Multi-context Deep Convolutional Features and Exemplar-SVMs for Scene Parsing

Xiaofei Cui[1], Hanbing Qu[1,2,3(✉)], Songtao Wang[2,3], Liang Dong[2,3], and Ziliang Qi[1]

[1] School of Computer Science and Engineering, Hebei University of Technology, Tianjin 300401, China
dlaicxf@126.com, quhanbing@gmail.com, 18233183254@163.com
[2] Beijing Institute of New Technology Applications, Beijing 100094, China
wangsongtao1983@163.com, dongliang@bionta.cn
[3] Beijing Academy of Science and Technology, Beijing 100089, China

Abstract. Scene parsing is a challenging task in computer vision field. The work of scene parsing is labeling every pixel in an image with its semantic category to which it belongs. In this paper, we solve this problem by proposing an approach that combines the multi-context deep convolutional features with exemplar-SVMs for scene parsing. A convolutional neural network is employed to learn the multi-context deep features which include image global features and local features. In contrast to hand-crafted feature extraction approaches, the convolutional neural network learns features automatically and the features can better describe images on the task. In order to obtain a high class recognition accuracy, our system consists of the exemplar-SVMs which is training a linear SVM classifier for every exemplar in the training set for classification. Finally, multiple cues are integrated into a Markov Random Field framework to infer the parsing result. We apply our system to two challenging datasets, SIFT Flow dataset and the dataset which is collected by ourselves. The experimental results demonstrate that our method can achieve good performance.

Keywords: Scene parsing · Convolutional neural network
Exemplar-SVMs · Markov Random Field

1 Introduction

In the computer vision field, scene parsing is one of the core problems. This paper researches the problem of scene parsing. Scene parsing aims at assigning a label to each pixel in images with the semantic category it belongs to. This work is very challenging, and it involves all kinds of problems, e.g., detecting object categories problems, segmentation problems, and multi-label recognition problems of the image, and so on. In the last decade, scene parsing had been comprehensive studied by several researchers.

J. Yang et al. (Eds.): CCCV 2017, Part I, CCIS 771, pp. 494–505, 2017.
https://doi.org/10.1007/978-981-10-7299-4_41

For lots of problems in computer vision field including scene parsing, it is necessary to extract features for each image. Features are the expressions of images for easy calculation and classification easily. Hence extracting good features is the key of computer vision tasks. For image representation, the conventional approaches use hand-crafted features by designing carefully, e.g., HOG [3], SIFT [14], GIST [16], etc. Since the introduction by LeCun [11], deep CNN has been applied to a wide range of computer vision tasks such as hand-written digit classification and face detection. Recently, the latest generation of CNNs have substantially outperformed traditional approaches in many recognition tasks [2,5,10,17,18]. Deep learning is advantageous for large image regions with complex variations, because its deep architectures can automatically learn the hierarchies of features representation.

The issue of scene parsing is improving the recognition rate of the object classes. Some of the previous methods [4,12,21] had low recognition rate for small objects which occupied only a few pixels. To increase the accuracy rate of thing samples, Tighe and Lazebnik [22] proposed using per-exemplar detector which is training a linear SVM classifier for every exemplar in the training set. In addition, Yang et al. [23] expanded the retrieval set using rare class exemplars to achieve more balanced superpixel classification results and combined both global and local semantic context information to boost the performance of their system. Our approach is inspired by the work of [22], so we use exemplar-SVMs [15] which are identical to Tighe and Lazebnik [22] for improving the accuracy rate of rare classes.

In this work, we present a scene parsing approach which combines the multi-context deep convolutional features with exemplar-SVMs. This method can improve the overall accuracy and recognize the rare classes better. Our system has the following steps. We first retrieves the most similar images to query images, as shown in Fig. 1, in an annotated dataset using the global features which are obtained from the convolutional neural network of the GoogLeNet [20]. Then, we segment the images into superpixels and compute the likelihood scores of superpixels of the query images using local features, which incorporate the exemplar Support Vector Machine for classification. Finally, our system integrates multiple cues into a Markov Random Field framework to infer the parsing result.

The rest of this paper is organized as follows. In Sect. 2, we review some relevant works for scene parsing. Section 3 presents our approach which combines the multi-context deep convolutional features with exemplar-SVMs for scene parsing. Section 4 provides the experimental results and compares the performance with other approaches, followed by a conclusion in Sect. 5.

