Video Instance Segmentation of Rock Particle Based on MaskTrack R-CNN

Man Chen, Maojun Li, and Yiwei Li

Abstract Video instance segmentation (VIS) of rock particles in motion is the basis for revealing the laws of motion and quantitative analysis. It has important scientific and engineering value. Use an end-to-end network called MaskTrack R-CNN to complete the VIS task for rock particles. The network introduces a new tracking branch on Mask R-CNN. It integrates particle detection, segmentation, and tracking tasks into the framework. The tracking branch primarily uses appearance similarity cues to linearly combine cues such as semantic consistency and spatial correlation to improve tracking accuracy. To facilitate the study of rock particle visibility, we have created a set of experimental equipment for collecting rock particle datasets. We conducted training and testing experiments to verify the effectiveness of the algorithm and compared it to some baselines of our own dataset. Experimental results show that MaskTrack R-CNN uses ResNet-50 to get 33.1% AP. It better than other two-stage models. This work provides an intelligent solution for meso-analyzing particles.

Keywords Video instance segmentation · Rock particle · MaskTrack R-CNN

1 Introduction

Video instance segmentation is a challenging visual task. You need to track the instances across frames and segment objects in individual frames. Many video-based tasks have core applications, such as video editing, autonomous driving, and augmented reality. The pioneering work of VIS is MaskTrack R-CNN [1], which is an extension of Mask R-CNN [2]. In addition to the initial three branches of object classification, bounding box regression, and masking, there is a fourth branch with external storage for tracking object instances frame by frame. First, the use of Region Proposal Network (RPN) [3] of Faster R-CNN to generate a set of candidate recommendations. Then, motion-based RoI features are clipped and inserted at the beginning of each task for bounding box prediction, object masking, and object

M. Chen · M. Li (✉) · Y. Li
Changsha University of Science and Technology, Changsha 410114, China
e-mail: 19205060770@stu.csust.edu.cn

J. Yao et al. (eds.), *The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021)*, Lecture Notes in Electrical Engineering 813,
https://doi.org/10.1007/978-981-16-6963-7_63

tracking. It also recommends a large video dataset called YouTube-VIS to measure video version segmentation algorithms. The new dataset can be used as a useful benchmark for various pixel-level video comprehension tasks.

Particles are an important part of global geographic disasters and are used in construction projects and vehicles [4, 5]. The development of general theory of various materials was one of the 125 cutting-edge science projects. Interpreting the rock particles in motion is the basis for demonstrating the laws of motion and their size parameters, and can provide accurate guidance for construction work. This is also the reason for the insufficient use of verification studies and numerical modeling methods (such as Finite Element Method and Discrete Element Method) in engineering technology. When you generalize these models [6, 7], similar particles can also provide reliable data for the VIS particle mode. In short, the clarity of rock light research is of great value to scientific and technological research.

In this article, we will introduce VIS into digital technology and apply the rock particle video segmentation method on MaskTrack R-CNN. In summary, the main contributions of this work are as follows:

- An end-to-end approach is used to achieve particle visibility by integrating detection, segmentation and tracking operations into the video frames. As far as we know, this is the first VIS application in construction.
- We develop experimental procedures for capturing video and create a dataset with moving objects under external force, which includes 160 videos of rock particles.
- We conducted training and measurement experiments to determine the efficiency of the process and compared it with the different bases of our self-designed data set.

All our papers were designed this way. In Sect. 2 we briefly describe the work of VIS and the main parts that are developed. In Sect. 3 we officially discuss MaskTrack R-CNN algorithms. Section 4 introduces a new set of equipment and experimental results.

2 Related Work

Little research has been done on VIS, especially on the VIS components. However, particle separation has been thoroughly analyzed as the basis for VIS to create some new machine vision algorithms. In this section, we will look at the development.

