Combining local and global Bag-of-Words representations for semantic segmentation.

X. Boix  J. M. Gonfaus  F. S. Khan  J. van de Weijer
A. Bagdanov  M. Pedersoli  J. González  J. Serrat
Motivation

What’s inside a local segment?
Motivation

...and with context? [FulkersonICCV09]
Motivation

Global classifier

What's inside a local segment?
Contributions

- Novel segmentation method that jointly uses global and local information.
- Concatenating the description of a superpixel and its context.
- Learn a per class normalization of the classification scores.
Model

- Original Image
- Unsupervised Segmentation
- Superpixel Nodes
- Global Node
- Local Classification
- Global Classification
- Inference with Graph-Cuts
Model

- Original Image
- **Unsupervised Segmentation**
- Superpixel Nodes
- Global Node
- Local Classification
- Global Classification
- Inference with Graph-Cuts
Model

- Original Image
- Unsupervised Segmentation
- **Superpixel Nodes**
- Global Node
- Local Classification
- Global Classification
- Inference with Graph-Cuts
Model

- Original Image
- Unsupervised Segmentation
- Superpixel Nodes
- **Global Node**
- Local Classification
- Global Classification
- Inference with Graph-Cuts

N-order (binary) clique
Model

- Original Image
- Unsupervised Segmentation
- Superpixel Nodes
- Global Node
- **Local Classification**
- Global Classification
- Inference with Graph-Cuts
Model

- Original Image
- Unsupervised Segmentation
- Superpixel Nodes
- Global Node
- Local Classification
- **Global Classification**
- Inference with Graph-Cuts
Model

- Original Image
- Unsupervised Segmentation
- Superpixel Nodes
- Global Node
- Local Classification
- Global Classification
- **Inference with Graph-Cuts**
\[
\sum_{s \in S} \text{local} + \sum_{(p,q) \in \mathcal{N}_S} \text{smoothness} + \sum_{g \in \mathcal{G}} \text{global} + \sum_{(p,q) \in \mathcal{N}_{Sg}} \text{consistency}
\]
\[ \sum_{s \in S} \text{local} + \sum_{(p,q) \in N_S} \text{smoothness} + \sum_{g \in G} \text{global} + \sum_{(p,q) \in N_{SG}} \text{consistency} \]
Smoothness term

\[\text{smoothness}(s_i, s_j, c_{ij}) = \lambda \theta(c_{ij}) N_{ij} \delta(s_i, s_j)\]
Smoothness term

\[
\text{smoothness}(s_i, s_j, c_{ij}) = \lambda \theta(c_{ij}) N_{ij} \delta(s_i, s_j)
\]
Smoothness term

\[ \text{smoothness}(s_i, s_j, c_{ij}) = \lambda \theta(c_{ij}) N_{ij} \delta(s_i, s_j) \]
Smoothness term

\[ \text{smoothness}(s_i, s_j, c_{ij}) = \lambda \theta(c_{ij}) N_{ij} \delta(s_i, s_j) \]
\[ \sum_{s \in S} \text{local} + \sum_{(p,q) \in \mathcal{N}_S} \text{smoothness} + \sum_{g \in \mathcal{G}} \text{global} + \sum_{(p,q) \in \mathcal{N}_{SG}} \text{consistency} \]
Local term

Bag-of-Words:
- Inside Region (20%)
- Contextual Regions (27%)
- Concatenate Both Regions (29%)
Local term

Bag-of-Words:
- Inside Region (20.02%)
- Contextual Regions (27%)
- Concatenate Both Regions (29%)
Local term

Bag-of-Words:
- Inside Region (20%)
- **Contextual Regions** (27.14%)
- Concatenate Both Regions (29%)
Local term

Bag-of-Words:
- Inside Region (20%)
- Contextual Regions (27%)
- Concatenate Both Regions (29.53%)
Local term

Detector:
- Dense Grid with 50% of overlapping between patches.
- 4 different scales.

Description:
- Shape feature: SIFT. (28.34%)
- Color feature: RGB Histogram. (22.5%)
- Concatenate SIFT + Color histogram. (29.53%)
Local term

- 20 SVM with Intersection Kernel.
- 20,000 training samples for each class.

