Nonparametric Scene Parsing

103062588
呂立嫻
國立清華大學
新竹市光復路二段101號
lizzylu@hotmail.com.tw

103062543
黃家琬
國立清華大學
新竹市光復路二段101號
lisajwhl@hotmail.com

Abstract

Scene parsing is the problem of assigning a semantic label to every pixel in an image. This work adopts mid-level windows that are designed to capture entire objects, instead of low-level superpixels that tend to fragment objects. Rather than training a classifier for each class, we use a nonparametric method to tackle this problem. Besides, low per-class accuracy is a problem that most of scene parsing work faced. Through this project we aim to increase both per-pixel and per-class accuracy.

1. Introduction

Computer vision enables us to understand scenes at many different levels of abstraction. There are many nonparametric algorithms for scene parsing match superpixels in the query image with superpixels in similar database images. The advantage of superpixel based parsing is that it can label large groups of pixels at a time. Thus, superpixel based parsing tends to effectively label large regions of background, while doing less well on object classes. Because superpixel based parsing may fragment objects, it is necessary to apply some technique to combine them. Semantic label co-occurrence probabilities are commonly incorporated via a Markov random field model.

In this work so far, we follow the nonparametric image parsing method in CollageParsing [2]. To capture complete objects, CollageParsing uses mid-level, content-adaptive windows instead of low-level superpixels. First finds a retrieval set from global image features. Then content-adaptive windows are extracted from the query. The windows of the query are matched with the content-adaptive windows in the retrieval set to compute a unary energy for labelling each pixel. The unary potential is combined with a pairwise potential in a Markov random field to obtain an initial labelling, which is refined by aligning the labelling to the query image's superpixels. Figure 1 shows an overview of the pipeline.

2. Proposed Method

We use Caffe [1] to extract features of every image and regions in images. Caffe was created by Yangqing Jia during his PhD at UC Berkeley, it is a deep learning framework developed with cleanliness, readability, and speed in mind. Many works get impressive performance by using Caffe to extract features. This is why we use it to get features. In order to speed up, we extract all features for every image and region beforehand.

Our method can be divided to three steps. First, we retrieve images which are similar to query from training dataset. For every region in query image, we get similar regions that possibly contain correct labels from retrieval set. Then we resize retrieval regions and directly paste the labels to the corresponding position in query. For the last step, we smooth the labels with an MRF function.

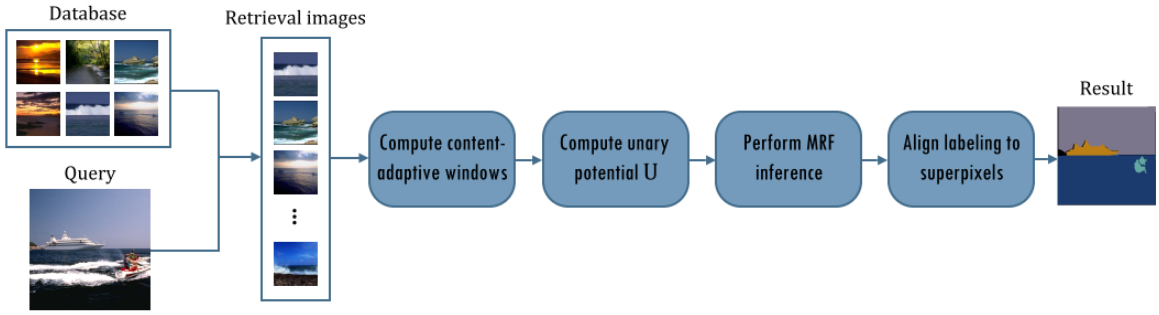

Figure 1: Flowchart of proposed method.

2.1. Image retrieval

In this part, we want to retrieve k most similar images to the query. Because the features are already off-lined extracted, we can directly use Euclidean distance to measure similarity between training image and the query. After that, we rank the similarity and then choose k most similar images as image retrieval set.