2 Related Work

Most recently, as the development of artificial intelligence, many researchers use the deep learning approach which is one of the machine learning algorithms to solve a lot of issues in computer vision field, including object recognition,

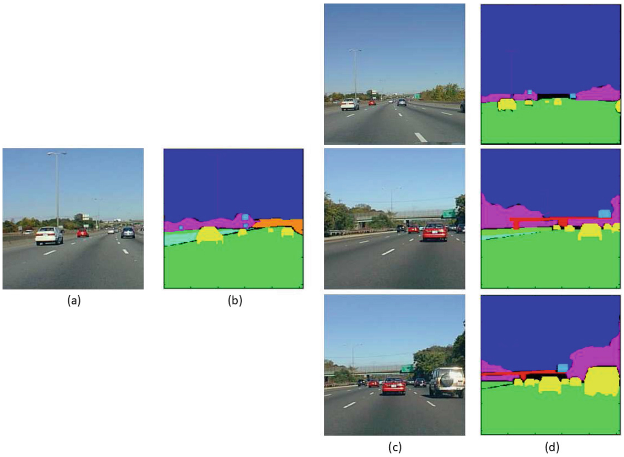

Fig. 1. For a query image (a), our system retrieves the most similar images (c) (three images are shown here) from training set using the global features which are obtained from the convolutional neural network. The ground-truth annotation of (a) and (c) are shown in (b) and (d), respectively.

object detection, simultaneous image segmentation and labeling. The studies have shown that convolution neural network has achieved good results in these works. In recent years, a lot of methods for scene parsing had been proposed. The goal of this work is to assign each pixel of the image a semantic label. Grangier et al. [8] presented a method that did not use any graphical model for scene parsing. The deep convolutional network was used for this paper to model the complex scene label structures, depending on a supervised greedy learning approach. Farabet et al. [5] learned the different levels features applying a multi-scale convolutional neural network handling on original pixels. Their system depended on those features representation to think about context for each decision. Another innovation point of this paper was that they employed the optimal cover to obtain the most consistent region from the tree, taking the place of graph cuts and other inference approaches. Pinheiro and Collobert [17] applied the recurrent convolutional neural networks to scene labeling. The advantage of this paper was that it was no need segmentation and task-specific features compared to other approaches. It was trained in an end-to-end way over raw pixels, and the inference cost was low. In order to improve the overall performance, Bu et al. [2] combined the convolutional neural network architecture with a graphical model. The neural network can learn the hybrid features which could provide the hierarchical information and the spatial relationship information.

From what has been discussed above, the deep learning method was applied to the scene parsing work had achieved impressive results. But training network model of deep learning need a lot of data. The following will introduce some other methods that they were the traditional and did not apply deep learning. For instance, Liu et al. [12] proposed a nonparametric approach for scene parsing using a dense scene alignment to label transfer. The SIFT flow algorithm was used to match the structures within two images. This system did not require any training, but inference via the SIFT flow was very complex and computationally expensive. With the consideration of these, Tighe and Lazebnik [21] presented a scalable nonparametric image parsing with superpixels. This method also required no training and it was easy to implement. It established a new benchmark for scene parsing. The system in Zhang et al. [24] proposed an ANN bilateral matching scheme to match the test images and retrieved the training images. Eigen and Fergus [4] used per-descriptor weights to minimize the classification error and to improve performance on rare classes, the paper used a context-driven adaptation of the training set for each query image. Different from other paper, this system did not use SVMs for classification. Tighe and Lazebnik [22] incorporated region-level features with per-exemplar sliding window detectors outputs for image parsing to achieve broad coverage across hundreds of object categories. Yang et al. [23] used thing class exemplars extending the retrieval set, to obtain more balanced superpixel classifications and combined the global with local semantic context information for refining the image retrieval and superpixel matching. The experimental results compared with previous methods have improved. To boost the label likelihood estimates at superpixels George [6] integrated the likelihood scores forming different probabilistic classifiers and in the parsing process incorporated the semantic context through global label costs. The newest paper, a large and open vocabulary was presented by Zhao et al. [25] for scene parsing, which aimed at parsing images in the wild. Image pixels and word concepts connecting with semantic relations were merged to the parsing framework. To parse a 2D image and recover the 3D scene structures, Liu et al. [13] presented an attribute grammar, which contained a set of production rules, each of which described a series of spatial relation among planar surfaces in 3D scenes.

