Recognition Using Visual Phrases

Ali Farhadi, Mohammad Amin Sadeghi
University of Illinois at Urbana-Champaign

CVPR'11, Best Student Paper

Object Recognition

Object Recognition

Parts, Poselets and Attributes

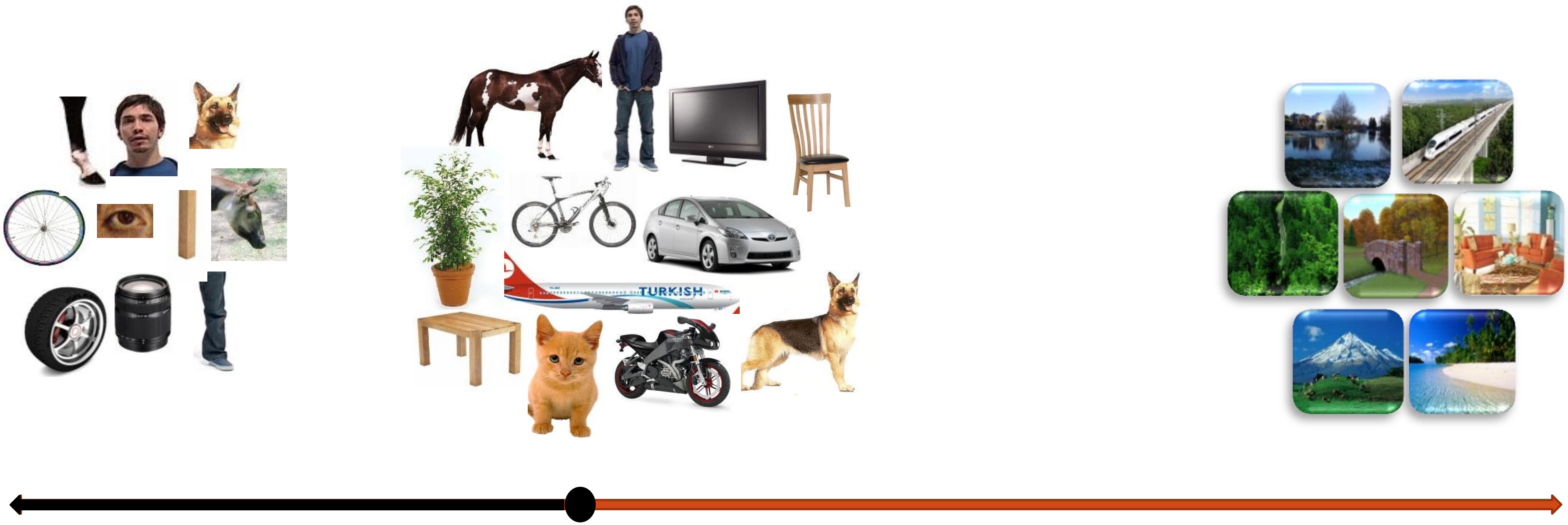
High literature, For example;
(Fergus, Perona, Zisserman, 2003) ,
(Bourdev, Malik 2009),..

Object Recognition

Scenes

Huge literature, For example;
{Oliva, Torralba 2001}
{SUN, 2010}

Object Recognition



Parts, Poselets and Attributes

Scenes

What is a Visual Phrase ?

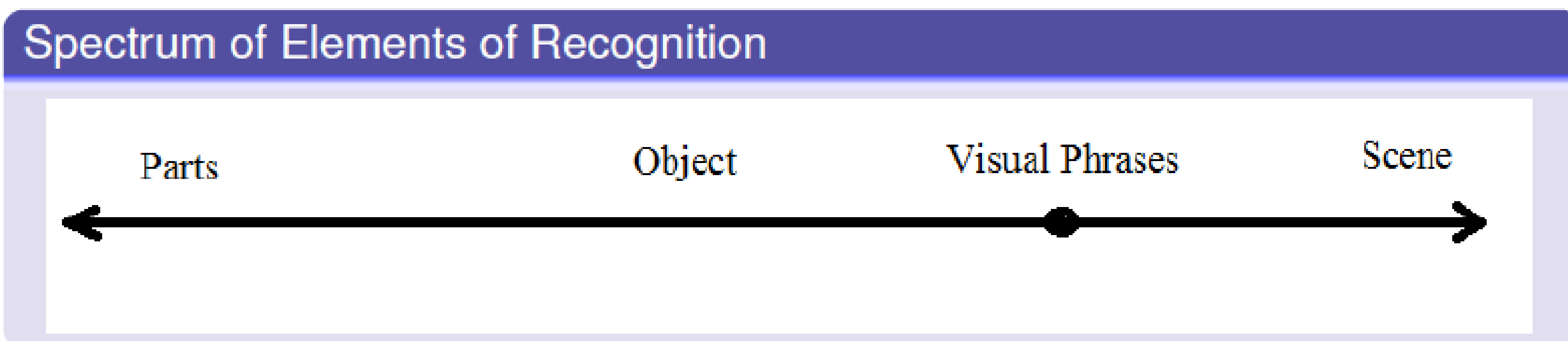
Objects

Visual Phrases

Scenes

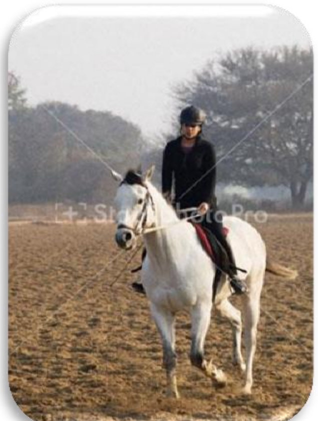
What is a Visual Phrase ?

- Part of image natural to cut out
- Corresponds to chunk of meaning **bigger than object and smaller than scene**
- Example: Person lying on a sofa, Dog jumping



Visual Phrases

- Corresponds to chunk of meaning **bigger than object and smaller than scene**








Visual Phrases

A person riding a horse

Objects + Interactions

A woman drinks from a water bottle

Visual Phrases

Dog Jumping

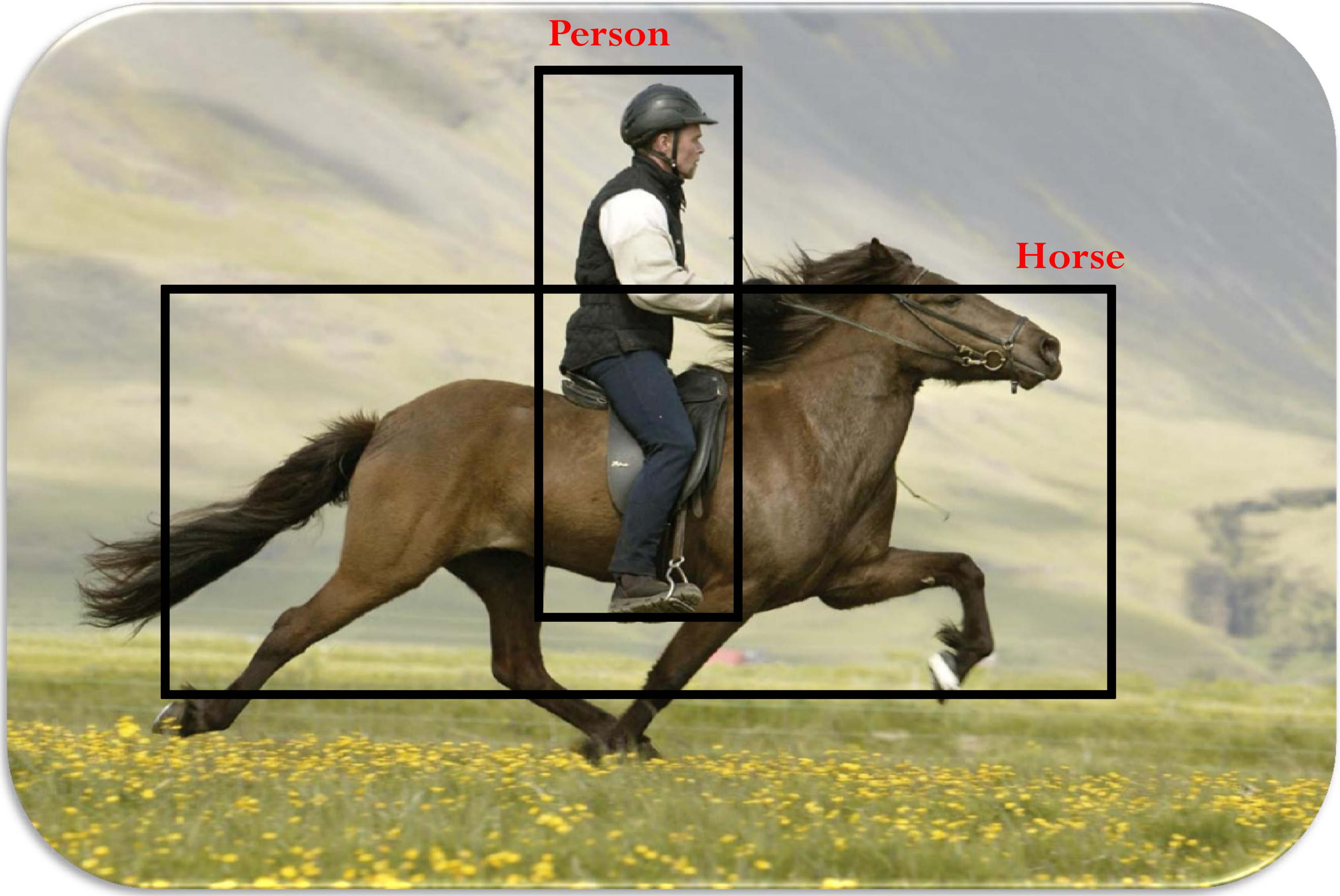
Object + Activity

Semantically Speaking

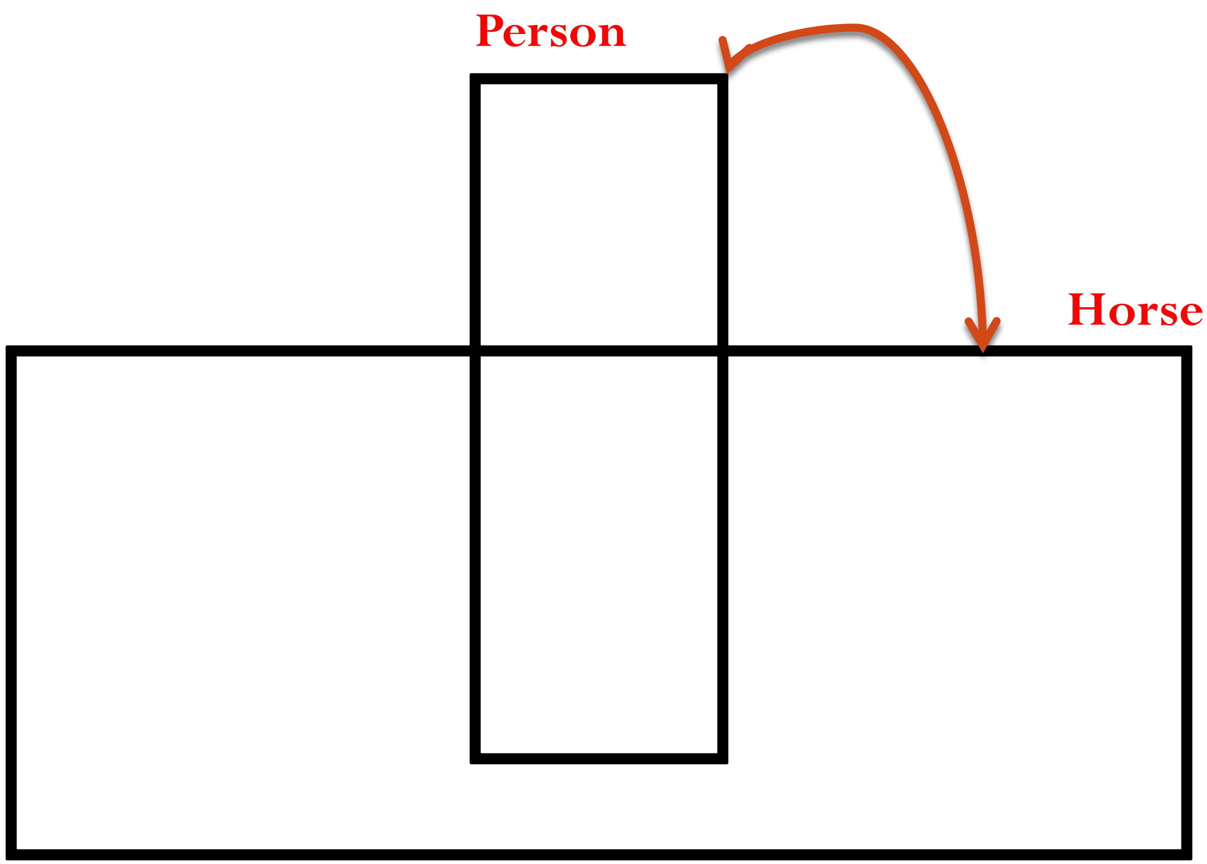
“a person riding a horse”?

