human-object interaction, i.e. the spatial relation between the human and the object. We compare experimentally to [1] and [2] on the action classification dataset from [1] and also present results on a new human-object interaction dataset.

Key-words: Action Recognition, Weak Supervision

* Computer Vision Laboratory at ETH Zurich, LEAR team at INRIA Grenoble
† LEAR team at INRIA Grenoble.
‡ Computer Vision Laboratory at ETH Zurich

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Résumé :

Mots-clés :

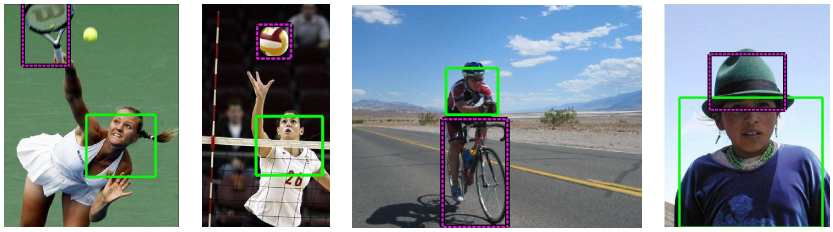


Figure 1: *Example results of our approach showing the automatically detected human (green) and the automatically detected object (pink).*

1 Introduction

Human action recognition is one of the most challenging problems in computer vision. It is important for a wide range of applications, such as video indexing and surveillance, but also image search. It is a challenging task due to the variety of human appearances and poses. Most existing methods for action recognition either learn a spatio-temporal model of an action [3, 4, 5] or are based on human pose [6, 7]. Spatio-temporal models measure the motion characteristics for a human action. They are, for example, based on bags of space-time interest points [3, 8, 9] or represent the human action as a distribution over motion features localized in space and time [5, 4, 10]. Pose-based models learn the characteristic human poses from still images. The pose can, for example, be represented by a histogram-of-gradient (HOG) [7, 11] or based on shape correspondences [6].

Our approach, in contrast, defines an action as the interaction between a human and an object. Interactions are often the main characteristic of an action (fig. 1). For example, the action 'tennis serve' can be described as a human holding a tennis racket in a certain position. Characteristic features are the object *racket* and its spatial relation to the human. Similarly, the actions 'riding bike' and 'wearing a hat' are defined by an object and its relation to the human.

In this paper we introduce a weakly supervised approach for learning interaction models between humans and objects from a set of images depicting an action. We automatically localize the relevant object as well as its spatial relation to the human (fig. 1). Our approach is weakly supervised in that it can learn from images annotated only with the action label, without being given the location of humans nor objects.

Most related to our approach are the works of Yao et al. [2] and Gupta et al. [1] who also learn human-object spatial interactions. However, these approaches operate in a fully supervised setting, requiring training images with annotated object locations as well as human silhouettes [1] or limb locations [2]. Another work by Yao et al. [12] deals with a somewhat different formulation of the human action recognition problem. Their goal is to discriminate subtle situations where a human is holding an object without using it versus a human performing a particular action with the object (e.g. 'holding a violion' vs 'playing a violin'). Moreover this model requires manually localized humans both at training and testing time .

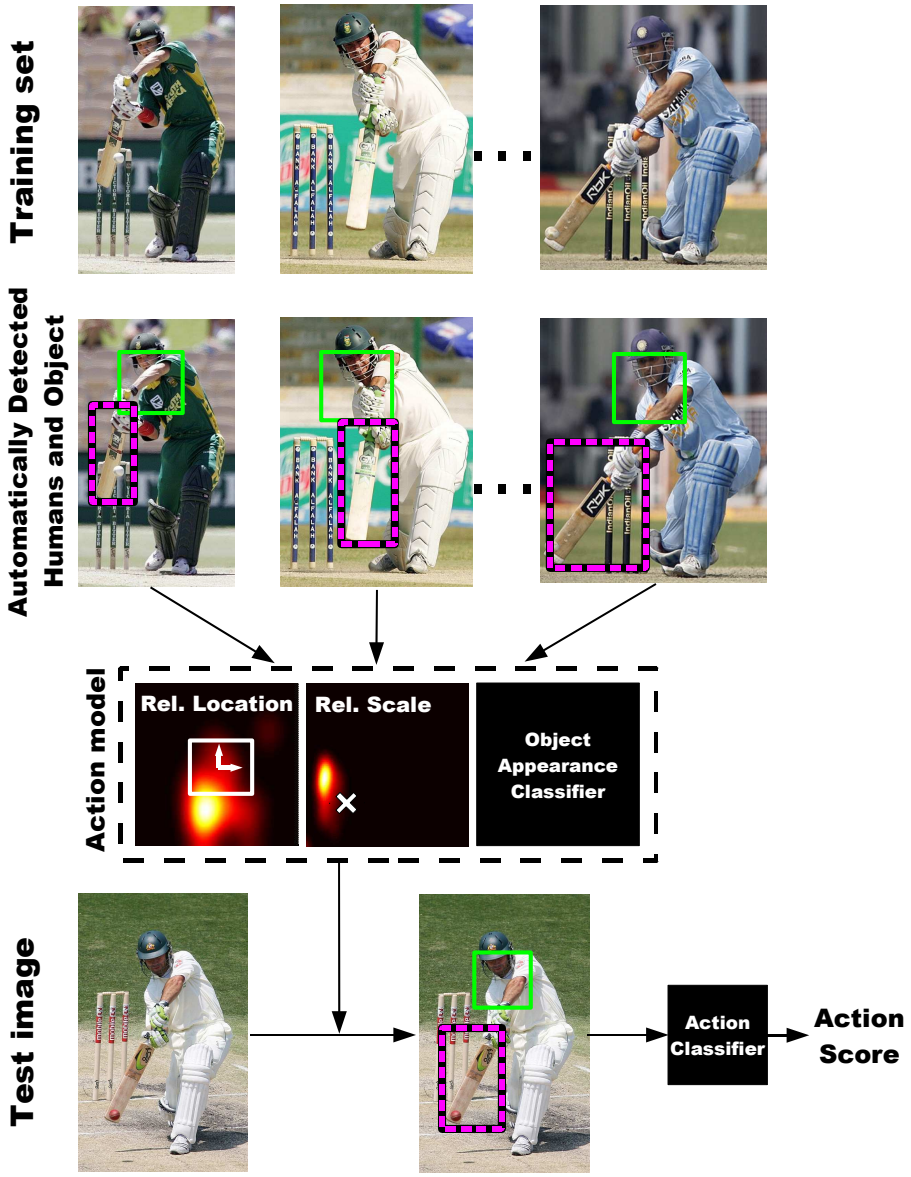


Figure 2: *Overview of our approach. See main text for details.*

Also related is the work of [13], who model spatial relations between object classes such as cars and motorbikes, but in a fully supervised setting, and for object localization rather than action recognition.