One class against all classes.
20 SVM with Intersection Kernel.
20,000 training samples for each class.

One class against its background. Similar to [CsurkaBMCV08].
\[
\sum_{s \in S} \text{local} + \sum_{(p,q) \in \mathcal{N}_S} \text{smoothness} + \sum_{g \in \mathcal{G}} \text{global} + \sum_{(p,q) \in \mathcal{N}_{SG}} \text{consistency}
\]
Consistency term

- $g_i \in \{0, 1\}$
- All global nodes are connected to each superpixel node.

$$\text{consistency}(s_i, \mathcal{G}) = \beta M_i \prod_{g_j=1 \in \mathcal{G}} (1 - \delta(s_i, j))$$
Consistency term

Equivalent problem:

- Substitute $g_i$ with ONE node $g \in \{\mathcal{L}_{comb}\}$.
- Each label in $\{\mathcal{L}_{comb}\}$ represents a **combination** of classes in the image.
- Thus, $g$ has a total amount of $2^N$ possible labels.

Too many labels to be solvable in reasonable time.
Approximate problem:

- Use only the most likely $\mathcal{L}_{comb}$:
  - Discard objects with very low global classification rate ($\leq 0.05$).
  - Possible combinations of objects in the same image.
- Solvable with standard graph-cuts (less than 2 seconds).
\[ \sum_{s \in S} \text{local} + \sum_{(p,q) \in N_S} \text{smoothness} + \sum_{g \in G} \text{global} + \sum_{(p,q) \in N_{SG}} \text{consistency} \]
Global term

[KahnlCCV09]
Global term

Feature Detection

- Grid Sampling
- Harris-Laplace
- Boosted Harris-Laplace
- Blob detector
- Boosted Blob detector
- Spatial Pyramid (2x2)
- Spatial pyramid (1x3)

[KahniCCV09]
Global term

Feature Detection
- Grid Sampling
- Harris-Laplace
- Boosted Harris-Laplace
- Blob detector
- Boosted Blob detector
- Spatial Pyramid (2x2)
- Spatial pyramid (1x3)

Descriptors
- SIFT
- Color Name
- Hue
- Color SIFT
- Gist
- Spatial Pyramid (2x2)
- Spatial Pyramid (1x3)

[KahnlCCV09]
Global term

Feature Detection
- Grid Sampling
- Harris-Laplace
- Boosted Harris-Laplace
- Blob detector
- Boosted Blob detector
- Spatial Pyramid (2x2)
- Spatial Pyramid (1x3)

Descriptors
- SIFT
- Color Name
- Hue
- Color SIFT
- Gist
- Spatial Pyramid (2x2)
- Spatial Pyramid (1x3)

Codebook Model
- Bag-of-words
- Bag-of-words
- Bag-of-words
- multiple Bag-of-words

[KahnICCV09]
Global term

Feature Detection:
- Grid Sampling
- Harris-Laplace
- Boosted Harris-Laplace
- Blob detector
- Boosted Blob detector
- Spatial Pyramid (2x2)
- Spatial pyramid (1x3)

Descriptors:
- SIFT
- Color Name
- Hue
- Color SIFT
- Gist

Codebook Model:
- Bag-of-words
- Color Attention
- Bag-of-words

[KahnlCCV09]
Global term
Learning the parameters

- The best configuration maximizes the geometric mean of the performance of all classes.
- We obtain new configurations in a **Gibbs sampler** manner:
  \[ x_{i}^{t+1} \sim \mathcal{N}(x_{i}^{t}, f(t)) \]
- 2-fold cross validation.
- Learning stages:
  1. Weights of the graphical model. (29.53%)
  2. Per class normalization of the local term. (31.25%)
  3. Per class normalization of the global term. (35.1%)
Conclusions

- We propose a novel segmentation method that jointly uses global and local information.
- Using as negative examples only the segments that appear in the same image of positive samples decreases the variability of the data.
- Concatenating both the description of a superpixel and its context is helpful for classification. (+2.5%)
- We empirically prove that a per class normalization of the observed terms is able to efficiently equalize classification scores. (+5.6%)
Gràcies!

Thank you!

Arigato!

Centre de Visió per Computador