Semantically Speaking

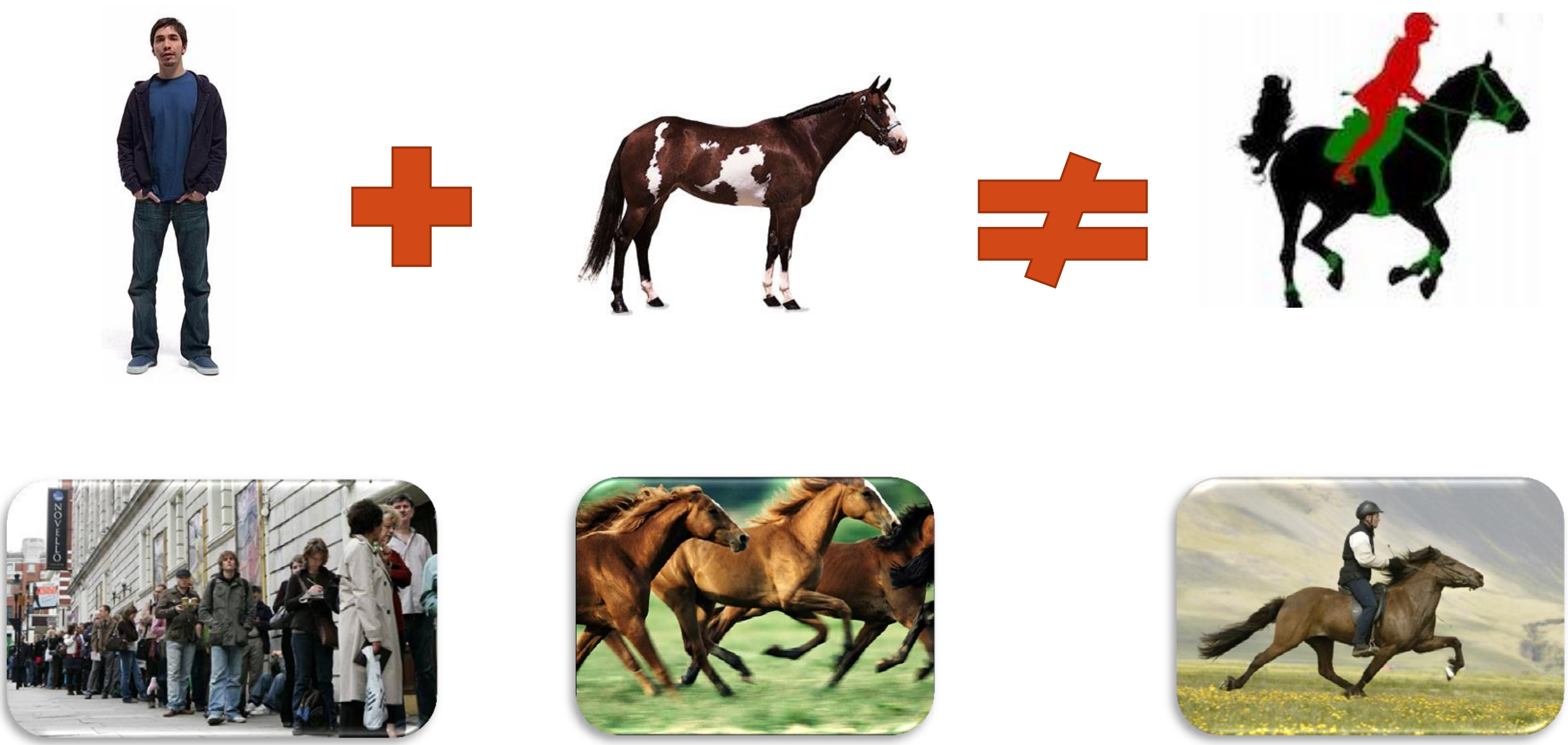
Semantically Speaking

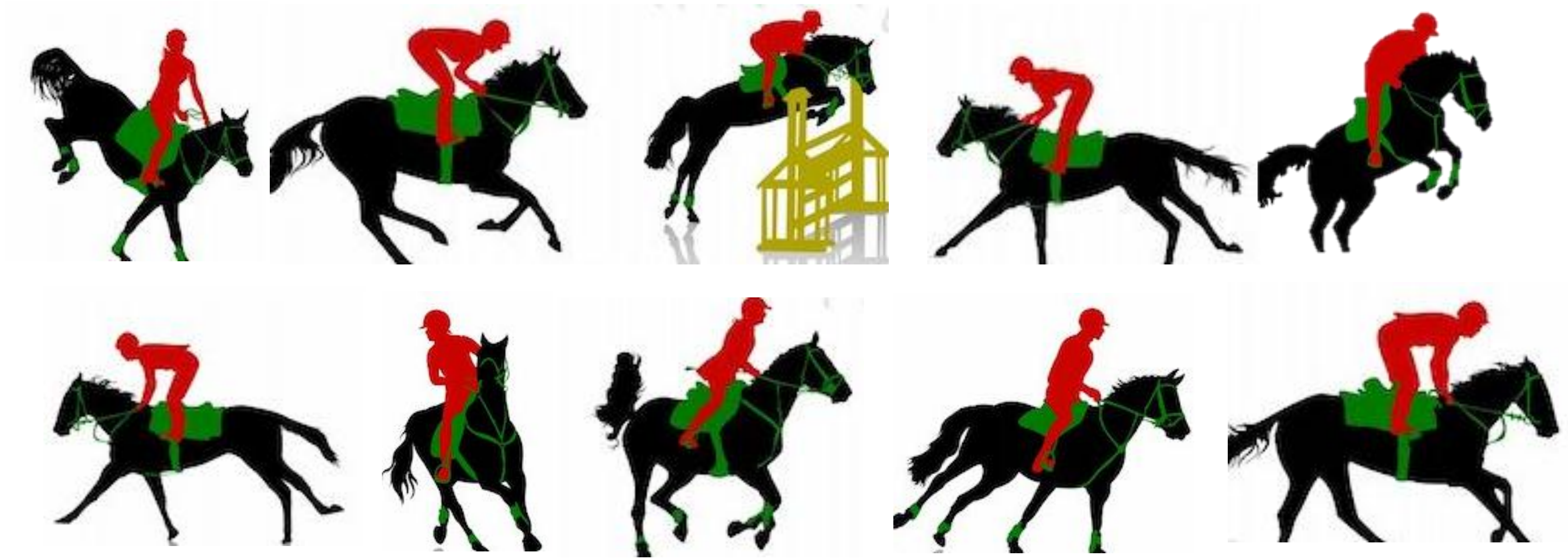


Semantically Speaking



Participating in Phrases affects the appearance of the objects





Change in Appearance

A few postures

One leg not visible

…

- Visual composites might be much easier to detect than their participant components.

Characteristic Appearance

Adding Visual Phrases to The Vocabulary of Recognition

- Learn to detect visual phrases
  - Person riding horse, dog lying on sofa

- Potential Concerns:
  - Combinatorial number of visual phrases
    - Not all possible combinations of words make a visual phrase
  - Lack of training data
    - No need for several training examples
    - Visual phrases are less complex, easy to detect.

Phrasal Recognition Dataset

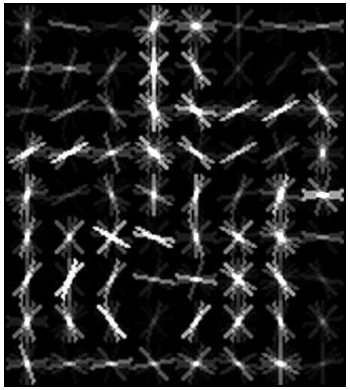
- Individual Objects that are well studied
  - Pascal Objects
  - Person, bike, car, dog, horse, bottle, sofa, and chair
- Phrases
  - person riding horse; person sitting on sofa; person sitting on chair; person lying on sofa; person lying on beach; person riding bicycle; horse and rider jumping; person next to horse; person next to bicycle; bicycle next to car; person jumping; person next to car; dog lying on sofa; dog running; dog jumping; person running; and person drinking from a bottle

- 8 Objects
  from Pascal
- 17 visual phrases
- 2769 images
  ‘120 per categ.
- 5067 examples
  1796 visual phr.
  +
  3271 objects

Training the Detectors

- Visual Phrases :
  - Deformable part models [P. F. Felzenszwalb et. al. 2010 v4]
  - On Phrasal Recognition Dataset
  - 50 examples per visual phrase