1.1 Overview of the method

1.1.1 Training

Our method takes as input a set of training images showing the humans performing the action. Our approach runs over the following stages (fig. 2):

(1) Detect humans in the training set (sec. 2). Our overall detector combines several detectors for different human parts, including face, upper-body, and fully body. This improves coverage as it can detect human at varying degrees of visibility. The detector provides the human reference frame necessary for modeling the spatial interaction with the object in stages (2) and (3).

(2) Localize the action object on the training set (sec. 3.1). The basic idea is to find an object recurring over many images at similar relative positions with respect to the human and with similar appearance between images. Related to our approach are weakly supervised methods for learning object classes [14, 15, 16], which attempt to find objects as recurring appearance patterns.

(3) Given the localized humans and objects from stages (1) and (2), learn the probability distribution of human-object spatial relations, such as relative location and relative size. This defines the human-object interaction model (sec. 3.4). Additionally we learn an object appearance classifier based on the localized objects from (2). This appearance classifier together with the human-object interaction model constitute the action model.

(4) Based on the information estimated in steps 1-3, we train a binary action classifier to decide whether a novel test image contains an instance of this action class (sec. 4).

1.1.2 Testing

Given a novel test image $\mathcal{I}$ and n different action models learned in the previous subsection, we want to assign one of the n possible action labels to $\mathcal{I}$ (fig. 2 bottom):

(1) Detect the single most prominent human in $\mathcal{I}$.

(2) For each action model, find the best fitting location for the action object given the detected human, the human-object interaction model and the object appearance classifier.

(3) Compute different features based on the information extracted in (1) and (2).

(3) Classify $\mathcal{I}$ in an action class, based on the information estimated in steps (1) and (4) (sec. 4). This uses the n classifiers trained in sec. 1.1.1 stage (4).

1.2 Overview of the experiments

In sec. 5 we present experiments on the dataset of Gupta et al. [1] and on a new human-object interaction dataset. The new dataset and the corresponding annotations will be made available online upon acceptance of this paper. The experiments show that our method, learning with weak supervision only, obtains classification performance comparable to both [1] and [2]. This despite using only action labels for training, which is a far less supervision than what required by [1] and [2]. Moreover, our model learns meaningful human-object spatial relations.

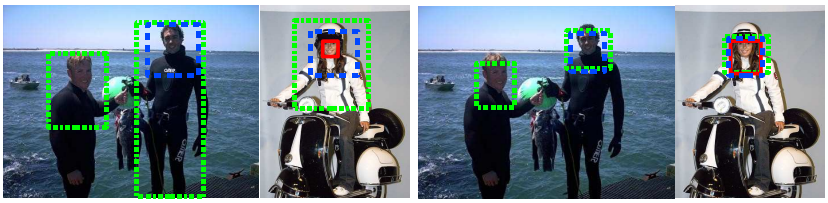


Figure 3: *Left: Detection windows returned by the individual detectors (Green: FB + UB1, Blue: UB2, Red: F). Right: corresponding regressed windows.*

2 A part-based human detector

In real world images of human actions the person can be fully or partially visible (fig. 6, 10 and 11). In this context a single detector (full person, an upper-body or face) is insufficient. Our detector build on the one by Felzenszwalb et al. [17]; it trains several detectors for different human parts, adds a state-of-the-art face detector and learns how to combine the different part detectors. Our combination strategy goes beyond the maximum score selection strategy of [17] and is shown experimentally to outperform their approach (sec. 2.5). Furthermore, it provides the human reference frame necessary for modeling the spatial interaction with the object.

2.1 Individual part detectors.

We use four part detectors: one for the full human body (FB), two for the upper-body (UB1, UB2) and one for the face (F). For the fully body detector (FB) and the first upper-body detector (UB1) we use the two components of the human detector by [17][1] learnt on the the PASCAL VOC07 training data [18]. Note that we use the two components as two separate part detectors. For the second upper body detector (UB2) we train [17] on another dataset of near-frontal upper-bodies [19][2]. Therefore, UB2 is specialized to the frontal case, which occurs frequently in real images. Our experiments show UB2 to provide detections complementary to UB1 (sec. 2.5).

For the face detector (F) we use the technique of [20], which is is similar to the popular Viola-Jones detector [21], but replaces the Haar features with local binary patterns, providing better robustness to illumination changes [22]. The detector is trained for both front and side views.

2.2 Mapping to a common reference frame.

As the detection windows returned by different detectors cover different areas of the human body, they must be mapped to a common reference frame before they can be combined. Here we learn regressors for this mapping (fig. 3).

For each part detector we learn a linear regressor $\mathrm{R}(w, p)$ mapping a detection window w to a common reference frame. A regressor R is defined by

$$\mathrm{R}(w, p) = (x - Wp_1,\ y - Hp_2,\ Wp_3,\ Wp_3p_4) \tag{1}$$

[1] Code available at http://people.cs.uchicago.edu/∼pff/latent.
[2] Data available at http://www.robots.ox.ac.uk/∼vgg/software/UpperBody.

where $w = (x, y, W, H)$ is a detection window defined by the top-left co-ordinates (x, y), its width W and its height H. The regression parameters $p = (p_1, p_2, p_3, p_4)$ are determined from the training data as follows.

We have a set of n training pairs of detection windows w^i and corresponding manually annotated ground-truth reference windows h^i. We find the optimal regression parameters p^* as

$$p^* = \arg\max_p \sum_{i=1}^{n} \mathrm{IoU}(h^i, \mathrm{R}(w^i, p)) \tag{2}$$

where $\mathrm{IoU}(a, b) = |a \cap b|/|a \cup b|$ is the intersection-over-union between two windows a, b. The optimal parameters p^* assure the best overlap between the mapped detection windows $\mathrm{R}(w^i, p)$ and the ground-truth references h^i.

Fig. 4 shows an example of the original stickman annotation and the common reference frame derived from it. The height of the reference frame is given by the distance between the top point of the head stick and the mid point of the torso stick. The width is fixed to 90% of the height.

2.3 Clustering part detections.

After mapping detection windows from the part detectors to a common reference frame, detections of the same person result in similar windows. Therefore, we find small groups of detections corresponding to different persons by clustering all mapped detection windows for an image in the 4D space defined by their coordinates.

Clustering is performed with a weighted dynamic-bandwidth mean-shift algorithm based on [23]. At each iteration the bandwidth is set proportionally to the expected localization variance of the regressed windows (i.e. to the diagonal of the window defined by the center of the mean-shift kernel in the 4D space). This automatically adapts the clustering to the growing error of the part detectors with scale.

To achieve high recall it is important to set a very low threshold on the part detectors. This results in many false-positives which cause substantial drift in the traditional mean-shift procedure. To maintain a robust localization, at each iteration we compute the new cluster center as the mean of its members *weighted* by their detection scores. The final mean-shift location in the 4D space also gives a weighted average reference window for each cluster, which is typically more accurately localized than the individual part detections in the cluster.

2.4 Discriminative score combination.

Given a cluster C containing a set of part detections, the goal is to determine a single combined score for the cluster. Each cluster C has an associated representative detection window computed as the weighted mean of the part detection windows in C.

