# Introspective Semantic Segmentation

Gautam Singh Jana Košecká
George Mason University
Fairfax, VA
{gsinghc,kosecka}@cs.gmu.edu

## Abstract

*Traditional approaches for semantic segmentation work in a supervised setting assuming a fixed number of semantic categories and require sufficiently large training sets. The performance of various approaches is often reported in terms of average per pixel class accuracy and global accuracy of the final labeling. When applying the learned models in the practical settings on large amounts of unlabeled data, possibly containing previously unseen categories, it is important to properly quantify their performance by measuring a classifier's introspective capability. We quantify the confidence of the region classifiers in the context of a non-parametric $k$-nearest neighbor ($k$-NN) framework for semantic segmentation by using the so called* strangeness measure. *The proposed measure is evaluated by introducing confidence based image ranking and showing its feasibility on a dataset containing a large number of previously unseen categories.*

## 1. Introduction

The problem of semantic segmentation requires simultaneous segmentation of an image into regions and categorization of all the image pixels. Traditional approaches for semantic segmentation typically require large training sets and performance of different approaches is evaluated by reporting per pixel average class accuracy and global accuracy. In practical settings, we would like to use the obtained models on unlabeled data which contain novel previously unseen semantic categories. It is therefore important to quantify the performance of the existing models on this unlabeled data.

Numerous techniques have been proposed in the past which tackled the issues of estimating an uncertainty of a given model learned in a supervised setting. Majority of the previous proposals have looked at these problems in the context of object detection or whole image categorization. In this paper, we present an approach for quantifying the uncertainty of semantic labels associated with the image regions obtained by semantic segmentation. The existing approaches to semantic segmentation differ in the choice of elementary regions, features characterizing them, classifiers used to predict the final labels and an optional regularization stage. In this paper, we consider a non-parametric approach for semantic segmentation which is based upon using a $k$-nearest neighbor classifier as was shown in [15, 5]. Using the output from the $k$-NN classifier, we compute a measure called $strangeness$ [13], the computation of which utilizes nearest neighbor distances. We follow the observation that a single global distance metric is often not sufficient for handling the large variations within a class. Instead, we compute weights for the individual feature channels by adopting a locally adaptive method [4] for feature relevance computed at the test time. The strangeness measure characterizes an instance's uncertainty with respect to its own label. The computation of the strangeness of an image region then helps to provide a confidence for a classification decision and regions with high uncertainty of belonging to familiar categories can be labeled $unseen$. In the experiments section, we display the efficacy of this measure for ranking images containing unfamiliar semantic categories on the large scale SUN09 dataset [3].

## 2. Related Work

With the increasing sizes of datasets and an increasing number of labels, the use of non-parametric approaches have shown notable progress for semantic segmentation and classification [11, 15, 5]. They are appealing as they do not need to be retrained as newer categories or images are added, can utilize efficient approximate nearest neighbor search techniques e.g. $k$-d trees [12] and contextual cues. Often, context is captured using a retrieval set of images which are similar to the query and methods developed for establishing matches between image regions (at pixel or superpixel level) for labeling the image. Authors in [15] formulate semantic labeling at the superpixel level. They retrieve similar images using global image features which is followed by superpixel-level matching using a wide variety of superpixel features and a Markov random field (MRF)

to incorporate neighborhood context. This work was extended by [5] by training per superpixel per feature weights and also by incorporating superpixel level semantic context. Our approach for semantic segmentation is related to work of [15, 5] as we also pursue a non-parametric approach. However, we differ in the choice of features by also utilizing geometry based statistics which have displayed success in computing the geometric layout of scenes [7] and adopt a test time feature channel relevance learning method.