Appearance Models

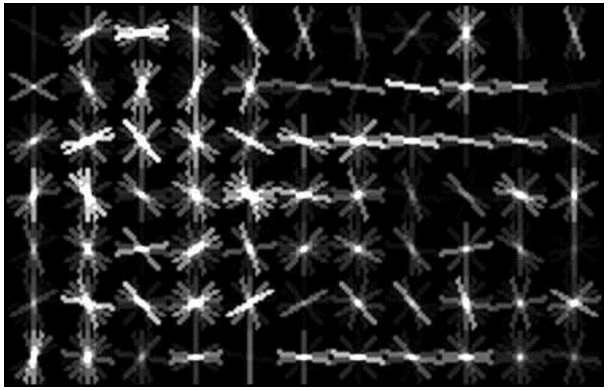


person riding horse

person riding bicycle

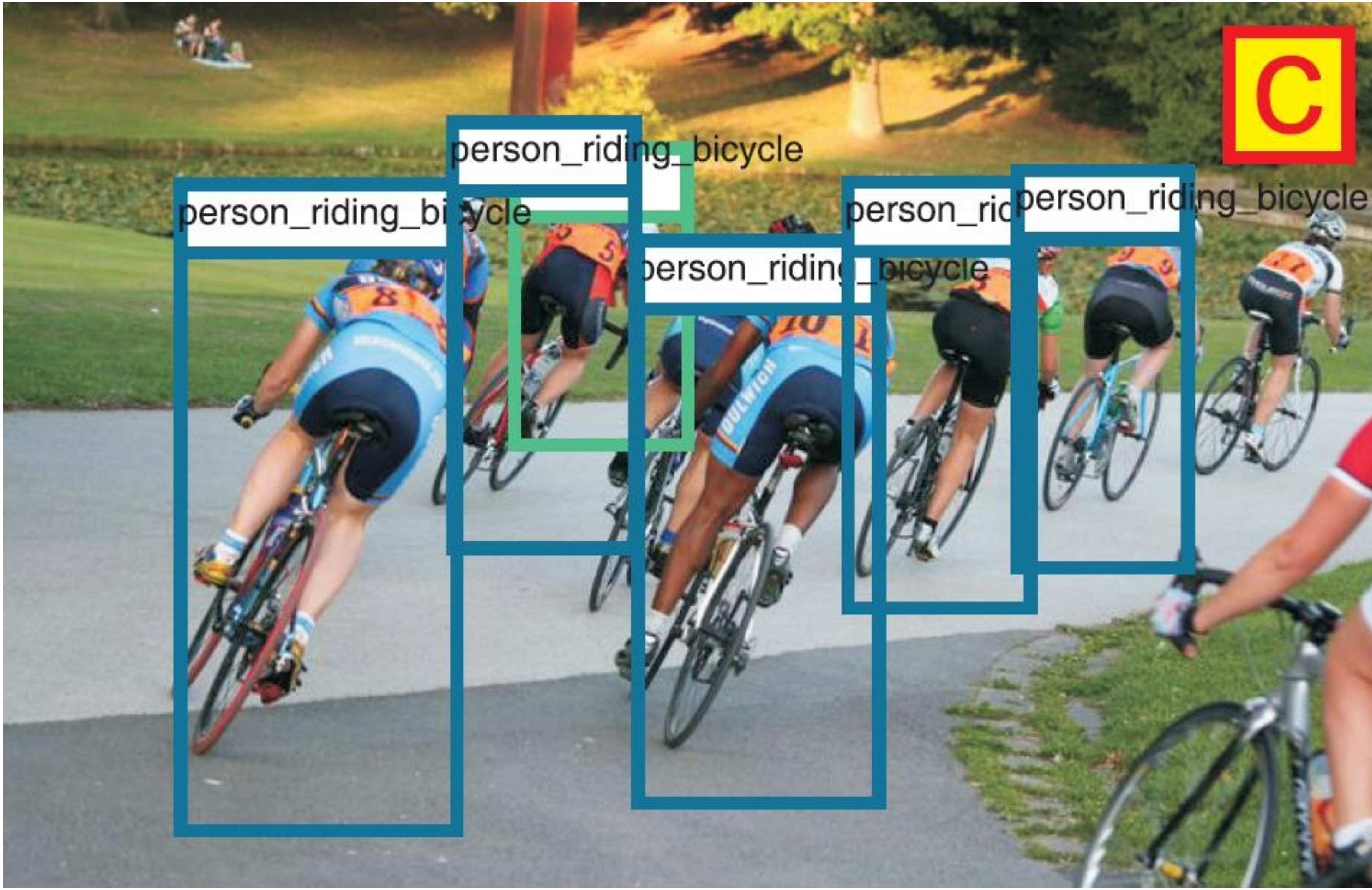
person jumping

person drinking bottle



person sitting on sofa

Visual Phrase Detectors



Visual Phrase Detectors

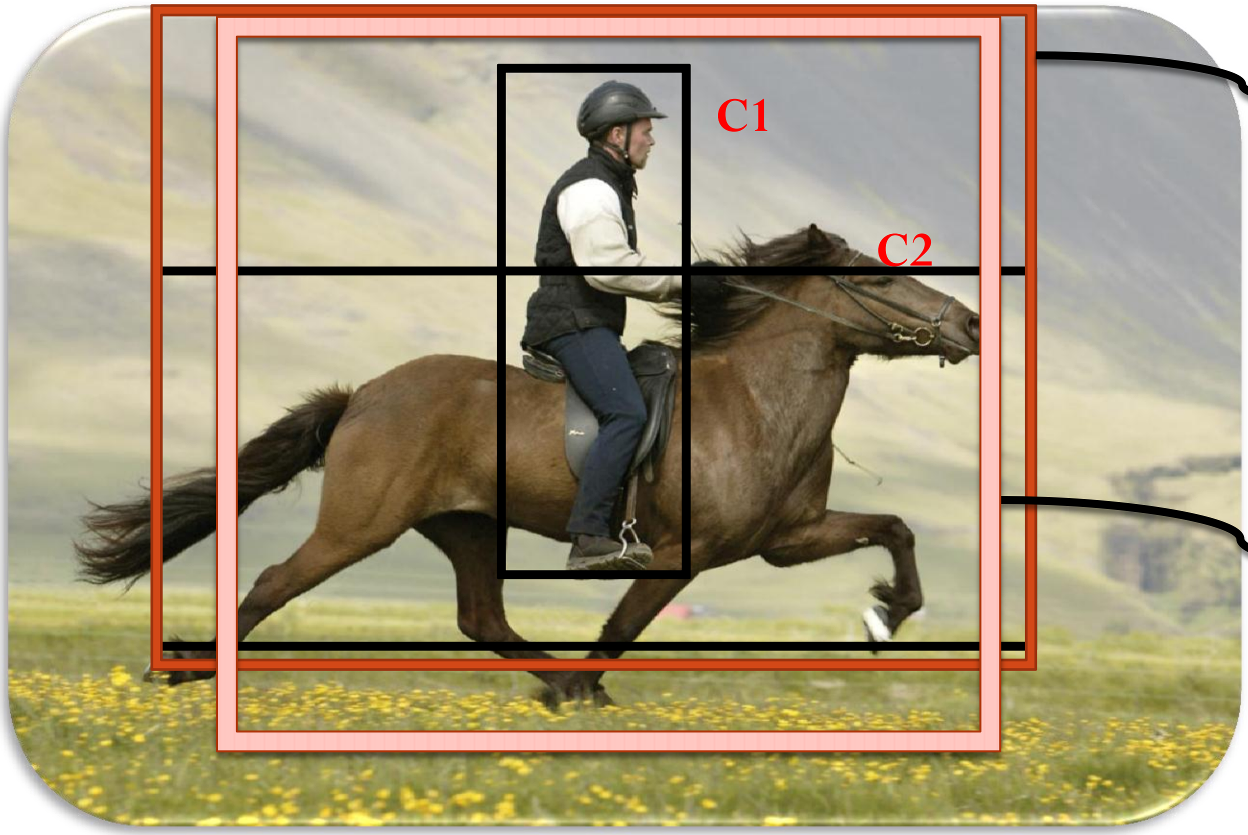
Baseline

- Baseline:
  - Upper bound on how well one can detect a visual phrase by detecting participating objects
- Fine tune the **baseline** to perform as best as it could potentially do
- Unfair Advantages to the baseline

Training the Detectors

- Objects:
  - State of the art detectors
    - V 4.0 of deformable part models
    - Trained on thousands of examples
    - Heavily fine tuned
  - Train deformable part models on Phrasal Recognition dataset

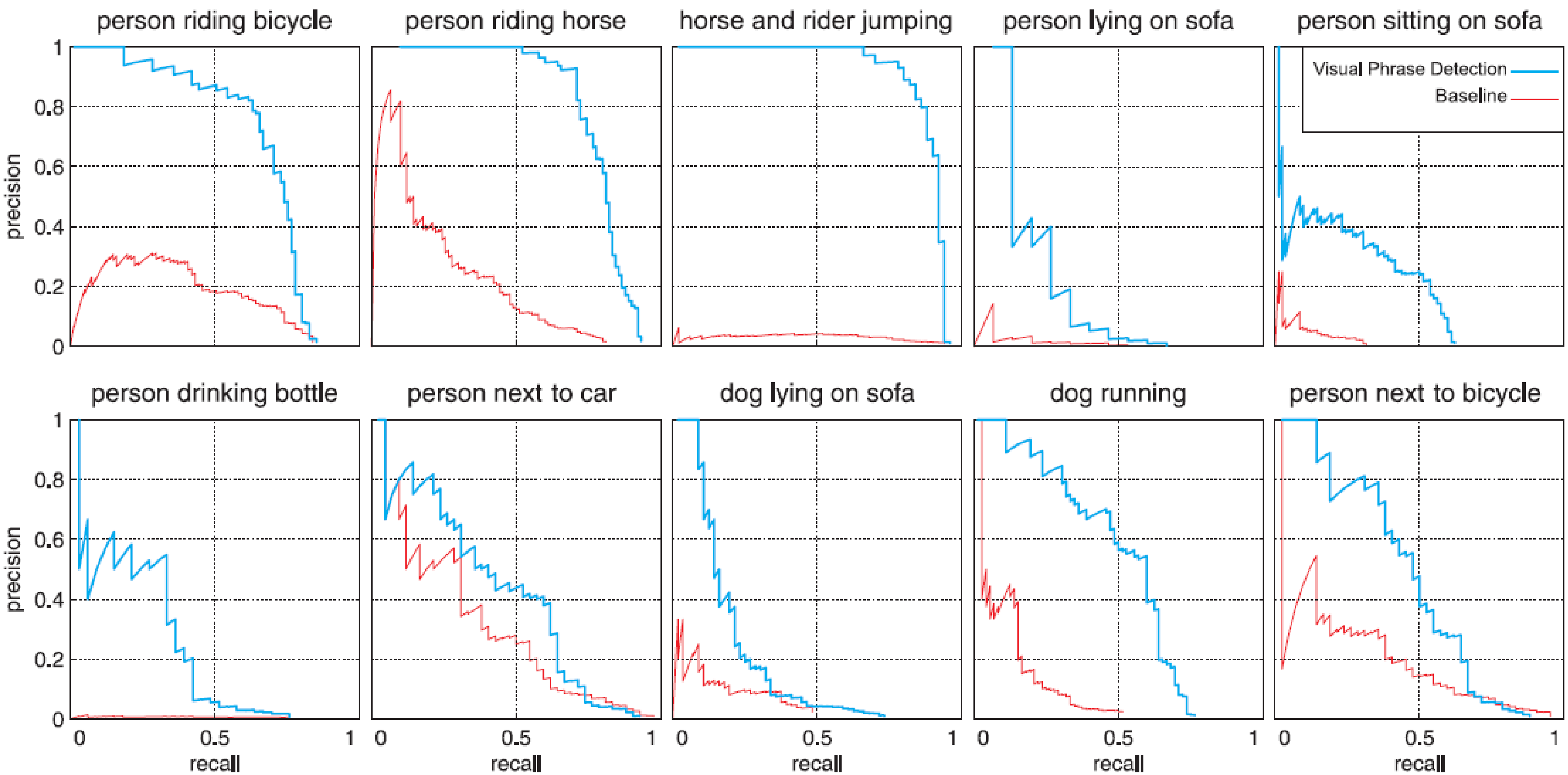
Baseline: From Detected Objects to Visual Phrase Detections



Min(C1,C2)
Max(C1,C2)
Mean(C1,C2)

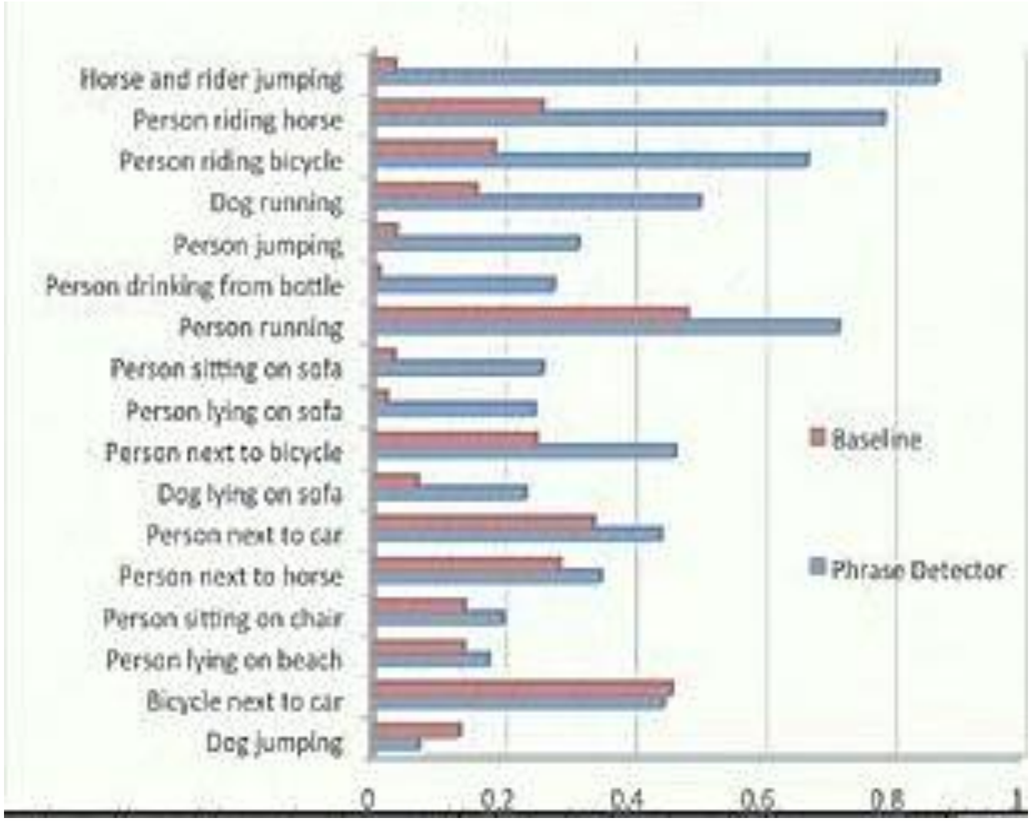
Regress(C1,C2)

Quantitative Results



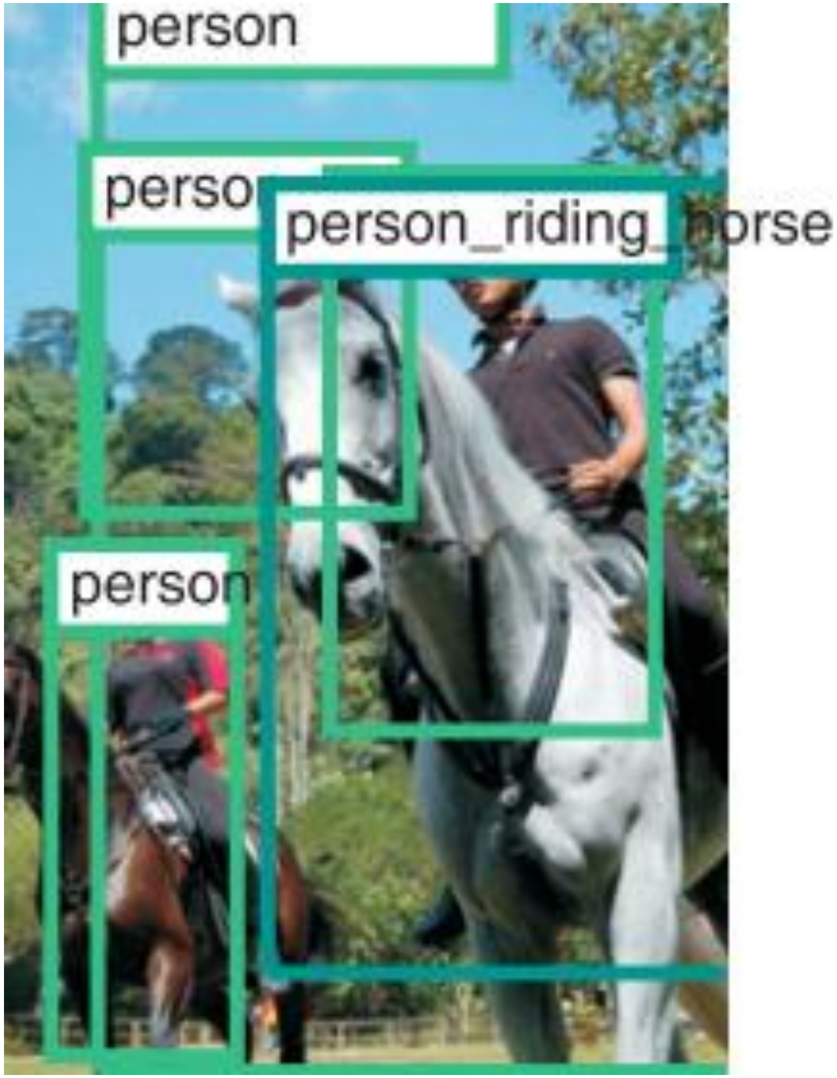
Average Precision

Phrases (Trained with 50 positive images)	Phrase (AP)	Baseline (AP)	Gain (AP)
Person next to bicycle	0.466	0.252	0.214
Person lying on sofa	0.249	0.022	0.227
Horse and rider jumping	0.870	0.035	0.835
Person drinking from bottle	0.279	0.010	0.269
Person sitting on sofa	0.262	0.033	0.229
Person riding horse	0.787	0.262	0.525
Person riding bicycle	0.669	0.188	0.481
Person next to car	0.443	0.340	0.103
Dog lying on sofa	0.235	0.069	0.166
Bicycle next to car	0.448	0.461	-0.013
Dog Jumping	0.072	0.134	-0.062
Person sitting on chair	0.201	0.141	0.060
Person running	0.718	0.484	0.234
Person lying on beach	0.179	0.140	0.039
Person jumping	0.317	0.036	0.281
Person next to horse	0.351	0.287	0.064
Dog running	0.504	0.160	0.344



