Semantic segmentation refinement based on image time series context

Yao Wang, Yisong Chen, Guoping Wang
School of Electronics Engineering and Computer Science, Peking University, Beijing, China

***Abstract*— At present, deep convolutional neural networks (DCNNs) own state-of-art performance in semantic segmentation tasks. The prerequisite for training a DCNN is to have a large enough pixel level annotated training set. However, in practical applications, if the task target is semantic of insufficient sample size or large data sets with too large workloads, it is not possible to fine tune some pre-trained models of neural networks. This paper proposes a method combing a DCNN and a superpixel segmentation algorithm which gives a better accuracy of segmentation result of non-annotated dataset called key frame labeling voting (KFLV). KFLV requires the prediction of a pre-trained model and the segmentation result of a superpixel algorithm with proper parameters as input. Combining the semantic and edging information, KFLV gives 64.9% IOU (intersection over union) accuracy in the test set. KFLV proved to be a reasonable supplement of DCNN models in cases where fine tuning is not able to be done.**

I. Introduction

Image segmentation plays an important role in the field of computer vision since its naissance. Traditional methods like graph cut[1], random walks[2], grabcut[3] and SLIC[4], just barely reveal middle level of features[5]. However, understanding the semantic content of images is a high-level task that requires global and integrated information.

Recently, with the help of powerful deep convolution neural network(DCNN) tools, semantic segmentation is solved as a black-box task and the accuracy of segmentation escalates to a level never reached before[6]. In the field of semantic segmentation, many approaches are proposed in recent years such as deeplab[6] , VGGnet[7] and FCN[8]. The standard process of semantic segmentation requires the application of deep convolution neural networks (DCNNs), using the images as input, and the pixel level annotated semantic labels as output to train an end-to-end network. DCNN, in comparison with traditional segmentation method, DCNN reveals a higher level of information, and results in higher accuracy of semantic segmentation.

Training a DCNN with high accuracy on the training set and preventing over fitting requires sufficient pixel level annotated training data, which is laborious and time consuming. For individual researchers, it is hardly possible to annotate the entire dataset by hand. To find an alternative approach of obtaining more excellent results without annotation is the focus of this subject. Based on the end-to-end depth neural network, this topic is burnished by the cluster-based super-pixel segmentation method, and two street video sequences are selected for training and testing our algorithm.

In Section 2, we introduce the related work of this research. In Section 3, we propose key frame label voting (KFLV) algorithm, a semantic segmentation optimization algorithm based on temporal context, and to analyzes the feasibility of the optimization algorithm. Section 4 presents the experiment results. Section 5 is the conclusion.

II. Related work

Recently, some methods have shown promising results in image semantic segmentation[6][8]. Most of these works concentrate on the fully supervised setting, in which each pixel in the training image need to be annotated in advance[9]. However, acquiring such data is an expensive, time-consuming mission. Since that, there are quite a lot researchers turning into weekly supervised setting methods. These methods usually use superpixel algorithm as a powerful grouping tool. Superpixel algorithm groups pixels into perceptually meaningful atomic regions[4]. FeaBoost[10] attempts to label refinement over superpixels on weekly supervised setting. A framework called Weakly-Supervised Dual Clustering (WSDC) aims to cluster superpixels and assign a suitable label to each cluster[11].

The concept of key frame is widely used in computer vision[9][12]. It is a small subset of the video series, representative in general, that can stand for the entire image sequence. For frames in one view, key frames are often selected and used to compute binocular matching problems of 3D points[12]. In this study, we redefined key frames and proposed a new concept called normal frames. Key frames are chosen artificially from a dataset as the ground truth of label, while normal frames are the rest of the frames. The stereovision matching is implemented between one key frame and one normal frame.

III. Key frame label voting(KFLV)

In this section, we propose a novel framework in order to refine semantic segmentation result based on image time series context. As shown in Fig.1, the proposed framework consists of several steps.

