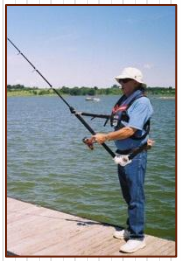
Vincent Delaitre - Ecole Normale Superieure
Josef Sivic - INRIA Paris - Rocquencourt
Ivan Laptev - INRIA Paris - Rocquencourt

Learning person-object interactions for action recognition in still images


riding a bike


taking photos


playing guitar


fishing


running

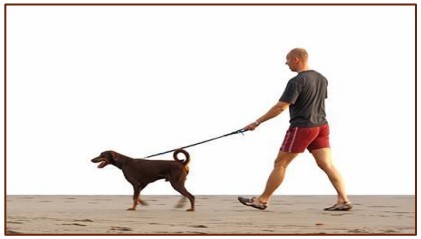

brushing teeth

by Fadime Sener

Human actions in still images

- Actions describe many images : Human actions are a natural description of many images.

- Still images has information to **understand** the content of many still images



consumer photographs

news images

surveillance videos

Automatic recognition of human actions and interactions

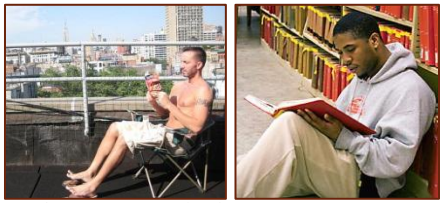
- A very **challenging** problem
- Vary significantly due to many factors such as

person's clothing

occlusions

object appearance

camera viewpoint



layout of the scene

variation of body pose

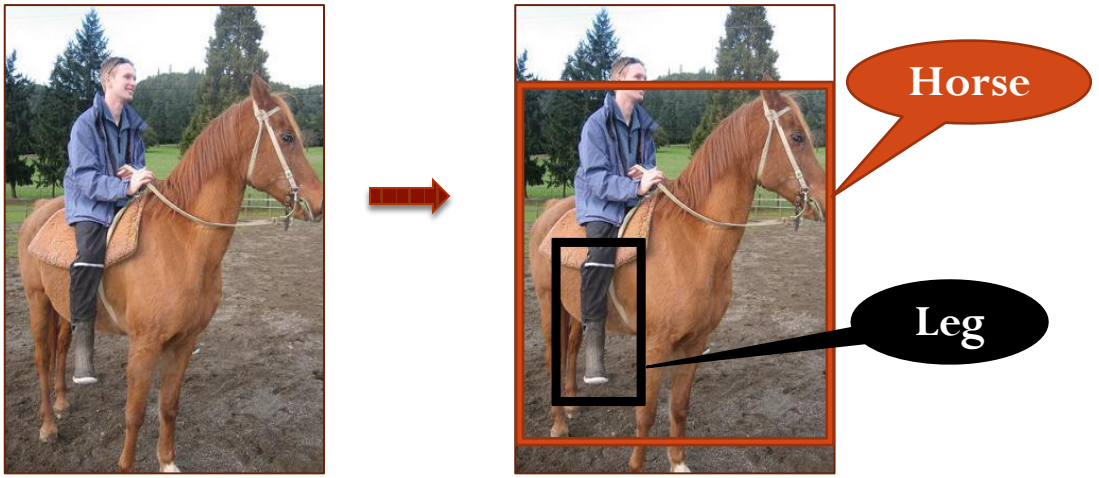
- Motion cues in videos

not available in still images

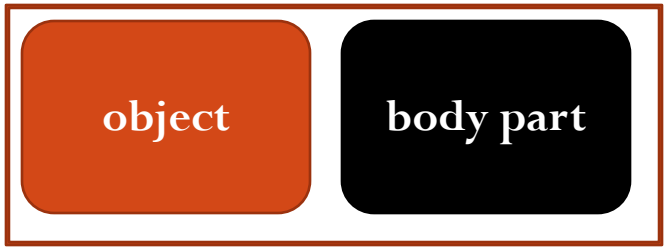
Some actions are static and video may not help.

In this work ;

- Models of interactions between objects and human body parts

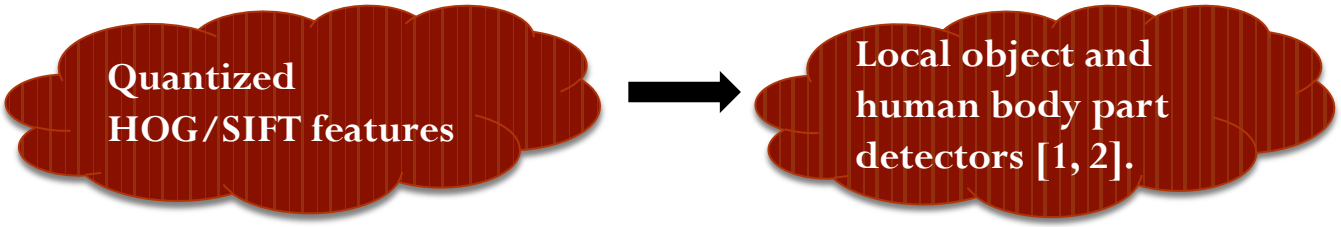


- Challenges:
  - (i) What should be the representation of appearance
  - (ii) How to model interaction
  - (iii) How to choose suitable interaction pairs in the huge space of all possible combinations of