In this paper, we are interested in quantifying the uncertainty of the predicted semantic labels. Different approaches have been proposed in the literature to associate classifier uncertainty on query data. An area where it has received significant attention is that of active learning where the task is to sequentially add labeled examples for training through human input on unlabeled samples but doing it with the added goal of minimizing the human effort. Hence, a critical component is to provide *informative* unlabeled samples for the human to label and classification uncertainty estimates are employed for selecting such samples. A commonly used uncertainty measure is entropy [8] which is an information-theoretic criterion computed over label likelihoods for an instance. An alternative measure is the best versus second best heuristic [9] computed as the difference between the probability values of the two labels having the highest estimated probability value with a low difference value indicating more uncertainty. This particular measure relates to the uncertainty between the most confusing labels in classification instead of using all the labels including ones which have low likelihoods as is done when computing entropy. In margin-based methods, the uncertainty of an instance has been characterized by its distance to the decision boundary between the classes [16]. An example which lies the closest to the boundary can be viewed as the one with the highest uncertainty. There has also been work [10] on utilizing Gaussian process classifiers for associating uncertainty based on the variance in the posterior. In our work, we propose to utilize the *strangeness* measure introduced by [13] to associate a confidence with the output of our non-parametric approach for semantic segmentation.

# 3. Semantic Segmentation Approach

In this section, we first review our non-parametric approach for semantic segmentation of images. This is followed by a description of the method for the computation of the strangeness measure in Section 4.

## 3.1. Problem Formulation

We formulate the semantic labeling of an image segmented into superpixels. The output of the semantic segmentation is a labeling $\mathbf{L} = (l_1, l_2, \ldots l_S)^\top$ with hidden variables assigning each superpixel $s_i$ a unique label, $l_i \in \{1, 2, \ldots, L\}$, where $L$ is the total number of the semantic categories and $S$ is the number of superpixels in the image. The posterior probability of a labeling $\mathbf{L}$ given the observed appearance feature vectors $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_S]$ computed for each superpixel can be expressed as:

$$P(\mathbf{L}|\mathtt{A}) = \frac{P(\mathtt{A}|\mathbf{L})\, P(\mathbf{L})}{P(\mathtt{A})}. \tag{1}$$

We estimate the labeling $\mathbf{L}$ as a Maximum A Posteriori Probability (MAP),

$$\underset{\mathbf{L}}{\operatorname{argmax}}\, P(\mathbf{L}|\mathbf{A}) = \underset{\mathbf{L}}{\operatorname{argmax}}\, P(\mathbf{A}|\mathbf{L})\, P(\mathbf{L}). \tag{2}$$

The appearance likelihood $P(\mathbf{A}|\mathbf{L})$ is obtained using a non-parametric $k$-NN method described in Section 3.3. While many approaches to semantic segmentation model the joint prior $P(\mathbf{L})$ using a pairwise smoothness term in a Markov random field (MRF), we forgo the use of an MRF in our approach and utilize the appearance likelihood for the semantic labeling of the image.

## 3.2. Superpixels and features

To perform the semantic labeling of an image, we extract SLIC superpixels [1, 2] using the publicly available VLFeat library [17]. To characterize the image superpixels, we use both geometric as well as appearance features to capture the statistics of individual regions. The choice of features has been adopted from [7] where each superpixel is characterized by color (color histograms of RGB and HSV values and saturation value), texture (mean absolute response of the filter bank of 15 filters and histogram of maximum responses), location and shape (position of the centroid, relative position, number of pixels and area in the image), and perspective cues computed from long linear segments and lines aligned with different vanishing points. These features were computed using the publicly available code provided by the authors of [7]. In addition to these features, we endow each superpixel region with a histogram of SIFT descriptors computed densely at each image location and quantized into 100 clusters. The entire feature vector is of 194 dimensions.

## 3.3. Appearance Likelihood

In order to compute the appearance likelihood for the entire image, we approximate it with a Naive Bayes assumption yielding

$$P(\mathbf{A}|\mathbf{L}) \approx \prod_{i=1}^{S} P(\mathbf{a}_i|l_i). \tag{3}$$

Such an approximation assumes independence between appearance features of the superpixels given their labels. We follow a non-parametric approach to obtain the individual

label likelihood $P(\mathbf{a}_i|l_j)$ for a superpixel $s_i$ which is obtained using a $k$-NN method. Since a superpixel is uniquely represented by its feature vector, we use the symbols $s_i$ and $\mathbf{a_i}$ interchangeably. For each class $l_j$ and every superpixel $s_i$ of the query image, we compute a label likelihood score:

$$\tau(\mathbf{a}_i, l_j) = \frac{n(l_j, N_{ik})/n(l_j, G)}{n(\bar{l_j}, N_{ik})/n(\bar{l_j}, G)} \tag{4}$$

where

- $\bar{l_j}$ is the set of all labels excluding $l_j$;
- $N_{ik}$ is a neighbourhood around $\mathbf{a}_i$ with exactly $k$ points in it;
- $n(l_j, N_{ik})$ is the number of superpixels of class $l_j$ inside $N_{ik}$;
- $n(l_j, G)$ is the number of superpixels of class $l_j$ in the entire dataset $G$.

