Scholar Publishing Group
Kinetic Mechanical Engineering
https://doi.org/10.38007/KME.2021.020106
ISSN 2790-2676 Vol. 2, Issue 1: 46-54

Visual Intelligent Recognition System based on Visual Thinking

Umma Kavita

Faculty of Engineering, Department of Renewable Energy Engineering, Amman Arab University, Amman 11953, Jordan

Keywords: AI Technology, Deep Learning, Construction Machinery, Image Segmentation Technology

Abstract: Image semantic segmentation plays an important role and has application value in robot arm object capture, automatic driving, medical image analysis, geographic information system, etc. Aiming at the semantic segmentation method of deep learning(DL), this paper makes some attempts in the direction of weak supervised semantic segmentation, proposes AI and DL technology, and applies them to the Image segmentation technology(IST) in the construction machinery(CM) grasping task for analysis and exploration. The improvement of image weak location based on depth learning, network segmentation graph and image classification objective function are briefly analyzed. Finally, the experimental test analysis shows that a better initial attention map is not only helpful for direct segmentation, but also can promote the joint optimization performance, which verifies the effectiveness and feasibility of IST in the CM grasping task relying on AI and depth learning in this paper.

1. Introduction

With the development of e-commerce and logistics, the demand for packaging and transportation of goods has increased significantly. Highly automated cargo transportation is very important for large-scale commodity sales. Using robots to sort goods independently rather than manually may greatly improve the efficiency of the warehousing system and shorten the delivery time of goods. In this paper, the problem of image semantic segmentation in CM grasping task is studied, and AI and DL technology are proposed. Image semantic segmentation is an important part of image understanding in computer vision, and plays a key role in automatic driving, geographic information system, medical image analysis and other practical applications.

Many scholars at home and abroad have studied and analyzed the IST in engineering machinery grasping tasks that rely on AI and DL. Rusyn B P is committed to developing segmentation

Copyright: © 2021 by the authors. This is an Open Access article distributed under the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (https://creativecommons.org/licenses/by/4.0/).

methods for atmospheric cloud images obtained through remote sensing using aircraft or satellite airborne systems. To some extent, this method is a further improvement of the U-shaped network convolution neural network [1]. Using the known segmentation quality standards, the proposed method is compared with the known methods in the field of atmospheric cloud image segmentation. A large number of experiments on real images show that it is feasible to use the proposed segmentation method for automatic processing according to the requirements of real-time operation [2].

This paper relies on AI and DL technology to research and analyze the IST in the CM grasping task. Firstly, we introduce IST, weak supervised image segmentation based on depth learning, and how to achieve segmentation through classification with weakened image category labels. In this process, it introduces how AI technology and DL technology can combine the classification and segmentation of CM grasping tasks to obtain the discriminant regions of images at the same time. Next, we introduce the latest image segmentation models inspired by image pyramids, which are respectively spatial pyramid pooling and empty space pyramid pooling. PSPNet and DeepLab, which are based on these two pyramid models respectively, are the benchmarks of recent academic semantic segmentation tasks, and their ideas are worth learning from [3-4].

2. Research on Image Segmentation of CM Grabbing Task Based on AI and DL

Because VGGnet has a simple structure and is easy to modify, we conduct experiments on the basis of VGGnet GAP modified by this method. First, we introduce the aperture convolution used in the strongly supervised image segmentation network into the network, which increases the receptive field of the convolution core, making the performance of the modified VGGnet GAP equal to that of the original GooLeNet GAP [5-6]. Then the Dropout layer is removed from the network, and it is found that this has significantly improved the weak supervision positioning performance. In addition, the GoogLeNet with the best performance in the basic network structure in this article does not include the Dropout layer, as shown in Table 1.