3 Approach

In this section, we propose an overview of our scene parsing system which combines the convolutional neural network with exemplar-SVMs. Our approach builds on the framework of [22]. Firstly, the GoogLeNet [20] is used to extract the global image features for computing the similar images of training set to the query image. Then we divide the image into superpixels. The features of superpixels are extracted to obtain the probability of class labels for each query image. Finally, we incorporate the superpixel parsing results with exemplar-SVMs parsing results and integrate them to the Markov Random Field framework for obtaining the final parsing. The common architecture of the system which our presented is shown in Fig. 2.

3.1 Feature Extraction with CNN

The deep Convolutional Neural Network (CNN) [11], which showed its powerfulness in extracting high-level feature representations [7] recently, can solve

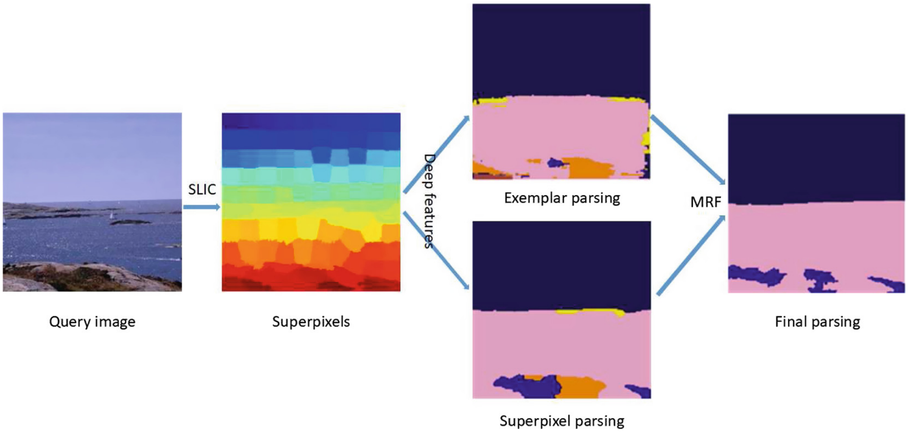

Fig. 2. The general framework of our system. Given a query image, we extract superpixels using the SLIC method. Then, we use the multi-context deep convolutional features including global and local deep features to describe superpixels and compute the superpixel likelihood ratio scores to obtain the superpixel parsing. We use exemplar detectors to obtain the exemplar parsing. Finally, in order to obtain the final parsing, we combine the exemplar parsing results with superpixel parsing results into the MRF framework.

previous some problems well. Deep CNN aims to mimic the functions of neocortex in human brain as a hierarchy of filters and nonlinear operations. The major goal of deep learning is learning the feature representations. The good network structure types should be selected to study features. In this paper, we choose the convolutional neural network of GoogLeNet [20]. The multi-context deep convolutional features which are introduced in this paper refer to the global features and the local features.

Global Features. For each query image I_q, we extract the 1024-dimensional feature vector of GoogLeNet [20] the pool5 layer output for computing similar images which include similar classes and spatial layouts of the training dataset. Then we rank all the training images in increasing order according to the Euclidean distance between training images and the query image. It is similar to the previous works. We select the top K similar images to constitute the retrieval set. The number of images in retrieval set is less than that in training dataset. So we process images in retrieval set instead of those in training set for reducing the computational cost. The process of extracting image features using the convolutional neural network of GoogLeNet is presented in Fig. 3.

Local Features. According to the computer vision applications, superpixels are becoming increasingly popular for use. To reduce the problem space and complexity, we also use superpixels as computing element, which can provide

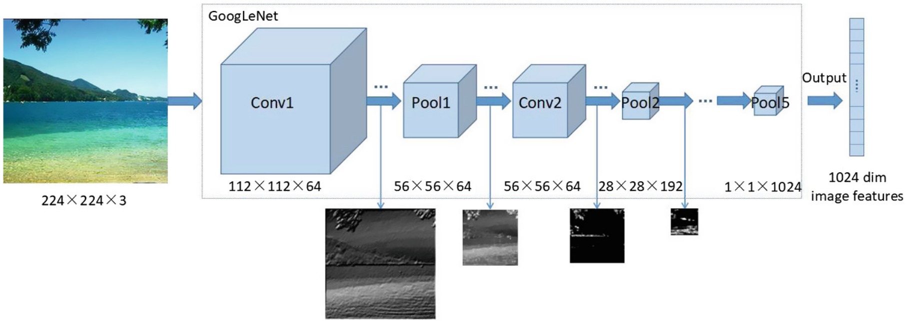

Fig. 3. The process of extracting image features using the convolutional neural network of GoogLeNet. For a query image, the input size of the network is 224 × 224 × 3. Though convolution and pooling operation, the visualization results of some layers are shown in the figure. We extract the pool5 layer output of GoogLeNet as our deep global features which are 1024-dimensional.