We compute the normalized label likelihood score using the individual label likelihood:

$$P(\mathbf{a}_i|l_j) = \frac{\tau(\mathbf{a}_i, l_j)}{\sum_{l_k=1}^{L} \tau(\mathbf{a}_i, l_k)} \tag{5}$$

The basic approach to compute the neighborhood $N_{ik}$ is to use the concatenated feature $\mathbf{a}_i$ (Section 3.2) and retrieve the $k$ nearest points by computing its distance to superpixels in $G$. Such a retrieval can be efficiently performed by the use of approximate nearest neighbour methods like $k$-d trees [12].

### 3.4. Weighted $k$-NN

The basic $k$-NN approach for semantic labeling uses Euclidean distance over the concatenated region feature to compute the neighborhood around a point. To handle the variations between different semantic categories and better exploit the contribution of the different feature channels, we adopt a non-parametric distance learning approach where now, a weighted $k$-NN method is used to compute the neighborhood around a query point. To compute a weighted distance between two superpixels, we split their individual feature vectors into six feature channels which characterize the following properties of an image segment: (1) color (2) texture filter bank responses (3) location and size (4) statistics of lines in the region (5) statistics of intersecting lines in the region and (6) dense SIFT histogram yielding a vector of distances:

$$d_f = [d_{color}, d_{tex}, d_{loc}, d_{line}, d_{per}, d_{sift}]^\top \tag{6}$$

where $d_{color}, d_{tex}, d_{loc}, d_{line}, d_{per}, d_{sift}$ are the Euclidean distances between the six feature channels (described above) of the feature vectors of the two superpixels respectively. We define a weighted distance between the two superpixels as

$$d_w = w^\top d_f \tag{7}$$

where $\mathbf{w} \in \Re^6$ defines the weights for the individual feature distances. Using the weighted distance from Eq. (7), we can now obtain the neighborhood around the query point.

We use the locally adaptive metric approach of [4] for the weight computation. It is a query-based technique which computes a global metric to select neighbors for a test point which are then used to refine the feature weights. In our approach, the test points correspond to the individual superpixels of an image. The approach estimates the relevance of a feature channel by evaluating its ability to predict class posterior probabilities locally at a query point. This is done by computing the expectation of the posterior conditioned at a test point $\mathbf{a_0}$ along this feature channel. This is estimated by considering local neighborhoods around the test point and is described in detail by [4]. This constitutes our non-parametric approach for the semantic segmentation of an image. Its distinguishing characteristic is the use of a single oversegmentation of an image described by geometric and appearance features which are used in a weighted $k$-NN framework where the weights are computed at test time by analyzing the local neighborhood around a superpixel.

## 4. Strangeness Measure

While evaluating the semantic labeling output of an image is often the final goal of traditional approaches for semantic segmentation, in this work, we are also interested in computing the uncertainty of the semantic labels that we associate with an image. This is motivated by our desire to perform semantic segmentation of images with an introspective capability. In the previous section, we presented a non-parametric approach for the semantic labeling of an image. We now extend this method to help analyze a query image and identify image regions which instead of being associated with one of the known semantic categories can be characterized as *unfamiliar*. Our approach for this analysis is based on the concept of transduction in which an estimate about the properties of a query point of interest is made directly from the training data as opposed to induction where first a general rule is inferred from the training data and then applied to the query point.

