COPS: A Real-Time Cross-Domain Object Part Segmentation System

Xueqing He

COPS: A Real-time Cross-domain Object Part Segmentation System

He Xueqing [1(✉)]

[1] China Agricultural University, No. 17 Qinghua East Road, Beijing 100083, P.R. China
2015314060617@cau.edu.cn

Abstract. Although the object part segmentation is widely applied to surveillance video analysis and smart recommendation and so on, however, it does not show a good performance in cross-domain testing. This means the segmentation model has to label various data in different scenarios and it is costly due to the time and labor cost. Accordingly, in the paper, we would like to propose a real-time cross-domain object part segmentation system (COPS) based on the work of *Cross-domain Human Parsing via Adversarial Feature and Label Adaptation* [2]. Several vital techniques are applied in this real-time cross-domain object part segmentation system, including object detection, object tracking, and cross-domain adaptation object part segmentation. Taking an unconstrained benchmark dataset with rich pixel-wise labeling as the source domain, the real-time cross-domain object part segmentation system aims to segment frames of target domain videos without any additional manual labeling in real-time. Compared with the traditional approaches, this system is demonstrated to be a highly efficient and useful one among most practical applications, and the exploration on the challenging issue will contribute our real-time cross-domain object part segmentation system and push human parsing into next step. Therefore, we would like to present the details of our real-time cross-domain object part segmentation system in the following parts.

Keywords: Cross-domain · Object part segmentation ·Real-time system

1 Introduction

Semantic segmentation is understanding an image at pixel level i.e., and we want to assign each pixel an object class in the image [8]. Scene parsing and object parsing, as types of semantic segmentation, are widely used in autonomous driving and surveillance video analysis. However, pixel-level labeling is costly, and semantic segmentation models tend to be poorly generalized across domains. A well-trained model based on the database of one scene always performs poorly in the data from other scenes.

Object part segmentation, which refers to an object decomposed into several semantic parts, enables a computer to understand an image deeply as well as automatically. It takes a significant part in various practical product applications such as face beauty fication, virtual reality and so on. However, we generally use to train a model using a lot of data with annotations, and it will cost a high annotation expense both in

time and labor [1]. Consequently, we expect to create a superior model trained with labeled data from source domain to segment data of any target domain. Nevertheless, a trained object part segmentation model often turns out to be bad results in application due to the shift between source domain and target domain. As a consequence, we are focused on solving this new problem: taking an unrestrained benchmark dataset with rich pixel-wise labels as the source domain and our real-time cross-domain object part segmentation system seems to be necessary and promising. But how to acquire an ideal parser that can annotate automatically by the system itself for the target domain is the remaining problem.

In this paper, we would like to put forward a real-time cross-domain object part segmentation system so as to solve the problem. This system consists of four modules, namely video capturing module, object capturing module, object segmentation module and result displaying module. Object segmentation module, the core module, applies the cross-domain human parsing method. It consists of two new compensation components, adversarial feature adaptation component and adversarial structured label adaptation component. The feature adaptation component is aimed at minimizing the feature distinctions between different domains, while the structured label adaptation is used for maximizing the label map universalities across the domains. The outcomes conformably corroborate our system an ideal one with efficient data and excellent performance for the challenging cross-domain human parsing problem. Meanwhile, we will continue to extend this approach to cross-domain object part segmentation as well. In addition, we presume that the exploration of the challenging issue will contribute our real-time cross-domain object part segmentation system to be more useful one among various practical applications.

The superiorities of this created system are summarized as follows. For one thing, the real-time cross-domain object part segmentation system can segment any target domain without any annotation just through transforming a source domain with rich pixel-wise annotation. For another, the capability of the system outperforms the state-of-art cross-domain approaches by two novel compensation components. Moreover, the system can segment 25 frames per second using NVIDIA GeForce GTX 1080, almost in real-time, and it can also be applied among the practical scenes in many fields.

2 Related Work

Although for many situations, we can find sufficient good performance on some correlating domains, extending the semantic segmentation model to a large variety of areas has many advantages, especially the real-time cross-domain area. Current statistical parsers tend to perform well only on their training domain and nearby genres [5]. We explore this issue as a new mission---taking an unconstrained benchmark dataset with rich pixel-wise labeling as the source domain. With the following techniques object detection, object tracking and cross-domain adaptation object part segmentation, our system aims to segment frames of target domain videos without any additional manual labeling in real-time. When the system is applied to a certain scene, we

only need to collect the data of that scene, and then it can be used as the training data of our system without any labeling. Once the training is completed, it can be applied to this scene, which can save a lot of time and labor costs. Compared with the traditional approaches, this resulting system shows obvious advantageous and presents to be a highly effective one among most practical applications, and the exploration on the challenging problem will contribute our real-time cross-domain object part segmentation system to be a more useful one among practical applications. Therefore, we are able to display the value of the real-time cross-domain object part segmentation system.

Object parsing. The semantic object parsing problem has aroused people's interest in exploring general object parsing, person part segmentation, as well as human parsing. Combining CNNs and CRFs is one method for capturing the rich structure information based on the advanced CNN architecture [4].

Domain adaptation. Deep domain adaptation focuses on transferring model learnt in one labeled source domain to one target domain in the deep learning framework. The exploration on this topic has been carried out along three various dimensions: unsupervised adaptation, supervised adaptation and semi-supervised adaptation. Unsupervised domain adaptation refers to the setting when the labeled target data is not available [9].