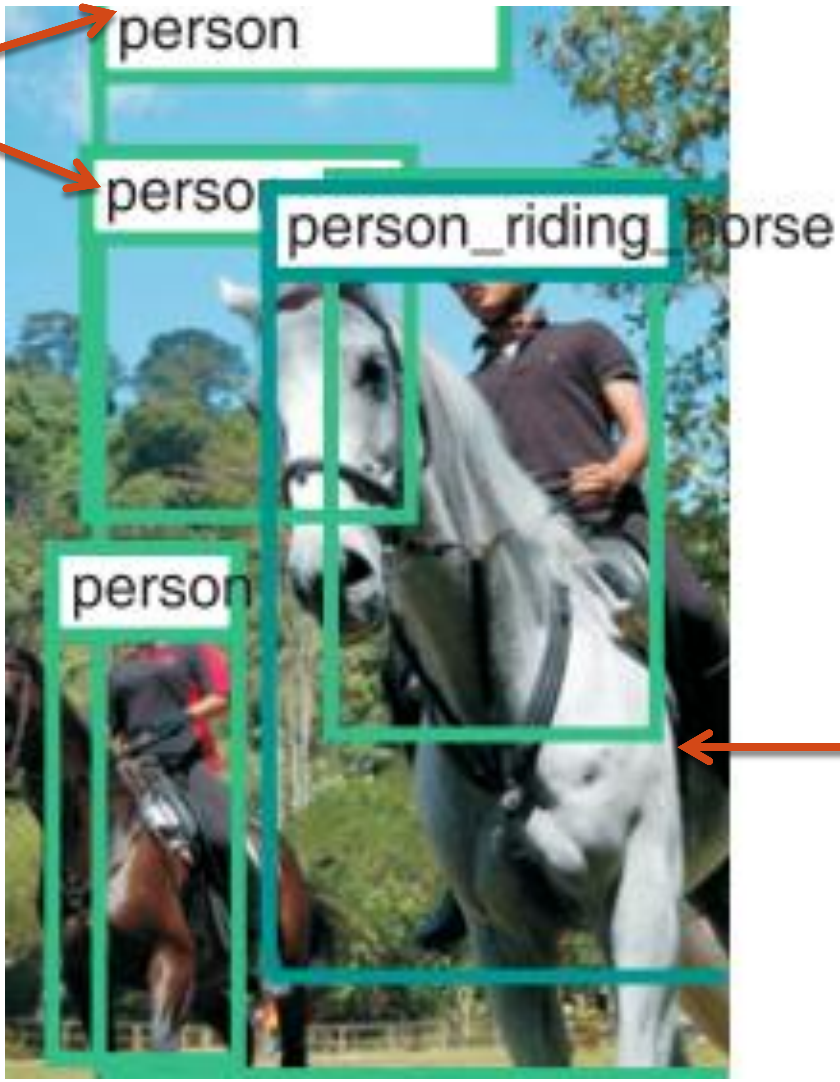
Optimistic upper-bound on how well one can detect visual phrases by individually detecting participating objects then Modeling the relation.

Multiple Independent Detectors



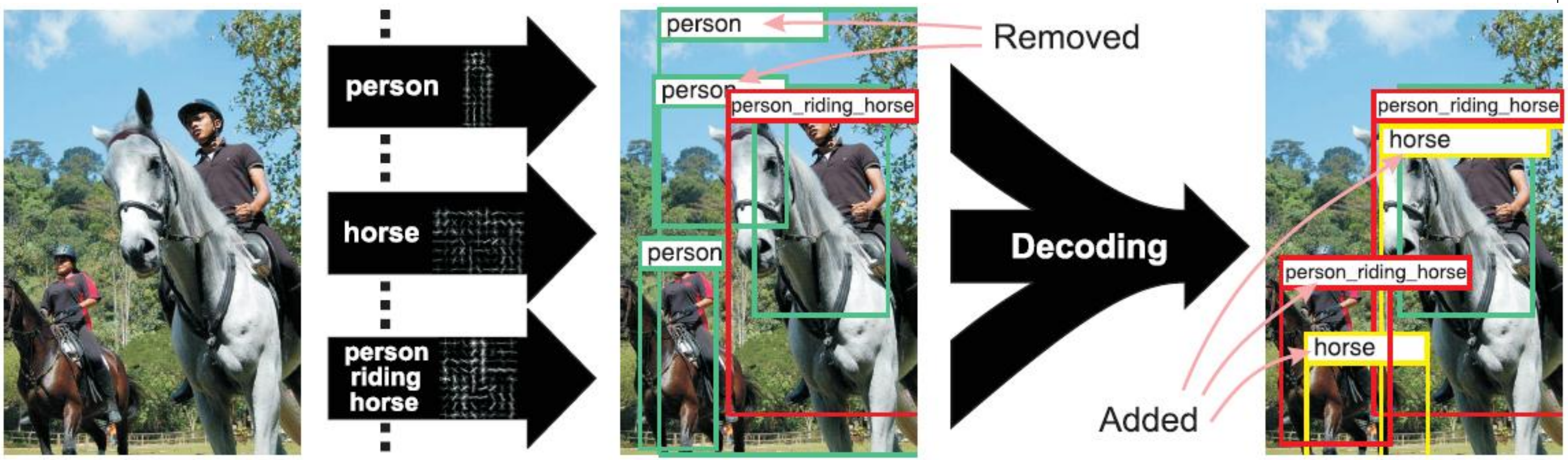
Multiple Independent Detectors

Discourage Predictions

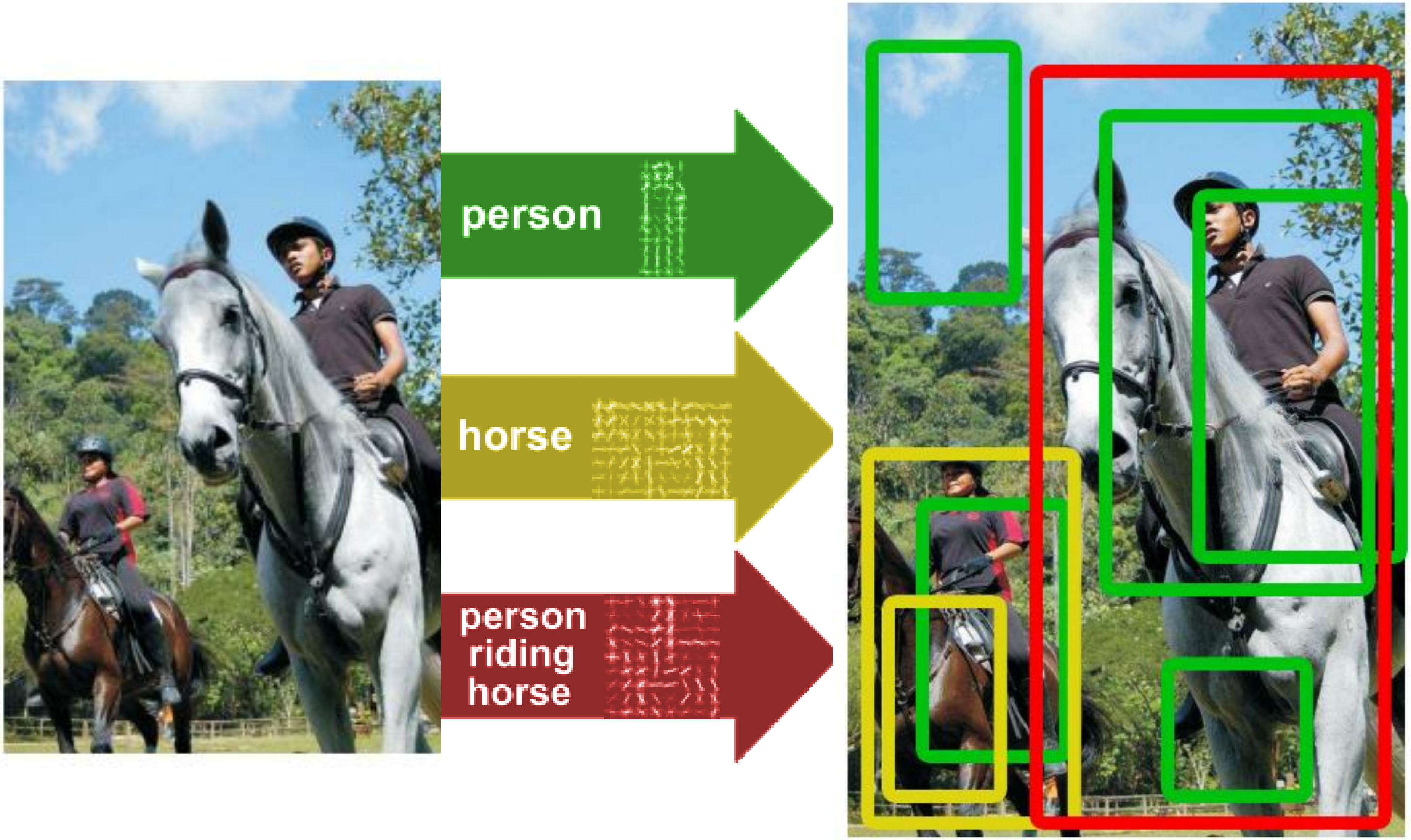


Encourage Predictions

Decoding Multiple Detectors



Design a Visual Phrase Detector

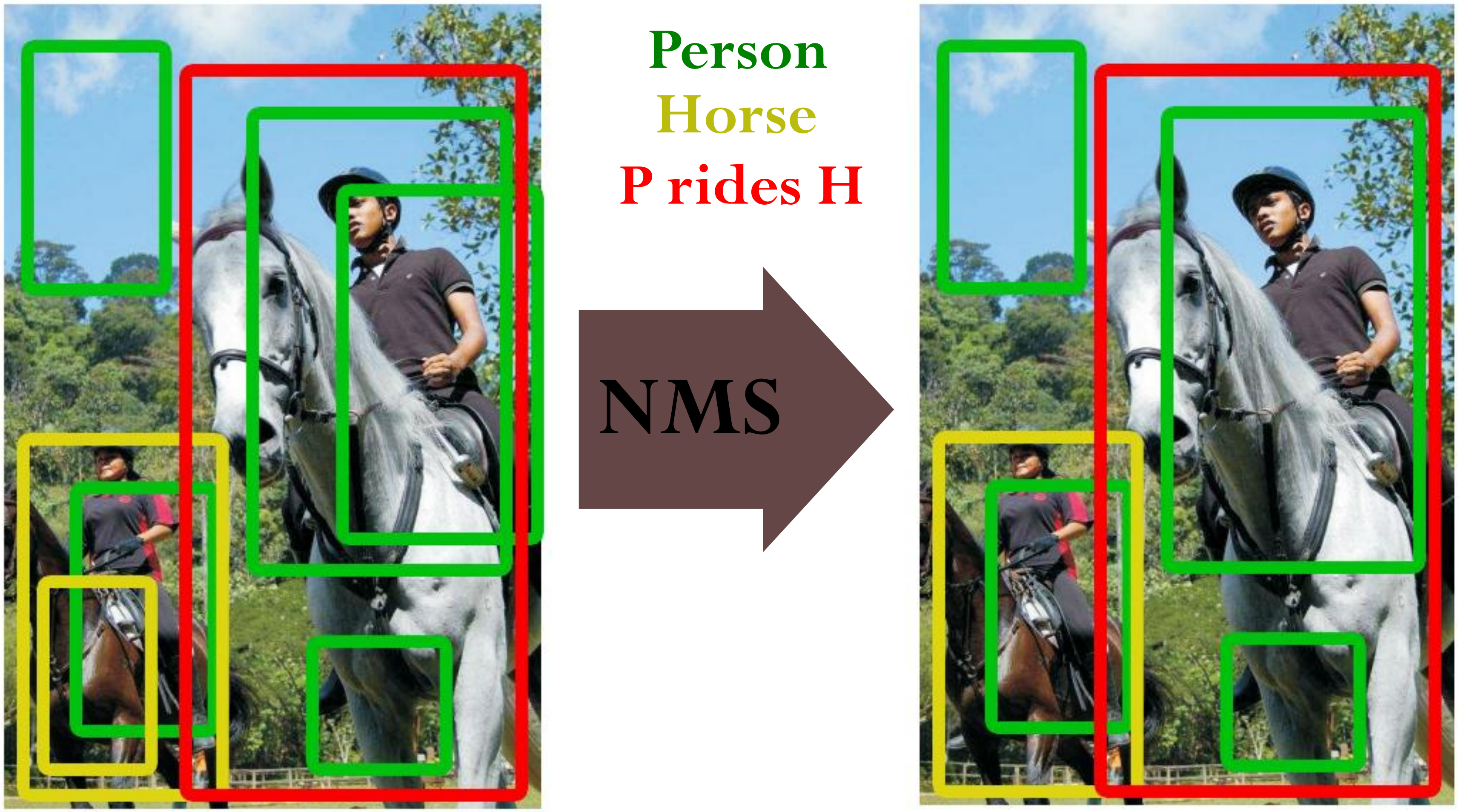


Feature Representation

- Well designed feature representations should make it unnecessary to account for pairwise interactions