2.1 Video Instance Segmentation

VIS requires simultaneous classification, distribution and tracking in the video. Depending on how the sequence originates, it can be divided into two types. One type divides tracking and search into two parts [1, 8, 9]. In the form of research,

examples are set up in a frame-by-frame manner using existing image-level example segmentation methods. And the detected positions can be linked to different frames of the tracking component. The second type is abbreviated as 'Clip-Match' methods [10, 11]. Breaks the whole video into several short clips and creates a separate VIS for each by scattering or space-time embedding. It then links the clips to other related clips.

2.2 Particle Image Processing

Particle photo processing involves the classification, detection and segmentation of particles. Extracting particle mask is the basis of VIS, so we summarize it mainly in this single copy. Particle images often have density and adhesion properties which make the distribution process very difficult [12]. It focuses on solving the problem of pollution caused by interactions and shadows in metal images and designs a method of distributing metal images based on holistic nested edge detection [13]. The light-weight U-Net deep training network is designed to automatically detect particles from photographs and obtain potential particle change maps. This method can be used to monitor the particle product quality. Liu et al. [14] introduced a photo sharing method based on U-Net and its improved network. Pre-processed real-time images from open cast mines to reduce noise and capture the object area with applicable traditional vision techniques.

3 MaskTrack R-CNN

The Mask R-CNN is the basis of the MaskTrack R-CNN, so we show the Mask R-CNN first and then explain the new parts in detail.

3.1 Mask R-CNN

Mask R-CNN [2] can distribute pixel level images by combining the advantages of object detection network and semantic distribution network. The RoI Align is used to place the RoI Pooling on the R-CNN, which solves the problem of field misalignment. Backbone, RPN, RoI Align and classification: The general network structure of the Mask R-CNN is shown in Fig. 1, which is actually made up of four components.

The backbone is a series of mixed layers that can map features. It has several constitutional layers. Samples are reliable, consistent, and dynamic at every level and receive maps displayed at different sizes. The feature pyramid network (FPN)

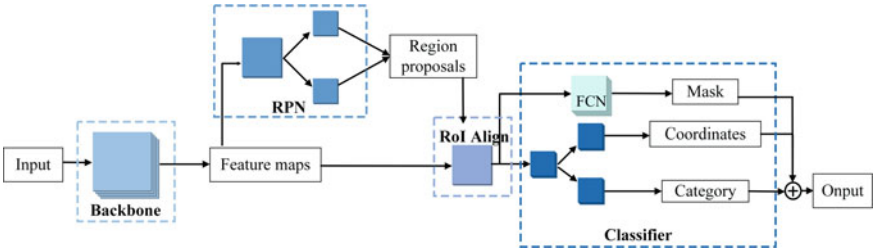

Fig. 1 Mask R-CNN network structure

[15] can be used to obtain multi-layer semantic fuses and post-fusion mapping of different sizes.

RPN is a network that can issue regional proposals for follow-up tasks. Inputs of RPN are the results of the final FPN list. First, it produces a certain number of anchors per pixel on these maps. The probability that these anchors are placed in front or in the background is calculated and the migration is between the combination of these anchors and the corresponding fact on earth. Redistribution and regression can also be tested in terms of RPN loss function. Finally, you will find the appropriate area recommendations and weight parameters after multiple repetitions.

RoI alignment is applied to the instead of approximation in the original RoI section. It can change the location of the search to make a map size without losing location information, so that each pixel remains precisely pointed.

Classifier has three similar branches. Two of these components and layouts can get the more accurate box. One branch uses the Fully Convolution Net (FCN) [16] to predict the mask.

3.2 The Tracking Branch

The network adopts a framework in two stages. In the second step, we added a fourth branch to assign a sample label to each candidate box. As shown in Fig. 2, this branch is parallel to the three branches (the above three branches). Let the number of cases recognized by the model from the previous frame is n. If the new candidate box is one of the previous copies, it will only be entered for n identities. If this is a new instance, it will be given a new identity. This is generally a distribution problem with several classes. There is a $n + 1$ class digits number that already identifies the n instance. A new example is represented by the number 0. The probability of assigning label m to a candidate box i is defined as

$$C_i(n) = \begin{cases} \frac{e^{t_i^T t_m}}{1+\sum_{j=1}^{n} e^{t_i^T t_j}}, & 1 < m < n \\ \frac{1}{1+\sum_{j=1}^{n} e^{t_i^T t_j}}, & m = 0 \end{cases} \tag{1}$$