Given a source set $T_s$ of fully annotated images, we segment the images of this set and compute the features for the corresponding segments yielding the dataset of superpixels $G$ - $(\mathbf{a}, y)$ where $y$ is the segment label. For each segment $s_i$ in this dataset, an individual strangeness measure $\alpha_i$ is computed

$$\alpha_i = \frac{\sum_{r=1}^{K} d_{ir}^{c}}{\sum_{r=1}^{K} d_{ir}^{\bar{c}}} \tag{8}$$

where $c$ is the semantic label $y_i$ for instance $s_i$, $d_{ir}^{c}$ is the $r$-th shortest distance between $s_i$ and an instance of class $c$, $d_{ir}^{\bar{c}}$ is the $r$-th shortest distance between $s_i$ and an instance not belonging to class $c$ and $K$ is the number of nearest

neighbors considered for each sum. In this work, we use the weighted $k$-NN method of Section 3.4 for computing the nearest neighborhood around a point. The strangeness measure is the ratio of sum of $K$ nearest distances from the same class to the sum of the $K$ nearest distances from all other classes and it measures how "strange" an instance in question is with respect to its semantic category. An example closer to other class instances in comparison to its own class instances has higher strangeness and vice versa. To quantify the confidence of association with a semantic category, we count the number of examples of the category in the dataset which have a larger strangeness value and compute the p-value statistic proposed by [13]:

$$t_i = \sum_{\substack{r=1 \\ y_i=y_r=c}}^{|c|} 1\{\alpha_i > \alpha_r\}/|c| \tag{9}$$

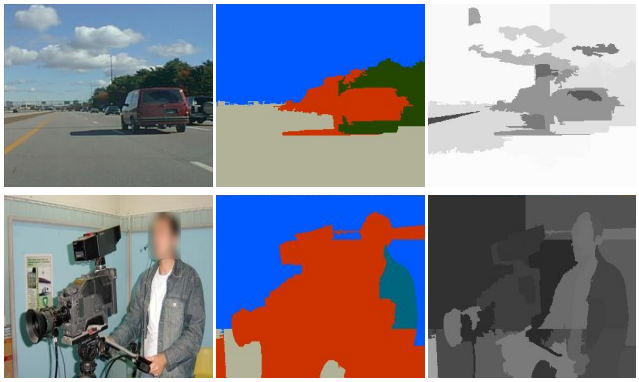
where $|c|$ is the number of instances in $G$ with the label $c$ and $1\{.\}$ is the indicator function. The value $t_i$ can be viewed as a measure of the probability of having instances in the class with strangeness greater than or equal to that of $s_i$. Using Eq. (8) and Eq. (9), the strangeness and p-values are computed for all the instances in $G$.

Now, given a set $T_u$ of query images on which we wish to investigate the introspective capacity of our non-parametric approach for semantic labeling (with $D_u$ the segments corresponding to these query images), we wish to discover regions which do not belong to any of the semantic categories in $T_s$. In doing so, we also want to associate a measure of confidence for it to not belong to any of the familiar semantic categories. For this purpose, we utilize the strangeness measure from Eq. (8). When computing the strangeness for instances in $G$, we already know the semantic label for the instance and the strangeness computation is straightforward. However, for the instances in $D_u$, we do not know the category of the instances. But note that our primary interest is determining if the instance belongs to *any* of the known semantic categories or not.

Therefore, we compute the strangeness and p-value for a query image region by assuming its putative label to be each of the known semantic categories $\{1, 2, \ldots, L\}$ one by one i.e. given a query instance $s_i$, we compute $\alpha_i^l$ and $t_i^l \; \forall \; l \in \{1, 2, \ldots, L\}$. The uncertainty for the region to belong to the known semantic categories is now defined as:

$$u_i = \min_{l=1}^{L} \left(1 - t_i^l\right) \tag{10}$$

Above, we compute the uncertainty of belonging to familiar semantic categories as a complement of $t_i^l$ as a lower $t_i^l$ corresponds to higher uncertainty for $s_i$ with respect to class $l$. The subsequent minimum function will select the class which is the least *strange* in comparison to $s_i$. Sample outputs for the strangeness based uncertainty are shown in Figure 1. In the figure, images have been labeled using the SiftFlow [11] dataset with semantic categories - *sky, building, tree, mountain, road, sea, field*. In the top row, a majority of the image regions have low uncertainty except for the road divider and the vehicle. The bottom row is an indoor scene where most of the image regions are associated with high uncertainty values.