- All detectors should be aware of responses of other detectors in a vicinity

Design a Visual Phrase Detector



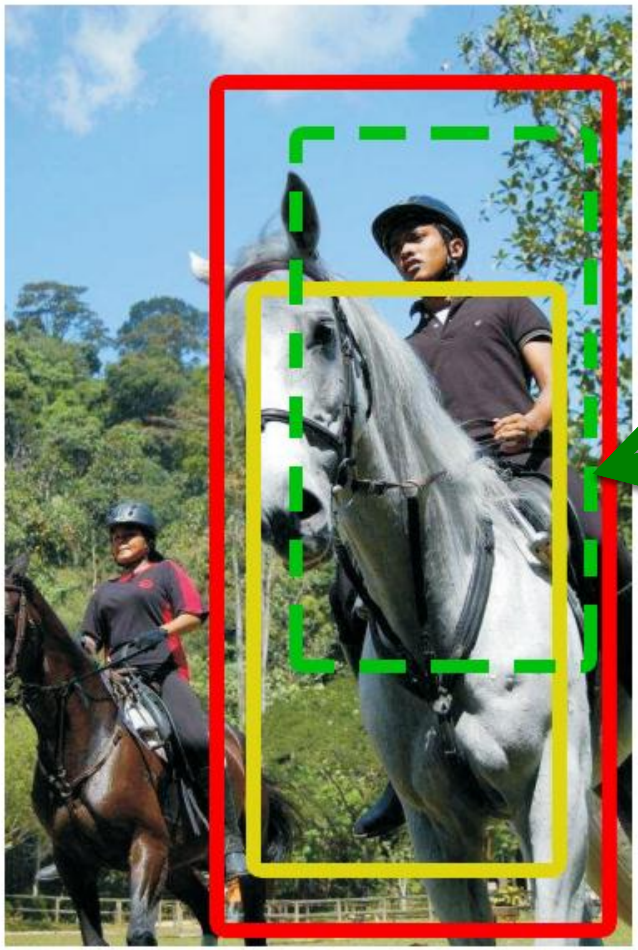
Non-maximum suppression

What's wrong with NMS

We could have done better if visual phrase plays a role

Maybe remove this because some person is riding a horse and there shouldn't be another person under the horse

What's wrong with NMS



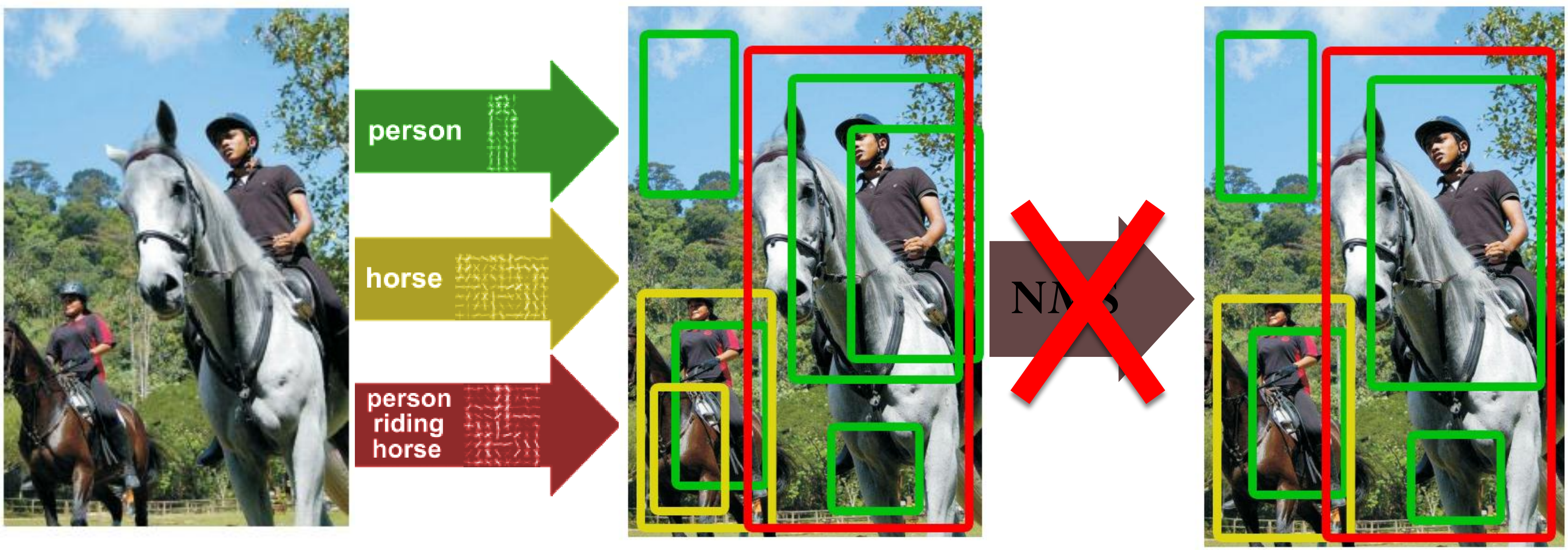
We could have done better
if visual phrase plays a role

If person detector gives a low confidence, but we are pretty sure there are horse and person riding it, confidence for this person should go up

Need a better method that take into account the relationship between objects

NMS to Decoder

Our current pipeline

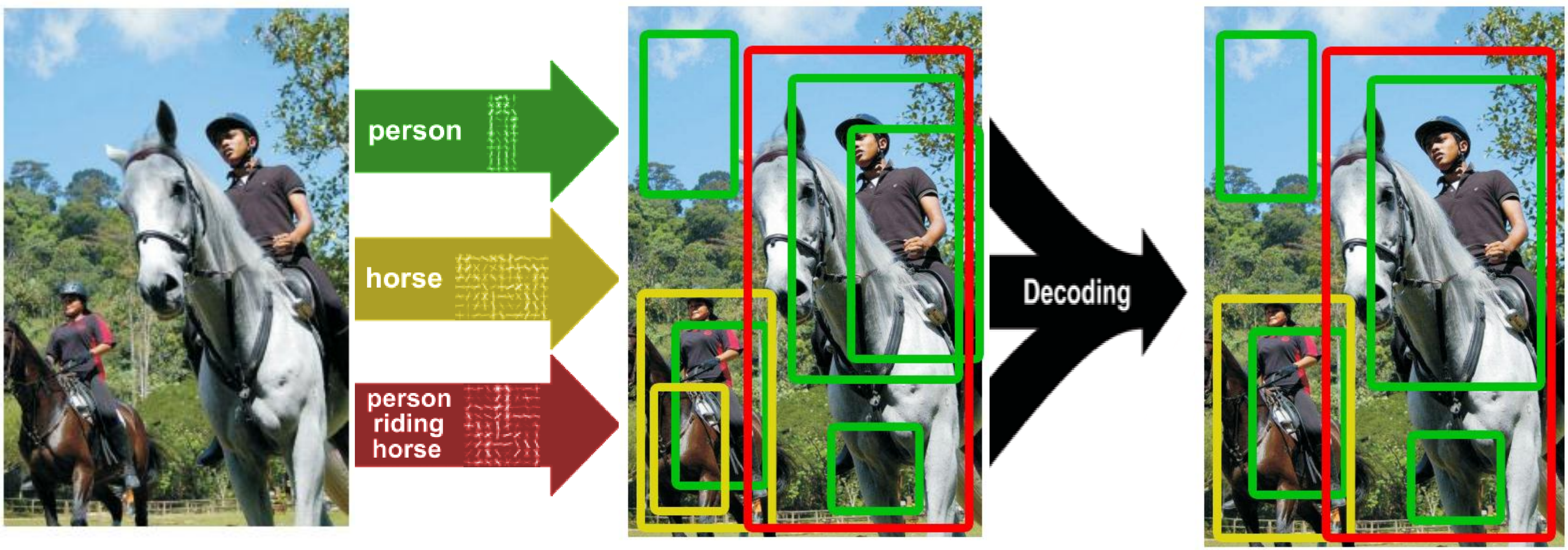


Novel decoding procedure

"Recognition Using Visual Phrases"
Mohammad Sadeghi, Ali Farhadi

NMS to Decoder

Our current pipeline

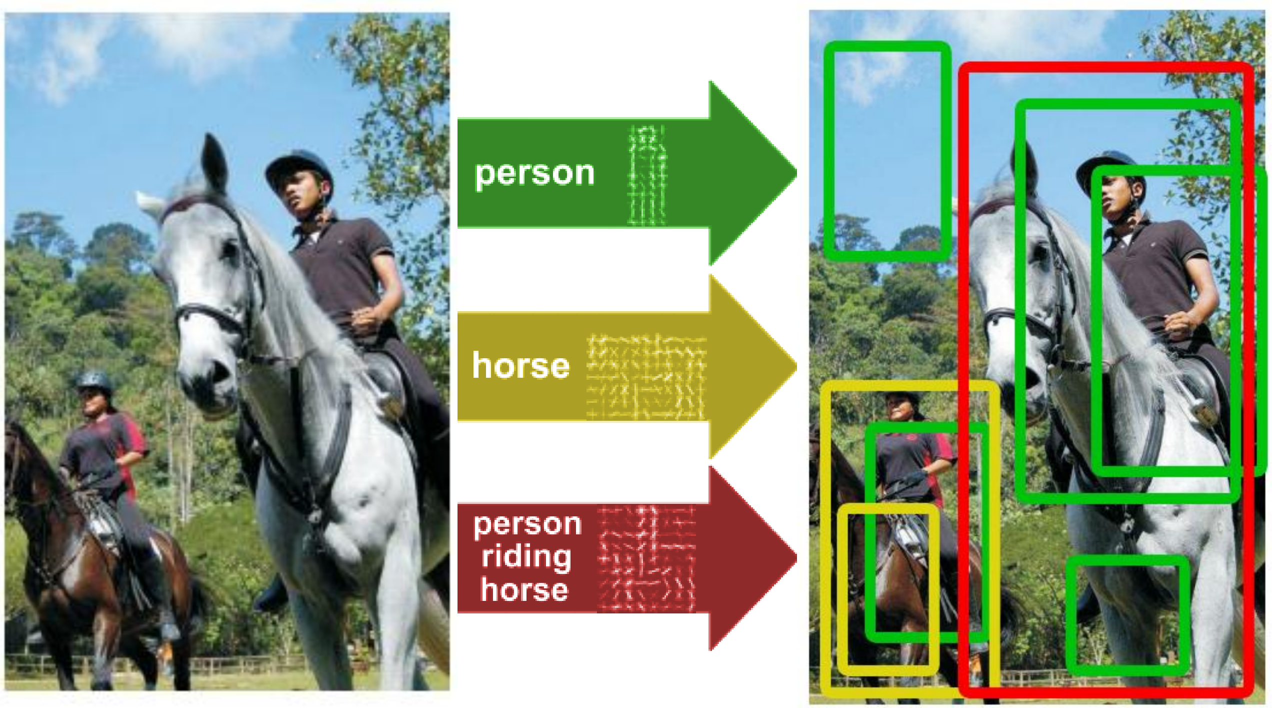


Novel decoding procedure

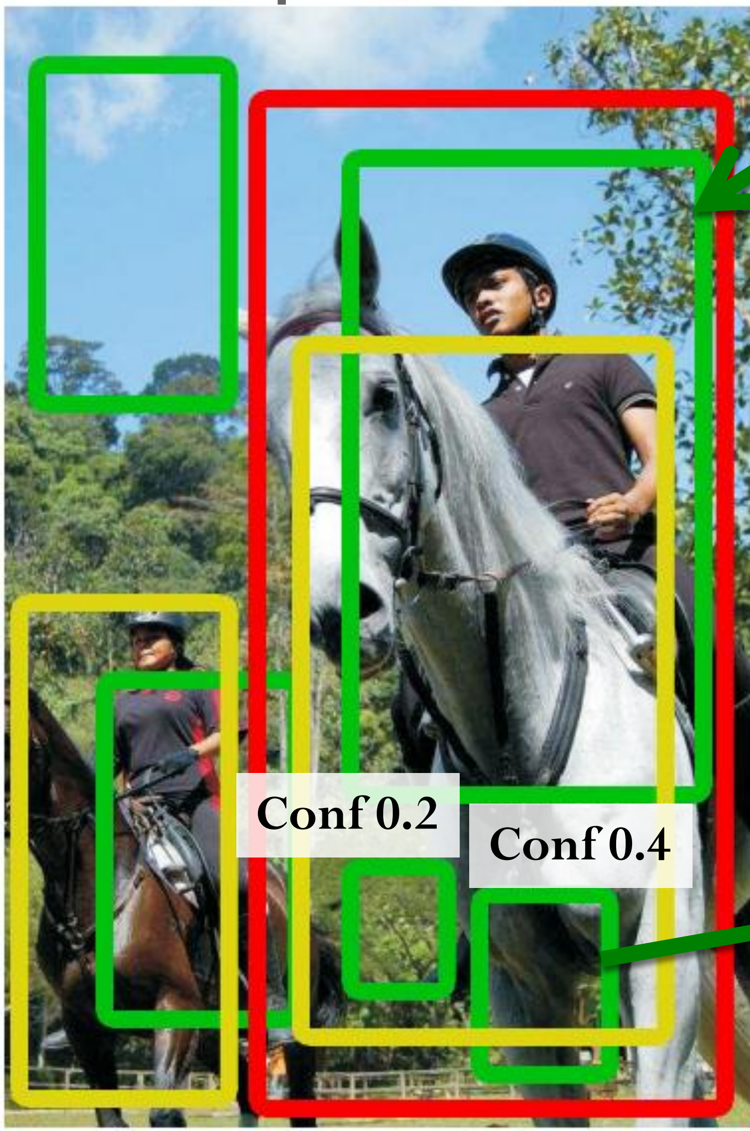
"Recognition Using Visual Phrases"
Mohammad Sadeghi, Ali Farhadi