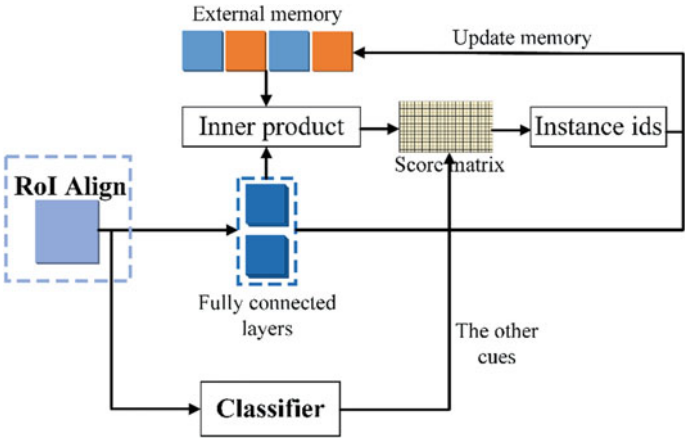

Fig. 2 The tracking branch

where t_i and t_j are the new features extracted by the tracking branch. Our tracking branch has two layers that are fully connected. These layers can project the feature maps drawn by RoI Align into new features. The two fully connected layers transform the input function cards into 1-D 1024 dimensions. Cross entropy loss is used for a tracking branch, which can be expressed as follows:

$$L_{track} = -\sum_i \log(C_i(r_i)) \tag{2}$$

where r_i the ground truth instance label.

It also uses external memory for storage for greater efficiency. External memory is dynamically updated as the instance label is assigned to the new candidate frame. If the selected frame belongs to an existing instance, the instance characteristics stored in memory are updated with the new candidate characteristics. If the candidate is given 0 points, the candidate's characteristics are stored in memory and increase by 1 depending on the number of examples identified.

The tracking branch mainly uses the appearance similarity to accomplish the tracking task. However, there are also other information such as semantic consistency, spatial correlation and detection confidence which could be leveraged to determine the instance labels. MaskTrack R-CNN also combines all these cues together to improve the tracking accuracy in a post-processing way. It completes the matching of instances by calculating the score of assigning label m to the candidate box i and the calculation formula is shown as follows:

$$L = L_{cls} + L_{box} + L_{mask} + L_{track} \tag{3}$$

3.3 The Other Cues

The control branch uses visual equations to perform tasks. However, there are other information that can be used to identify sample tags, such as semantic similarity, positional relationship, and recognition reliability. MaskTrack R-CNN collects all these signals to confirm the accuracy of post-processing. Now complete the random adjustment by calculating the given point of the corner m of candidate i, the calculation formula is as follows:

$$S_i(m) = \log C_i(m) + \alpha \log(d_i) + \beta IoU(b_i, b_m) + \gamma \varphi(c_i, c_m) \tag{4}$$

where i is the sequence number of the candidate box. b_i, c_i and d_i denote the bounding box prediction, category label and detection score. c_m is the bounding box prediction and category label associated with the saved features in the memory. $\varphi(c_i, c_m)$ is a Kronecker delta function which equals 1 when c_i and c_m are same and 0 otherwise. α, β and γ are hyperparameters which can balance the effect of different cues.

4 Experiments

4.1 Experimental Equipment and Dataset

We designed a set of test tools to collect data from the parts needed for the experiment. This mainly includes test bench, motion providing device, sight sensor and object to be measured. The motion providing device moves the objects and gives some scrolling action. Test bench is used to secure the testing process. The visual sensor can capture video of the experimental environment. Figure 3 shows a video of the experimental environment captured by the visual sensor.