To compute to the score of a cluster, we use the 4D vector c where each dimension corresponds to one of the detectors. The value of an entry c_d is set to the maximum detection score for detector d within the cluster. If the cluster does not contain a detection for a detector d, we set $c_d = \tau_d$, with τ_d the threshold at which the detector is operating (see sec. 2.5). Given the 4D

score vector for each cluster, we learn a linear SVM to separate positive (human detections) from negative examples. The score for a test image is then the confidence value of the SVM. Section 2.5 explains how we collect positive ($\mathcal{T}^+$) and negative ($\mathcal{T}^-$) training examples. The training set for this score-combiner SVM is the same used to train the regressors.

2.5 Experimental evaluation.

The experimental evaluation is carried out on the ETHZ PASCAL Stickmen dataset [24] [3]. It contains 549 images from the Pascal VOC 2008 person class. In each image, one person is annotated by line segments defining the position and orientation of the head, torso, upper and lower arms (fig. 4). As we want the common reference frame to be visible in most images, we set it as a square window starting from the top of the head and ending at the middle of the torso (fig. 3). Note that this choice has no effect on the combined human detector.

We build our positive training set $\mathcal{T}^+$ out of the first 400 images and use the remaining 149 as a positive test set $\mathcal{S}^+$. The negative examples are obtained from Caltech-101 [25] as well as from PASCAL VOC [26] [27]. We end up with 5158 negative images: 3956 are randomly selected as the negative training set $\mathcal{T}^-$ while the remaining form the negative test set $\mathcal{S}^-$.

The optimal regressor parameters p^* are learnt on the positive training set $\mathcal{T}^+$ (as described in sec.2.2).

Figure 4: *Example of an annotated image from the ETHZ PASCAL Stickmen dataset. Left: the original stickman annotation. Right: the common reference frame we derived from the sticks.*

The score-combiner SVM is trained on the clusters obtained from the entire training set $\mathcal{T}^+ \cup \mathcal{T}^-$. All clusters from $\mathcal{T}^-$ are labeled as negative examples. Clusters from $\mathcal{T}^+$ are labeled as positive examples if their IoU with a ground-truth person is greater than 50%. All other clusters from $\mathcal{T}^+$ are discarded, as their ground-truth label is unknown (although an image in ETHZ PASCAL Stickmen might contain multiple persons, only one is annotated). Note that before clustering we only keep detections scoring above a low threshold τ_d, such as to remove weak detections likely to be false positives.

Fig. 5 shows a quantitative evaluation on our test set $\mathcal{S}^+ \cup \mathcal{S}^-$ as a precision-recall curve. The recall axis indicates the percentage of annotated humans that were correctly detected (true positives, IoU with the ground-truth greater than

[3] Available at http://www.vision.ee.ethz.ch/∼calvin/datasets.html.

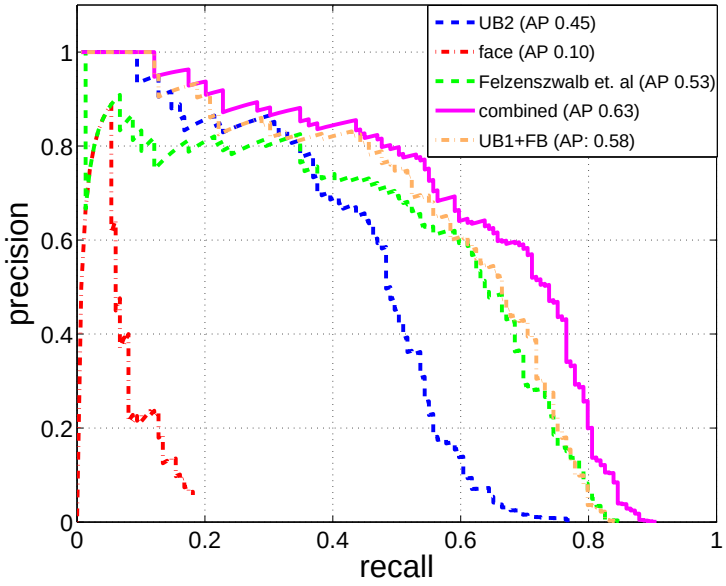


Figure 5: *Precision-recall curve for the individual detectors and the combined ones. We consider a detection as correct when the Intersection-Over-Union (IoU) with a ground-truth annotation is at least 50%. In parenthesis are average precision values (AP), defined as the area under the respective curve.*

50%). All detections in $\mathcal{S}^-$ are counted as false positives. Notice how in $\mathcal{S}^+$ only one human per image is annotated. Hence, only true positives in $\mathcal{S}^+$ are counted in the evaluation and all other detections are discarded, as their ground-truth label is unknown. Precision is defined as the ratio between the number of true positives and the total number of detections at a certain recall value.

Our combined human detector UB1+FB+UB2+F brings a considerable increase in average precision compared to the state-of-the-art human detector of [17], which it incorporates. For a fair comparison, its detection windows are also regressed to a common reference frame (using the same regressor as in our combined detector).

Note that the person model of [17] uses its two components (FB and UB1) in a 'max-score-first' combination: if two detections from the two different components overlap by more than 50% IoU, then the the lower scoring one is discarded. In the experiment UB1+FB we use our novel combination strategy to combine only the two components UB1 and FB. This performs significantly better than the original model [17], further demonstrating the power of our combination strategy. In all experiments all detection windows are regressed to the same common reference frame as ours.

Although the face detector performs much below the other detectors, it is valuable in close-up images, where the other detectors do not fire.

3 Learning human-object interactions

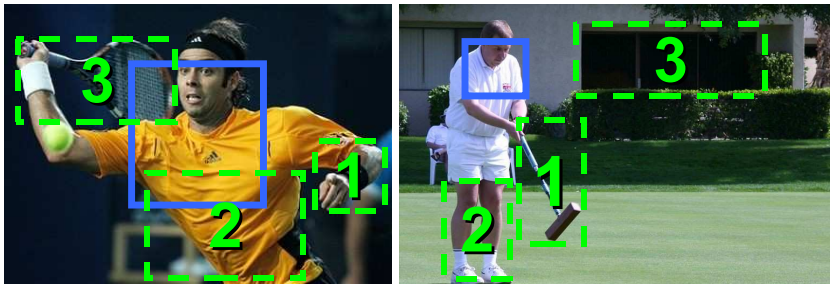


Figure 6: *Two images with three candidate windows each. The blue boxes indicate the location of the human calculated by the detector. The green boxes show possible action object locations.*

This section presents our human-object interaction model and how to learn it from weakly supervised images. The goal is to automatically determine the object relevant for the action as well as its spatial relation to the human. The intuition behind our human-object model is that some geometric properties relating the human to the action object are stable across different instances of the same action. Let's imagine a human playing a trumpet: the trumpet is always at approximately the same relative distance with respect to the human. We model this intuition with spatial cues involving the human and the object. We measure them relative to the position and scale of the reference frame provided by the human detector from sec. 2. This makes the cues comparable between different images.