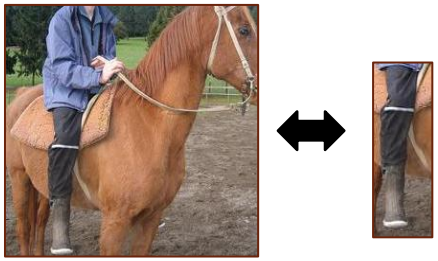


Contributions

- Replace



- Develop a part interaction representation;
  - Capture pair-wise relative **position and scale** between object/body parts



- Choosing interacting parts
  - ✗ Do not choose manually
  - Select suitable pair-wise interactions in a discriminative way from a large pool of hundreds of thousands of candidate interactions

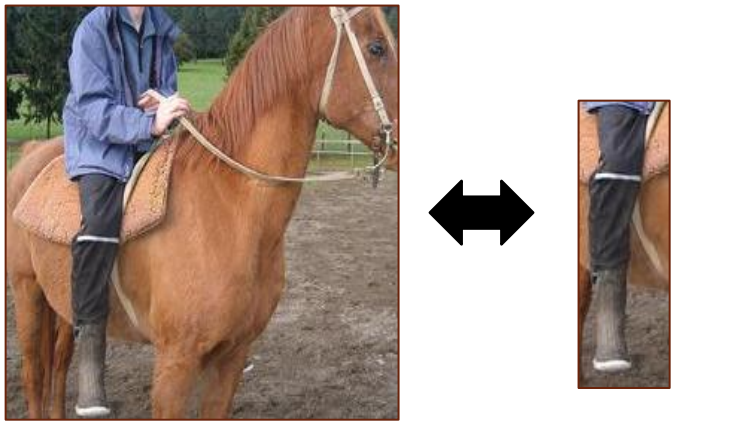
[1] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In ICCV, 2009
[2] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In CVPR, 2011

Challenges

(i) What should be the representation of object and body part appearance?

(ii) How to model object and human body part interactions?

(iii) How to choose suitable interaction pairs in the huge space of all possible combinations and relative configurations objects and body parts ?



object **body part**

Representing body parts and objects

- Detectors d_1, d_2, …, d_n

pre-trained for different body parts and object classes

- Each i^{th} detector produces

  A map of dense 3D responses $d_i(I, p)$

  over locations and scales of a given image I.

  - **I** : image
  - **p** : the positions of detections ; $p = (x, y, \sigma)$
  - (x, y) : spatial location
  - σ : an additive scale parameter

d_i →

Representing body parts and objects

- Two types of detectors;
  - For objects
    - LSVM detector [1]
    - Trained on PASCAL VOC images
    - Ten object detector classes: bicycle, car, chair, cow, dining table, horse, motorbike, person, sofa, tv/monitor

  - For body parts
    - Implement the method of [2]
    - Train ten body part detectors
    - Torso, left, right {fforearm, upper arm, lower leg, thigh}

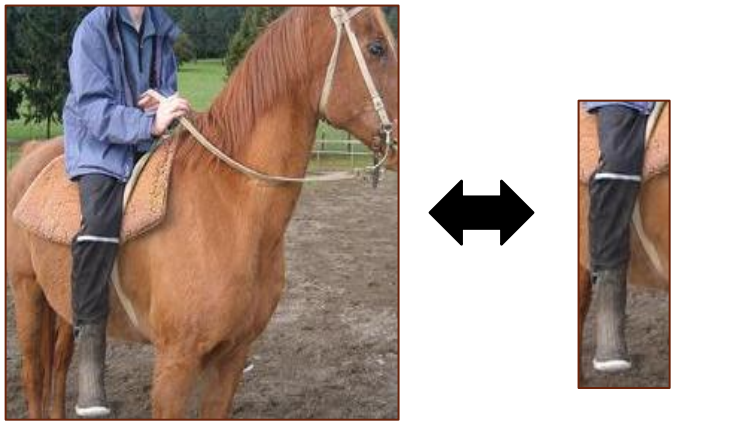
[1] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. IEEE PAMI, 2009
[2] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In CVPR, 2011

Challenges

(i) What should be the representation of object and body part appearance?

(ii) How to model object and human body part interactions?

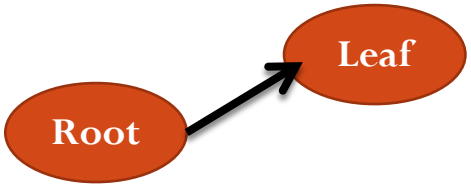
(iii) How to choose suitable interaction pairs in the huge space of all possible combinations and relative configurations objects and body parts ?