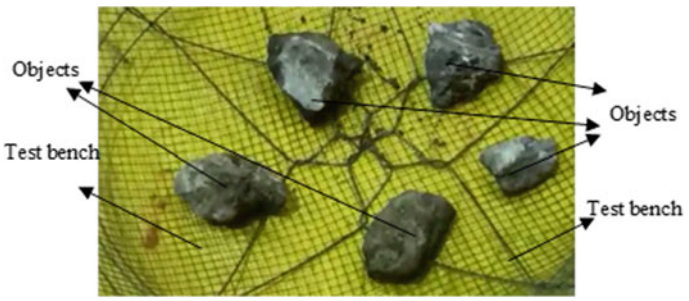

Fig. 3 Experimental equipment

The collected videos are used to support our approach, and are divided into different three categories according to the complexity of the data. General categories 9. Each category has 20 different compressed videos. Like the YouTube VIS design guidelines, some objects are defined by manually tracking the boundaries of every 5 frames, and each video has a rate of 30 frames per second. We also change the original frame sizes to 640 × 360 in training. In particular, we divide the data into training and validation functions according to a certain proportion.

4.2 Implementation Details

Model training and testing experiments were performed on Ubuntu 18.04. The processor is Intel Core i7-8700 K CPU @ 3.7 GHz and the GPU is NVIDIA GeForce RTX 2080Ti. Use the original MaskTrack R-CNN weights as direct weights to increase the convergence speed. The model is ready by the end of the eighth epoch. The primary learning level is 0.05. The hyperparameters α, β and γ are chosen to be 1, 2 and 10 respectively.

4.3 Main Results

Baselines In this experiment, we set up three baselines to compare with MaskTrack R-CNN. They are IoUTrack+ [1], OSMN [17] and DeepSORT [18] respectively. For the frame-by-frame examples generated by the Mask R-CNN, the baseline has the same segmentation effect. Convert the generated video data set into an image. Then create an image data set to train a Mask R-CNN. The structure of the Mask R-CNN is like a network. In addition to the branches of the track.

Quantitative Results MaskTrack R-CNN baseline was compared against a self-derived data set. Table 1 shows the results of the MaskTrack R-CNN comparison, achieved across all metrics. Specifically, the MaskTrack R-CNN consistently surpasses baselines by a significant margin in AP (26.4% vs. 33.1% with IoUTrack+,

Table 1 Quantitative evaluation of the proposed algorithm and baselines

Method	Backbone	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
IoUTrack+ [1]	ResNet-50	26.4	43.8	27.9	22.3	27.8
OSMN [17]	ResNet-50	29.6	48.0	30.3	21.9	27.2
DeepSORT [18]	ResNet-50	28.7	47.1	32.5	22.4	29.8
MaskTrack R-CNN [1]	ResNet-50	33.1	54.3	36.9	25.4	31.7
MaskTrack R-CNN [1]	ResNet-101	**34.8**	**56.7**	**38.2**	**27.5**	**33.9**

The best results are highlighted in bold

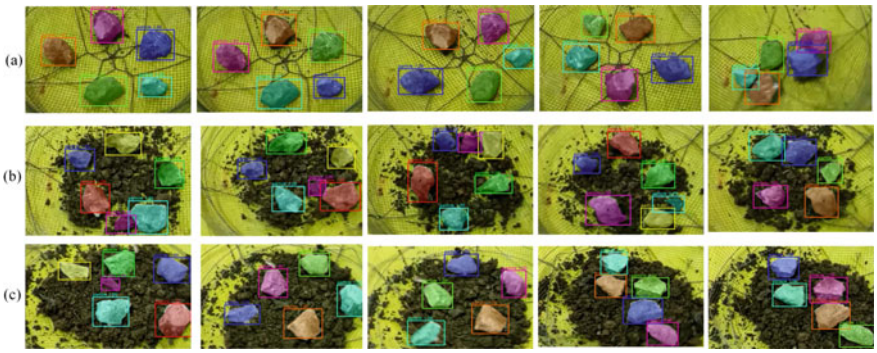

Fig. 4 Sample results of VIS. Each row have six sampled frames from a video sequence

29.6% vs. 33.1% with OSMN and 28.7% versus 33.1% with DeepSORT). In the case of AR, the results of all methods are very low. This may be due to the obvious difference between the background and foreground in the video. These conditions can cause difficulty in segmentation and can also cause particles to disappear. Among them, MaskTrack R-CNN is at least 3.5% higher than its starting lineup. In general, MaskTrack R-CNN can achieve better results than baseline in AR. Moreover, we observe that the ResNet-101 is better than ResNet-50 (33.1% vs. 34.8%).