2.2. Region retrieval

Since we want to get more complete objects, we use windows to locate possible regions that may contain the target. We use selective search to obtain object proposals both in retrieval images and the query, then use RCNN to extract features. This step is also done off-lined.

For each region in the query, we get k most similar regions from retrieval set. Euclidean distance is used as similarity measurement.

2.3. Label Propagate

The final label result will be determined by minimizing the following energy function:

$$\mathrm{E(Y)} = \sum_{p} U(y_p = c) + \lambda \sum_{pq} V(y_p = c, y_q = c') \quad (1)$$

The first part of the energy function is unary term. It can be written as:

$$U(y_p = c) = -\emptyset_{sim}(qw_p, rw_p)\emptyset_{idf}(c)\emptyset_{winsize}(rw_p) \quad (2)$$

where

$$\emptyset_{sim}(qw_p, rw_p) = e^{-(\gamma \left\| f_{qw_p} - f_{rw_p} \right\|)^2} \quad (3)$$

$$\emptyset_{idf}(c) = \frac{1}{N(c)^{\alpha}} \quad (4)$$

$$\emptyset_{winsize}(rw_p) = \frac{1}{N(rw_p)} \quad (5)$$

$\emptyset_{sim}$ is window similarity, formed by RBF distance between color features of qw_p and rw_p, where qw_p is the query window that pixel p locates in, rw_p is the retrieval windows, and γ is a parameter. The higher similarity between the query and retrieval region, it is more possible that these two windows share the same labels.

$\emptyset_{idf}$ is term frequency for class c appearing in the retrieval images, where $\mathrm{N}(.)$ is the number of pixels, and α is a parameter.

$\emptyset_{winsize}$ is the size of retrieval window. The reason why we use this term is that the smaller window tends to precisely locate the object at center.

The second part of the energy function is binary term. First we get label co-occurrence, which means the probability that two labels are neighbors. For two adjacent pixels, if they belong to different labels and two labels seldom appear together in reality, the binary term will impose a penalty.

To obtain the final labeling, we can simply take the highest scoring label at each pixel, but this produces noisy results. As a result, we smooth the labels with an MRF energy function.

3. Experiments

4. Conclusion

5. Future work

This work still remains some problem, such as low per-class accuracy. Therefore, we want to combine the ideas of other method from [3].

In [3], this work follows a hybrid framework, which combines parametric and nonparametric method to solve this problem. First, they retrieved knn images for every query. Next, they divide query and training images into multi-level superpixels. The goal is to find the best labeling for each superpixel, and MRF is used to achieve the task. The parametric and nonparametric parts are embedded in the unary term of MRF energy function. For the nonparametric part, the k most similar superpixels of a query region are chosen. Therefore, the label of query region is determined by referencing those k neighboring regions. For the parametric part, they train a linear SVM for every label. As a result, the cost of belonging to certain label for a query region could be determined. After getting initial labeling, the authors further extract global and local labeling context. The purpose of this is to refine the results of image retrieval and superpixel matching.

There are some points about this paper worth mentioning. First, since a single pixel does not contain sufficient information for recognition, the authors chose to recognize pixels in proper neighboring regions, i.e. superpixels. Second, there are some rare classes causing data imbalance but somehow important. The authors expand data of rare classes to achieve better performance for parametric part. Third, they create a feedback loop to refine the results, and show that it really affects the performance.

6. Milestone achieved so far

- Image retrieval

- Region retrieval
- Label Propagate
- Region-based parsing

7. Remaining milestone

2014/12/31:

Add the idea of [3], such as replacing per-pixel labeling with per-superpixel labeling or add parametric idea to the work

2015/01

Add new ideas to refine labeling results.

References

[1] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093, 2014.

[2] F. Tung, and J. J. Little, CollageParsing: Nonparametric scene parsing by adaptive overlapping windows. In *ECCV*, 2014.

[3] J. Yang, B. Price, S. Cohen, and M.-H. Yang, "Context driven scene parsing with attention to rare classes," In *CVPR*, 2014.