Table 1. Classification and localization error on the ILSVRC validation set (%)

	top-k	Classification positioning error rate	Classification error rate
VGGnet-GAP	1	57.20	33.4
	5	45.14	12.2
GoogLeNet-GAP	1	56.40	35.0
	5	43.00	13.2
VGGnet-GAP+AtrousConv	1	55.85	33.2
	5	44.15	12.1
VGGnet-GAP+AtrousConv - Dropout	1	52.85	30.7
	5	41.80	11.2

In this paper, the research and analysis of IST in CM grasping tasks that rely on AI and DL are conducted under the PASCALVOC 2012 dataset. PASCALVOC is one of the most popular computer vision data sets, and is still regarded as the authoritative verification data set of models in the academic community. The image segmentation task includes 20 interested target categories, such as people, trains, cats and chairs, and the other one is background. The comparison of Pascal VOC data sets is shown in Figure 1 of Table 2 [7].

Table 2. Pascal VOC dataset comparison table

method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow
Color+Depth	91.1	78.6	33.6	79.6	57.3	66.1	87.1	77.7	83.1	32.0	69.1
Color+Gray	90.8	76.6	31.4	78.5	55.3	64.2	86.0	77.4	82.6	30.4	67.4
method	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU
Color+Depth	52.3	75.7	71.2	69.2	76.7	44.5	70.2	42.7	79.7	60.7	66.6
Color+Gray	51.5	75.7	71.0	68.7	76.1	44.6	69.7	40.3	79.2	59.0	65.5

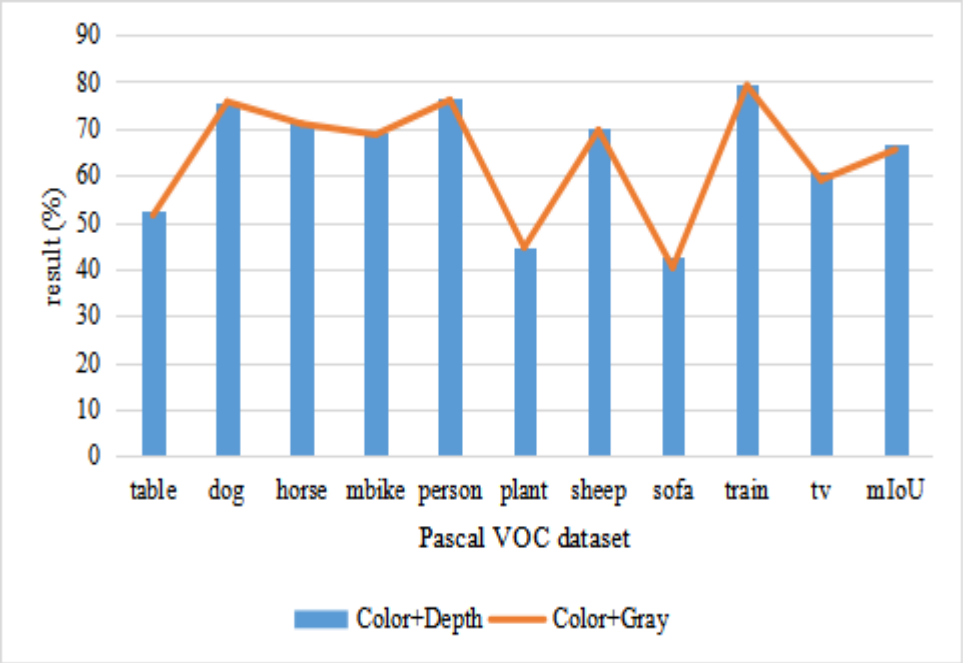

Figure 1. Pascal VOC data chart

The current RGB-D data sets are mainly collected by the above sensors. Although the depth neural network can train and predict the depth on these data sets, it is difficult to apply them to many high-level visual tasks due to the limited scene. Although humans can estimate the depth of the environment, they are not good at estimating the measured value of the depth, but are good at judging the relative depth order, such as judging which of the two points in the image is closer. The end-to-end training from color image to depth image can be realized by using a full convolution network. The key problem is to design a loss function for relative depth annotation [8-9]. This loss function follows the principle that for two points in the image, when the true depth order is "equal", the smaller the difference between the predicted depth values is, the better; otherwise, the larger the difference is, the better.