object **body part**

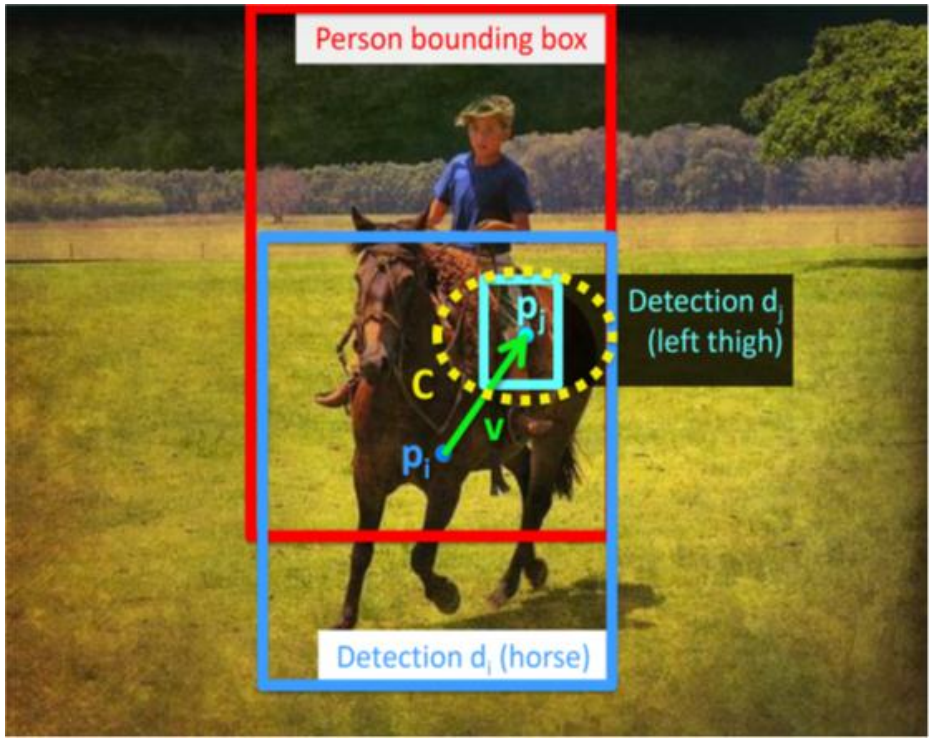
Representing pairwise interactions

An interaction pair of detectors (d_i , d_j) is defined by a



quadruplet $\boldsymbol{q = (i, j, v, C) \in \mathbb{N} x \mathbb{N} x \mathbb{R}^3 x \mathcal{M}_{3,3}}$

Position and the scale of the leaf are related to the root by scale-space offset and a spatial deformation cost.



Representing person-object interactions

p_j body part detectors ----- p_i object detectors.

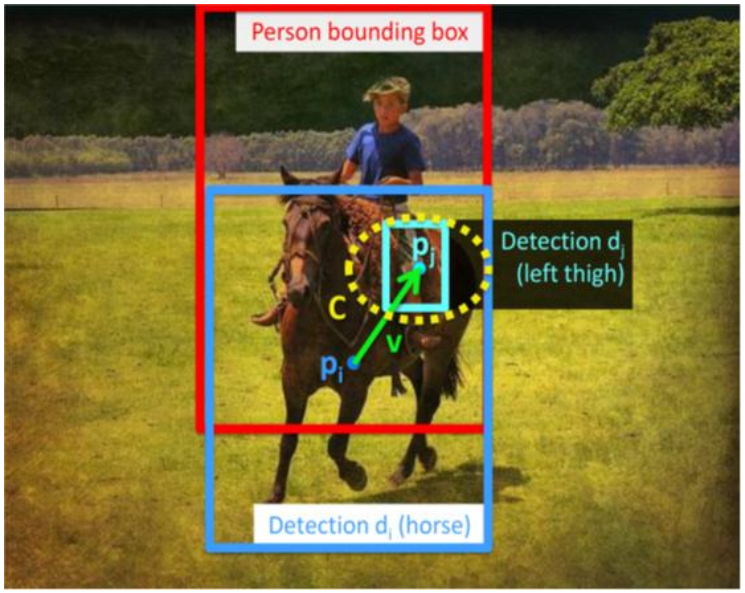
- The pair of detectors p_j and p_i
- 3D scale-space displacement **v**
  Offset (x, y, scale space)
- Scale-space displacement cost **C**
  Fixed displacement cost "C" of leaf covariance matrix penalizing distance between actual object location and expected object location

Representing pairwise interactions

- The response of the interaction $\mathbf{q}$ located at the root position $\mathbf{p_1}$

  How strongly is this interaction represented ?

$$\mathbf{r(I, q, p_1)} = \max_{\boldsymbol{p_2}} (\boldsymbol{d_{i(I,p_1)}} + \boldsymbol{d_{j(I,p_2)}} - \boldsymbol{u^T C u})$$

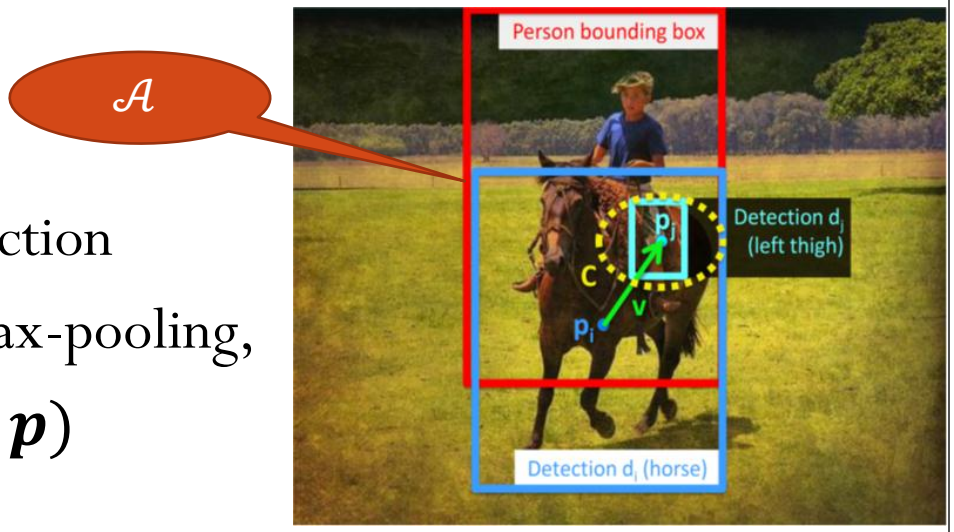


- $u = p_2 - (p_1 + v)$: the displacement vector
- Maximizing over p_2
  → optimal trade-off between the detector score and the displacement cost.
- For any interaction q ,
  → compute its responses for all pairs of node positions p_1, p_2. *

* P. Felzenszwalb and D. Huttenlocher. Distance transforms of sampled functions. Technical report, Cornell University CIS, Tech. Rep. 2004-1963, 2004.