better spatial support for aggregating image features that may pertain to a single object than pixels. The image superpixels are extracted by the SLIC [1] method which is easy to implement. For each superpixel, we use the convolutional network of GoogLeNet to extract features for describing it. For each query image, we pad the image and select a superpixel-centered context window. The input of the local-context network is the superpixel-centered context window which is upsample to 224 × 224 × 3. The local feature shares the same deep structure with the global feature. Other deep structures can also be used to extract the local image features. The process is shown in Fig. 4. Those local features produce a log-likelihood ratio score $L(r_i, c)$ for class label c at superpixel r_i:

$$L(r_i, c) = \log \frac{P(r_i|c)}{P(r_i|\bar{c})} = \sum_k \log \frac{P\left(f_i^k|c\right)}{P\left(f_i^k|\bar{c}\right)} \tag{1}$$

where f_i^k is the feature; $P\left(f_i^k|c\right)$ (resp. $P\left(f_i^k|\bar{c}\right)$) is the likelihood of feature type k for region r_i given class c (resp. all classes but c) and here $k = 1$.

3.2 Exemplar-SVMs

In order to improve the rare class recognition rate, we incorporate the per-exemplar framework of [15], which is similar to the previous work [22]. In our dataset, the per-exemplar detector is trained for each labeled class instance. As the same with the detector training procedure of [15], all objects which include "thing" classes and "stuff" classes are trained. Each exemplar-SVMs creates the detection windows in a sliding fashion, and an exemplar co-occurence-based mechanism is applied for suppressing the redundant responses.

The leave-one-out method is employed to train SVM classifier on the training set. Each feature dimension is normalized by its standard deviation and the

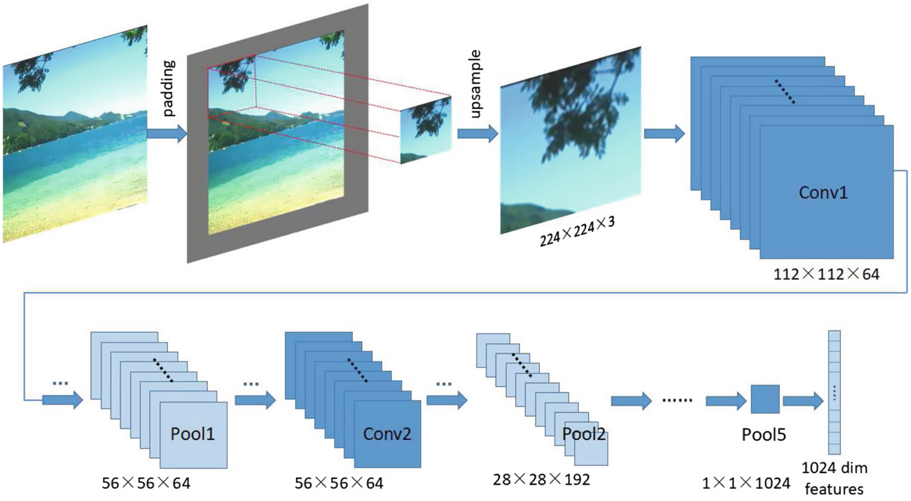

Fig. 4. Process of extracting local features using GoogLeNet. For a query image, we pad regions exceeding boundaries with mean pixel value. Then a superpixel-centered large context window is selected, and warped to $224 \times 224 \times 3$ as input. We extract the pool5 layer output of GoogLeNet as our deep local features.

regularization constant is computed using fivefold cross validation. The RBF kernel is applied for SVM training. At each pixel, the output results of SVMs produce Y responses. The $D_{SVM}(c_i, r_i)$ is defined to denote the response of the SVM for class c_i at superpixel r_i.