Through the above experimental performance analysis, it is proved that MSPP improves the performance of weakly supervised image segmentation significantly. This paper continues to analyze the segmentation results, and MSPP improves the effect of weak supervised image segmentation to a certain extent, such as small target recognition, image relationship correction, remote target positioning and edge detail optimization [10-11]. However, due to the complexity of the image scene, there are still some directions to be improved. The following shows the specific results and makes corresponding analysis on the above four advantages and unresolved problems of MSPP [12].

3. IST Dependent on AI and DL

3.1. Network Partition Diagram

The classification segmentation network outputs the attention map, and the saliency detection network outputs the saliency map. The combination of the attention map and the saliency map can generate the network segmentation map, and the directly generated network segmentation map can be used as the final image segmentation result [13]. Here we describe how to get the network segmentation map by combining the attention map and the saliency map. Suppose that the attention map containing all categories is represented by tensor A, and the saliency map is represented by C, where the attention map is standardized:

$$A = \frac{A - \min(A)}{\max(A)} \tag{1}$$

In this way, the attention map value range is also 0 to 1, which can be compared with the significance map. Secondly, drawing on the hierarchical optimization structure of DHSNet, and considering that there may be multiple objects of different sizes in an image, the saliency map is also layered. Specifically, a pair of images is obtained through the first saliency detection network to obtain the first saliency map, and then the part of the first saliency map whose probability is greater than 0.7 is erased again into the saliency detection network to obtain the second saliency map, and finally the part of the second saliency map whose probability is greater than 0.8 is erased again into the saliency detection network to obtain the third saliency map [14-15]. The maximum value of the cubic significance map is taken at each position to get the final significance map. In this paper, C directly represents the final significance graph.

Then the joint information graph K is obtained by the harmonic average of the attention graph and the saliency graph:

$$K = \frac{2}{\frac{1}{A} + \frac{1}{C}} \tag{2}$$

After the reconciled joint information map is segmented, the part with threshold value less than 0.4 is marked:

$$B = I(K < 0.4) \tag{3}$$

The labeled background map is added to the joint information map to represent the background dimension, and other category maps that do not appear in the image are embedded with all 0 values according to the image category level to obtain a 21 dimensional information map. The 20 dimension represents the interest category in PASCALVOC, and the other one represents the background. The network segmentation map can be obtained by taking argmax from the information map pixel by pixel.

The process of generating a network segmentation map from the attention map obtained by the classification segmentation network and the significance map obtained by the significance detection network. The feature map does not display target information, and the color of the segmentation map is converted according to the known category color correspondence for convenience of observation [16].

3.2. Objective Function

According to the previous introduction, the classification and segmentation network needs two stages of training. The first stage is the image classification training with the tag as the image target category, and the second stage is the image segmentation training with the tag as the network segmentation map generated by the attention map and saliency map.

3.2.1. Image Classification Objective Function

This paper shows that T2 regularization is applied to the objective function, so the objective function in the image classification stage is:

$$T = t_{cls} + \eta \| \gamma \|_2 \tag{4}$$

among γ Represents network parameters, | |. | | 2 represents 2 norms, η Represents the regular strength.

The image classification stage is carried out in the way of supervised learning, but the tags used only provide information about whether the image contains the corresponding target, which is quite different from the pixel level category information, so the final image segmentation method belongs to weak supervised learning.

3.2.2. Classification and Segmentation Joint Objective Function

The network segmentation graph generated by weak information guides the network learning at the segmentation feature graph. The loss function of this part is the segmentation loss function, which is represented by tseg. Assuming that the network segmentation feature map is the probability tensor of the network segmentation feature map after Softmax conversion, then tseg is equivalent to the average value of multi class cross entropy loss at all locations:

$$t_{seg} = \sum_{x=1}^{\gamma} \sum_{y=1}^{h} \sum_{i=1}^{s+1} Seg_{x,y,i} \cdot \log(P_{x,y,i}) / (h\gamma) \tag{5}$$

Among them, Seg is coded by one pot in the category dimension, with the highest probability of 1 and others of 0.