Representing images by response vectors of pair-wise interactions

- M interaction pairs : $q_1, q_2, \ldots q_M$
  aggregate responses



- Define score $s(I, q, \mathcal{A})$ of an interaction pair q within $\mathcal{A}$ of an image I by max-pooling,

$$\boldsymbol{s(I, q, \mathcal{A}) = \max_{p \in \mathcal{A}} \; r(I, q, p)}$$

- An image region $\mathcal{A}$ is represented by a M-vector of interaction pair scores

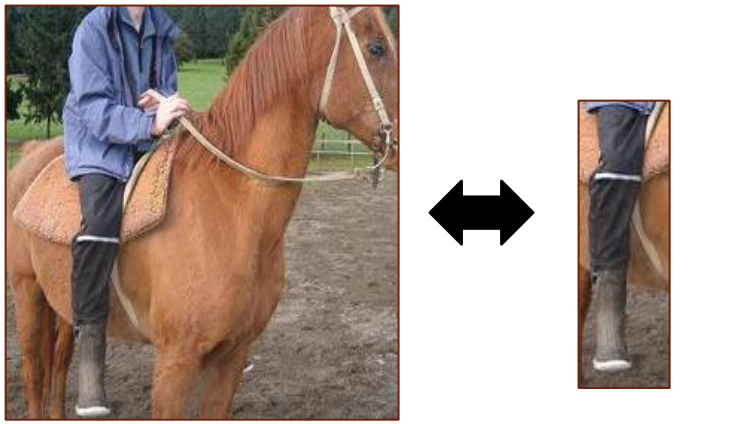
$$\boldsymbol{z = (s_1, s_2, \ldots, s_M) \;\; with \;\; s_i = s(I, q_i, \mathcal{A})}$$

Challenges

(i) What should be the representation of object and body part appearance?

(ii) How to model object and human body part interactions?

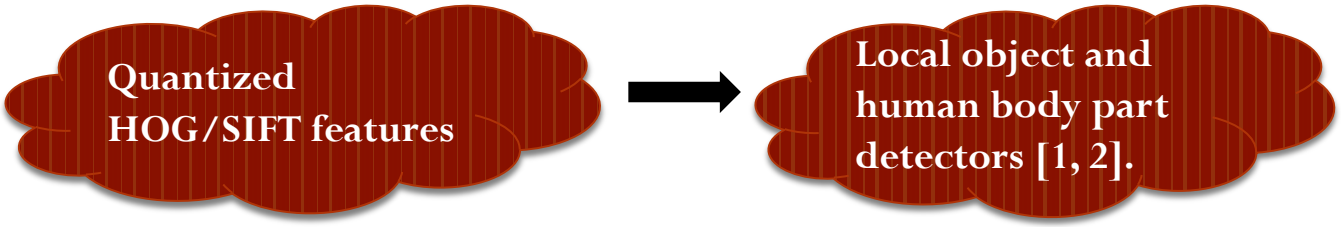
(iii) How to choose suitable interaction pairs in the huge space of all possible combinations and relative configurations objects and body parts ?



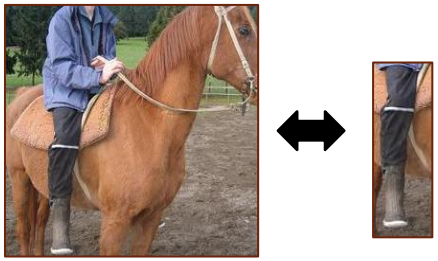
object **body part**

Contributions

- Replace



- Develop a part interaction representation;
  - Capture pair-wise relative **position and scale** between object/body parts



- Choosing interacting parts
  - ✗ Do not choose manually
  - Select suitable pair-wise interactions in a discriminative way from a large pool of hundreds of thousands of candidate interactions

[1] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In ICCV, 2009
[2] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In CVPR, 2011

Learning person-object interactions

object and body-part interaction pairs q

Use for action classification in still images

- A brute-force approach
  - ✗ Analyzing all possible interactions.

    Space of all possible interactions

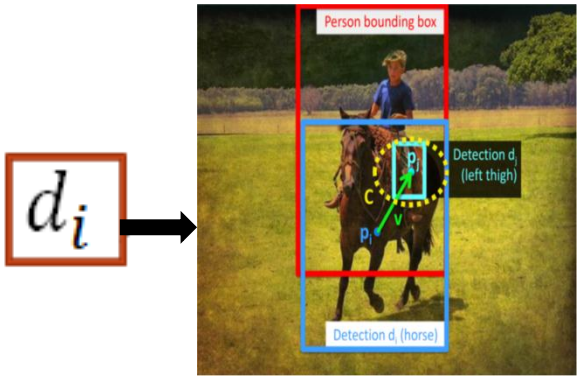
Number of detectors and scale-space relations among them

- Learning procedure :
  - **For each action;**
    - Select candidate detection pairs which frequently occur from pool of candidate interactions
    - Select a set of M discriminative interactions which best separate the particular action class from other classes in our training set.
    - Combine discriminative interactions across classes.