A. General Process

Firstly, a DCNN model is trained on an annotated dataset. Secondly, the pre-trained DCNN model is adopted to generate a noisy label prediction result. Since it is not fine tuned, the average pixel accuracy is naturally low. But to some level, there are also some reasonable frames.

Then, the superpixel segmentation algorithm is implemented to over-segment the image into atomic pixel groups. After

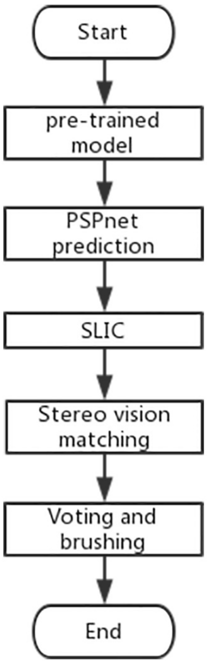


Figure 1: General procedure of KFLV

that, we compute the transformational matrix by stereo vision matching algorithm. Finally, an algorithm called voting and brushing is proposed to correct the noisy label.

1) Training: The dataset that DCNN is trained on must be similar to the test dataset. The percentage of each class should roughly be corresponded. Then the preliminary prediction results of semantic labels are prepared for superpixel algorithm. Since there are no annotated images (ground truth) in our test dataset, it is necessary to select several well segmented images as ground truth, or to label several frames manually.

2) Superpixel Segmentation: After we obtain the prediction results of normal frames and key frames, we should use superpixel segmentation function on normal frames.

The following concern is to combine the semantic and edge information together. Considering the shift between the neighbor frames is small and can be calculated by stereo vision matching[14], we try to combine the instance information of normal frames and the semantic label of key frames to optimize the semantic label of normal frames.

Taking this into account, we proposed voting and brushing algorithm to adjust the semantic labels of normal frame. The details of algorithm is explained in the next section.

3) Voting and Brushing Algorithm: The key frame gives a benchmark of semantic labels, while the normal frame contains the edge and instance contour information. The combination can be implemented using a variety of strategies, among which majority vote[15] is by far the simplest, and yet it has been found to be just as effective as more complicate schemes in optimizing. The winner-take-all mechanism[16] also inspires us, which substantially implements in free market competition theory, president election of America, etc.

We notice that for gradual changes of camera views, the relationship between normal and key frame is like the coordinate transformation relationship in 3-D reconstruction problems. We can calculate the shift between the two frames and acquire camera transformation matrix (**Algorithm 1**, line 7).

After mapping the points of normal frame to the key frame by stereo vision matching, we can find the label of most pixels in each mapped superpixel (maxLabel, **Algorithm 1**, line 6-10). Under the assumption that every superpixel is only contained in one instance, it is a proper approximation to set all the labels in one superpixel to be maxLabel (**Algorithm 1**, line 11-15).

IV. Experiment Result

A. *Experiment of SLIC Parameters*

The parameters of SLIC have a great influence on pixel accuracy. Since it is the foundation of KFLV, we have to optimize its parameters to meet our satisfaction. The main dependencies are the number of blocks(n) and edge complexity(c), and it is noteworthy that the regression function is not monotonically increasing or decreasing. We set up several sets of values to get the proper combination of parameters(Table I).

B. *Experiment of KFLV*

To validate the feasibility and performance of our method, we used cityscapes[17] fine dataset, which includes 5000 annotated images(frames), as the training set of PSPnet. For test, we choose KingsCollege dataset, street scenes collected from Kings college of London University. It is similar to cityscape scenes from the aspect of instance type and percentage. KingsCollege dataset has 8 video sequences with 1565 non-annotated frames.

The deep-learning framework used in this method is caffe[18]. First, we trained a model in cityscapes dataset. Then we propagate forward the test picture into PSPnet[13] network in order to get the preliminary prediction results.