Redefine Feature

- Decoding needs more info from features
- Goal: a new representation of feature that is aware of the surrounding features



Representation of Feature x_1

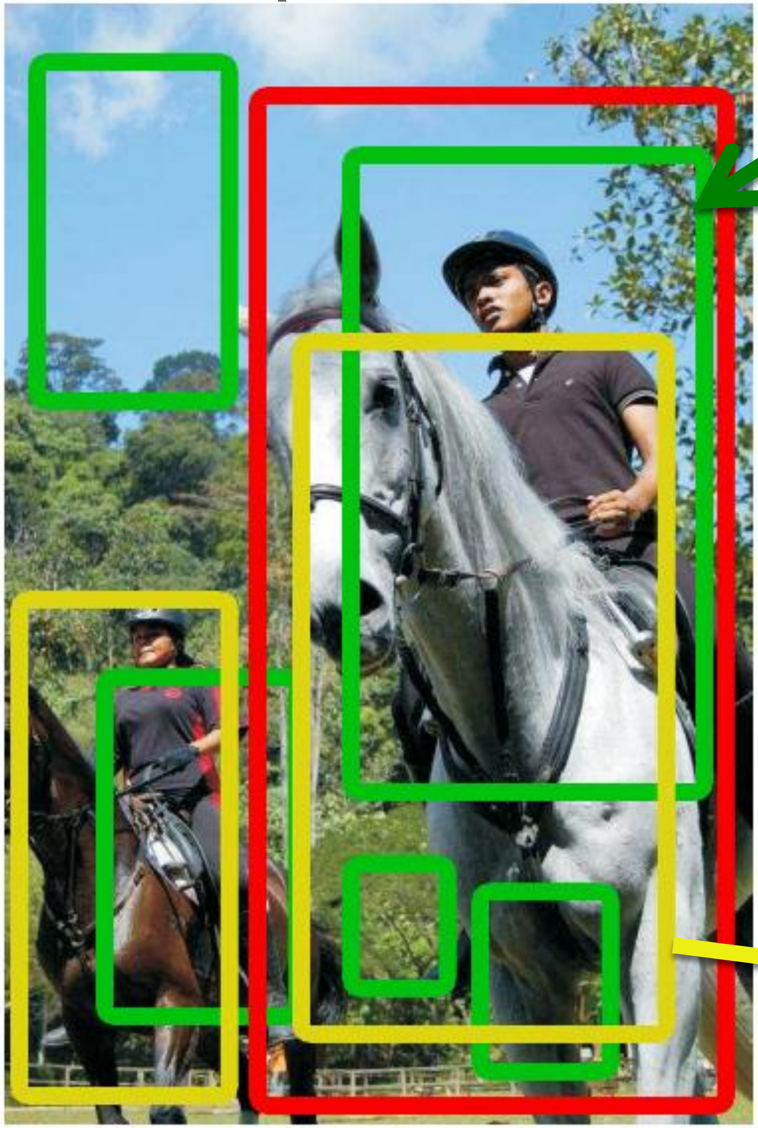


Consider this "person" bounding box
Suppose this is feature x_1

Now let's consider x_1 in relation with other surrounding "person"

	Confidence	Overlap	Size ratio
Above	0	0	0
Below	0.4	0	0.2
Overlap	0	0	0

Representation of Feature x_1

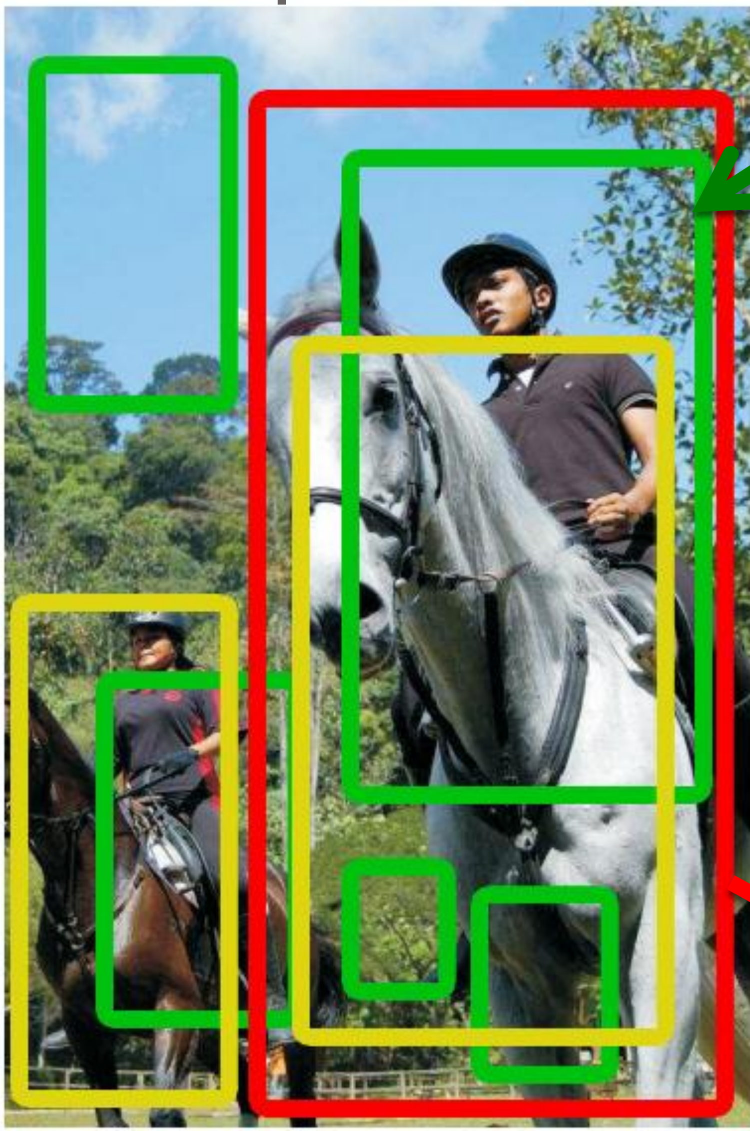


Consider this "person" bounding box
Suppose this is feature x_1

Now let's consider x_1 in relation with other surrounding "horse"

	Confidence	Overlap	Size ratio
Above	0	0	0
Below	0	0	0
Overlap	0.8	0.7	1.2

Representation of Feature x_1

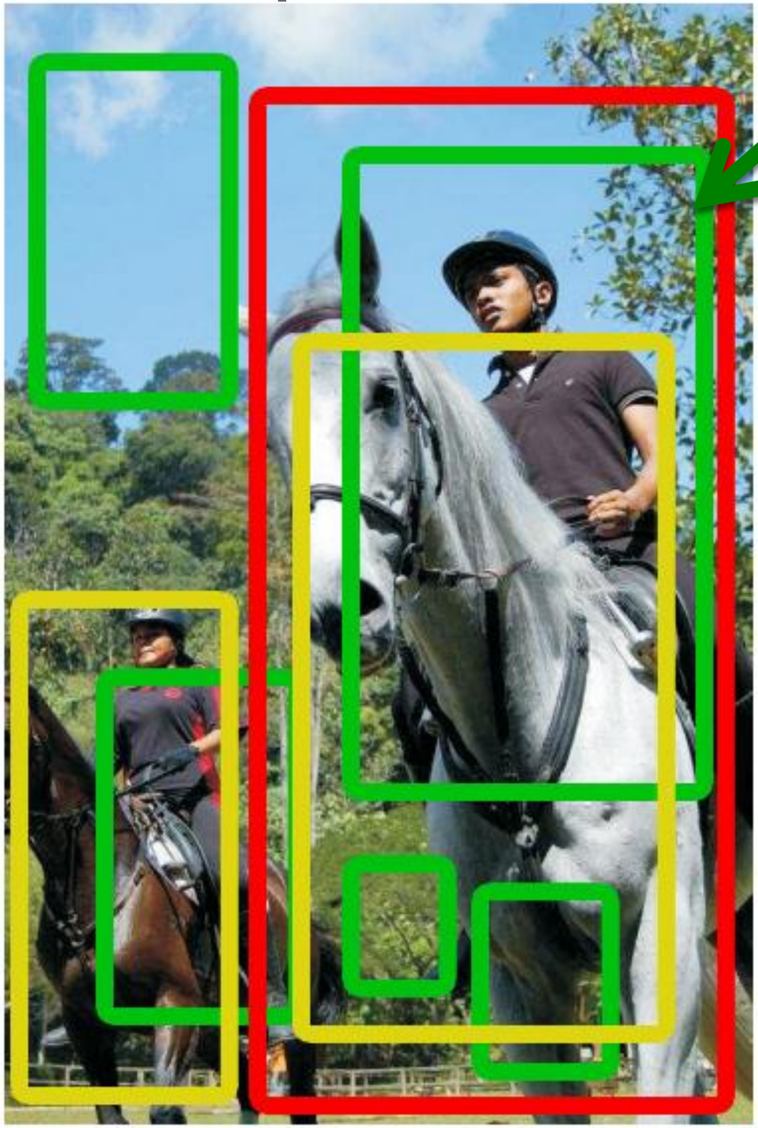


Consider this "person" bounding box
Suppose this is feature x_1

Now let's consider x_1 in relation with other surrounding "P rides H"

	Confidence	Overlap	Size ratio
Above	0	0	0
Below	0	0	0
Overlap	0.9	0.6	1.8

Representation of Feature x_1



feature vector $\mathbf{x}_1$ (class "person")

0	0	0
0.4	0	0.2
0	0	0

Interaction of $\mathbf{x}_1$ with "person"

0	0	0
0	0	0
0.8	0.7	1.2

Interaction of $\mathbf{x}_1$ with "horse"

0	0	0
0	0	0
0.9	0.6	1.8

Interaction of $\mathbf{x}_1$ with "P rides H"

Representation of Feature x_1

"person"

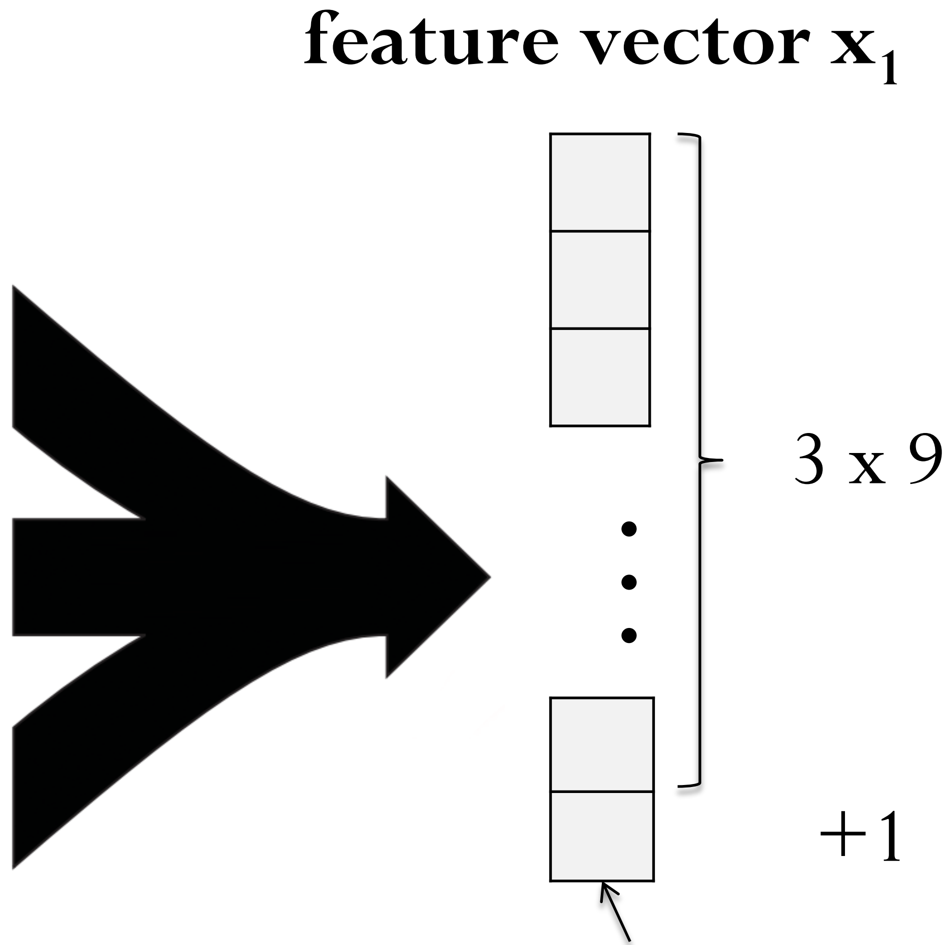
0	0	0
0.4	0	0.2
0	0	0

"horse"