(a) (b) (c)

Figure 1. Example uncertainty outputs (best viewed in color). (a) Query image (b) Predicted semantic segmentation (c) Strangeness based uncertainty. Darker intensity pixel implies higher uncertainty of belonging to the known semantic categories. Color coding for the semantic labeling: sky - blue, building - red, road - grey, tree - green, mountain - light blue.

## 5. Experiments

In this section, we present an evaluation of the approach proposed in this work. We first report the accuracy of our non-parametric approach for semantic segmentation. This is followed by an evaluation of the strangeness measure for confidence based ranking of a set of query images.

**Semantic Segmentation** We first evaluate the efficacy of the non-parametric approach for the problem of cross dataset semantic segmentation. Unlike the standard approach to semantic labeling where the evaluation is carried out by splitting a dataset into train and test sets, we consider source data from one benchmark set and evaluate models learned from it on another dataset. The motivation for doing so is to first establish the efficacy of the labeling approach for familiar semantic category segmentation before evaluating any confidence based ranking. For this purpose, we consider three datasets of varying sizes which are commonly used by the research community for semantic labeling experiments. The details for the datasets are summarized in Table 1.

For the source dataset, we consider the two smaller datasets - Stanford and SiftFlow i.e. these constitute $T_s$. In both of these datasets, we select the seven most frequent background categories - *sky, building, tree, mountain,*

| Dataset | Images | Categories | Scene Type |
|---|---|---|---|
| Stanford BG [6] | 715 | 8 | Outdoor |
| SiftFlow [11] | 2688 | 33 | Outdoor |
| SUN09 [3] | 8662 | 107 | Indoor + Outdoor |

Table 1. Details for dataset used in our experiments.

*road, sea, field*. Evaluation of the non-parametric approach is carried out on the large scale SUN09 dataset. Pixels in SUN09 which do not belong to any of the seven categories are labeled *void*. As a baseline, we train boosting classifiers which have previously shown success in geometric layout computation [7] and semantic segmentation of urban environments [14]. Within the boosting framework, we use decision trees as the weak learners since they automatically provide feature selection. We learn separate classifiers for each of the seven semantic categories classes in a one vs. all fashion. Given a query image, the separate classifiers are run on the individual feature vectors of the superpixels of the image and output confidence scores. The class with the maximum confidence score is assigned to be a superpixel's label. In our implementation, each strong classifier is composed of 25 decision trees with the tree size limited to 8 nodes. Table 2 reports the performance of our approach against the baseline boosting method. The evaluation criterion for the methods is the per pixel accuracy (percentage of pixels correctly labeled) and per category accuracy (the average of semantic category accuracies).

| $T_s$ | Labeling Method | Per Pixel | Per Category |
|---|---|---|---|
| Siftflow | Boosting | 73.9 | 61.7 |
| Siftflow | UKNN | 70.4 | 58.5 |
| Siftflow | UKNN | 73.4 | 61.9 |
| Stanford | Boosting | 68.0 | 59.4 |
| Stanford | UKNN | 64.1 | 54.4 |
| Stanford | WKNN | 67.6 | 58.9 |

Table 2. Results for semantic labeling on SUN09 dataset. $T_s$ is the dataset which provides training instances. UKNN is the uniform weight $k$-NN classifier while WKNN is learned weight $k$-NN.

It can be observed that the results for our non-parametric approach and the boosting classifier are similar. While the uniformly weighted $k$-NN lags behind in comparison to the boosting output, utilizing feature channel relevance at query time leads to an improvement of more than 3% with both SiftFlow and Stanford datasets. The results for using SiftFlow are better than the output for Stanford due to the fact that SiftFlow and SUN09 share a few images as they are drawn from the large scale SUN database [18].