Algorithm 1 Voting and Brushing Algorithm

```
1:  for every normal frame $p_c$, choose key frame $p_{key}$
2:  for SLIC result pixel block $S_i$ do
3:      for every label class $j$ do
4:          $label[j] = 0$
5:      end for
6:      for $pixel(x, y)$ in $S_i$ do
7:          require the corresponding coordinate in key frame
            $(x', y')$, denoted as $label_{key}(x', y')$
8:          $label[label_{key}(x', y')] = label[label_{key}(x', y')] + 1;$
9:      end for
10:     $maxLabel = max(label)$
11:     for $pixel(x, y)$ in $S_i$ do
12:         if $(label_c(x, y)) \neq label_{key}(x', y')$ or $label_{key}(x', y')$
            doesnt exist then
13:             $labelc(x, y) = maxLabel$
14:         end if
15:     end for
16: end for
```

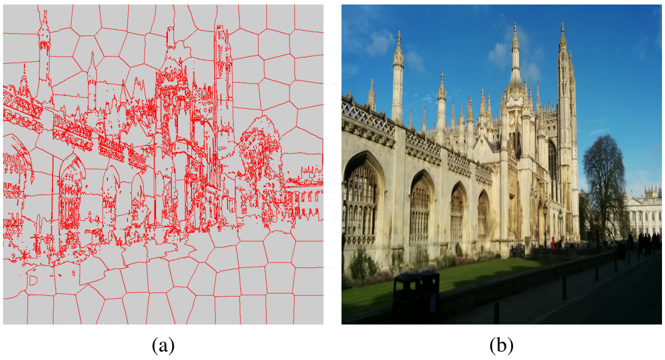

(a) (b)

Figure 2: SLIC result of kingsCollege frame
(a):SLIC result. (b):original kingsCollege frame

We start with our observation and analysis of the performance of DCNNs on street view datasets. We trained PSPnet by cityscapes fine dataset. After finishing training on cityscapes fine annotation data set, we use the model to predict KingsCollege. By observing in the experiment, the performance of the model on KingsCollege data set is polarized.

For instance, Fig.3(a) is the prediction result of Fig.3(b)(key frame). It has high accuracy over the prediction result of Fig.3(d)(normal frame). Furthermore, segmentation results of some frames is fine in general, but some frames have large false-segmented superpixels. Comparing those consecutive frames which have a wide difference in correct rate, the difference of scenes between the original pictures is not that large.

For the preliminary prediction results of PSPnet pre-training model, we first implement SLIC algorithm to require several superpixels, then choose the key frames artificially. For each key frame and its adjacent normal frames, we use SIFT detector[19] to match across views and calculate rotation and transformation. Then we implement **Algorithm 1** to acquire new semantic labels on normal frames.

The statistic value of cityscapes is given by reference[17], while that of KingsCollege data set is 1/25 by overall random sampling comparing with manual calibration of ground truth. This algorithm is useful for data set without ground truth. As for those with ground truth, training is obviously a more efficient way.

TABLE I: Comparison between datasets with different parameters of SLIC

Number of block(n)	Edge complexity(c)	Pixel Acc.(%)
200	1/10	87.4
200	1/100	69.2
200	1/20	88.2
100	1/20	63.7
300	1/20	87.6