Qualitative Results Figure 4 shows the qualitative results of both videos. Most rock particles are clear, but some are incomplete. Particles, which are mostly blocked, have more serious errors. The reason for this phenomenon is that objects hidden surrounding impurities make classification, segmentation and tracking difficult. In general, MaskTrack R-CNN can segment the most viewed particles.

5 Conclusion

In this study we used a holistic method to obtain VIS particles. We also set up an experimental kit for collecting video. And create a dataset about particles moving under vibration. It includes 180 videos. We conducted training and testing experiments to prove the effectiveness of MaskTrack R-CNN on VIS particles. Below, we compare the effect of this final model with the baseline data. This is the first VIS application in civil engineering as far as we know in our own dataset. We believe the new application will innovate research ideas and provide a new direction for video awareness to the research community. We believe that this research will provide new ideas for microscopic analysis of particles.

References

1. Linjie, Y., Yuchen, F., Ning, X.: Video instance segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5188–5197 (2019)
2. Kaiming, H., Georgia, G., Piotr, D., Ross G.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
3. Shaoqing, R., Kaiming, H., Ross, G., Jian. S.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99. (2015).
4. Jerónimo, P., Resende, R.: Fortunato, E: An assessment of contact and laser-based scanning of rock particles for railway ballast. Transp. Geotech. 100302 (2020)
5. Gao, G., Meguid, M.A., Chouinard, L.E., Xu, C.: Insights into the transport and fragmentation characteristics of earthquake-induced rock avalanche: numerical study. Int. J. Geomech. 04020157 (2020)
6. Liu, G.Y., Xu, W.J., Sun, Q.C., Govender, N.: Study on the particle breakage of ballast based on a GPU accelerated discrete element method. Geosci. Front. 461–471 (2020)
7. Bagherzadeh, H., Mansourpour, Z., Dabir, B.: Numerical analysis of asphaltene particles evolution and flocs morphology using DEM-CFD approach. J. Petrol. Sci. Eng. 108309 (2021)
8. Jiale, C., Rao Muhammad, A., Hisham, C., Fahad Shahbaz, K., Yanwei, P., Ling, S.: Sipmask: Spatial information preservation for fast image and video instance segmentation. In: ECCV (2020)
9. Jonathon, L., Idil Esen, Z., Bastian, L.: Unovost: Unsupervised offline video object segmentation and tracking. In: WACV (2020)
10. Ali, A., Sabarinath, M., Aljosa, O., Laura, L., Bastian, L.: Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In: ECCV (2020)
11. Gedas, B., Lorenzo, T.: Classifying, segmenting, and tracking object instances in video with mask propagation. In: CVPR (2020)
12. Yuan, L., Duan, Y.: A method of ore image segmentation based on deep learning. In: Proceedings of the International Conference on Intelligent Computing (ICIC), pp. 508–519 (2018)
13. Duan, J., Liu, X., Wu, X., Mao, C.: Detection and segmentation of iron ore green pellets in images using lightweight U-net deep learning network. Neural Comput. Appl. 1–16 (2019)
14. Liu, X., Zhang, Y., Jing, H., Wang, L., Zhao, S.: Ore image segmentation method using U-Net and Res_Unet convolutional networks. RSC Adv. 9396–9406 (2020)
15. Lin, Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
16. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
17. Yang, L., Wang, Y., Xiong, X., Yang, J., Katsaggelos, A.: Efficient video object segmentation via network modulation. In: CVPR (2018)
18. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649 (2017)