Generating a candidate pool of interaction pairs

Find clusters of frequently co-occurring detectors (d_i, d_j) in specific relative configurations from pool of candidate interactions

- For each detector i and an image I
  - Collect a set of positions of all positive detector responses $P_i^I = \{p | d_i(I,p) > 0\}$,
    - $d_i(I,p)$ response of detector i at position p in image I.

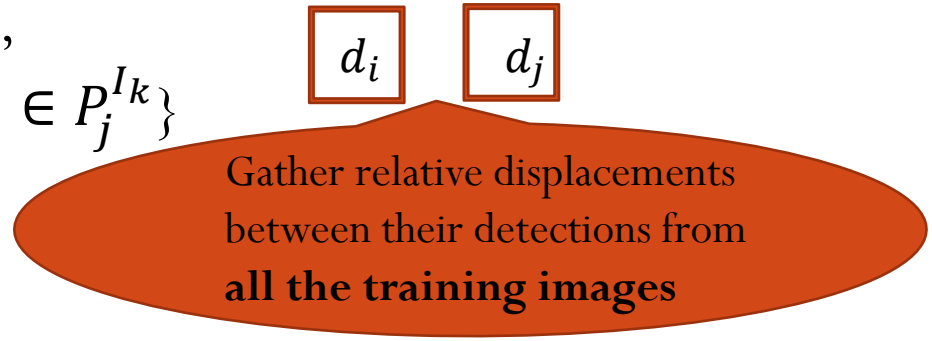


- Apply a standard non-maxima suppression (NMS) step to eliminate multiple responses of a detector in local image neighborhoods and then limit P_i^I to the L top-scoring detections.
  - The intuition behind this step is that a part/object interaction is not likely to occur many times in an image

Generating a candidate pool of interaction pairs

- For each pair of detectors $(d_i ; d_j)$,

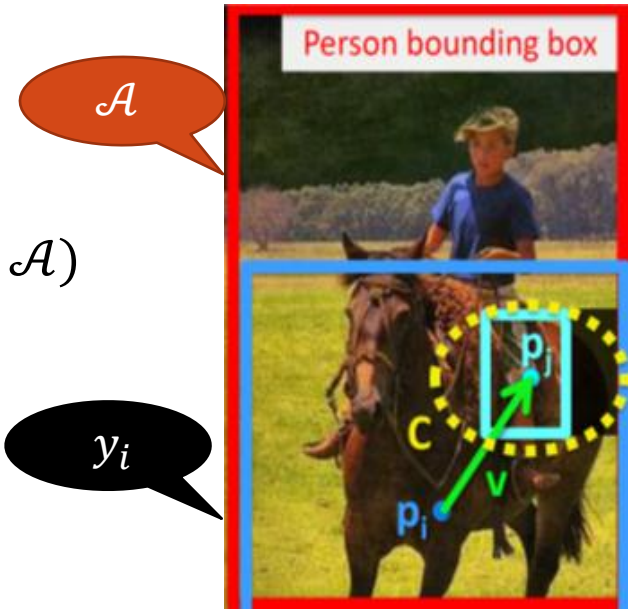
$$I_k\colon D_{ij} = U_k\{p_j - p_i \mid p_i \in P_i^{I_k}\ p_j \in P_j^{I_k}\}$$



- Discover potentially interesting interaction pairs, performing a mean-shift clustering over D_{ij} .
  - Discard clusters which contribute to less than η percent of the training images.
- The set of m resulting candidate pairs $(\boldsymbol{i, j, v_1, C})$, … , $(\boldsymbol{i, j, v_m, C})$ is built from the centers $v_1, \ldots, v_m$ of the remaining clusters.
- Apply this procedure to all pairs of detectors and generate a large pool of potentially interesting candidate interactions

Discriminative selection of interaction pairs

- Select a smaller number of M discriminative interactions :
  - Large number of candidate interactions,
  - Many of them, may not be informative



- Given a set of N training images,

  response vector $z = (s_1, s_2, \ldots, s_M)$ $with$ $s_i = s(I, q_i, \mathcal{A})$

  $\mathcal{A}$: extended person bounding box

  y_i : binary label (in a 1-vs-all setup for each class),

- The learning problem for each action class

  Binary SVM cost function:

$$J(w, b) = \lambda \sum_{i=1}^{N} \max\{0, 1 - y_i(w^T z_i + b)\} + \|w\|_1$$

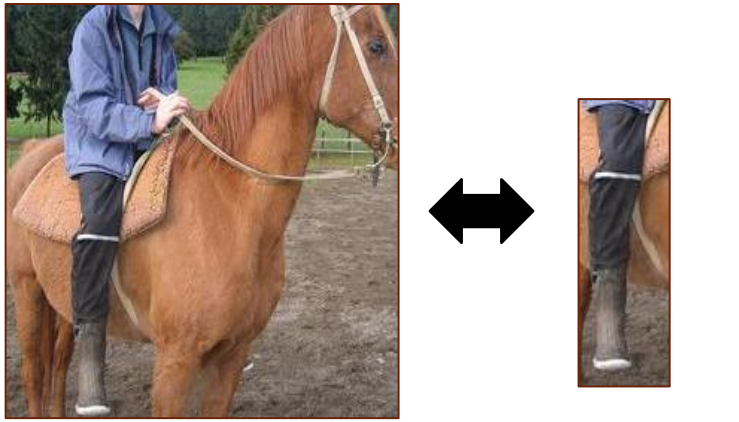
  - w, b are parameters of the classifier .
  - Selection of M interaction pairs corresponding to non-zero elements of w gives M most discriminative interaction pairs per action class.

