The Most Telling Window for Image Classification

Contributors:
Jasper Uijlings
Koen van de Sande
Arnold Smeulders
Theo Gevers
Nicu Sebe
Cees Snoek

Under submission at TPAMI. Please contact jrr@disi.unitn.it before using this work.
Image Classification

CAR?
Bag-of-Words
Is Global Optimal?

~ Relevant conclusions:

~ The object alone yields significantly more accuracy than the whole image

~ Once the object location is known, context contributes very little.

Visualising Bag of Words
Demo @ ICCV
Tuesday 17:20 – 20:00

We need an explicit object location

It has been shown that object localisation can improve classification:


Joint Winner Pascal 2008 Detection Challenge
Is Exact Localisation Optimal?
Is Exact Localisation Optimal?
Is Exact Localisation Optimal?

Parts were earlier used in “visual identification” to distinguish Bob's from Mary's Mercedes

*Learning to Locate Informative Features for Visual Identification*, IJCV 2008,
A. Ferencz, E. Learned-Miller, J. Malik
Is Exact Localisation Optimal?

Parts may be more discriminative because of pose change, often caused by interaction.
Is Exact Localisation Optimal?

For occluded objects only the non-occluded part is informative.
Is Exact Localisation Optimal?

In crowded scenes, compared to an individual object: a collection is both more easy to find and may be more discriminative
Is Exact Localisation Optimal: NO

~ Parts may be more discriminative for some classes.

~ Interacting objects may change pose, retaining typical appearance only for object part.

~ Occluded objects are hard to find when searching for complete objects.

~ In crowded scenes groups are more easy to recognize.

The Windows that Tell the Story of an Image, J.R.R. Uijlings and A.W.M. Smeulders. Under submission at TPAMI. Please contact jrr@disi.unitn.it before using this work.
The Most Telling Window

May focus on:

- Object Parts
- Complete Objects
- Object Collections

*The Windows that Tell the Story of an Image*, J.R.R. Uijlings and A.W.M. Smeulders. Under submission at TPAMI. Please contact jrr@disi.unitn.it before using this work.
Methodology Most Telling Window

~ Object Location

~ General framework training/classification
Methodology: Object Location

Most Dominant: Sliding Windows.

But yields 100,000 – 1,000,000 windows: infeasible for powerful Bag-of-Words implementation.

Solution: Selective Search
Methodology: Object Location

We introduce Selective Search

Which uses multiple, complementary, hierarchical segmentations.

More details in ILSVRC presentation

Segmentation as Selective Search for Object Recognition, ICCV 2011,
K.E.A. van de Sande, J.R.R. Uijlings, T. Gevers, and A.W.M. Smeulders,
Poster #42, Wednesday 17:20-20:00

Matlab pcode for selective search will be released soon.
Methodology: Object Location

- Small set of class-independent locations
- Captures parts, objects, and collections

Example Windows generated by our method:

_Segmentation as Selective Search for Object Recognition, ICCV 2011, K.E.A. van de Sande, J.R.R. Uijlings, T. Gevers, and A.W.M. Smeulders, Poster #42, Wednesday 17:20-20:00_
Methodology: Framework

Normal Bag-of-Words

Descriptor Extraction → Visual Word Assignment

Training

Use Complete image → Train SVM model

Classification

Use Complete image → Classification
Methodology: Framework

Descriptor Extraction → Visual Word Assignment

Training
- Use ground truth windows
- Train SVM model

Classification
- Selective Search Locations
- Classification

Most Telling Window
Methodology: Framework

Descriptor Extraction → Visual Word Assignment → Training

Use ground truth windows → Train SVM model

Extra Negatives

Selective Search Locations → Classification

Most Telling Window

Retraining: e.g. Laptev 2009, Felzenszwalb et al. 2010
Localisation vs Most Telling Window

