Optik 156 (2018) 791–796

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Original research article

Semantic Constraint Based Target Object Recognition

Hao Wu[a], Rongfang Bie[a], Junqi Guo[a], Xin Meng[b], Shenling Wang[a,*]

[a] College of Information Science and Technology, Beijing Normal University, China
[b] Electric Power Planning & Engineering Institute, China

ARTICLE INFO

Article history:
Received 1 November 2017
Accepted 13 December 2017

Keywords:
Semantic constraint
Object recognition
Wordnet subtree
Candidate learning instance joint entropy

ABSTRACT

With the growth of deep learning, object recognition has received increasing interests and its accuracy has been improved significantly in the past few years, However, high-quality recognition largely depends on a large number of learning instances. If the number of learning instances is reduced, it's difficult to maintain realistic recognition accuracy. Moreover, traditional methods usually don't consider the semantic relationship between different regions. Actually, semantic constraint would contribute to improve the recognition accuracy effectively.

Aiming at the problems above, we proposed one semantic constraint based object recognition method. On the one hand, instance-based transfer learning model could make use of learning instances of other categories to maintain realistic recognition accuracy. On the other hand, semantic constraint between different regions simulated as joint entropy is used to recognize target object more accurately. At last, adequate experiments using a large number of images show that our model not only could reduce the number of learning instances but also could achieve realistic recognition.

1. Introduction

Object recognition [1,2] is used to recognize specific object in the image. In the last few years, this technology in the field of computer vision contributed to find and identify objects in an image or video sequence. In the traditional recognition process, some classic feature descriptors, such as SIFT [3], GIST [4] and HOG [5], could extract the features of images effectively. Based on them, many optimized feature descriptors [6–11] could extract the features more fast or adequately. For a long time, SVM combined with optimized feature descriptors are used to estimate the specific category of each object. Moreover, above-mentioned models are also effective for mage classification [12,13], image retrieval [14], and image annotation [15,16]. Within a long duration, recognition methods are on the foundation of improvement for feature extraction and classification. In recent years, with the development of hardware, deep-seated relationship between different pixels could be extracted more adequately, the overwhelming experimental results also verify the performance of deep learning. The essence of deep learning is to extract deep level information through complicated structure and parameters using a large number of learning instances. From all of them, CNN-based methods [17–19], RBM-based methods [20–22], Autoencoder-based methods [23–25] and Sparse coding-based Methods [26–28] are often considered as four classic categories.

As discussed above, quite a few methods have been used to maintain or improve the recognition accuracy. Especially for deep learning based models, they nearly have replaced all traditional methods using overwhelming experimental results.

* Corresponding author.
E-mail address: shenlingwangbnu@163.com (S. Wang).

https://doi.org/10.1016/j.ijleo.2017.12.033
0030-4026/

More importantly, further improvements based on deep learning models are still promising. Promising experimental results indeed cover up a lot of shortcomings but shortcomings really exist obviously. Firstly, a large number of learning instances are one non-ignorable burden. If adequate learning instances from website and some other databases are needed, collection process would waste lots of human resource. In some other cases, it's difficult for us to retrieve enough learning instances even if we would like to spend lots of human resource. For instance, even if we collect all red-crowned crane images of websites, they still couldn't fulfill the requirements of deep learning model. Although some previous methods [29–32] have already reduced the number of learning instances through decision model optimization or learning instance replacement, the majority of learning instance reduced models are still too complicated which has become one new burden for human resource and computing resource.

Moreover, the majority of recognition methods ignore the semantic relationship between different regions. Actually, the sematic constraint between different regions could contribute to recognize the target object effectively. For instance, we can't ensure whether the target object is building, but the possibility of building existence is increased through recognizing its neighbor region as street. Even if some methods have taken the semantic relationship between different regions into consideration, the models are still not suitable for complicated images. If they are applied to complicated images concluding more regions, the effectiveness is reduced obviously.

Aiming at the problems above, we combined instance-based transfer learning model and semantic constraint model to achieve realistic recognition using relatively few learning instances. The main contributions of this paper are in the following:

(1) Instance-based transfer learning model is used to reduce the number of learning instances through thought of learning instance substitution.
(2) Semantic constraint based model is used to improve the object recognition accuracy through semantic relationship constraint between different regions simulated as joint entropy.