**Confidence based Ranking** The next evaluation focuses on the introspective capacity of the proposed approach. Similar to the evaluation of the semantic segmentation, the introspection of the semantic labeling is also carried out on the large scale SUN09 dataset. As mentioned previously in Table 1, both Stanford and SiftFlow are composed of outdoor scenes only while SUN09 consists of both outdoor and indoor scenes. For evaluation purposes, the category of the pixels which do not correspond to one of the seven aforementioned categories is set to the *void* label. With this processing, 3,321 images out of the 8,662 images in SUN09 dataset do not have a single pixel sharing a semantic label with the data from SiftFlow or Stanford datasets. Therefore, this provides an ample set of images on which we can evaluate our confidence based ranking method.

We compare the strangeness measure presented in our paper to two baseline methods which have been previously utilized for computing classifier confidence in active learning experiments [8, 9]. For these, we utilize the boosting classifier evaluated in the previous section. Given a query image, for each region $s_i$, the boosting classifier provides the probability $p_i^l$ for assigning a semantic label $l$ to the region. Given this probability, we compute two metrics to characterize the uncertainty of labeling the region:

- Normalized entropy of the boosting output (NEP)

$$N_i = \frac{\sum_{l=1}^{L} -p_i^l \log\left(p_i^l\right)}{\log\left(L\right)} \tag{11}$$

  Higher values for the normalized entropy implies more uncertainty in the labeling by the boosting classifier.

- Best versus Second Best probability (BvSB)

$$B_i = 1 - \left(p_i^{l_{m1}} - p_i^{l_{m2}}\right) \tag{12}$$

  where $p_i^{l_{m1}}$ and $p_i^{l_{m2}}$ are the highest two probability outputs from the boosting classifier. The lower the difference between $p_i^{l_{m1}}$ and $p_i^{l_{m2}}$, greater the uncertainty of labeling an instance.

The uncertainty measure presented in Eq. (10) is computed at the image region level. Using this uncertainty measure, we compute an image level uncertainty score. Given a query image $X_j$ composed of superpixels $\{s_1^j, s_2^j, \ldots, s_q^j\}$ where $q$ is the number of superpixels, the image level uncertainty of associating its regions with the familiar classes is computed as:

$$U(X_j) = \sum_{i=1}^{q} u_i \times size(s_i^j) \tag{13}$$

where $u_i$ is the uncertainty defined in Eq. (10) and $size(s_i^j)$ is the percent of image pixels corresponding to superpixel $s_i^j$ in $X_j$. When evaluating the NEP and BvSB methods, we substitute $u_i$ by $N_i$ and $B_i$ from Eq. (11) and Eq. (12) respectively.

Having obtained the image level uncertainty score, we sort the images of the query image set in a descending order. For any metric which is being evaluated for confidence

based ranking, the goal is to obtain a higher number of images with unfamiliar categories in the (numerically) lower ranks e.g. in our evaluation, a metric performs better if an indoor scene of a bedroom with floor, walls and tables is assigned higher uncertainty (and subsequently lower rank) with respect to the familiar semantic categories than an outdoor scene of a beach. As part of this evaluation, we divide the resultant rankings into subsets and compute the percentage of *void* pixels present in each subset of the rankings. The results for the different measures when using SiftFlow and Stanford dataset are presented in Figure 2 and 3 respectively.

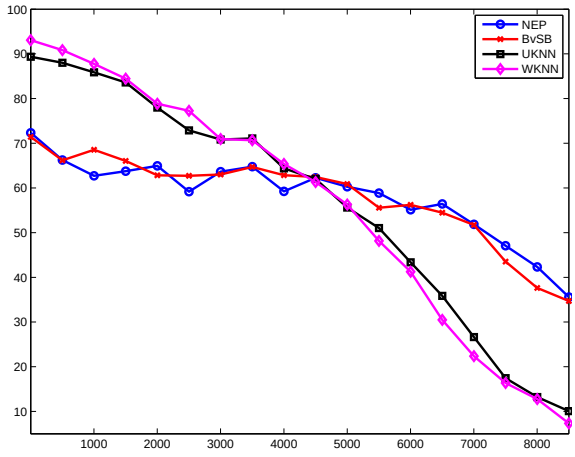



Figure 2. Comparison of uncertainty measures for confidence ranking in the SUN09 dataset using SiftFlow dataset. The y-axis denotes the percent of void pixels in the images present in a particular ranking subset (each of size 500 images) when using an uncertainty measure e.g. there are 93.1% void pixels in images ranked 1-500 using WKNN based strangeness. Higher percent of void pixels in lower ranks implies a better performance.