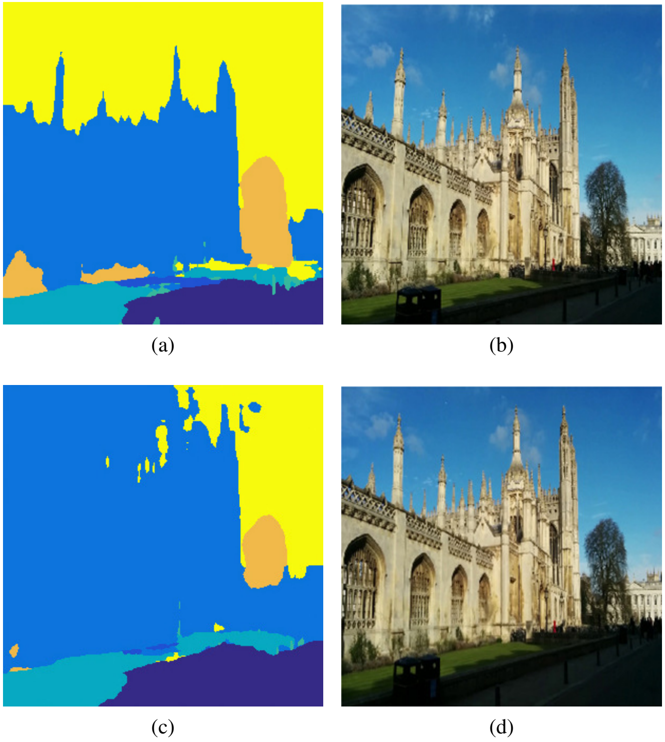

(a) (b) (c) (d)

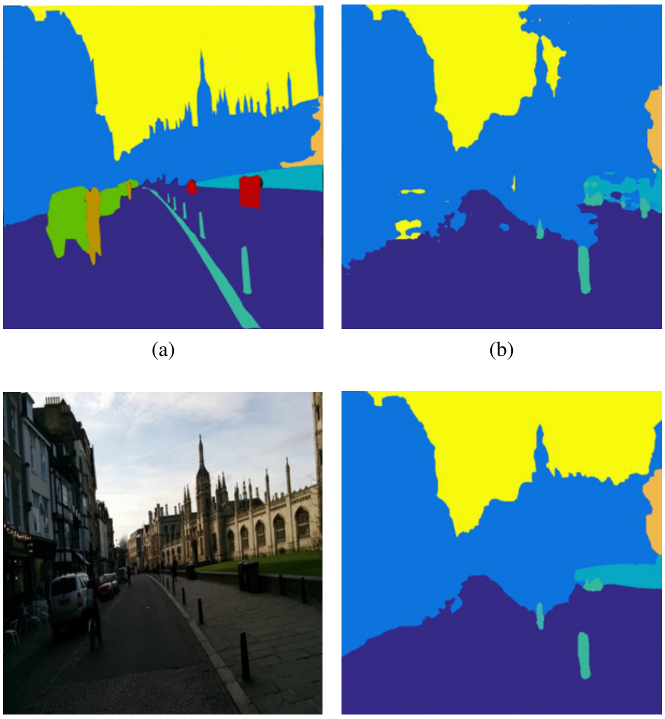
Figure 3. Predictions on kingsCollege by PSPnet.
(a): semantic segmentation result of (b) (b): normal frame
(c): semantic segmentation result of (d) (d): key frame

C. *Result*

The parameters of SLIC algorithm have a huge influence on the results (Table I). The edge complexity shows the description accuracy level of object edges, and the number of pixels indicates the average size of pixels. The size of superpixels and the edge complexity should be moderate so as to receive a high accuracy.

For general evaluation of KFLV, we compared the original result of PSPnet on Cityscapes, the raw result of PSPnet on KingsCollege and the result of KFLV. The performance of PSPnet in cityscapes and KingsCollege data set and the results of KingsCollege data set after doing KFLV algorithm are shown in Table II. Our method (KingsCollege with KFLV) gives 64.9% IOU (intersection over union) accuracy in the test set, which exceeds the pre-trained model 16.1%. It is reasonable that our result is still far behind the state-of-art method, because we are under the supposed situation that there is no enough data for fine tune. From our observation, the main reason for the improvement of correct rate is the rectify of segmented pixel blocks of large areas, such as sky and building.

Fig.4 represents a general process result of key frame label voting(KFLV). Fig.4(a) is the normal frame we arbitrarily picked and the ground truth is labeled by hand for comparison. We could see that Fig.4(b) recognizes a large area of sky as building. Fig.4(d) corrects a lot of the misinterpreted areas.