The classification and segmentation stage is also conducted in the way of supervised learning. However, since the supervision information of the classification network and the significance network is still weakened compared with the true segmentation graph, the whole process still belongs to weak supervised learning [17-18].

4. Research on IST in Engineering Machinery Grabbing Task Relying on AI and DL

4.1. Experimental Setup

First, introduce the super parameter settings of MSPP, mainly covering the proportion β Area signal masking coefficient ρ In the actual experiment, β and ρ Set to 0.125 and 0 respectively. Because the number of pooled output regions is different, the final number of pooled masks at each level is obtained by multiplying the total number of regions by the mask ratio and then taking an integer upward. The purpose of signal masking coefficient of 0 is to close the most discriminant

region and mine the secondary discriminant region. No more experiments have been conducted for other values, but it does not hinder the corresponding expansion. In addition, the competition logic function in this experiment is always max.

The second is the selection of the regularization strength of the objective function β Take 0.0005 for all and do not try other options. Then the learning optimization algorithm is introduced. The optimization methods in this paper are all based on random gradient descent, in which the learning rate is gradually declining.

In the image classification stage, the initial learning rate is set to 0.001, the total learning step size is 30000 steps, the batch size of random gradient decline is 7, and the network has been pre trained in ImageNet. In the joint optimization stage of classification and segmentation, the initial learning rate is still 0.001, the total learning step size is 10000 steps, and the batch size is 7. The classification and segmentation learning is carried out immediately after completing the image classification task under the same network. In addition, the significance map is output by DHSNet after training, and no network optimization is conducted. It is worth noting that even compared with the original model with a batchsize of 10, the model in this paper has achieved some improvement in both the verification set and the test set due to the introduction of MSPP.

4.2. Image Segmentation Performance

Table 3 summarizes and compares the weakly supervised image segmentation methods, where * represents the results obtained by reproducing with a batch size of 7. It can be seen that MSPP achieves 61.1% and 62.8% mIoU in the validation set and test set respectively, which is closer to the new state of the art. Another fact is that compared with the original DCSP, the performance of the DCSP reproduced in this paper decreases by 1% on average due to the reduction of the batch size from 10 to 7. The mIoU in the verification set and the test set are 59.8% and 60.9% respectively. Only replace global pooling with MSPP to increase the same setup model by 1.3% and 1.9% in the verification set and test set, and increase the state of the art by 0.3% and 0.9%, respectively. MSPP is expected to perform better if it is not affected by the batchsize. The performance of MSPP shows that it is feasible to consider weakly supervised image segmentation from the perspective of secondary discriminant region feature mining, which also has great inspiration for other weakly supervised tasks. It should be noted that DCSP has a greater advantage over AE-PSL in that the quality of the saliency map obtained through DHSNet is much higher. This paper continues to use the saliency map in DCSP.

Table 3. Comparison of weakly supervised semantic segmentation methods

Methods	mIOU(val)	mIoU(test)
MIL-FCN(ICLR2015)	25.7	24.9
CCNN(ICCV2015)	35.3	35.6
EM-Adapt(ICCV2015)	38.2	39.6
DCSM(ECCV2016)	44.1	45.1
STC(PAMI2016)	49.8	51.2
SEC(ECCV2016)	50.7	51.7
AE-PSL(CVPR2017)	55.0	55.7
DCSP(BMVC2017)	60.8	61.9
DCSP*	59.8	60.9
MSPP(ours)	61.1	62.8