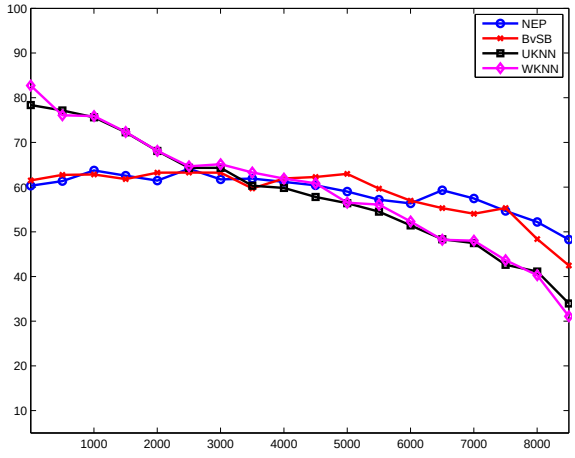



Figure 3. Results for SUN09 using Stanford as source dataset.

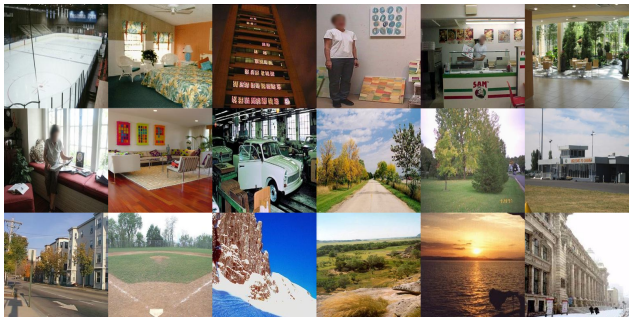


Figure 4. Visualization of the first image from each ranking subset of SUN09 using SiftfFlow as source dataset and WKNN strangeness for uncertainty computation.

As can be observed, the normalized entropy and best vs second best probability difference measures based on probabilistic output of the boosting classifier perform inferiorly in comparison to the strangeness measure. In particular, we obtain a higher percentage of void pixels when using strangeness instead of normalized entropy or best vs second best probability in the images with high uncertainty scores e.g. when using SiftFlow on SUN09, images ranked 1-500 had 93.1% void pixels while NEP had 72.3%. As the ranks increase, there is a drop in the ratio of void pixels indicating that images with familiar categories are associated with lower uncertainty scores. The UKNN metric outperforms both methods indicating the efficacy of using the transductive strangeness measure over the probabilistic output of the boosting classifier. When the strangeness is computed using a weighted $k$-NN, the performance improves further thereby highlighting the utility of feature relevance in a nearest neighbour framework. In Figure 4, we visualize the top ranked image in each ranking subset where the subsets are of size 500 images each i.e. image in the top left corner is ranked 1 while the image in right bottom corner is ranked 8501. It can be observed that the lower ranked sets are composed more from indoor scenes while the later ranks typically include outdoor scenes. Some examples of the associated confidences are provided in Figure 5.

## 6. Conclusions

We present a non-parametric approach for an introspective semantic segmentation of images. The approach is formulated over a single oversegmentation of an image which is labeled using a weighted $k$-NN approach where the weights are computed for individual feature channels at query time. The output of the non-parametric approach provides an introspective capability to analyze query image data by quantifying the uncertainty of semantic labels associated with the image regions. This is based on using the transductive *strangeness* measure which utilizes nearest neighbor distances. We presented results for confidence