Our model (subsec. 3.1) incorporates several cues (subsec. 3.3). Some relate the human to the object while others are defined purely by the appearance of the object. Once the action objects have been localized in the images, we use them together with the human locations to learn probability distributions of human-object spatial relations (subsec. 3.4). Experimental results show that these relations are characteristic for the action, e.g. a bike is below the person riding it, whereas a hat is on top of the person wearing it (sec. 5). These distributions constitute our human-object interaction model.

3.1 The Human-Object model

Our model inputs a set of training images $\{\mathcal{I}^i\}$ showing an action (e.g. 'tennis forehand' (fig. 6 left) and 'croquet' (fig. 6 right)). We retain for each image i the single highest-scored human detection h^i, and use it as an anchor for defining the human-object spatial relations. Furthermore, for each $\mathcal{I}^i$ we have a set $\mathcal{X}^i = \{b^i_j\}$ of candidate windows potentially containing the action object (fig. 6). We use the generic object detector [28] to select 500 windows likely to contain an objects rather than background (sec. 3.2).

Our goal is to select one window $b^i_j \in \mathcal{X}^i$ containing the action object for each image $\mathcal{I}^i$. We model this selection problem in energy minimization terms. Formally, the objective is to find the configuration $\mathcal{B}^*$ of windows (one window per image), so that the following energy is minimized

$$E(\mathcal{B}|\mathcal{H},\Theta) = \sum_{b_j^i \in \mathcal{B}} \Theta_U(h^i, b_j^i) + \sum_{(b_j^i, b_m^l) \in \mathcal{B}\times\mathcal{B}} \Theta_H(b_j^i, b_m^l, h^i, h^l) + \sum_{(b_j^i, b_m^l) \in \mathcal{B}\times\mathcal{B}} \Theta_P(b_j^i, b_m^l) \tag{3}$$

We give here a brief overview of the terms in this model, and explain them in more detail in sec. 3.3.

Θ_U is a sum of unary cues measuring (i) how likely a window b_j^i is to contain an object of any class ($\theta_o(b_j^i)$); (ii) the amout of overlap between the window and the human ($\theta_a(h^i, b_j^i)$)

$$\Theta_U(h^i, b_j^i) = \theta_o(b_j^i) + \theta_a(h^i, b_j^i) \tag{4}$$

Θ_H is a sum of pairwise cues capturing spatial relations between the human and the object. They encourage the model to select windows with similar spatial relations to the human across images (e.g. Δ_d measures the difference in relative distance between two human-object pairs). These cues are illustrated in fig. 8.

$$\begin{aligned}\Theta_H(b_j^i, b_m^l, h^i, h^l) &= \Delta_d(b_j^i, b_m^l, h^i, h^l) + \Delta_s(b_j^i, b_m^l, h^i, h^l) \\ &+ \Delta_l(b_j^i, b_m^l, h^i, h^l) + \Delta_o(b_j^i, b_m^l, h^i, h^l)\end{aligned} \tag{5}$$

Finally, Θ_P is a sum of pairwise cues measuring the appearance similarity between pairs of candidate windows in different images. These cues prefer $\mathcal{B}^*$ to contain windows of similar appearance across images. They are χ^2 distances on color histograms (Δ_c) and bag-of-visual-words descriptors (Δ_i).

$$\Theta_P(b_j^i, b_m^l) = \Delta_c(b_j^i, b_m^l) + \Delta_i(b_j^i, b_m^l) \tag{6}$$

We normalize the range of all cues to $[0, 1]$ but do not perform any other reweighting beyond this.

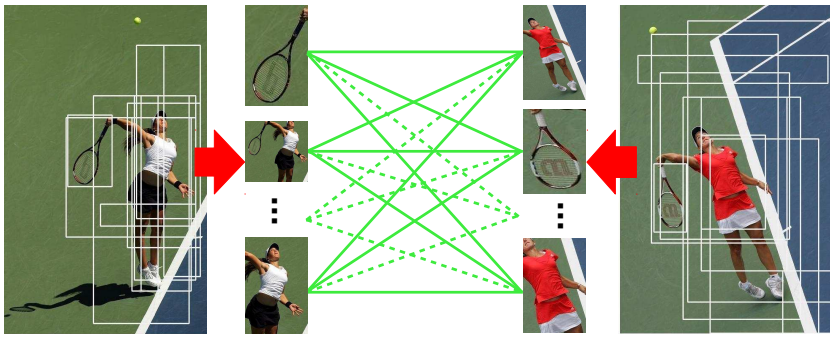


Figure 7: *A pair of training images from the 'tennis serve' action. Candidate windows are depicted as white boxes. We employ a fully connected model, meaning that pairwise potentials (green lines) connect each pair of candidate windows between each pair of training images.*

As the pairwise terms connect all pairs of images, our model is fully connected. Every candidate window in an image is compared to every candidate window in another. Fig. 7 shows an illustration of the connectivity in our model. We perform inference on this model using the TRW-S algorithm [29] obtaining a very good approximation of the global optimum $\mathcal{B}^* = \arg\min E(\mathcal{B}|\mathcal{H}, \Theta)$.

3.2 Candidate Windows

To obtain the candidate windows $\mathcal{X}$ and the unary cue θ_o we use the objectness measure of [28], which quantifies how likely it is for a window to contain an object of *any* class rather than background. Objectness is trained to distinguish windows containing an object with a well-defined boundary and center, such as cows and telephones, from amorphous background windows, such as grass and road. Objectness combines several image cues measuring distinctive characteristics of objects, such as appearing different from their surroundings, having a closed boundary, and sometimes begin unique within the image.

We use objectness as a location prior in our model, by evaluating it for all windows in an image and then sampling 500 windows according to their scores. These form the set of states for a node, i.e. the candidate windows the model can choose from.

This procedure brings two advantages. First, it greatly reduces the computational complexity of the optimization, which grows with the square of the number of windows (there are millions of windows in an image). Second, the sampled windows and their scores θ_o attract the model toward selecting objects rather than background windows.

For the experiments we used the code of [28] available online [4] without any modifications or tuning. It takes only about 3 seconds to compute candidate windows for one image.

3.3 Cues

Unary cues.

Each candidate window b is scored separately by the unary cues θ_o and θ_a.

The cue $\theta_o(b) = -\log(p_{obj}(b))$, where $p_{obj}(b) \in [0, 1]$ is the objectness probability [28] of b which measures how likely b is to contain an object of any class (sec. 3.2).

The cue $\theta_a(h, b) = -\log(1 - \text{IoU}(h^i, b^i_j))$ measures the overlap between a candidate window and the human h (with $\text{IoU}(\cdot, \cdot) \in [0, 1]$). It penalizes windows with a strong overlap with the human, since in most images of human-object interactions the object is near the human, but not on top of it. This cue proved to be very succesful in suppressing trivial outputs such as selecting a window covering the human upper-body in every image, i.e. is the most frequently recurring pattern in human action datasets.

Human-object pairwise cues.

