Part I
Unsupervised Feature Learning with Convolutional Neural Networks

Thomas Brox
Computer Vision Group
University of Freiburg, Germany

Research funded by ERC Starting Grant VideoLearn and Deutsche Telekom Stiftung
Status quo: CNNs generate great features

Do we need these massive amounts of class labels to learn generic features?

ILSVRC 2012 classification
Krizhevsky et al. 2012

PASCAL VOC object detection
Girshick et al. 2014
Unsupervised feature learning

• Dominant concept: reconstruction error + regularization

• Existing frameworks:
  – Autoencoders (dimensionality reduction)
    (Hinton 1989, Vincent et al. 2008,…)
  – Sparse coding (sparsity prior)
  – Slowness prior
    (Wiscott-Sejnowski 2002, Zou et al. 2012,…)
  – Deep belief networks (prior in contrastive divergence)
    (Ranzato et al. 2007, Lee et al. 2009,…)

• Reconstruction error models the input distribution
  → dubious objective
Exemplar CNN: discriminative objective

- Train CNN to discriminate **surrogate classes**

- Take data augmentation to the extreme (translation, rotation, scaling, color, contrast, brightness)

- Transformations define invariance properties of the features to be learned

Acknowledgements to caffe.berkeleyvision.org
Application to classification

- Pooled responses from each layer used as features
- Training of linear SVM

<table>
<thead>
<tr>
<th></th>
<th>STL-10</th>
<th>CIFAR-10</th>
<th>Caltech-101</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolutional K-means network</td>
<td>60.1</td>
<td>70.7</td>
<td>-</td>
</tr>
<tr>
<td>View-invariant K-means</td>
<td>63.7</td>
<td>72.6</td>
<td>-</td>
</tr>
<tr>
<td>Multi-way local pooling</td>
<td>-</td>
<td>-</td>
<td>77.3</td>
</tr>
<tr>
<td>Slowness on video</td>
<td>61.0</td>
<td>-</td>
<td>74.6</td>
</tr>
<tr>
<td>Hierarchical Matching Pursuit (HMP)</td>
<td>64.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Multipath HMP</td>
<td>-</td>
<td>-</td>
<td>82.5</td>
</tr>
<tr>
<td>Exemplar CNN</td>
<td>72.8</td>
<td>75.3</td>
<td>85.5</td>
</tr>
</tbody>
</table>

Outperforms all previous unsupervised feature learning approaches
Which transformations are most relevant?
How many surrogate classes?
How many samples per class?
Application to descriptor matching

Descriptor matching between two images
CNNs won’t work for descriptor matching, right?

Descriptors from a CNN outperform SIFT

Mikolajczyk dataset

New larger dataset

Philipp Fischer

Alexey Dosovitskiy
Supervised versus unsupervised CNN

Mikolajczyk dataset

New larger dataset

Unsupervised feature learning advantageous for descriptor matching
Improvement of Examplar CNN over SIFT is as big as SIFT over color patches
Exemplar CNN: Unsupervised feature learning by discriminating surrogate classes

Outperforms previous unsupervised methods on classification

CNNs outperform SIFT even on descriptor matching

Unsupervised training advantageous for descriptor matching
Part II
Benchmarking Video Segmentation

Thomas Brox
Computer Vision Group
University of Freiburg, Germany

Contains joint work with
Fabio Galasso, Bernt Schiele (MPI Saarbrücken)

Research funded by DFG and ERC
Motion segmentation
Benchmarking motion segmentation

Freiburg-Berkeley Motion Segmentation Dataset (FBMS-59)
59 sequences split into a training and a test set
Pixel-accurate ground truth

Ground truth mostly every 20 frames
Precision-recall metric

Region \( c_j \) to ground truth \( g_i \) assignment with Hungarian method

\[
P = \frac{1}{n} \sum_{i=1}^{n} \frac{c_j \cap g_i}{c_j} \quad R = \frac{1}{n} \sum_{i=1}^{n} \frac{c_j \cap g_i}{g_i} \quad F = \frac{2PR}{P + R}
\]

Under-segmentation

- Machine: P=1, R=0
- Ground truth: P=0.94, R=0.67, F=0.78

Over-segmentation

- Machine: P=1.00, R=0.56, F=0.72
- Ground truth: P=0.98, R=0.80, F=0.88
Results on the test set
Benchmarking general video segmentation

VSB-100: Benchmark based on Berkeley Video Segmentation Dataset
100 HD videos (40 training, 60 test)
Four human annotations per video

Fabio Galasso
Naveen S. Nagaraja
Bernt Schiele

Galasso et al. ICCV 13
Metric for supervoxels

$$P = \frac{1}{H} \sum_{i=1}^{H} \left( \left( \sum_{c \in C} \max_{g \in G_i} |c \cap g| \right) - \max_{g \in G_i} |g| \right)$$

$$\frac{\sum_{c \in C} |c| - \frac{1}{H} \sum_{i=1}^{H} \max_{g \in G_i} |g|}{\sum_{c \in C} |c|}$$

Evaluated pixels in the video minus the largest ground truth region

$$R = \frac{\sum_{i=1}^{H} \left( \sum_{g \in G_i} \max_{c \in C} |c \cap g| - 1 \right)}{\sum_{i=1}^{H} \left( \sum_{g \in G_i} |g| - |G_i| \right)}$$

Size of all ground truth regions minus size of the largest ground truth region

- Many-to-one matching (important for supervoxels)
- Normalization penalizes extreme segmentations
Results

- Xu et al. (ECCV 12)
- Corso et al. (TMI 08)
- Galasso et al. (ACCV 12)
- Arbelaez et al. (image segmentation) (TPAMI 11)
- Grundmann et al. (CVPR 10)
- Ochs-Brox (ICCV 11)

Graph showing precision and recall with annotations for different methods and their respective years.
Motion segmentation subtask

Human performance

Arbelaez et al. + oracle

Simple baseline

Ochs-Brox ICCV 11

Galasso et al. ACCV 12

Grundmann et al. CVPR 10

Precision

Recall
1. Take superpixel hierarchy from Arbelaez et al.
2. Propagate labels to next frame using optical flow
3. Next frame: label determined by voting

Image segmentation + optical flow < video segmentation

There is work to do
Balanced graph reduction

Edge reweighting necessary for weight balancing in spectral clustering
Balancing clearly improves results

Simple baseline

Reweighted graph reduction

Galasso et al.
ACCV 12
Summary of part II

FBMS-59:
Motion segmentation benchmark

VSB-100:
General video segmentation benchmark

Spectral clustering with superpixels:
Don’t forget to rebalance