Challenges

(i) What should be the representation of object and body part appearance?

(ii) How to model object and human body part interactions?

(iii) How to choose suitable interaction pairs in the huge space of all possible combinations and relative configurations objects and body parts ?

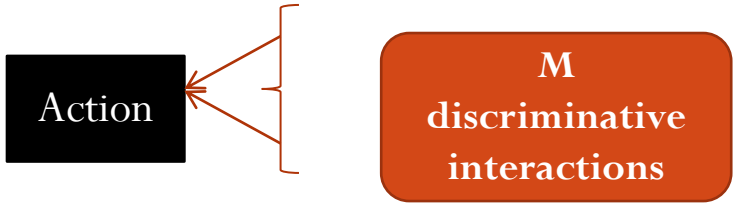


object **body part**

Classification

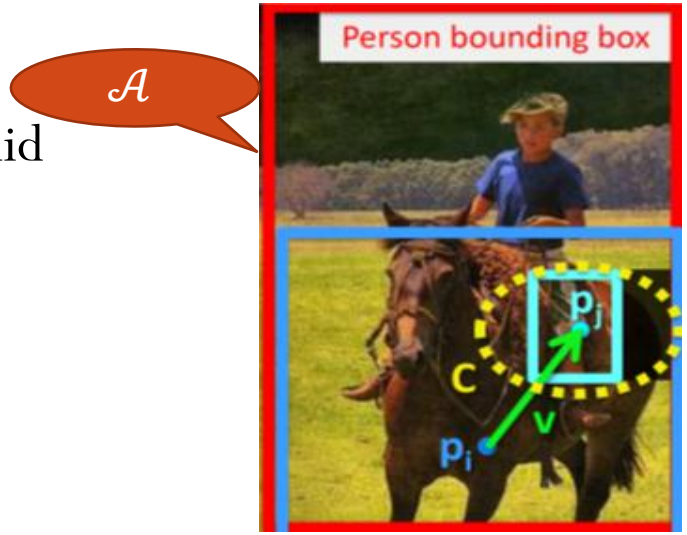
Using interaction pairs for classification

- M discriminative **interactions** for each **action** class,



- Use spatial pyramid-like representation,
  aggregating responses in each cell of the pyramid using max-pooling as described by eq.
  $\mathcal{A}$ is one cell of the spatial pyramid.

$$s(I, q, \mathcal{A}) = \max_{p \in \mathcal{A}} r(I, q, p)$$



- Extend the standard 2D pyramid representation to scale-space resulting in a 3D pyramid with $D = 1 + 2^3 + 4^3 = 73\ cells.$
- Train a non-linear SVM with RBF kernel

Experiments - - - Dataset

- Dataset ;
  - Willow-action dataset
    - A dataset for human action classification in still images.
    - *Interacting with computer, Photographing, Playing Instrument, Riding Bike, Riding Horse, Running,Walking*

    - More than 900 images , more than 1100 labeled person detections
    - The training set contains 70 examples of each action class

      The testing set contains at least 39 examples per class.

  - PASCAL VOC 2010 dataset
    - 7 above classes + Phoning and Reading.
    - Similar number of images.
    - Annotated with smallest bounding box

Results:

- Per-class average-precision for different methods on the Willow-actions dataset.
  - Reported using average precision for each class

Action / Method	a. BOF [11]	b. LSVM	c. Detectors	d. Interactions
(1) Inter. w/ Comp.	**58.15**	30.21	45.64	56.60
(2) Photographing	35.39	28.12	36.35	**37.47**
(3) Playing Music	**73.19**	56.34	68.35	72.00
(4) Riding Bike	82.43	68.70	86.69	**90.39**
(5) Riding Horse	69.60	60.12	71.44	**75.03**
(6) Running	44.53	51.99	57.65	**59.73**
(7) Walking	54.18	55.97	**57.68**	57.64
Average (mAP)	59.64	50.21	60.54	**64.12**

- **BOF** → The bag-of-features classifier [1]
- **LSVM** →Latent SVM classifier trained in a 1-vs-all fashion for each class. Take the maximum LSVM detection score from the detections overlapping the extended bounding box with the standard overlap score higher than 0.5.
- **Detectors** → SVM classifier with an RBF kernel trained on max-pooled responses of the entire bank of body part and object detectors in a spatial pyramid representation but without interactions.
- **Interactioms** → Proposed method

[1]V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bagof-features and part-based representations. In Proc. BMVC., 2010.
[2] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. IEEE PAMI, 2009

Results:

- Per-class average-precision on the Pascal VOC 2010 action classification dataset.
  - Reported using average precision for each class

Action / Method	a. BOF [11]	b. LSVM	c. Detectors	d. Interactions
(1) Inter. w/ Comp.	**58.15**	30.21	45.64	56.60
(2) Photographing	35.39	28.12	36.35	**37.47**
(3) Playing Music	**73.19**	56.34	68.35	72.00
(4) Riding Bike	82.43	68.70	86.69	**90.39**
(5) Riding Horse	69.60	60.12	71.44	**75.03**
(6) Running	44.53	51.99	57.65	**59.73**
(7) Walking	54.18	55.97	**57.68**	57.64
Average (mAP)	59.64	50.21	60.54	**64.12**

outperforms all baselines

Results:

- Per-class average-precision on the Pascal VOC 2010 action classification dataset.