2. Algorithm

In this process, we mainly presented how to achieve high-quality recognition. Compared to traditional recognition methods, instance-based transfer learning and semantic constraint are essential contributions, then we would introduce them in details.

2.1. Instance-based transfer learning

As is known to us, the quality of deep learning model largely depends on a large number of learning instances. If the number of learning instances is reduced, it's difficult to construct one CNN model, not to mention maintain realistic accuracy. Aiming at this problem, we drew on the experience of instance-based transfer learning idea to reduce the number of learning instances. Based on this idea, the selection of candidate learning instance has become one pending problem. Although some complicated models have been used to select the candidate learning instances effectively, excessive consumption of resources are special burden for us. So in this paper, a relatively simple model of Wordnet subtree [33] is used as reference to select the suitable candidate learning instances.

The similarity of two categories is defined by the number of nodes shared by their parent branches, divided by the length of the longer of the two branches (Fig. 1). The similarity could be defined by:

$$S_{ij} = \text{intersect}(\text{par}(i), \text{par}(j))/\max(\text{length}(\text{par}(i)), \text{length}(\text{par}(j))), \quad (1)$$

For instance, the similarity between "tabby cat" and "felis domesticus" is 0.93, while the similarity between "tractor trailer" and "felis domesticus" is 0.21. We could judge whether the unknown learning instances are candidate learning instances through Wordnet subtree based similarity. In most cases, if the value of Wordnet subtree based similarity is over 0.5, we consider it as a candidate learning instance. Based on the theory of reinforcement learning, we give each candidate learning instance as different enhancement weights referenced by the value of Wordnet subtree based similarity. More concretely, if the candidate learning instances are more semantically similar to target object, we give them more weights. Otherwise, they are given as less weights.

In the process of recognition model construction, similar to structure and parameters in the paper [34], we extract a 4096-dimensional feature vector from the target region using Caffe [35] model. The model is introduced by Krizhevsky and features are computed by forward propagating a mean-subtracted 227*227 RGB image through five convolutional layers and two fully connected layers. The network architecture details could be learnt from papers [35,36].

2.2. Semantic constraint

As discussed above, traditional methods usually recognize the target object without considering the semantic relationship between different regions. Actually, the semantic relationship would contribute significantly to improve the recognition

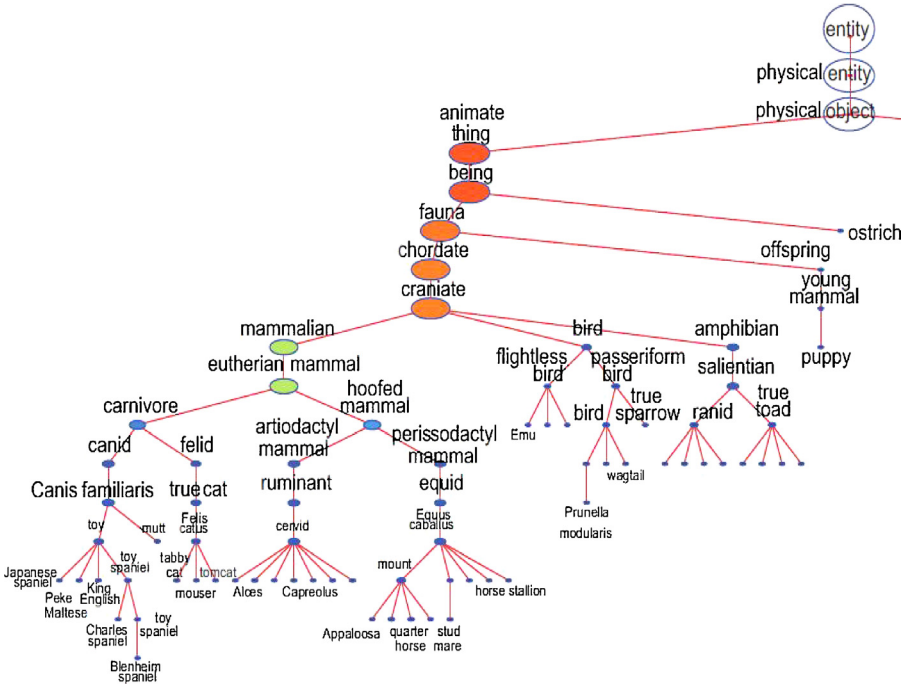

Fig. 1. One portion of the Wordnet Subtree.