Localisation:
- Ferrari cars
- Trains
- Motorcycles

Most Telling Window:
- Positive images
- No negative examples from positive images!
Localisation vs Most Telling Window

~ Large difference in motivation:
   ~ Parts
   ~ Complete objects
   ~ Collections of objects

~ Subtle difference in training windows

~ Significant difference in final results

~ (Of course, it would be better to also obtain new positive examples in retraining loop)
Implementation details

- Pixel-wise sampling
- (Colour) SIFT descriptors (Lowe04, Sande2010)
- K-means visual vocabulary
- Hard assignment.
- Store “Visual Word Images”
- Spatial Pyramid (Lazebnik06). BoW: 1x1, 2x2, 1x3. MTW: 2x2/4x4
- Bag-of-Words GPU acceleration (Sande2011)
- Selective Search (Sande 2011, Poster #42, Wednesday 17:20-20:00)
- Support Vector Machine with Histogram Intersection kernel. Fast additive classification (Maji 2009)
Results

Comparable with top scores reported in e.g. Chatfield et al. BMVC 2011
- We: Pixel-wise sampling, 5 Colour SIFT (Sande 2010), kmeans vocabulary 4096
- Chatfield et al.: dense sampling, grey-SIFT only, Fisher/Sparse coding
Significant improvement by using not the whole image but its Most Telling Window
Most Telling Window consistently outperforms Exact Localisation (using same basic framework)
Scores Detection Task: Felzenszwalb: 0.253 MTW: 0.317, Our localisation: 0.336, Discrepancy in results on detection and classification suggests that exact localisation tends to hallucinate objects that are not there while Most Telling Window finds object approximately.
Final combination by cross-validation using weighted addition of classifier output:
- 2 parts Most Telling Window SP 4x4
- 2 parts Localisation (Felzenszwalb 2010)
- 1 part Most Telling Window SP 2x2
- 1 part global Bag-of-Words

3 variations of global Bag-of-Words and our exact localisation were discarded. Location is crucial!
Visualising the Most Telling Window of top-ranked images

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

High-ranked Positives

High-ranked Negatives

Aeroplane
Visualising the Most Telling Window of top-ranked images

High-ranked Positives

High-ranked Negatives

Bicycle
Visualising the Most Telling Window of top-ranked images

High-ranked Positives

High-ranked Negatives

Cat
Visualising the Most Telling Window of top-ranked images
Visualising the Most Telling Window of top-ranked images

High-ranked Positives

High-ranked Negatives

Motorcycle
Visualising the Most Telling Window of top-ranked images

High-ranked Positives

High-ranked Negatives

Person
Pascal VOC 2011 Classification Challenge

The top-3 each has a different focus for boosting classification performance:

1st NUSPSL: Focus on combination of exact localisation and classification
(Song et al. CVPR 2011)

2nd NLPR: Focus on vocabulary: Semi-semantic, Salient and Supervector coding.
(Huang et al. CVPR 2011)

3rd UVA/DISI: Focus on location: The Most Telling Window
(Uijlings and Smeulders, submitted to TPAMI, Sande et al. ICCV 2011)
Conclusions Most Telling Window

The Most Telling Window is the window that is most discriminative for classifying the presence of an object. It can be an (1) Object Part. (2) Whole Object. (3) Object Collection.

First time that window within the image yields better results by itself than whole image?

The Most Telling Window works better than exact localisation.

Suboptimal positive windows suggest room for improvement.

Selective Search enables powerful, local Bag-of-Words

Segmentation as Selective Search for Object Recognition, ICCV 2011, K.E.A. van de Sande, J.R.R. Uijlings, T. Gevers, and A.W.M. Smeulders, Poster #42, Wednesday 17:20-20:00

Class independent parts, wholes, and collections.

The Windows that Tell the Story of an Image, J.R.R. Uijlings and A.W.M. Smeulders. Under submission at TPAMI. Please contact jrr@disi.unitn.it before using this work.