0	0	0
0	0	0
0.8	0.7	1.2

"P rides H"

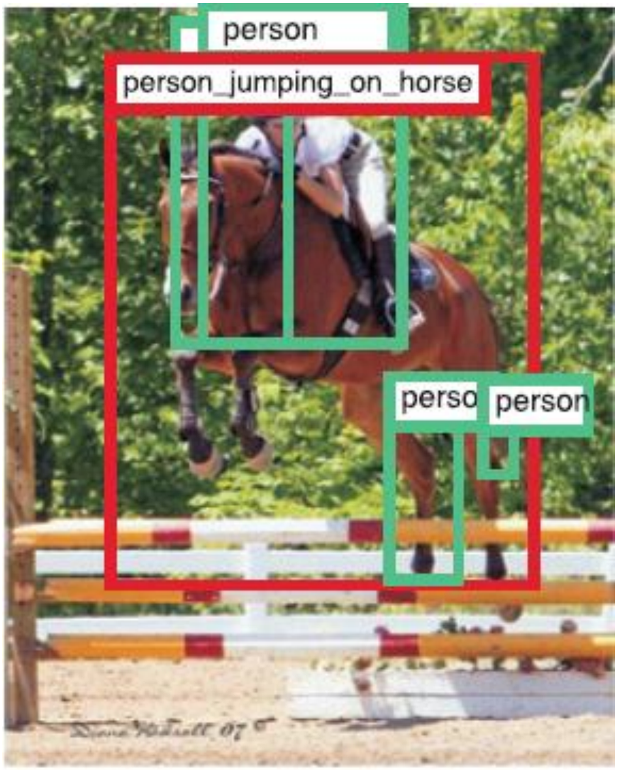
0	0	0
0	0	0
0.9	0.6	1.8



Confidence of this bounding box

More generally (K x 9) + 1, K=# of classes

Decoding

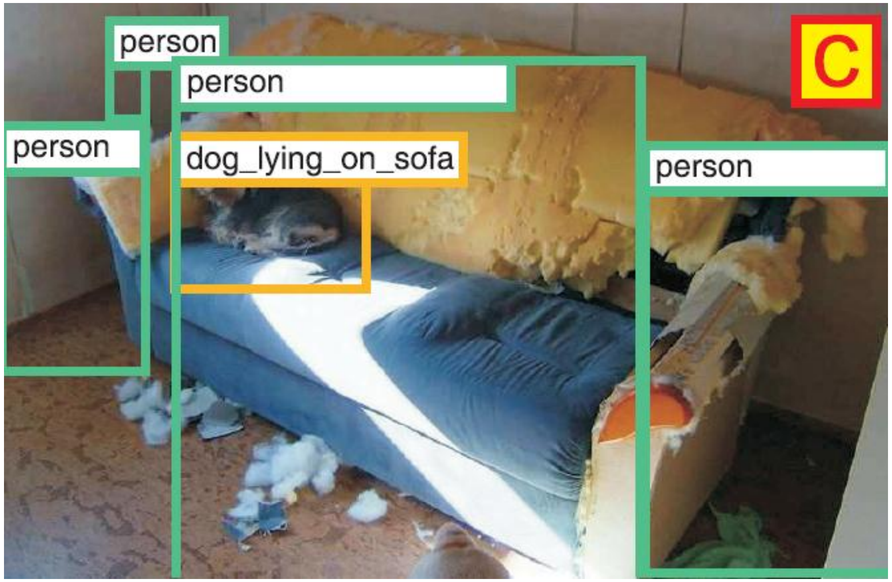


$$\min_{w} \sum_{c \in \{0,\ldots,K\}} \frac{1}{2} \| w_c \|_2^2 + \tag{3}$$

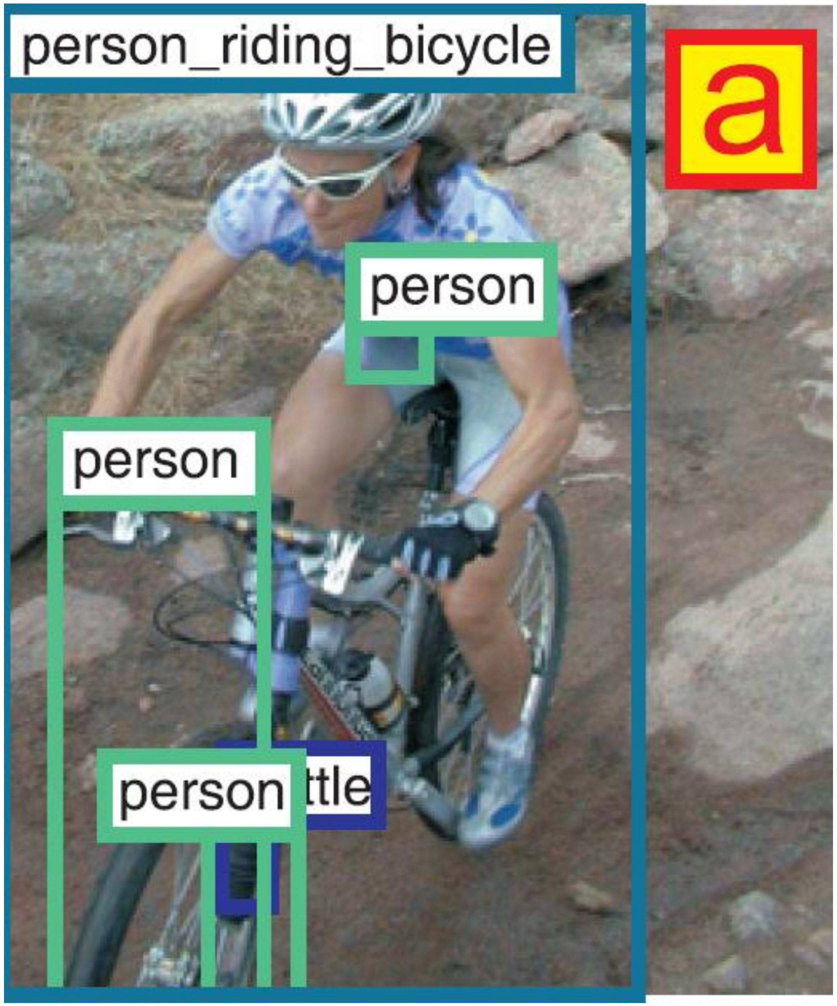
$$\lambda \sum_{n}^{N} \sum_{i}^{M} w_{c_i}^T (\phi(X_n, h_{n,i}^*) - \phi(X_n, y_{n,i})) + L(H_n^*, Y_n)$$

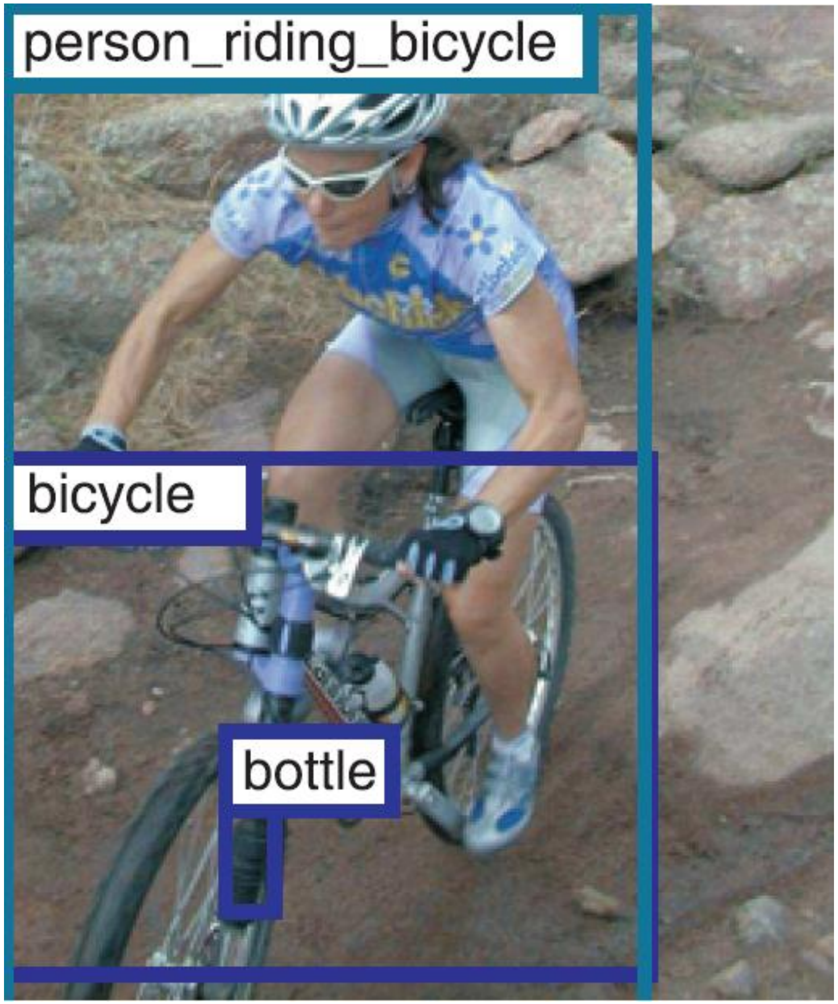
$$\text{s.t. } \; H_n^* = \arg\max_{H_n} \sum_{i}^{M} w_{c_i}^T \phi(X_n, h_{n,i}) + L(H_n, Y_n) \tag{4}$$