Candidate windows from two different images $\mathcal{I}^i, \mathcal{I}^l$ are pairwise connected as shown in fig. 7. Human-object pairwise cues compare two windows b^i_j, b^l_m ac-

[4] Source code at www.vision.ee.ethz.ch/∼calvin/software.html.

cording to different spatial layout cues. We define 4 cues measuring different

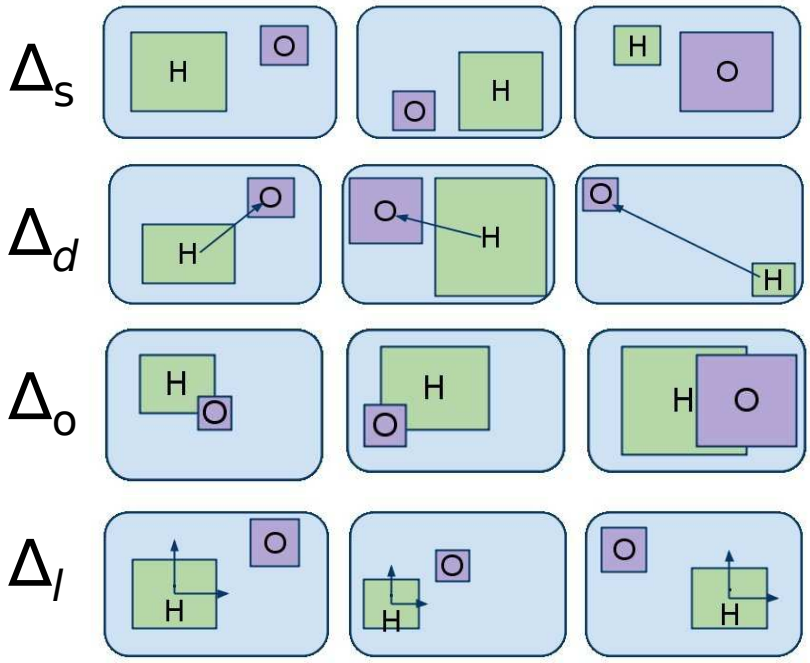


Figure 8: *For each human-object cue we show three possible configurations of human-object windows. The two left-most configurations have a low pairwise energy, while the right-most has a high energy compared to any of the first two.*

spatial relations between the human and the object (fig. 8). These cues prefer pairs of candidate windows with a similar spatial relation to the human in their respective images. Such recurring spatial relations are characteristic for the kind of human-object interactions we are interested in (e.g. tennis serve).
Let

$$\mathrm{l}(b_j^i, h^i) = ((x_j^i - x^i)/W^i \ , \ (y_j^i - y^i)/H^i) \tag{7}$$

be the 2D location $\mathrm{l}(b_j^i, h^i)$ of a candidate object window $b_j^i = (x_j^i, y_j^i, W_j^i, H_j^i)$ in the reference frame defined by the human $h^i = (x^i, y^i, W^i, H^i)$ in image $\mathcal{I}^i$. With this notation, the four cues are

1) The difference in the relative scale between the object and the human in the two images

$$\begin{aligned}\Delta_s(b_j^i, b_m^l, h^i, h^l) = \max(\mathrm{a}(h^i, b_j^i)/\mathrm{a}(h^l, b_m^l), \\ \mathrm{a}(h^l, b_m^l)/\mathrm{a}(h^i, b_j^i)) - 1\end{aligned} \tag{8}$$

where

$$\mathrm{a}(h^i, b_j^i) = \mathrm{area}(b_j^i)/\mathrm{area}(h^i) \tag{9}$$

is the ratio between the area (in pixels) of a candidate window and the human window.

2) The difference in the Euclidean distance between the object and the human

$$\Delta_d(b_j^i, b_m^l, h^i, h^l) = \text{abs}\left(\left|\left|\text{l}(b_j^i, h^i)\right|\right| - \left|\left|\text{l}(b_m^l, h^l)\right|\right|\right) \tag{10}$$

3) The difference in the overlap area between the object and the human (normalized by the area of the human)

$$\Delta_o(b_j^i, b_m^l, h^i, h^l) = \text{abs}\left(\frac{b_j^i \cap h^i}{\text{area}(h^i)} - \frac{b_m^l \cap h^l}{\text{area}(h^l)}\right) \tag{11}$$

where $a \cap b$ indicates the overlapping area (in pixel) between two windows a and b.

4) The difference in the relative location between the object and the human

$$\Delta_l(b_j^i, b_m^l, h^i, h^l) = \left|\left|\text{l}(b_j^i, h^i) - \text{l}(b_m^l, h^l)\right|\right| \tag{12}$$

Object-only pairwise cues.

The similarity $\Theta_P(b_j^i, b_m^l)$ between a pair of candidate windows b_j^i, b_m^l from two images is computed as the χ^2 difference between histograms describing their appearance. We use two descriptors. The first is a a color histogram $\Delta_c(b_j^i, b_m^l)$. The second is a bag-of-visual-words on a 3-level spatial pyramid using SURF features [30] $\Delta_i(b_j^i, b_m^l)$ (whose vocabulary is learnt from the positive training images and is composed of 500 visual words). These cues prefer object windows with similar appearance across images.

3.4 Learning Human-Object interactions

Given the human detections $\mathcal{H}$ and the object windows $\mathcal{B}^*$ minimizing equation (3), we learn the interactions between the human and the action object as two relative spatial distributions. More precisely, we focus on relative location (eq. (7)) and relative scale (eq. (9)).

We estimate a 2D probability density function for the location of the object with respect to the human (eq. 7) as:

$$\text{k}_l(\mathcal{B}^*, \mathcal{H}) = \sum_i \frac{1}{\sqrt{2\sigma}} e^{-\text{l}(b^i, h^i)(1/2\sigma^2)} \tag{13}$$

where $b^i \in \mathcal{B}^*$ is the selected object window in image $\mathcal{I}^i$, $h^i \in \mathcal{H}$ is the reference human detection in that image, and the scale σ is set automatically by a diffusion algorithm [31].
A second density is given by the scale of the object relative to the human (eq. (9)):

$$\text{k}_s(\mathcal{B}^*, \mathcal{H}) = \sum_i \frac{1}{\sqrt{2\sigma}} e^{-\text{a}(b^i, h^i)(1/2\sigma^2)} \tag{14}$$

The learnt spatial relations for various actions are presented in subsec. 5.4.

Additionally we train an object appearance classifier θ_t. This classifier is a SVM on a bag-of-words representation [32] using dense SURF descriptors [30].

As positive training samples we use the selected object windows $\mathcal{B}^*$. As negative samples we use random windows from images of other action classes.

The spatial distributions k_l and k_s together with the object appearance classifier θ_t constitute the action model $\mathcal{A} = (k_l, k_s, \theta_t)$.

4 Action recognition

The previous section described how we automatically learn an action model from a set of training images $\{\mathcal{I}\}$. Given a test image $\mathcal{T}$ and n action models $\{\mathcal{A}^a\}_{a=1,\ldots,n}$, we want to determine which action is depicted in it.