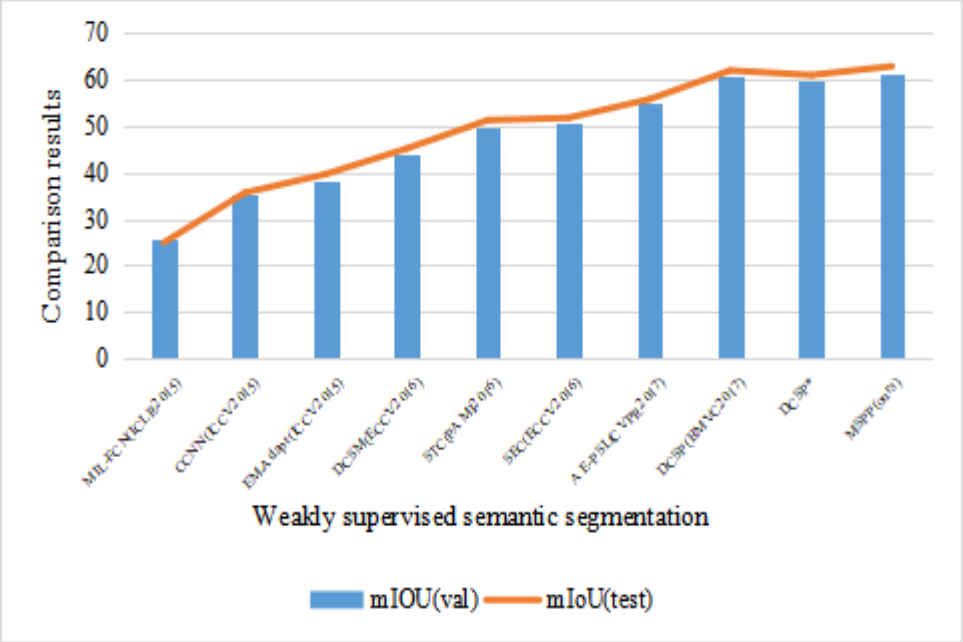

Figure 2. Data graph of weakly supervised semantic segmentation methods

From the data analysis in Figure 2 above, it can be seen that for the performance of outputting segmentation results directly with attention map and saliency map without classification and segmentation joint training; The classification network and significance detection network are consistent respectively. It can be seen that MSPP directly improves the results by 0.8% through the attention map. It shows that a better initial attention map not only contributes to direct segmentation, but also promotes joint tuning performance.

5. Conclusion

Aiming at AI technology and DL image segmentation method, this paper has made some attempts in the direction of weak supervised image segmentation, but has not achieved good results. It is very difficult to segment images because of the lack of location supervision information in image level annotation. The next improvement direction for this is cross category semi supervised segmentation, that is, using some categories containing pixel level labels and some categories containing only image level labels to train, so that the categories containing only image level labels are as close to the segmentation performance of pixel level labels as possible. For the strongly supervised image segmentation direction, the depth image predicted from the color image is used to fuse it into the original color image segmentation network. However, the method of predicting the depth image from the color image requires additional processing, and there is still room for improvement in the network structure and output resolution. The next step is to use a single network model to predict the depth image and segment the image, and optimize the loss function and network structure of the image segmentation model according to the characteristics of the depth image.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] *Rusyn B P , Lutsyk O A , Kosarevych R Y , et al. Image segmentation of clouds based on deep learning. Information Extraction and Processing, 2020, 2020(48):72-78. https://doi.org/10.15407/vidbir2020.48.072*

[2] *Li M X , Yu S Q , Zhang W , et al. Segmentation of retinal fluid based on deep learning: application of three-dimensional fully convolutional neural networks in optical coherence tomography images. International Journal of Ophthalmology, 2019, 12(6):1012-1020. https://doi.org/10.18240/ijo.2019.06.22*

[3] *Utomo T W , Cahyadi A I , Ardiyanto I . Suction-based Grasp Point Estimation in Cluttered Environment for Robotic Manipulator Using Deep Learning-based Affordance Map. International Journal of Automation and Computing, 2021, 18(2):277-287. https://doi.org/10.1007/s11633-020-1260-1*

[4] *Lee K , Kim J H , Lee H , et al. Boundary-Oriented Binary Building Segmentation Model With Two Scheme Learning for Aerial Images. IEEE Transactions on Geoscience and Remote Sensing, 2021, PP(99):1-17. https://doi.org/10.1109/TGRS.2021.3089623*