accuracy. In this paper, we used joint entropy to simulate the semantic constraint between each region. The joint entropy of one image could be expressed using the following probabilities.

$$H(I) = \sum_{(n_1, n_2 \ldots) \in N} p(n_1, n_2 \ldots n_u \ldots n_n | I) \log(p(n_1, n_2 \ldots n_u \ldots n_n, I)) \tag{2}$$

Where $p(n_1, n_2 \ldots n_u \ldots n_n I)$ is the probabilities of all possible class label assignments to different regions in one image while taking into consideration the contextual semantic relations between them. Image I can be segmented into different regions $(R_{1\ldots}R_{u\ldots}R_n)$ such that each region is connected with a noun node and $(n_1, n_2 \ldots) \in$ N represents the noun connected with regions $(R_{1\ldots}R_{u\ldots}R_n)$. In this paper, if we recognize some portion of the whole regions successfully, we could make use of these convinced regions to estimate the unknown target object though joint entropy. For instance, we can't ensure whether the target object is lion, if its neighbor region is mountain or forest, it's more likely to be lion. If its neighbor region is street or ocean, it's almost impossible to be a lion. So after traditional recognition process, joint entropy based joint entropy could be used to check the correctness of recognition. In the process of checking, some uncertain results will be determined objectively.

Except for making full use of classic CNN model, instance-based transfer learning model and semantic constraint model would contribute to achieve realistic recognition. In the following step, further experiments would be used for checking our model's validness and robustness.

3. Experiments

To our knowledge, there are quite a few databases concluding a large number of images which are adequate for our experiments. However, considering the special target of this paper, we need to collect more complicated images concluding more regions to set up a new database. Based on this target, we select 143,134 images from google, yahoo and some other websites. In most cases, the majority of images conclude over 3 regions. Special notes: in this paper, the location of object is not the main consideration

Firstly, we used instance-based transfer learning model to achieve recognition. Compared to other traditional methods, in Fig. 2, we could see that the recognition accuracy could be maintained in a realistic level even if there are no enough learning instances. Then Fig. 3 shows that semantic constraint improves the recognition accuracy obviously. Under the same experimental environment, there is obvious improvement between using semantic constraint or not. In order to verify our model's generalization ability, GoogLeNet [37],VGG [38]. SPP [39] and AlexNet [40] are used as baselines. Figs. 4 and 5 show that our model contributes to improve other deep learning models' power effectively evaluated by AP value and AUC value respectively. At last, Figs. 6–9 show some groups of recognition results using our model.

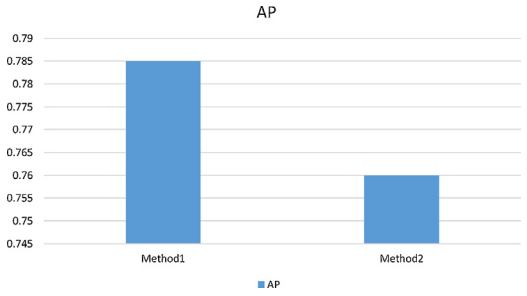

Fig. 2. The mean average precision(mAP) by different methods. Method1: Common method [34]. Method2: Instance-based transfer learning model.

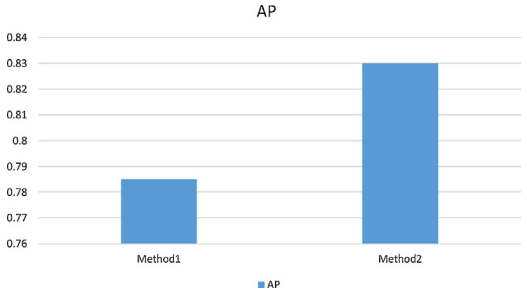

Fig. 3. The mean average precision(mAP) by different methods. Method1: Common method [34]. Method2: Semantic constraint based model.

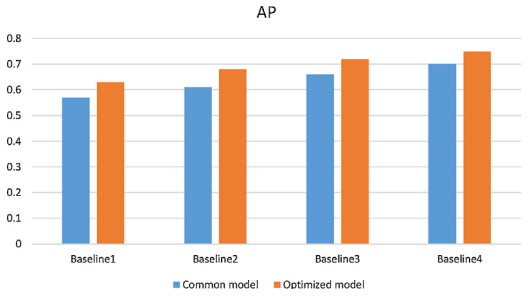

Fig. 4. The mean average precision(mAP) of recognition by different methods. The orange histograms show the mAP of different methods combined with our model. The blue histograms show the mAP of different methods. Baseline 1: GoogLeNet [37].Baseline 2:VGG [38]. Baseline 3:SPP [39]. Baseline 4:AlexNet [40].