Action / Method	d. Interactions	e. BOF+LSVM+Inter.	Poselets[30]	MK-KDA[1]
(1) Phoning	42.11	48.61	49.6	**52.6**
(2) Playing Instr.	30.78	53.07	43.2	**53.5**
(3) Reading	28.70	28.56	27.7	**35.9**
(4) Riding Bike	**84.93**	80.05	83.7	81.0
(5) Riding Horse	89.61	**90.67**	89.4	89.3
(6) Running	81.28	85.81	85.6	**86.5**
(7) Taking Photo	26.89	**33.53**	31.0	32.8
(8) Using Computer	52.31	56.10	59.1	**59.2**
(9) Walking	**70.12**	69.56	67.9	68.6
Average (mAP)	56.30	60.66	59.7	**62.2**

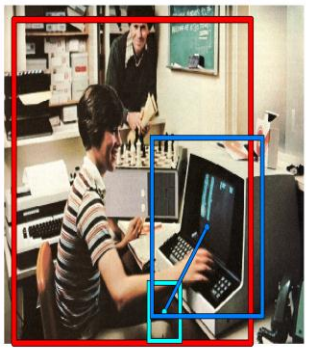
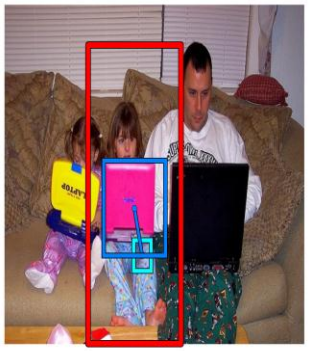
- **Interactioms** → Proposed method
- **BOF+LSVM+Inter** → Interaction and its combination with the baselines : BOF (51.25 mAP) and LSVM (44.08 mAP)
- **"Poselet" method** → [1],.
- **MK-KDA** → Kernel-level fusion with Spatial Pyramid Grids, Soft Assignment and Kernel Discriminant Analysis using spectral regression. 18 kernels have been generated from 18 variants of SIFT.

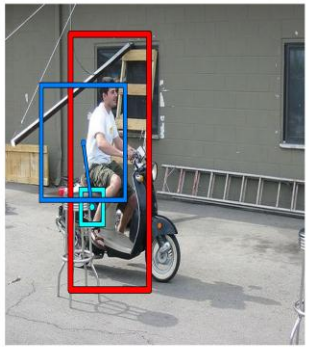
[1] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In CVPR, 2011.

Example detections of discriminative interaction pairs.

- These body part interaction pairs are chosen as discriminative for indicated on the left.

Inter. w/ Comp.
Blue: Screen
Cyan: L. Leg



action class

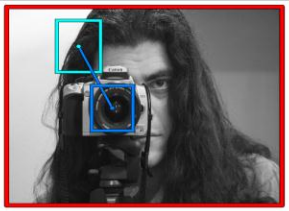
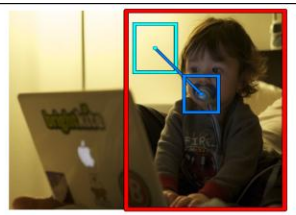
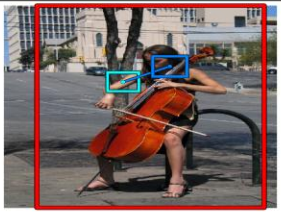
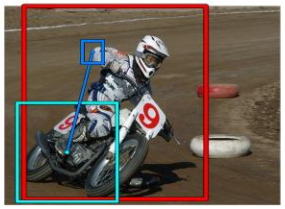
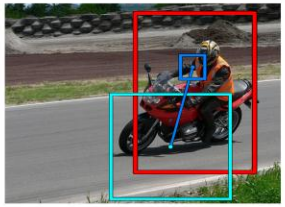
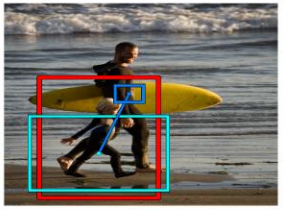
First three images show detections on the correct action class.

Shows a high scoring detection on an incorrect action class.

- The leg detector seems to consistently fire on keyboards (third image)

Example detections of discriminative interaction pairs.