[5] *Brutti F , Fantazzini A , Finotello A , et al. Deep Learning to Automatically Segment and Analyze Abdominal Aortic Aneurysm from Computed Tomography Angiography. Cardiovascular Engineering and Technology, 2021, 13(4):535-547. https://doi.org/10.1007/s13239-021-00594-z*

[6] *Park J , Kwon D , Choi B W , et al. Small object segmentation with fully convolutional network based on overlapping domain decomposition. Machine Vision and Applications, 2019, 30(4):1-10. https://doi.org/10.1007/s00138-019-01023-x*

[7] *Imagawa H , Iwata M , Kise K . Digital Watermarking Method for Printed Matters Using Deep Learning for Detecting Watermarked Areas. IEICE Transactions on Information and Systems, 2021, E104.D(1):34-42. https://doi.org/10.1587/transinf.2020MUP0004*

[8] *Tsai C Y , Chou Y S , Wong C C , et al. Visually Guided Picking Control of an Omnidirectional Mobile Manipulator Based on End-to-End Multi-Task Imitation Learning. IEEE Access, 2019, PP(99):1-1.*

[9] *Li J , Yin J , Deng L . A robot vision navigation method using deep learning in edge computing environment. EURASIP Journal on Advances in Signal Processing, 2021, 2021(1):1-20. https://doi.org/10.1186/s13634-021-00734-6*

[10] *Hariyani Y S , Eom H , Park C . DA-CapNet: Dual Attention Deep Learning based on U-Net for Nailfold Capillary Segmentation. IEEE Access, 2020, PP(99):1-1. https://doi.org/10.1109/ACCESS.2020.2965651*

[11] *Utomo T W , Cahyadi A I , Ardiyanto I . Suction-based Grasp Point Estimation in Cluttered Environment for Robotic Manipulator Using Deep Learning-based Affordance Map. International Journal of Automation and Computing, 2021, 18(2):277-287.*

https://doi.org/10.1007/s11633-020-1260-1

[12] Tavousi P , Shahbazi Z , Shahbazmohamadi S . AI-based Brain Image Segmentation Using Synthesized Data. Microscopy and Microanalysis, 2020, 26(S2):694-697. https://doi.org/10.1017/S143192762001555X

[13] Park I , Kim K , Choi G , et al. AI-based Automatic Spine CT Image Segmentation and Haptic Rendering for Spinal Needle Insertion Simulator. The Journal of Korea Robotics Society, 2020, 15(4):316-322. https://doi.org/10.7746/jkros.2020.15.4.316

[14] Gautam S , Singhai J . Cosine-similarity watershed algorithm for water-body segmentation applying deep neural network classifier. Environmental Earth Sciences, 2021, 81(9):1-16. https://doi.org/10.1007/s12665-022-10376-y

[15] Ahamad M , Maizul E . Digital Analysis of Geo-Referenced Concrete Scanning Electron Microscope (SEM) Images. Civil And Environmental Engineering Reports, 2020, 30(2):65-79. https://doi.org/10.2478/ceer-2020-0020

[16] Chetty G , Yamin M , White M . A low resource 3D U-Net based deep learning model for medical image analysis. International Journal of Information Technology, 2021, 14(1):95-103. https://doi.org/10.1007/s41870-021-00850-4

[17] Saravana K , Sivaprakasam S A , Rengasari N E , et al. Fast K-Means technique for hyper-spectral image segmentation by multiband reduction. Pollack Periodica, 2019, 14(3):201-212. https://doi.org/10.1556/606.2019.14.3.19

[18] Ghosh S , Pal A , Jaiswal S , et al. SegFast-V2: Semantic image segmentation with less parameters in deep learning for autonomous driving. International journal of machine learning and cybernetics, 2019, 10(11):3145-3154. https://doi.org/10.1007/s13042-019-01005-5