3.3 Inference

To assign labels to all of superpixels r_i in the query image I_q, the Markov Random Field model is incorporated into our system. The scene parsing problem is formulated as minimization of a standard MRF energy function, which combines the unary term with the smoothness term to obtain the final parsing result. We define the MRF energy function of the superpixel class labels $\mathbf{c}$ as:

$$\mathbf{E}(\mathbf{c}) = \sum_{r_i \in I_q} D(r_i, \mathbf{c}) + \sum_{r_i \in I_q} \max\left[0, M - D_{SVM}(c_i, r_i)\right] + \lambda \sum_{(r_i, r_j) \in N} \psi\left(c_{r_i}, c_{r_j}\right) \tag{2}$$

where $D(r_i, \mathbf{c})$ is the data term; once we obtain the likelihood $L(r_i, c)$, $D(r_i, \mathbf{c}) = -\omega_i logL(r_i, c)$; M is the highest expected value of SVM response; N is the set of adjacent pixels; The smoothing constant is λ. The smoothness term ψ is similar with which defined in [21,23] according to the probabilities of label c_i and c_j co-occurring in the training images:

$$\psi\left(c_{r_i}, c_{r_j}\right) = -\log\left[\left(\mathbf{P}\left(c_{r_i}|c_{r_j}\right) + \mathbf{P}\left(c_{r_j}|c_{r_i}\right)\right)/2\right]\delta\left[c_{r_i} \neq c_{r_j}\right] \tag{3}$$

where $\mathbf{P}\left(c_{r_i}|c_{r_j}\right)$ is the conditional probability of class label c_{r_i} of the pixel r_i, and its adjacent pixel r_j has class label c_{r_j}. We use the α-expansion algorithm [9] to optimize the MRF energy function.

4 Experiments

4.1 Dataset

The system results are reported on two datasets: the SIFT Flow dataset [12] containing 2688 images which were divided into 2488 training images, 200 test images, and 33 classes, and the per-pixel frequency counts of object categories in test images are presented in the left of Fig. 5. Another is video images dataset collected from the actual video surveillance scene, which included 500 images, among 450 training images, 50 test images, and 20 classes, and the per-pixel frequency counts of object categories are presented in the right of Fig. 5. As can be seen from the graph, some classes do not exist in the test set.

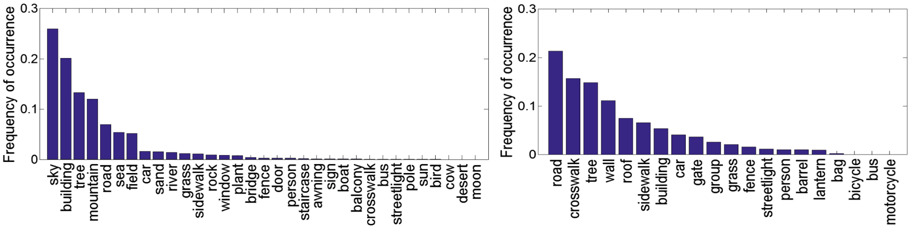

Fig. 5. The per-pixel frequency counts of the object categories on the SIFT Flow dataset (left) and video images dataset of we gathering in the actual video surveillance scene (right).

4.2 The Experimental Results

There are two measurement function for evaluating our scene parsing system: the per-pixel classification accuracy rate (the percentage of pixels that were correctly labeled of test images) and per-class classification rate (the average of all classes recognition rates). For SIFT Flow dataset, the size of the retrieval set is set as 200, namely $K = 200$. For the video images which are collected by ourselves, the size of the retrieval set is set as 50. The smoothing constant λ is 1. The highest expected value of SVM response M is approximately 10. The per-class recognition rates of SIFT Flow dataset and our collected video images dataset are presented in the left and right of Fig. 6, respectively. On the SIFT Flow dataset, the top five categories of recognition rate are sun, sky, building, sea and mountain.

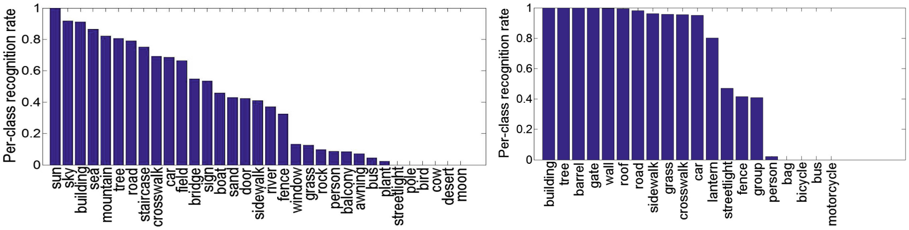

Fig. 6. Left: the per-class recognition rate on the SIFT Flow dataset. Right: the per-class recognition rate on the video images dataset of we gathering in the actual video surveillance scene.