In sections 4.1 to 4.3 we present three descriptors, each capturing a different aspect of an image. The human-object descriptor (sec. 4.1) exploits the spatial relations and the object appearance model in $\mathcal{A}$ (sec. 3) to localize the action object and then describes the human-object configuration. Sec. 4.2 and 4.3 present two descriptors capturing contextual information both at a global (sec. 4.2) and a local (sec. 4.3) level. Finally, in sec. 4.4, we show how we combine the different descriptors for classifying $\mathcal{T}$.

4.1 Human-object descriptor

We compute a low-dimensional descriptor for an image (the same procedure is applied equally to either a training or a test image): (1) detect humans and keep the highest scoring one h as anchor for computing Human-Object relations; (2) compute a set of candidate object windows $\mathcal{B}$ using [28] (sec. 3.2); (3) for every action model $\{\mathcal{A}^a\}_{a=1,\ldots,n}$ select the window $b^a \in \mathcal{B}$ minimizing the energy

$$E(\mathcal{B}|h, \mu^a) = \theta^a_t(b) + \theta^a_{k_l}(h, b) + \theta^a_{k_s}(h, b) \tag{15}$$

where $\theta^a_{k_l}(h, b_j)$ and $\theta^a_{k_s}(h, b_j)$ are unary terms based on the probability distributions k_l and k_s learned during training (sec. 3.4); $\theta^a_t(b^i)$ is the object appearance classifier, also learned during training. The optimal window can be found efficiently as the complexity of this optimization is linear in $|\mathcal{B}|$.

For each action model μ^a we create a descriptor vector containing the energy of the three terms in eq. (15), evaluated for the selected window b^a. The overall human-object descriptor for the image is the concatenation over all n actions and has dimensionality $3n$.

4.2 Whole-image descriptor

As shown by [1], describing the whole image using GIST [33] provides a valuable cue for action classification. This descriptor can capture the context of an action, which is often quite distinctive [34].

4.3 Pose-from-gradients descriptor

Both [1] and [2] use human pose as a feature for action recognition. In those approaches pose is represented by silhouettes [1] or limb locations [2], which are expensive to annotate manually on training images. In the same spirit of leveraging on human pose for action classification, but avoiding the additional

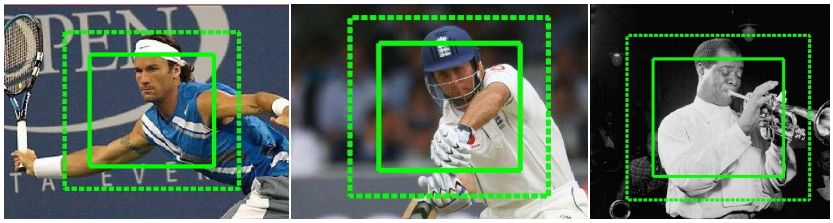


Figure 9: *Human pose has a high discriminative power for distinguishing actions. The solid window is the original human detection, while the dashed window shows the area from which the pose-from-gradients descriptor is extracted.*

annotation effort, we propose a much simpler descriptor to capture pose information.

Given an image and the corresponding human detection h we extract the GIST descriptor [33] from an image window obtained by enlarging h by a constant factor so as to include more of the arm pose. Fig. 9 shows example human detections and the corresponding enlarged windows. While this descriptor does not require any additional supervision on the training images, it proved successful in discriminating difficult cases (see results in sec. 5.3). Moreover, it takes further advantage of using a robust human detector, such as the one in sec. 2.

4.4 Action classifiers

For training, we extract the descriptors of sections 4.1-4.3 from the same training images $\{I^i\}$ used for learning the human-object model (notice how only the action class label is necessary as supervision, and not human or object bounding-boxes [1, 2], human silhouettes [1], or limb locations [2]). We obtain a separate RBF kernel for each descriptor and then compute a linear combination of them. Given the resulting combined kernel we learn a multi-class SVM. The combination weights are set by cross validation to maximize the classification accuracy [35].

Given a new test image $\mathcal{T}$, we compute the three descriptors and average the corresponding kernels according to the weights learned at training time. Finally we classify $\mathcal{T}$ (i.e. assign $\mathcal{T}$ an action label) according the multi-class SVM learned during training.

5 Experimental Results

We present action recognition results on two datasets: the 6 sports actions of [1] and a new dataset of 3 actions we collected, called the *Trumpets, Bikes and Hats* (TBH) dataset. The TBH dataset and the corresponding annotations will be released online upon acceptance of this paper. Section 5.1 describes the datasets. Section 5.2 presents the experimental setup, namely the two levels of supervision we evaluate on. Section 5.3 reports quantitative results and comparisons to [1] and [2]. The learned human-object interactions are illustrated in sec. 5.4.

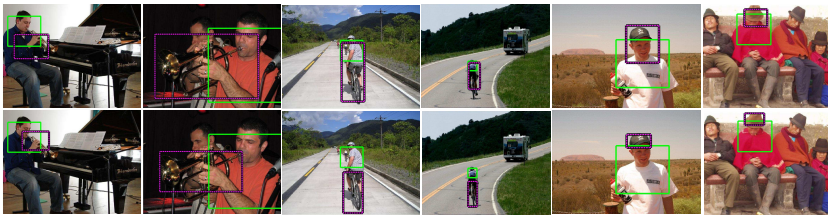


Figure 10: *Example results on the TBH dataset for test images that were correctly classified by our approach. Two images are shown for each action class (from left to right, 'playing trumpet', 'wearing hat' and 'riding bike'). First row: results from the weakly supervised setting WS. Second row: results from the fully supervised setting FS.*

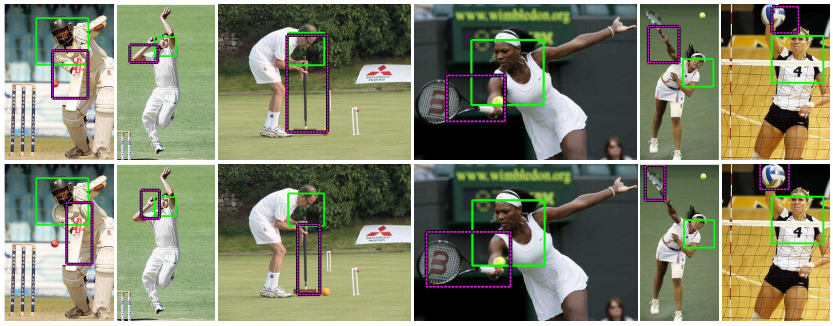


Figure 11: *Example results from the sports dataset of [1] for test images that were correctly classified by our approach.. One image per class is shown (from left to right: 'cricket batting', 'cricket bowling', 'croquet', 'tennis forehand', 'tennis serve' and 'volleyball'). First row: weakly supervised setting setting. Second row: fully supervised setting.*

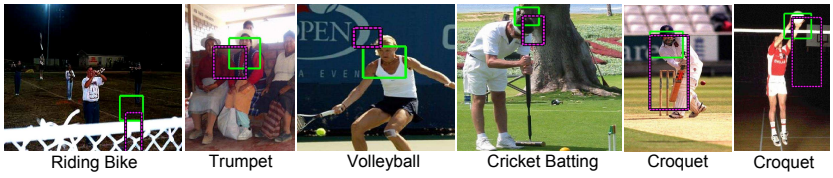


Figure 12: *Example failures of our method in the weakly supervised setting. The action labels indicate the (incorrect) classes the images were assigned to. The main reasons are: missed humans due to tilted pose or poor visibility (first, fourth and sixth image), similarities between different action classes (fifth image), truncation or poor visibility of the action object (second and third image).*

5.1 Datasets

TBH dataset.