Before and After

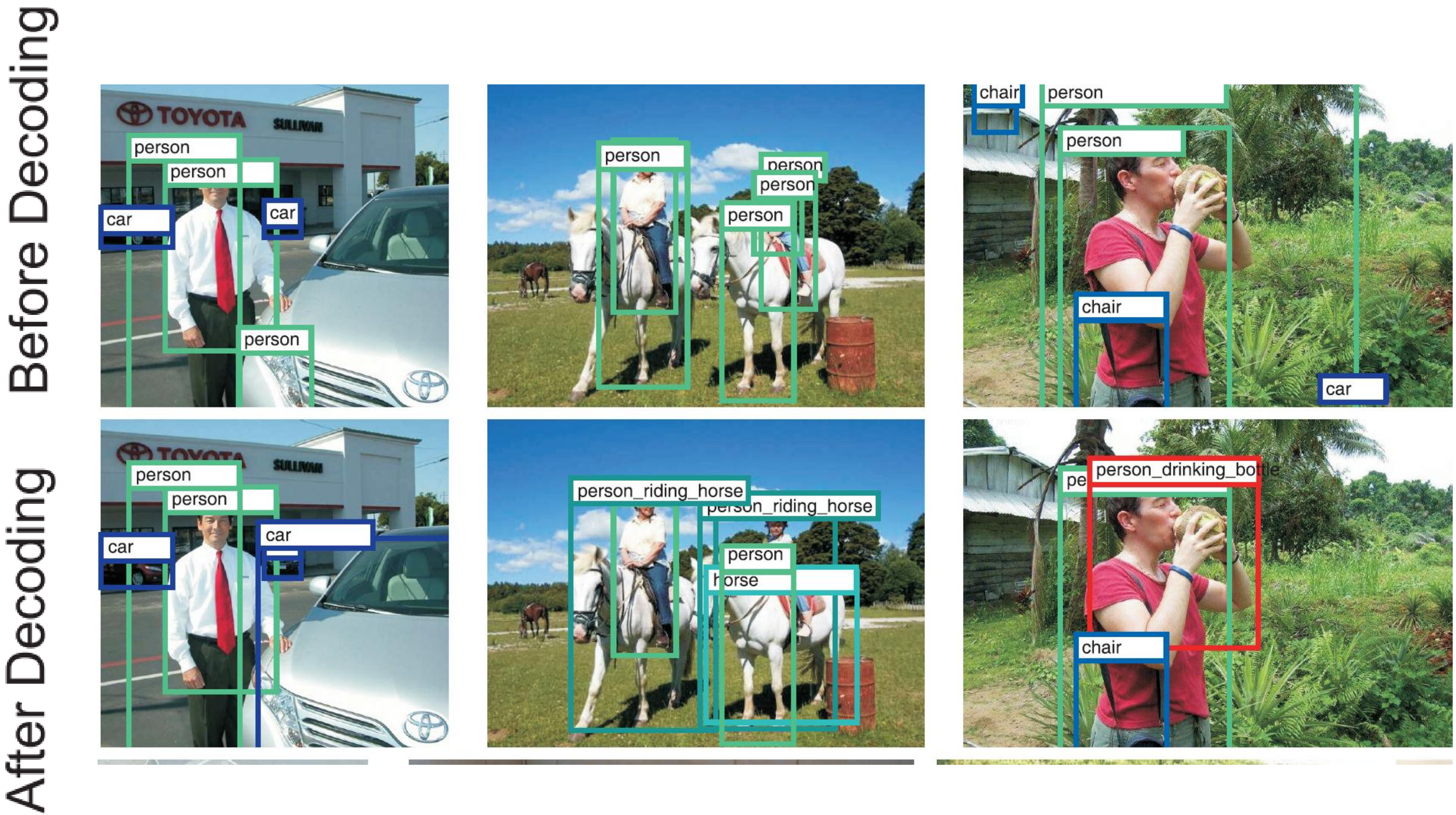


Before and After



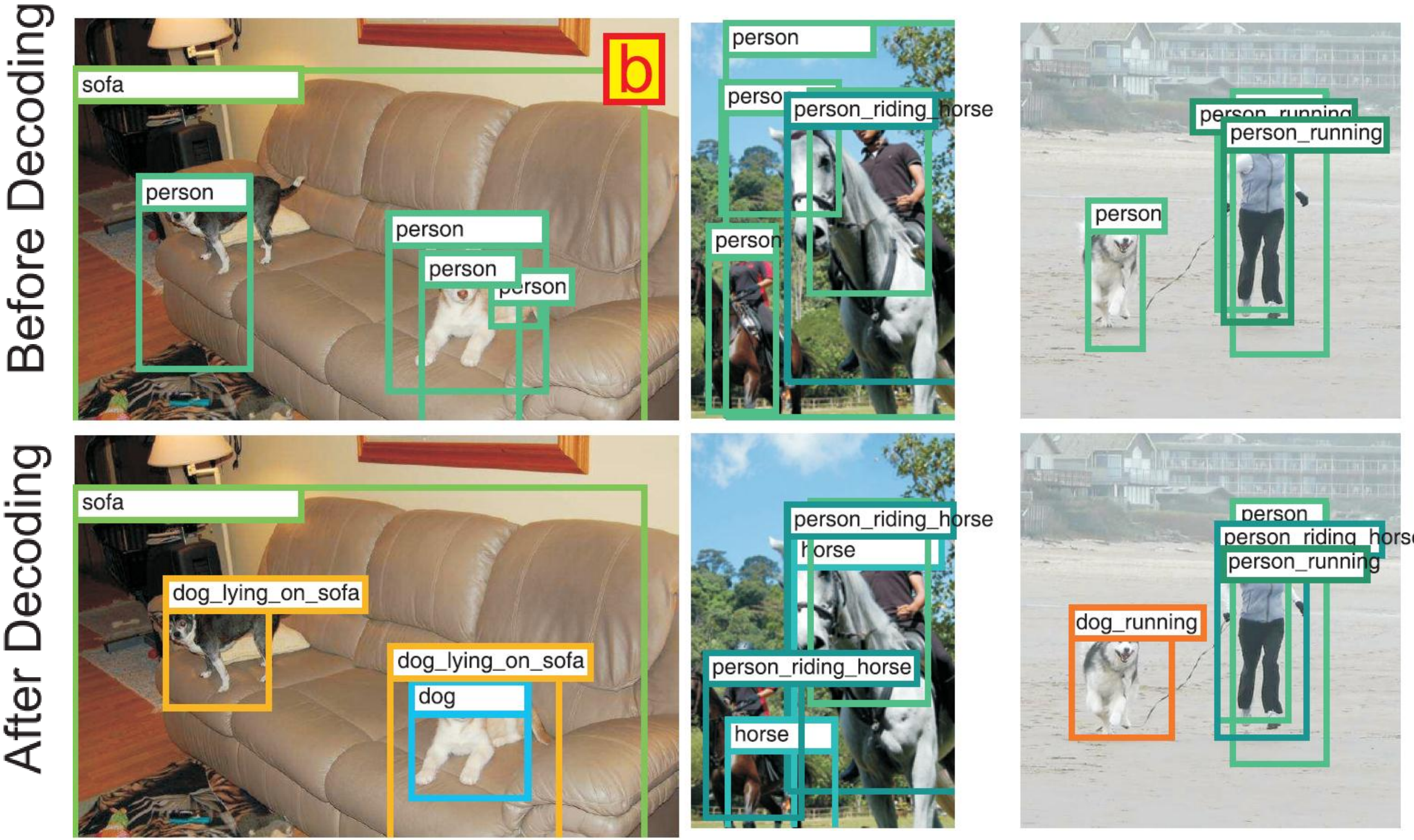


Results



Before Decoding
After Decoding
TOYOTA
SULLIVAN
person
car
person_riding_horse
horse
chair
person_drinking_bottle

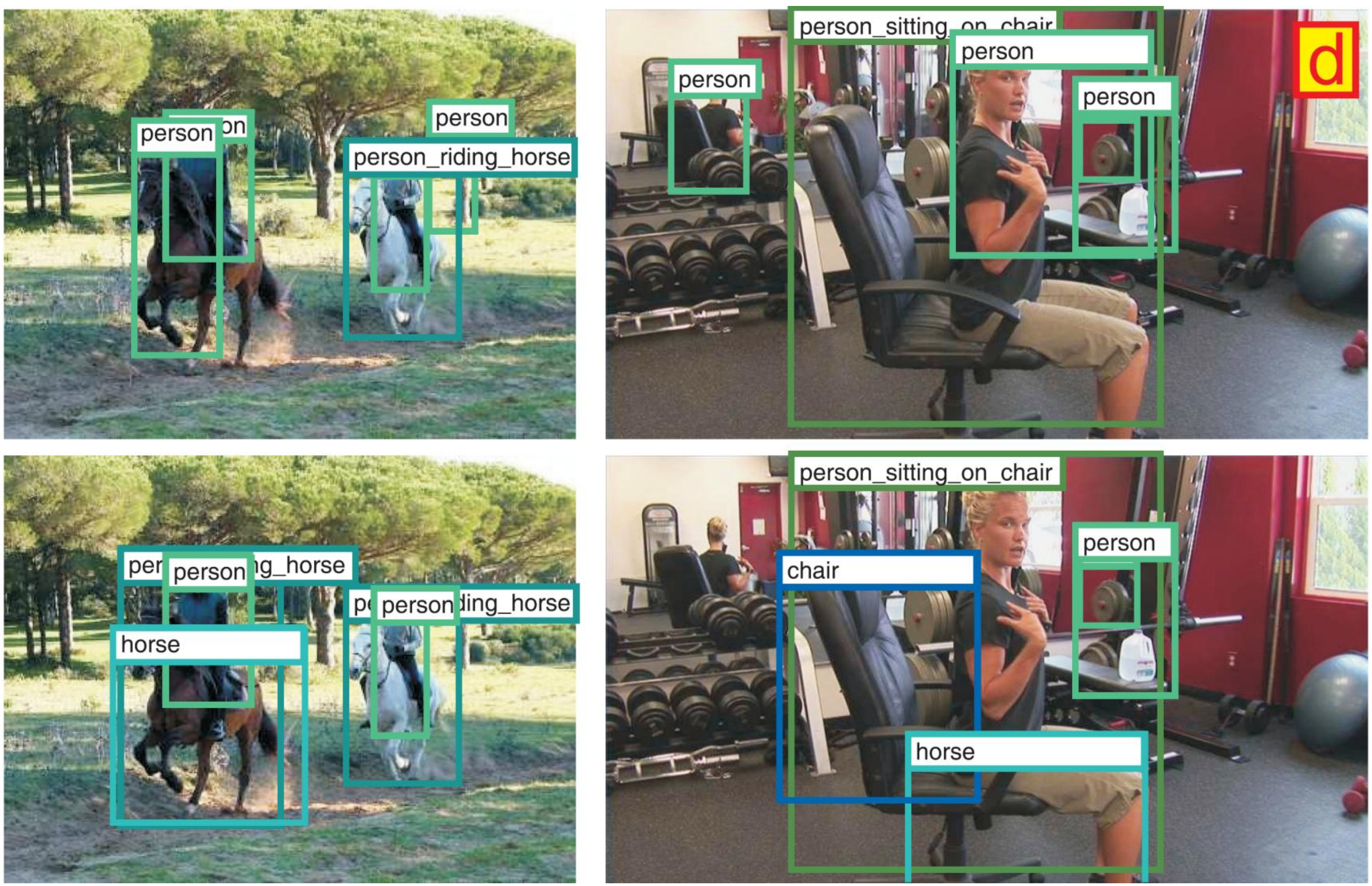
Results


Before Decoding
After Decoding
b
sofa
person
person
person
person
person
person
person_riding_horse
person
person_running
person_running
person
sofa
dog_lying_on_sofa
dog_lying_on_sofa
dog
person_riding_horse
horse
person_riding_horse
horse
person
person_riding_horse
person_running
dog_running

Results

Before Decoding

After Decoding



Results

	bicycle	bottle	car	chair	dog	horse	person	sofa
detectors of [8]	0.434	0.429	0.329	0.213	0.316	0.438	0.295	0.204
[2] without phrases	0.431	0.425	0.191	0.225	0.297	0.475	0.204	0.167
[2] with phrases	0.449	**0.435**	0.228	0.217	0.316	0.462	0.286	0.204
Our decoding without phrases	0.437	0.434	0.330	0.216	0.329	0.440	0.297	0.218
Our decoding with phrases	**0.457**	**0.435**	**0.344**	**0.227**	**0.335**	**0.485**	**0.302**	**0.260**

This method outperforms state-of-the-art object detector and state-of-the-art multiclass recognition method of C. F. C. Desai, D. Ramana.

[2] C. F. C. Desai, D. Ramanan. Discriminative models for multi-class object layout. In *ICCV*, 2010. 1348, 1349, 1351, 1352, 1353

Results

	bicycle	bottle	car	chair	dog	horse	person	sofa
detectors of [8]	0.434	0.429	0.329	0.213	0.316	0.438	0.295	0.204
[2] without phrases	0.431	0.425	0.191	0.225	0.297	0.475	0.204	0.167
[2] with phrases	0.449	**0.435**	0.228	0.217	0.316	0.462	0.286	0.204

[2] C. F. C. Desai, D. Ramanan. Discriminative models for multi-class object layout. In *ICCV*, 2010. 1348, 1349, 1351, 1352, 1353

Results

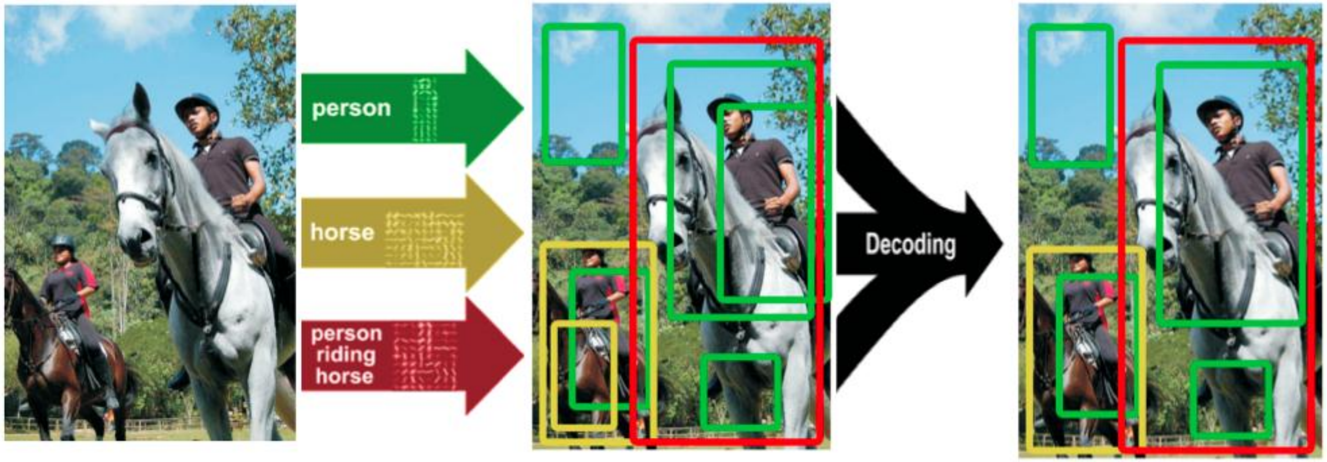
	bicycle	bottle	car	chair	dog	horse	person	sofa
detectors of [8]	0.434	0.429	0.329	0.213	0.316	0.438	0.295	0.204
Our decoding without phrases	0.437	0.434	0.330	0.216	0.329	0.440	0.297	0.218
Our decoding with phrases	**0.457**	**0.435**	**0.344**	**0.227**	**0.335**	**0.485**	**0.302**	**0.260**

[2] C. F. C. Desai, D. Ramanan. Discriminative models for multi-class object layout. In *ICCV*, 2010. 1348, 1349, 1351, 1352, 1353

Results

	bicycle	bottle	car	chair	dog	horse	person	sofa
detectors of [8]	0.434	0.429	0.329	0.213	0.316	0.438	0.295	0.204
[2] without phrases	0.431	0.425	0.191	0.225	0.297	0.475	0.204	0.167
[2] with phrases	0.449	**0.435**	0.228	0.217	0.316	0.462	0.286	0.204
Our decoding without phrases	0.437	0.434	0.330	0.216	0.329	0.440	0.297	0.218
Our decoding with phrases	**0.457**	**0.435**	**0.344**	**0.227**	**0.335**	**0.485**	**0.302**	**0.260**

This method outperforms state-of-the-art object detector and state-of-the-art multiclass recognition method of C. F. C. Desai, D. Ramana.



[2] C. F. C. Desai, D. Ramanan. Discriminative models for multi-class object layout. In *ICCV*, 2010. 1348, 1349, 1351, 1352, 1353

Conclusion

- Visual Phrases
  - Bigger than objects and smaller than scenes
  - Substantial gain in understanding images
- Phrasal recognition help object recognition
  - Including to the vocabulary of recognition
  - Decoding
- What should we recognize
  - Semantic spectrum of elements of recognition
- Visual phrases in practice, limitations

Images used in presentation are taken from web and UIUC Phrasal Recognition Dataset,
Slides based on authors' presentation