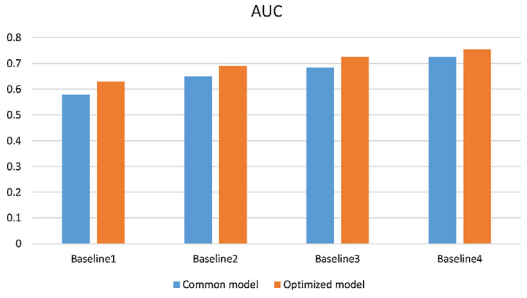

Fig. 5. Receiver Operating Characteristic (AUC) of recognition by different methods. The orange histograms show the AUC of different methods combined with our model. The blue histograms show the AUC of different methods. Baseline 1: GoogLeNet [37]. Baseline 2:VGG [38]. Baseline 3:SPP [39]. Baseline 4:AlexNet [40].

From the experimental results above, we could see that the experimental results coincide well with the theoretical predictions. On the one hand, recognition accuracy could be maintained in a realistic level by instance-based transfer learning model. On the other hand, semantic constraint is one effective model to improve the recognition accuracy. More importantly, our model could be applied to enhance the power of traditional deep learning models effectively.

Fig. 6. Recognition results by our model. (swallow).

Fig. 7. Recognition results by our model. (Squirrel).

Fig. 8. Recognition results by our model. (Desert).

Fig. 9. Recognition results by our model. (Windmill).

4. Conclusion

In this paper, we proposed one semantic constraint based target object recognition method. In the process, instance-based transfer learning model could contribute to maintain the recognition accuracy using some candidate learning instances. More importantly, semantic constraint is considered as an important reference which could improve the recognition accuracy significantly. After adequate experiments on a large number of images, we could see that our model outperforms the traditional methods obviously.

In spite of that, there is still a lot of work to be done. For the majority of images, semantic constraint could contribute to recognize the target objects through semantic relationship between different regions. However, for some special images, the semantic constraint is not effective because the constraint effectiveness largely depends on the quality of segmentation. Over-segmentation and under-segmentation would have negative influence on the final recognition results. Moreover, it's just a relatively simple and effective method to measure the semantic similarity using a Wordnet subtree. More further work still needs to be done in the future.

Acknowledgments

This research is sponsored by National Natural Science Foundation of China (No. 61571049, 61371185, No. 61601033, No. 61401029), Fundamental Research Funds for the Central Universities (No. 2016NT14), China Postdoctoral Science Foundation Funded Project (No. 2016M591109) and Beijing Advanced Innovation Center for Future Education (BJAICFE2016IR-004).

