Semantic Segmentation

Scene Understanding

Bebe Shi
* Assign a class label to each pixel of an image

* Formalized Task
Semantic Segmentation

- sky
- water
- grass

Object Detection

- me
- bag

*Comparisons*
http://people.csail.mit.edu/jxiao/StreetSeg/

* Ambiguities
* Arbitrary viewpoints
* Severe occlusion
* Texton: feature that jointly models shape appearance and context
* Focus on speed and larger datasets

* J. Shotton, J. Winn, C. Rother, and A. Criminisi. TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation. ECCV 2006
* 591 photos of 21 object classes: building, grass, tree, cow, sheep, sky, aeroplane, water, face, car, bike, flower, sign, bird, book, chair, road, cat, dog, body, boat
* All hand-labeled
* All images are approximately 320 × 240 pixels.
* 45% training, 10% validation and 45% test sets
* For comparison with previous work:
* 7-class Corel database (180×120 pixels)
* 7-class Sowerby database (96 × 64 pixels)

* LabelMe Outdoor (LMO) database
* SUN database

*Datasets*
* J. Shotton, M. Johnson, R. Cipolla. Semantic Texton Forests for Image Categorization and Segmentation

* weight each training example by the inverse class frequency

$$w_i = \xi_{c_i} \text{ with } \xi_c = \left(\sum_{i \in I} [c = c_i]\right)^{-1}$$

* Datasets
\[
\log P(c \mid x, \theta) = \sum_i \underbrace{\psi_i(c_i, x; \theta_{\psi})}_{\text{shape-texture}} + \underbrace{\pi(c_i, x_i; \theta_{\pi})}_{\text{color}} + \underbrace{\lambda(c_i, i; \theta_{\lambda})}_{\text{location}} \\
+ \sum_{(i, j) \in \mathcal{E}} \underbrace{\phi(c_i, c_j, g_{ij}(x); \theta_{\phi})}_{\text{edge}} - \log Z(\theta, x)
\]

*Conditional Random Field (CRF)*
* Ideal case: maximize the conditional likelihood by using gradient ascent
* Piecewise training

* Learning the CRF Parameters

* Overcounting

* Added power parameters for the location and color potentials

* Learning the CRF Parameters
*Boosted Learning of Shape, Texture and Context

*Texton Map*
\( v(i_1, r, t) \approx A, \ v(i_2, r, t) \approx 0, \) and \( v(i_3, r, t) \approx A/2. \)

*Shape Filters*
\[ h(c_i) = \begin{cases} \alpha \delta(v(i, r, t) > \theta) + b & \text{if } c_i \in N \\ k_{c_i} & \text{otherwise} \end{cases} \]

\[ H(c_i) = \sum_{m=1}^{M} h_m(c_i) \]

\[ \tilde{P}_i(c_i \mid x) = \frac{\exp(H(c_i))}{\sum_{c'_i} \exp(H(c'_i))} \]

**Weak Classifier**

**Softmax Transformation**

*Joint Boosting*
* Filter responses are calculated on a $\Delta \times \Delta$ grid

* Shape filters are still calculated at full resolution at test time

* At each round of boosting, a fraction of $\tau << 1$ of the features are chosen
Sub-Sampling and Random Feature Selection

* Extremely fast to both train and test
* Improves both quantitative performance and execution speed
* Ensemble of $T$ decision trees

* Each node $n$ is a learned class distribution $P(c|n)$

* Recursively branches left or right down the tree according to a learned binary function of the feature vector, until a leaf node $l$ is reached.

$$P(c|L) = \frac{1}{T} \sum_{t=1}^{T} P(c|l_t)$$

**Randomized Decision Forests**
* Extremely Randomized Trees Algorithm
* Each tree is trained separately on a small random subset of the training data

\[
I_1 = \{ i \in I_n \mid f(v_i) < t \} \\
I_r = I_n \setminus I_1.
\]

\[
\Delta E = - \frac{|I_1|}{|I_n|} E(I_1) - \frac{|I_r|}{|I_n|} E(I_r)
\]

Shannon Entropy

* Randomized Learning
* Small random subsets of the data are used for training

* Only a small portion of the tree is traversed for each data point

* Advantages
* Augment the training data with image copies that are artificially transformed geometrically and photometrically.