Table 1 shows that the results of our system in contrast to the state-of-the art methods about per-pixel and per-class recognition rate on the SIFT Flow dataset. Although the recognition rate of presenting method is not the highest, the per-pixel accuracy increases by around 4% and the per-class accuracy increase about 17% over the method of [21]. The per-pixel accuracy of our approach is less than George [6] about 6% and the per-class recognition rate is less approximately 3%.

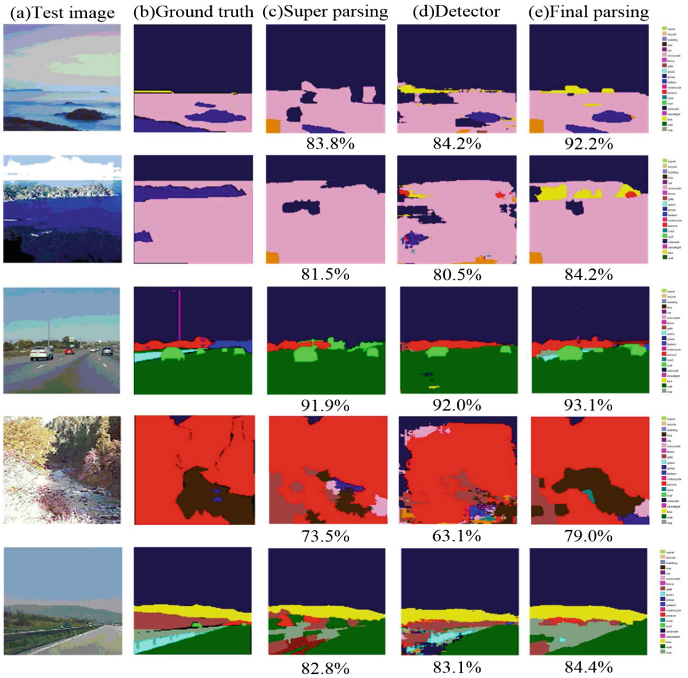

Fig. 7. The visual results of our approach on the SIFT Flow dataset. Column (a) is the test images; the ground truth of (a) is shown in (b); (c) is the superpixel parsing results; the exemplar parsing results are displayed in (d) and (e) is the final parsing result, which combines (c) with (d).

Table 1. A comparison of our system to the state-of-the-art methods about per-pixel and per-class classification accuracy rate (%) on the SIFT Flow dataset.

Approach	Per-pixel	Per-class
Liu et al. [12]	74.75	N/A
Tighe and Lazebnick [21]	76.90	30.10
Farabet et al. [5]	78.50	29.50
Tighe and Lazebnick [22]	78.60	39.20
Yang et al. [23]	79.80	48.70
Bu et al. [2]	80.40	35.80
Shuai et al. [19]	81.20	45.50
George [6]	81.70	50.10
Our approach	81.16	47.91

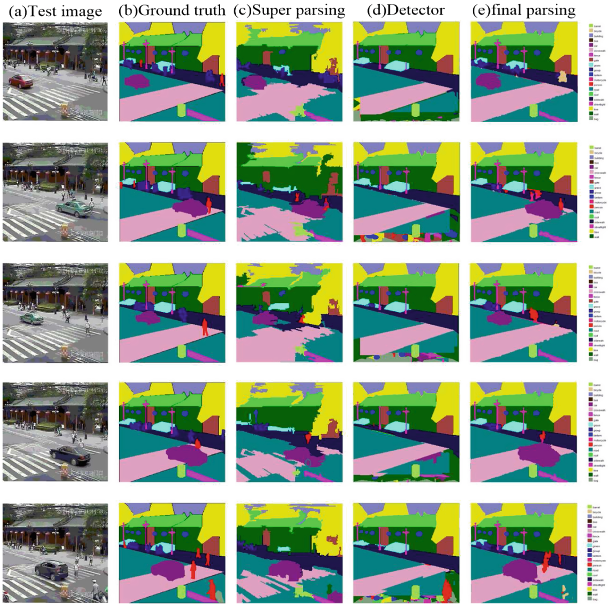

Fig. 8. The visual results of our approach on the video images. Column (a) is the test images; the ground truth of (a) is shown in (b); (c) is the superpixel parsing results; the exemplar parsing results are displayed in (d) and (e) is the final parsing result, which combines (c) with (d).