References

[1] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Comput. Vis. Pattern Recognit. (2014), arXiv preprint arXiv:1409.1556.
[2] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2016) 770–778.
[3] D.G. Lowe, Distinctive image features from scale-invariant keypoints, J. Comput. Vis. 60 (2) (2004) 91–110.
[4] Aude Oliva, Antonio Torralba, Building the gist of a scene: the role of global image features in recognition, Prog. Brain Res. 155 (2006) 23–36.
[5] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, ACM, New York, NY, Pro. ACM International Conference on Image and Video Retrieval (2007) 672–679.
[6] Yan-Tao Zheng, et al., Toward a higher-level visual representation for object-based image retrieval, Vis. Comput. 25 (1) (2009) 13–23.
[7] E. Bart, S. Ullman, Cross-generalization: learning novel classes from a single example by feature replacement, IEEE, San Diego, CA, Pro. IEEE Conference on Computer Vision and Pattern Recognition (2005) 672–679.
[8] A. Torralba, K.P. Murphy, Sharing visual features for multiclass and multiview object detection, IEEE Trans. Pattern Anal. Mach. Intell. 29 (5) (2007) 854–869.
[9] H. Wu, Y. Li, Z. Miao, et al., Creative and high-quality image composition based on a new criterion[J], J. Visual Commun. Image Represent. 38 (2016) 100–114.
[10] H. Wu, Y. Li, Z. Miao, et al., A new sampling algorithm for high-quality image matting[J], J. Visual Commun. Image Represent. 38 (2016) 573–581.
[11] Kai Kunze, et al., The wordometer–estimating the number of words read using document image retrieval and mobile eye tracking, in: Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE, 2013.
[12] S. Maji, A.C. Berg, Max-margin additive classifiers for detection, IEEE, Kyoto, Proc. IEEE International Conference on Computer Vision (2009) 40–47.
[13] N. Kumar, et al., Attribute and simile classifiers for face verification, IEEE, Kyoto, Proc. IEEE International Conference on Computer Vision (2009) 365–372.
[14] Z. Zha, et al., Joint multi-label multi-instance learning for image classification, Anchorage, AK, Proc. IEEE International Conference on Computer Vision (2008) 1–8.
[15] B.C. Russell, et al., LabelMe:a database and web-based tool for image annotation, Int. J. Comput. Vis. 77 (1–3) (2008) 157–173.
[16] A. Ulges, et al., Indentifying relevant frames in weakly labeled videos for training concept detectors Niagara Falls, Canada, Proceedings of International Conference on Content-Based Image and Video Retrieval (2008) 9–16.
[17] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1097–1105.
[18] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer International Publishing, 2014, pp. 818–833.
[19] K. He, X. Zhang, S. Ren, et al., Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.
[20] G.E. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[21] R. Salakhutdinov, G.E. Hinton, Deep Boltzmann machines, AISTATS 1 (2009) 3.
[22] J. Ngiam, Z. Chen, P.W. Koh, et al., Learning deep energy models, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 1105–1112.
[23] C. Poultney, S. Chopra, Y.L. Cun, Efficient learning of sparse representations with an energy-based model, Adv. Neural Inf. Process. Syst. (2006) 1137–1144.
[24] P. Vincent, H. Larochelle, Y. Bengio, et al., Extracting and composing robust features with denoising autoencoders, ACM, Proceedings of the 25th International Conference on Machine Learning (2008) 1096–1103.
[25] S. Rifai, P. Vincent, X. Muller, et al., Contractive auto-encoders: explicit invariance during feature extraction, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11)., 2011, pp. 833–840.
[26] R. Memisevic, U CA, D. Krueger, Zero-bias autoencoders and the benefits of co-adapting features, Stat (2014), 1050, 13.
[27] X. Zhou, K. Yu, T. Zhang, et al., Image Classification Using Super-Vector Coding of Local Image Descriptors European Conference on Computer Vision, Springer, Berlin, Heidelberg, 2010, pp. 141–154.
[28] S. Gao, I.W.H. Tsang, L.T. Chia, et al., Local features are not lonely–laplacian sparse coding for image classification, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 3555–3561.
[29] G. Wang, D. Forsyth, D. Hoiem, Comparative object similarity for improved recognition with few or no examples, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 3525–3532.
[30] H. Wu, Z. Miao, Y. Wang, et al., Optimized recognition with few instances based on semantic distance, Vis. Comput. 31 (4) (2015) 367–375.
[31] H. Wu, Z. Miao, Y. Wang, et al., Image completion with multi-image based on entropy reduction[J], Neurocomputing 159 (2015) 157–171.
[32] H. Wu, Z. Miao, J. Chen, et al., Recognition improvement through the optimisation of learning instances, IET Computer Vis. 9 (3) (2015) 419–427.
[33] Rob Fergus, et al., Semantic label sharing for learning with many categories, in: Computer Vision–ECCV 2010, Springer, Berlin, Heidelberg, 2010, pp. 762–775.
[34] S. Shankar, D. Robertson, Y. Ioannou, et al., Refining architectures of deep convolutional neural networks, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2016) 2212–2220.
[35] Y. Jia, Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding, 2013 http://caffe.berkeleyvision.org/.
[36] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, in: NIPS, 2012.
[37] C. Szegedy, W. Liu, Y. Jia, et al., Going deeper with convolutions, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2015) 1–9.
[38] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition arXiv preprint arXiv:1409.1556, 2014.
[39] K. He, X. Zhang, S. Ren, et al., Spatial pyramid pooling in deep convolutional networks for visual recognition, in: European Conference on Computer Vision, Springer International Publishing, 2014, pp. 346–361.
[40] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1097–1105.