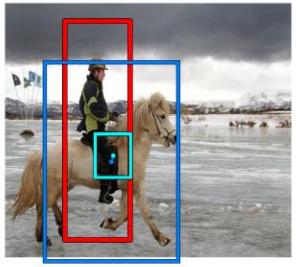
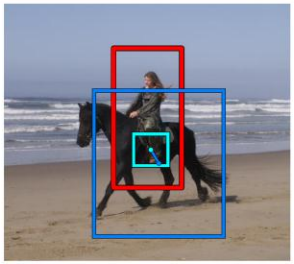
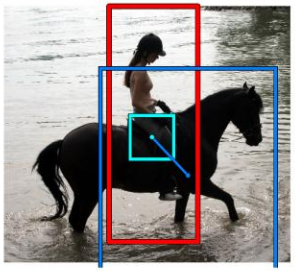
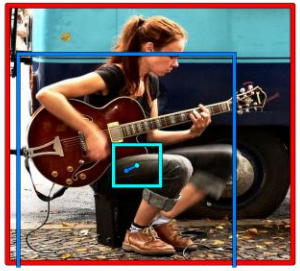
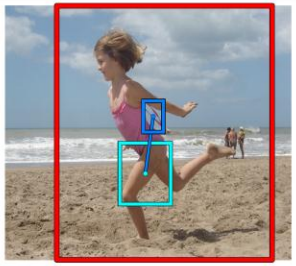
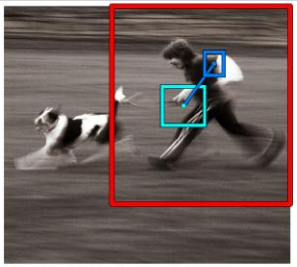
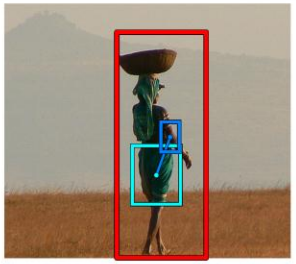
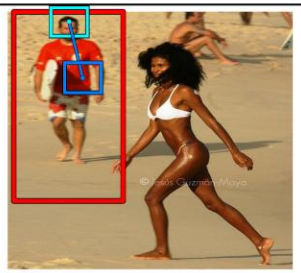
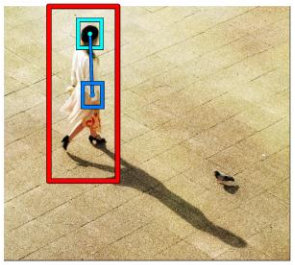
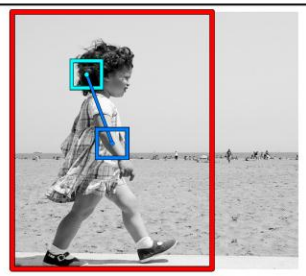
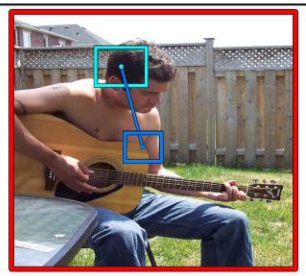
- In the examples shown, the interaction features capture either a body part and an object, or two body part interactions.

Photographing Blue: Head Cyan: L. Thigh				
Playing Instr. Blue: L. Forearm Cyan: L. Forearm				
Riding Bike Blue: R. Forearm Cyan: Motorbike				

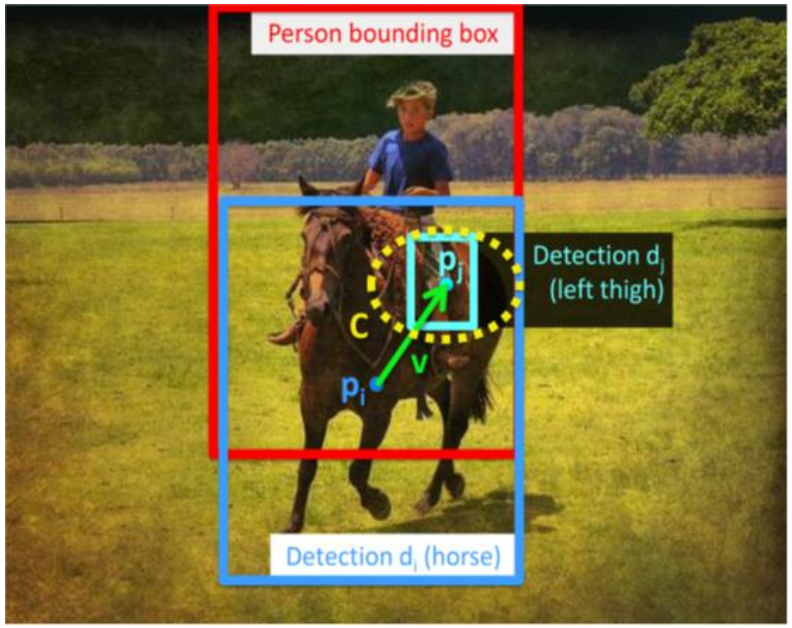
- These interaction pairs are found to be discriminative, due to the detection noise, they do not necessary localize the correct body parts in all images.
- They may still fire at consistent locations across many images as illustrated in the first row, where the head detector consistently detects the camera lens, and the thigh detector fires at the edge of the head.

Example detections of discriminative interaction pairs.

- In the examples shown, the interaction features capture either a body part and an object, or two body part interactions.

Riding Horse Blue: Horse Cyan: L. Thigh				
Running Blue: L. Arm Cyan: R. Leg				
Walking Blue: L. Arm Cyan: Head				

Conclusion



- Developed a body part/object interaction representation
- These features can be learnt in a discriminative fashion and can improve action classification performance over a strong bag-of-features baseline
- In addition, the learnt interaction features in some cases correspond to visually meaningful configurations of body parts, and body parts with objects

Images used in presentation are taken from web and Stanford 40 dataset