The part of visual samples results of our method on the SIFT Flow dataset and the video images are shown on Figs. 7 and 8, respectively. In Fig. 7, (a) expresses the test images; The ground truth of (a) is shown in (b); (c) is superpixel parsing results and the recognition rates are displayed under the images. The exemplar parsing results are shown in (d). The final parsing results are displayed in (e). The Fig. 8 is similar to Fig. 7, but does not have the accuracy of images. As can be seen from Fig. 8, the parsing results of video images are impressive.

The computing environment of our system is MATLAB 2012b and on a four-core 3.4 GHz CPU with Intel Core i7-6700 and 8G of RAM. Training time and parsing time are influenced by feature dimension. The higher the dimension and amount of features are, the longer the training time costs.

5 Conclusions

This paper has proposed a method for scene parsing to improve the quality of retrieval sets and the class recognition rate of combining the deep convolutional features with exemplar-SVMs. Our approach takes advantage of a retrieval set of similar objects and scenes for reducing computational complexity. The deep features we extracted can express the image well. Furthermore, multiple cues are integrated into a MRF framework to segment and parse the test image. On the SIFT Flow dataset, the experimental results show that the method achieve promising results. This system is used to actual scene images and obtained the ideal parsing results.

Acknowledgments. This research was supported by Beijing Municipal Special Financial Project (PXM2016_178215_000013), Beijing Urban System Engineering Research Center's Independent Project (2017C010).

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE TPAMI **34**(11), 2274–2282 (2012)
2. Bu, S., Han, P., Liu, Z., Han, J.: Scene parsing using inference embedded deep networks. PR **59**(C), 188–198 (2016)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. IEEE CVPR **1**, 886–893 (2005)
4. Eigen, D., Fergus, R.: Nonparametric image parsing using adaptive neighbor sets. In: IEEE CVPR, pp. 2799–2806 (2012)
5. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Scene parsing with multiscale feature learning, purity trees, and optimal covers. arXiv preprint arXiv:1202.2160 (2012)
6. George, M.: Image parsing with a wide range of classes and scene-level context. In: IEEE CVPR, pp. 3622–3630 (2015)
7. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE CVPR, pp. 580–587 (2014)
8. Grangier, D., Bottou, L., Collobert, R.: Deep convolutional networks for scene parsing. In: ICML 2009 Deep Learning Workshop, vol. 3 (2009)
9. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? IEEE TPAMI **26**(2), 147–159 (2004)

10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
11. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Comput. **1**(4), 541–551 (1989)
12. Liu, C., Yuen, J., Torralba, A.: Nonparametric scene parsing: label transfer via dense scene alignment. In: IEEE CVPR, pp. 1972–1979 (2009)
13. Liu, X., Zhao, Y., Zhu, S.C.: Single-view 3D scene reconstruction and parsing by attribute grammar. IEEE TPAMI **PP**(99), 1 (2017)
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2), 91–110 (2004)
15. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-SVMs for object detection and beyond. In: IEEE ICCV, pp. 89–96 (2011)
16. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV **42**(3), 145–175 (2001)
17. Pinheiro, P., Collobert, R.: Recurrent convolutional neural networks for scene labeling. In: ICML, pp. 82–90 (2014)
18. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
19. Shuai, B., Zuo, Z., Wang, G., Wang, B.: Scene parsing with integration of parametric and non-parametric models. IEEE Trans. Image Process. **25**(5), 2379–2391 (2016)
20. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE CVPR, pp. 1–9 (2015)
21. Tighe, J., Lazebnik, S.: SuperParsing: scalable nonparametric image parsing with superpixels. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 352–365. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15555-0_26
22. Tighe, J., Lazebnik, S.: Finding things: image parsing with regions and per-exemplar detectors. In: IEEE CVPR, pp. 3001–3008 (2013)
23. Yang, J., Price, B., Cohen, S., Yang, M.H.: Context driven scene parsing with attention to rare classes. In: IEEE CVPR, pp. 3294–3301 (2014)
24. Zhang, H., Fang, T., Chen, X., Zhao, Q., Quan, L.: Partial similarity based nonparametric scene parsing in certain environment. In: IEEE CVPR, pp. 2241–2248 (2011)
25. Zhao, H., Puig, X., Zhou, B., Fidler, S., Torralba, A.: Open vocabulary scene parsing. arXiv preprint arXiv:1703.08769 (2017)