We introduce a new action dataset called TBH. It is built from Google Images and the IAPR TC-12 dataset [36], and contains 3 actions: 'playing trumpet', 'riding bike', and 'wearing hat'.

Table 1: *Classification results on the sports dataset [1]: 1st row: our method with WS; 2nd row: our method with FS; 3rd row: Gupta [1] with FS-[1]; 4th row: Yao [2] with FS-[2] (they only report results for their full model). Each entry is the classification accuracy averaged over all 6 classes. Column 'Full model' in rows 1 and 2 includes our Human-Object spatial relations.*

	Human pose	Pose from Gradients	Object appearance classifier	Whole-scene	Pose from Grad. + Whole-scene + Obj. appear. class.	Full model
Ours WS	-	**54**	**32**	**67**	**76**	**81**
Ours FS	-	**58**	**46**	**67**	**80**	**83**
Gupta [1] FS-[1]	**58**	-	-	**66**	-	**79**
Yao [2] FS-[2]	-	-	-	-	-	**83**

Table 2: *Classification results on the TBH human action dataset: (first row) our method with weak supervision, (second row) our method with full supervision. See text for details.*

	Pose from Gradients	Object appearance classifier	Whole-scene	Pose from Grad. + Whole-scene + Obj. appear. class.	Full model
Ours WS	**54**	**58**	**58**	**71**	**74**
Ours FS	**58**	**61**	**58**	**74**	**79**

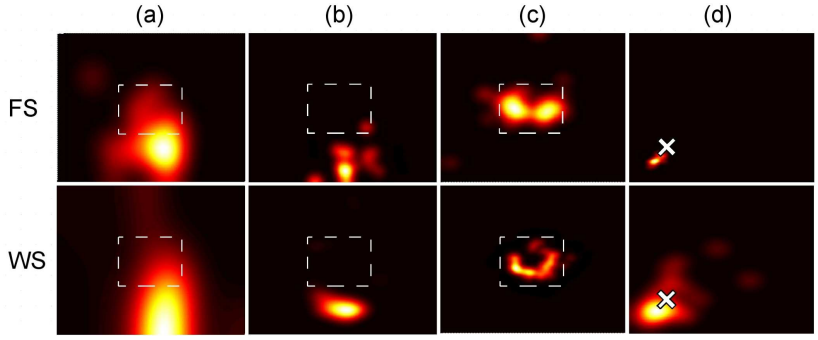


Figure 13: *Human-object spatial distributions learned in the FS setting (top) and in the WS setting (bottom) . (a)-(c): relative location of the action object wrt the human (k_l in sec. 3.4). Dashed boxes indicate the size and location of the human windows. (a) 'Cricket Batting', (b) 'Croquet', (c) 'Playing Trumpet'. (d): distribution of the object scale relative to the human scale for the action 'Volleyball' (k_s in sec. 3.4). The horizontal axis represents the x-scale and the vertical the y-scale. A cross indicates the scale of the human.*

We use Google Images to retrieve images for the action 'playing trumpet'. We manually select the first 100 images depicting the action in a set of images obtained by searching for "person OR man OR woman", followed by the action verb ("playing") and the object name ("trumpet"). The amount of negative images that have been manually discarded has been 25%. We split these 100

positive images into training (60) and testing (40), i.e. the same proportions as the sports dataset [1].

For the actions 'riding bike' and 'wearing hat' we collected images from the IAPR TC-12 dataset. Each image in this large dataset has an accompanying text caption describing the image. We run a natural language processor (NLP) [37] on the text captions to retrieve images showing the action. In detail, a caption should contain: (i) a subject, specified as either 'person', 'man', 'woman', or 'boy'; (ii) a verb-object pair. The verb is specified in the infinitive form, while the object as a set of synonyms (e.g 'hat' and 'cap'). Due to the high quality of the captions, this process returns almost only relevant images. We manually removed just 1 irrelevant image from each class. The resulting dataset contains 117 images for 'riding bike' (70 training, 47 testing) and 124 images for 'wearing hat' (74 training, 50 testing). In the resulting TBH dataset, images are only annotated by label of the action class they depict.

Sports dataset [1].

This dataset is composed of 6 actions of people doing sports. These actions are: 'cricket batting', 'cricket bowling', 'croquet', 'tennis forehand', 'tennis backhand' and 'volleyball smash'. Each action has 30 training images and 20 test images. These images come with a rich set of annotations. The approaches of [1] and [2] are in fact trained with full supervision, using all these annotations. More precisely, for each training image they need: (i) action label; (ii) ground-truth bounding-box for the action object; (iii) manually segmented human silhouette [1] or limb locations [2]. Moreover, [1] also requires (iv) a set of training images for each action object, collected from Google Images (e.g. by querying for 'tennis racket' and then manually discarding irrelevant images).

5.2 Experimental setups

Weakly supervised (WS).

Our method learns human actions from images labeled only with the action they contain, i.e. weakly supervised images (WS).

At training time we localize objects in the training set by applying the model presented in 3. Given the localized objects and the humans locations we learn spatial relations as well as an object appearance classifier (sec. 3.4).

At test time we recognize human actions in test images by applying the procedure described in 4.

Fully supervised (FS).

In order to fairly compare our approach with [1] and [2], we introduce a fully supervised variant of our model, where we use (i) and (ii). Instead of (iii) we just use ground-truth bounding-boxes on the human, which is less supervision than silhouettes [1] or limb locations [2]. It is then straightforward to learn the human-object relation models and the object appearance classifier (sec. 3.4) from these ground-truth bounding-boxes. We also train a sliding-window detector [17] for the action object using the ground-truth bounding-boxes (ii). This detector then gives the appearance cue θ_t in eq. 15.

In the following we denote with FS our fully supervised setting using one human bounding-box and one object bounding box per training image. Instead, we denote by FS-[2] the setting using (i)-(iii) and FS-[1] the setting using (i)-(iv).

In the FS setup, we recognize human actions in test images by applying the procedure described in 4. In step (2) of sec. 4.1 we run the action object detector to obtain candidate windows $\mathcal{B}$, i.e. all windows returned by the detector, without applying any threshold nor non-maxima suppression.