(a)

(b)

(c)

(d)

Figure 4. Performance of KFLV on kingsCollege dataset
(a): ground truth (b): prediction result of PSPnet
(c): original (d): prediction result refined by KFLV

V. Conclusion

This article raises KFLV(key frame label voting) algorithm, which is based on DCNN(Deep Convolution Neural Network) prediction results and super pixel segmentation algorithm. For data set of certain scene, some pre-training models already have high semantic segmentation accuracy. However, annotation is a time-consuming and expensive task. For models that are not fine tuned to a specific dataset, the prediction result must have some errors.

For datasets with ground truth, we should naturally fine tune the results. Another consideration is that since the image series is time related, it is naturally to introduce RNN(Recurrent Neural Network) to make use of sequential information hidden behind.

For datasets without ground truth, it is another story. DCNN can only give a very noisy and preliminary segmentation. With our method, we combine the semantic and edge information together by integrating DCNN and superpixel algorithm together. The KFLV algorithm can be applied as interactive semantic segmentation algorithm, even application in the future. Manual annotation of key frames uses the semantic and edge information of normal frame to optimize the adjacent normal frames. We proved that KFLV algorithm can optimize the image time series in a labor-saving way, in order to be popularized in many a field.

TABLE II: Performance of different methods

Dataset (and method)	MIoU cla.	iIoU cla.
Cityscapes [17]	78.4	90.6
KingsCollege	41.3	48.8
KingsCollege with KFLV	55.4	64.9

Acknowledgment

This research is supported by National Natural Science Foundation (NSF) of China under grant nos. 61232014, 61421062 and 61472010, National Basic Research Program of China (973 Program) under grant no. 2015CB351806, and National Key Technology R&D Program of China under grant no. 2015BAK01B06.

References

[1] Vicente S, Kolmogorov V, Rother C. Graph cut based image segmentation with connectivity priors[C]//Computer vision and pattern recognition, 2008. CVPR 2008. IEEE conference on. IEEE, 2008: 1-8.

[2] Grady L. Random walks for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2006, 28(11): 1768-1783.

[3] Rother C, Kolmogorov V, Blake A. Grabcut: Interactive foreground extraction using iterated graph cuts[C]//ACM transactions on graphics (TOG). ACM, 2004, 23(3): 309-314.

[4] Achanta R, Shaji A, Smith K, et al. Slic superpixels[R]. 2010.

[5] Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1717-1724.

[6] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062, 2014.

[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[8] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.

[9] Zhiwu Lu, Zhenyong Fu, Tao Xiang, Peng Han, Liwei Wang, Xin Gao. Learning from Weak and Noisy Labels for Semantic Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39(3): 486-500 (2017)

[10] Yulei Niu, Zhiwu Lu, Songfang Huang, Xin Gao, Ji-Rong Wen. FeaBoost: Joint Feature and Label Refinement for Semantic Segmentation. AAAI 2017: 1474-1480

[11] Liu Y, Liu J, Li Z, et al. Weakly-Supervised Dual Clustering for Image Semantic Segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2013:2075-2082.

[12] Klein G, Murray D. Improving the Agility of Keyframe-Based SLAM[C]// European Conference on Computer Vision. Springer-Verlag, 2008:802-815.

[13] Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network[J]. arXiv:1612.01105, 2016.

[14] Pajares G, Cruz J M D L. Stereovision matching through support vector machines[J]. Pattern Recognition Letters, 2003, 24(15):2575-2583.

[15] Lipton M. Majority Voting[J]. Venulex Legal Summaries, 2010.

[16] Maass W. On the computational power of winner-take-all[J]. Neural computation, 2000, 12(11): 2519-2535.

[17] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3213-3223.

[18] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]//Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 675-678.

[19] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.