Recognition Using Visual Phrases

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

CVPR'11, Best Student Paper
Object Recognition
Object Recognition

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

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

Parts, Poselets and Attributes

Scenes
What is a Visual Phrase?
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

- Correlates 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

Person

Horse
Participating in Phrases affects the appearance of the objects
Visual composites might be much easier to detect than their participant components.

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

person_drinking_bottle

person_drinking_bottle
person
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

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

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

Person
Horse
P rides H

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
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 |
Consider this “person” bounding box.

Suppose this is feature \( x_1 \).

Now let’s consider \( x_1 \) in relation with other surrounding “horse”.

<table>
<thead>
<tr>
<th>Confidence</th>
<th>Overlap</th>
<th>Size ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Above</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Below</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Overlap</td>
<td>0.8</td>
<td>0.7</td>
</tr>
</tbody>
</table>

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 $x_1$ (class “person”)

<table>
<thead>
<tr>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Interaction of $x_1$ with “person”

<table>
<thead>
<tr>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7</td>
<td>1.2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Interaction of $x_1$ with “horse”

<table>
<thead>
<tr>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Interaction of $x_1$ with “P rides H”
**Representation of Feature $x_1$**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>“person”</td>
<td>0</td>
</tr>
<tr>
<td>“horse”</td>
<td>0</td>
</tr>
<tr>
<td>“P rides H”</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Feature vector $x_1$:

```
[0, 0, 0, 0.4, 0, 0.2, 0, 0, 0, 0.8, 0.7, 1.2, 0, 0, 0, 0.9, 0.6, 1.8, 0, 0, 0]
```

More generally $(K \times 9) + 1$, $K=\#$ of classes.
Decoding

\[
\min_w \sum_{c \in \{0, \ldots, K\}} \frac{1}{2} \| w_c \|^2_2 + \\
\lambda \sum_n \sum_i w^T_{c_i} (\phi(X_n, h^*_n, i) - \phi(X_n, y_n, i)) + L(H^*_n, Y_n)
\]

s.t. \( H^*_n = \arg \max_{H_n} \sum_i w^T_{c_i} \phi(X_n, h_n, i) + L(H_n, Y_n) \)
Before and After
Before and After
Results
Results
Results
Results

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

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

## Results

<table>
<thead>
<tr>
<th></th>
<th>bicycle</th>
<th>bottle</th>
<th>car</th>
<th>chair</th>
<th>dog</th>
<th>horse</th>
<th>person</th>
<th>sofa</th>
</tr>
</thead>
<tbody>
<tr>
<td>detectors of [8]</td>
<td>0.434</td>
<td>0.429</td>
<td>0.329</td>
<td>0.213</td>
<td>0.316</td>
<td>0.438</td>
<td>0.295</td>
<td>0.204</td>
</tr>
<tr>
<td>[2] without phrases</td>
<td>0.431</td>
<td>0.425</td>
<td>0.191</td>
<td>0.225</td>
<td>0.297</td>
<td>0.475</td>
<td>0.204</td>
<td>0.167</td>
</tr>
<tr>
<td>[2] with phrases</td>
<td>0.449</td>
<td><strong>0.435</strong></td>
<td>0.228</td>
<td>0.217</td>
<td>0.316</td>
<td>0.462</td>
<td>0.286</td>
<td>0.204</td>
</tr>
</tbody>
</table>

Results

<table>
<thead>
<tr>
<th></th>
<th>bicycle</th>
<th>bottle</th>
<th>car</th>
<th>chair</th>
<th>dog</th>
<th>horse</th>
<th>person</th>
<th>sofa</th>
</tr>
</thead>
<tbody>
<tr>
<td>detectors of [8]</td>
<td>0.434</td>
<td>0.429</td>
<td>0.329</td>
<td>0.213</td>
<td>0.316</td>
<td>0.438</td>
<td>0.295</td>
<td>0.204</td>
</tr>
<tr>
<td>Our decoding without phrases</td>
<td>0.437</td>
<td>0.434</td>
<td>0.330</td>
<td>0.216</td>
<td>0.329</td>
<td>0.440</td>
<td>0.297</td>
<td>0.218</td>
</tr>
<tr>
<td>Our decoding with phrases</td>
<td><strong>0.457</strong></td>
<td><strong>0.435</strong></td>
<td><strong>0.344</strong></td>
<td><strong>0.227</strong></td>
<td><strong>0.335</strong></td>
<td><strong>0.485</strong></td>
<td><strong>0.302</strong></td>
<td><strong>0.260</strong></td>
</tr>
</tbody>
</table>

### Results

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

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

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
Any questions?