5.3 Experimental evaluation

Table 1 presents results on the sports dataset [1], where the task is to classify each test image into one of six actions. In the WS setup (first row), combining the object appearance classifier (sec. 3.4), the pose-from-gradients descriptor and the whole-image classifier improves over using any of them alone and already obtains good performance (76%). Importantly, adding the human-object interaction model ('Full model' column) raises performance to 81%, confirming that our model learns human-object spatial relations beneficial for action classification. Fig. 10 and fig. 11 show humans and objects automatically detected on the test images by our full method.An important point is that the performance of our model trained in the WS setup is 2% better than the FS-[1] approach of [1] and 2% below the FS-[2] approach of [2]. This confirms the main claim of the paper: our method can effectively learn actions defined by human-object interactions in a WS setting. Remarkably, it reaches performance comparable to state-of-the-art methods in FS settings which are very expensive in terms of training annotation.

The second row of table 1 shows results for our method in the FS setup. As expected, the object appearance classifier performs better than the WS one, as we can train it from ground-truth bounding-boxes. Again the combination with the pose-from-gradients descriptor and the whole-scene classifier significantly improves results (now to 80%). Furthermore, also in this FS setup adding the human-object spatial relations raises performance ('Full model'). The classification accuracy exceeds that of [1] and is on par with [2]. We note how [2, 1] use human body part locations or silhouettes for training, while we use only human bounding-boxes, which are cheaper to obtain. Interestingly, although trained with much less supervision, our pose-from-gradients decriptor performs on par with the Human pose descriptor of [1].

Table 2 shows results on the TBH dataset, which reinforce the conclusions drawn on the sports dataset: (i) combining the object appearance classifier, pose-from-gradients and whole-scene classifier is beneficial in both WS and FS setups; (ii) the human-object interaction model brings further improvements in both setups; (iii) the performance of the full model in the WS setup is only 5% below that of the FS setup, confirming our method is a good solution for WS learning.

We note that the performance gap of the object appearance classifier between FS and WS is smaller than on the sports dataset. This might be due to the greater difference between action objects in the TBH dataset, where a weaker object model already works well. Finally, we note how the whole-scene descriptor has lower discriminative power than on the sports dataset (67% across 6 classes vs. 58% across 3 classes). This is likely due to the greater intra-class variability of backgrounds and scenes in the TBH dataset. Fig. 10 and 11 show

example results for automatically localized action objects on the test data from the two datasets. Although, as to be expected, in the FS setup our method localizes the action objects more accurately, in many cases it detects it already well in the WS setup, in spite of having trained without any bounding-box. Failure cases are shown and discussed in fig. 12.

5.4 Learned human-object interactions

Fig. 13 compares human-object spatial relations obtained from automatically localized humans and objects in the WS setup to those derived from ground-truth bounding-boxes in the FS setup (sec. 3.4). The learnt relations are clearly meaningful. The location of the Cricket Bat (first column) is near the chest of the person, whereas the croquet mallet (second column) is below the torso. Trumpets are distributed near the center of the human reference frame, as they are often played at the mouth (third column). As the fourth column shows, the relative scale between the human and the object for the 'Volleyball' action indicates that a volley ball is about half the size of a human detection (see also rightmost column of fig. 11).

Importantly, the spatial relations learned in the WS setting are similar to those learnt in the FS setting, albeit less peaked. This demonstrates that our weakly supervised approach does learn correctly human-object interactions.

6 Conclusion

This paper has introduced a novel approach for learning human-object interactions automatically from weakly labeled images. Our approach automatically determines the spatial relations between human and action relevant objects. Obtained results are comparable to a state-of-the-art fully supervised approach [2]. Future work will extend our approach to videos, where temporal information can improve the human as well as objects detections. Furthermore, temporal information can help to model variations of action as well as action sequences in time.

References

[1] Gupta, A., Kembhavi, A., Davis, L.: Observing human-object interactions: Using spatial and functional compatibility for recognition. In: PAMI. (2009)

[2] Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR. (2010)

[3] Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: ICPR. (2004)

[4] Laptev, I., Perez, P.: Retrieving actions in movies. In: ICCV. (2007)

[5] Mikolajczyk, K., Uemura, H.: Action recognition with motion-appearance vocabulary forest. In: CVPR. (2008)

[6] Sullivan, J., Carlsson, S.: Recognizing and tracking human action. In: ECCV. (2002)

[7] Ikizler-Cinbis, N., Cinbis, G., Sclaroff, S.: Learning actions from the web. In: ICCV. (2009)

[8] Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS. (2005)

[9] Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008)

[10] Willems, G., Becker, J.H., Tuytelaars, T., van Gool, L.: Exemplar-based action recognition in video. In: BMVC. (2009)

[11] Thurau, C., Hlavac, V.: Pose primitive based human action recognition in videos or still images. In: CVPR. (2008)

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

[13] Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for multi-class object layout. In: ICCV. (2007)

[14] Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: CVPR. (2003)

[15] Winn, J., Criminisi, A., Minka, T.: Object categorization by learned universal visual dictionary. ICCV (2005)

[16] Deselaers, T., Alexe, B., Ferrari, V.: Localizing objects while learning their appearance. In: ECCV. (2010)

[17] Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI (2009)

[18] Everingham, M., van Gool, L., Williams, C., Zisserman, A.: (The PASCAL Visual Object Classes Challenge (VOC))

[19] Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR. (2008)

[20] Rodriguez, Y.: Face Detection and Verification using Local Binary Patterns. PhD thesis, EPF Lausanne (2006)

[21] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR. (2001)

[22] Heusch, G., Rodriguez, Y., Marcel, S.: Local binary patterns as an image preprocessing for face authentication. In: IEEE FG. (2006)

[23] Comaniciu, D., Ramesh, V., Meer, P.: The variable bandwidth mean shift and data-driven scale selection. In: ICCV. (2001)

[24] Eichner, M., Ferrari, V.: Better appearance models for pictorial structures. In: BMVC. (2009)

[25] Fergus, R., Perona, P.: Caltech object category datasets. http://www.vision.caltech.edu/html-files/archive.html (2003)

[26] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html (2007)

[27] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html (2008)

[28] Alexe, B., Deselaers, T., Ferrari, V.: What is an object ? In: CVPR. (2010)

[29] Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. PAMI **28** (2006) 1568–1583

[30] Bay, H., Ess, A., Tuytelaars, T., van Gool, L.: SURF: Speeded up robust features. CVIU **110** (2008) 346–359

[31] Botev, Z.: Nonparametric density estimation via diffusion mixing. The University of Queensland, Postgraduate Series, Nov (2007)

[32] Zhang, J., Marszalek, M., Lazebnik, S., C., S.: Local features and kernels for classification of texture and object categories: a comprehensive study. IJCV (2007)

[33] Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV **42** (2001) 145–175

[34] Li, L.J., Fei-Fei, L.: What, where and who? classifying event by scene and object recognition. In: ICCV. (2007)

[35] Gehler, P., Nowozin, S.: On feature combination for multiclass object classification. In: ICCV. (2009)

[36] Grubinger, M., Clough, P.D., Müller, H., Deselaers, T.: The IAPR benchmark: A new evaluation resource for visual information systems. In: LREC. (2006)

[37] Johansson, R., Nugues, P.: Dependency-based syntactic-semantic analysis with propbank and nombank. In: Computational Natural Language Learning. (2008)

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-0803