* Explored rotation, scaling and left-right flipping as geometric transformations and affine photometric transformations.

* Learning Invariances
*Bags of Semantic Textons (BoSTs)*
* Easy to implement
* Has few parameters
* Embeds contextual information naturally in the retrieval/alignment procedure
* Transfer the labels from existing samples in a large database to annotate an image

Nonparametric Scene Parsing
* LabelMe Outdoor (LMO) Database (2,688 fully annotated images, mostly outdoor scenes)
* Sun Database (9,566 fully annotated images, covers both indoor and outdoor scenes)
* Both annotated using the LabelMe online annotation tool
* 2,488 for training and 200 for testing
* Top 33 object categories with the most labeled pixels are chosen
*System Pipeline*
*Given a query image, find a set of nearest neighbors that share similar scene configuration with the query.

*Scene Retrieval*
$\mathcal{N}(x) = \{ y_i \mid \text{dist}(x, y_i) \leq (1 + \epsilon)\text{dist}(x, y_1)\},$

$y_1 = \arg \min_{i \leq K} \text{dist}(x, y_i).$
\*<K, \varepsilon>-NN
* Euclidean distance of GIST
* Spatial pyramid histogram intersection of HOG visual words
* Spatial pyramid histogram intersection of the ground-truth annotation
(a) Graph structure revealed by GIST distance
(b) The same as (a), showing ground-truth annotation
(c) The same as (d), showing RGB images
(d) Graph structure revealed by the distance on ground-truth annotation
\[ E(w) = \sum_p \min(\|s_1(p) - s_2(p + w(p))\|_1, t) + \]

\[ \sum_p \eta(|u(p) + |v(p)|) + \]

\[ \sum_{(p,q) \in \epsilon} \min(\lambda|u(p) - u(q)|, d) + \min(\lambda|v(p) - v(q)|, d), \]

*SIFT Flow for Dense Scene Alignment*
Coarse-to-Fine pyramid SIFT Flow Matching
\( \{ s_i, c_i, w_i \}_{i=1:M} \)

SIFT image, annotation and SIFT flow field

\( c_i(p) \in \{1, \ldots, L\} \)

*Label transfer*
Three components: likelihood, prior, spatial smoothness

\[
- \log P(c|I, s, \{s_i, c_i, w_i\}) = \sum_p \psi(c(p); s, \{s'_i\}) \\
+ \alpha \sum_p \lambda(c(p)) + \beta \sum_{\{p,q\} \in \mathcal{E}} \phi(c(p), c(q); I) + \log Z
\]
\[ \psi(c(p) = l) = \begin{cases} \min_{i \in \Omega_{p,l}} \| s(p) - s_i(p + w(p)) \|, & \Omega_{p,l} \neq \emptyset, \\ \tau, & \Omega_{p,l} = \emptyset, \end{cases} \]

\[ \Omega_{p,l} = \{ i; c_i(p + w(p)) = l \}, \ l = 1, \ldots, L \]

\[ \tau = \max_{s_1, s_2, p} \| s_1(p) - s_2(p) \| \]

*Likelihood*
\[ \lambda(c(p) = l) = -\log \text{hist}_l(p) \]
* The nonparametric scene parsing system outperforms TextonBoost in terms of both overall and per-class recognition rate.
* Color information is not included in the model.
* Similar performance to TextonBoost is achieved by matching color instead of dense SIFT descriptors.
* Smaller objects are better handled by detectors.
* Pixel-wise classifiers like TextonBoost can also be useful when good matching cannot be established or good nearest neighbors can hardly be retrieved.