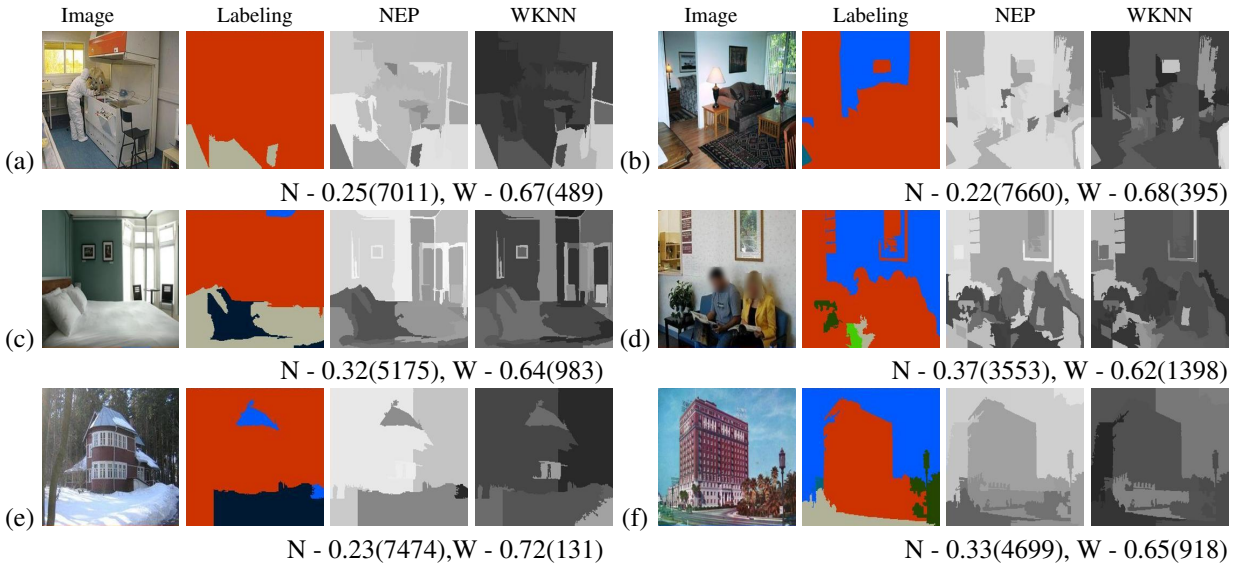



Figure 5. Results on SUN09 dataset using SiftFlow as the source dataset. Entries N and W denote the image level uncertainty score and corresponding image rank when using the NEP and WKNN strangeness measures respectively. (a)-(d) are examples of scenes where categories like person, wall, floor and furniture are associated with high certainty of being unfamiliar by WKNN strangeness but not necessarily by NEP. In example (e), there is an incorrect labeling of the trees as building and snow as sea. The NEP uncertainty score is low but WKNN strangeness correctly assigns a higher uncertainty score. Example-(f) is an instance of an incorrect high uncertainty score by WKNN strangeness for the familiar categories of sky and building. Color coding for semantic labeling: sky - blue, building - red, tree - dark green, mountain - light blue, road - grey, sea - dark blue, field - light green

ranking of images from the large scale SUN09 dataset. In future work, we would like to explore methods which can perform semantic category discovery using the highly uncertain image regions using region similarity and contextual cues from the high confidence familiar category regions.

**Acknowledgements** Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

## References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk. SLIC superpixels. Technical report, EPFL, 2010.

[3] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, pages 129–136, 2010.

[4] C. Domeniconi, J. Peng, and D. Gunopulos. Locally adaptive metric nearest-neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1281–1285, 2002.

[5] D. Eigen and R. Fergus. Nonparametric image parsing using adaptive neighbor sets. In *CVPR*, pages 2799–2806, 2012.

[6] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, pages 1–8, 2009.

[7] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007.

[8] A. Holub, P. Perona, and M. C. Burl. Entropy-based active learning for object recognition. In *CVPRW*, pages 1–8, 2008.

[9] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, pages 2372–2379, 2009.

[10] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. In *ICCV*, pages 1–8, 2007.

[11] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2368–2382, 2011.

[12] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP (1)*, pages 331–340, 2009.

[13] K. Proedrou, I. Nouretdinov, V. Vovk, and A. Gammerman. Transductive confidence machines for pattern recognition. In *ECML*, pages 381–390, 2002.

[14] G. Singh and J. Košecká. Acquiring semantics induced topology in urban environments. In *ICRA*, pages 3509–3514, 2012.

[15] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV (5)*, pages 352–365, 2010.

[16] S. Tong and E. Y. Chang. Support vector machine active learning for image retrieval. In *ACM Multimedia*, pages 107–118, 2001.

[17] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.

[18] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010.