Domain adaptation method for segmentation. The closely related work to ours is [2], where the adversarial feature and structured label adaptation method is firstly put forward and developed to learn to diminish the cross-domain feature distinctions and increase the label universalities across the two domains. Most mainstream fashion parsing model concentrates on parsing high resolution and clean images. However, parsers that directly apply benchmarks for high-quality samples to particular applications in the field often exhibit unsatisfactory performance because of the domain shifts. The authors propose a new cross-domain human parsing model to diminish the cross-domain distinctions in terms of visual appearance and environment conditions and fully increase universalities across domains. The model they proposed explicitly explores a feature compensation network, and it is focused on diminishing the cross-domain distinctions. The outcomes consistently confirm data efficiency and excellent performance of the proposed method for the challenging cross-domain human parsing issue. The superiorities of the cross-domain human parsing model can be concluded as follows. Not only does this model firstly explore the cross-domain human parsing problem, but it means putting forward a cross-domain human parsing framework with the new feature adaptation and structured label adaptation network as well. Since no manual labeling in the target domain is needed, the new method is very useful and practical in terms of fully considering both feature invariance and label structure regularization in their cross-domain work. We have made the authors' original method of the paper into a training-predictive integration system. And the authors' original model and work guarantee sufficient preparation for the reality of our real-time cross-domain object part segmentation system.

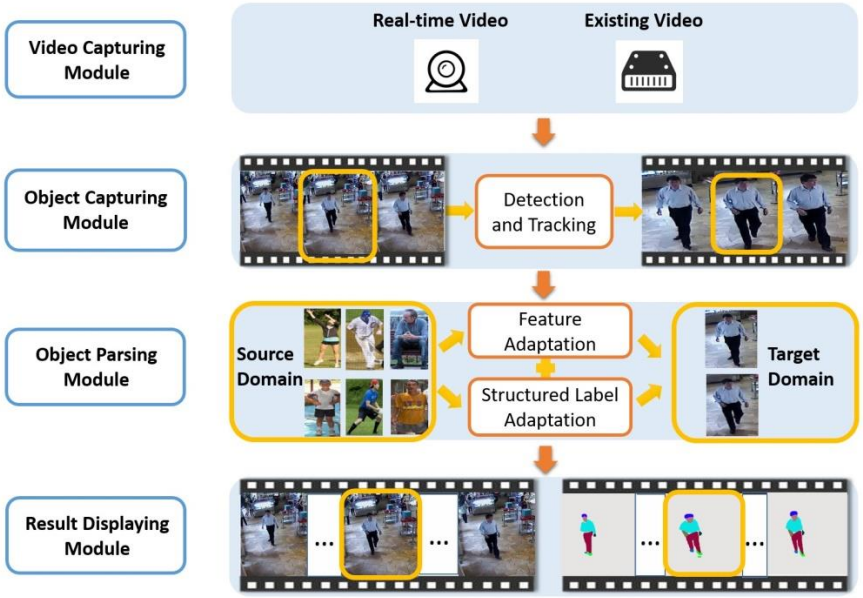

Fig. 1. The architecture of COPS system

3 System

Figure 1 displays the architecture of COPS system. As we can see, it contains four modules, including video capturing module, object capturing module, object part segmentation module and result displaying module. In the video capturing module, a video can be filmed by a real-time camera or loaded from the disk offline. After acquiring the video frames, detection [7], and tracking technique [3] are adopted. Detection technique is used in object capturing module to detect the main objects in the first frame, while tracking technique is used to track the detected objects in the following frames. Then, the core module, object part segmentation module, generates the segmentation results of the test video frames through a model trained in a source domain and transformed by adversarial feature adaptation and adversarial structured label adaptation. After finishing these steps, the result displaying module will finally show the live video and corresponding segmentation results on the screen at the same time. These four modules all work together and play their significant role in the whole system.

Moreover, we would like to highlight the third module, object part segmentation. From the third row of Figure 1, the following universalities and distinctions across the source domain and target domain can be found, where the universalities are expected to be fully utilized, and the distinctions are what we need to overcome and eliminate. As for distinctions, they own various viewpoint, tone, illumination, object scale, object posture, resolution and degree of blurry. For instance, the light in the target domain is much darker than the light in the source domain. Besides, object scale from

target domain is smaller than the source domain's. The people from both domains exist the intrinsic universalities, such as the similar structure of human parts. Taking a detail as an example, in both domains, according to human being's common sense, the body is below the head, but in the middle of both arms. The approach we adopted is to use the universalities and to overcome distinctions for domain adaptation. As a consequence, the distinctions of the features by adversarial feature adaptation are diminished, while the universalities of the structured labels by adversarial structured label adaptation are increased in this system. Plus, the system extremely improves the performance of segmentation.

We try to elaborate the details of adversarial feature adaptation component and adversarial structured label adaptation component. The real-time cross-domain object part segmentation system contains five networks, namely feature extractor, pixel-wise labeler, feature compensation network, feature adversarial network, and structured label adversarial network. Firstly, for feature adversarial network, we define one adversarial objective function that is adopted to measure the distance between the distribution of features of target domain and the distribution of the combined features including features of source domain and the output of feature compensation network. Secondly, for structured label adversarial network, another adversarial objective function is adopted for measuring the distance between the distribution of predicted label of target domain and the distribution of ground truth of source domain. Finally, the system can be trained jointly through the approach of LSGAN (least squares generative adversarial networks) that alternates between optimizing feature adaptation component and adversarial structured label adaptation component [6].

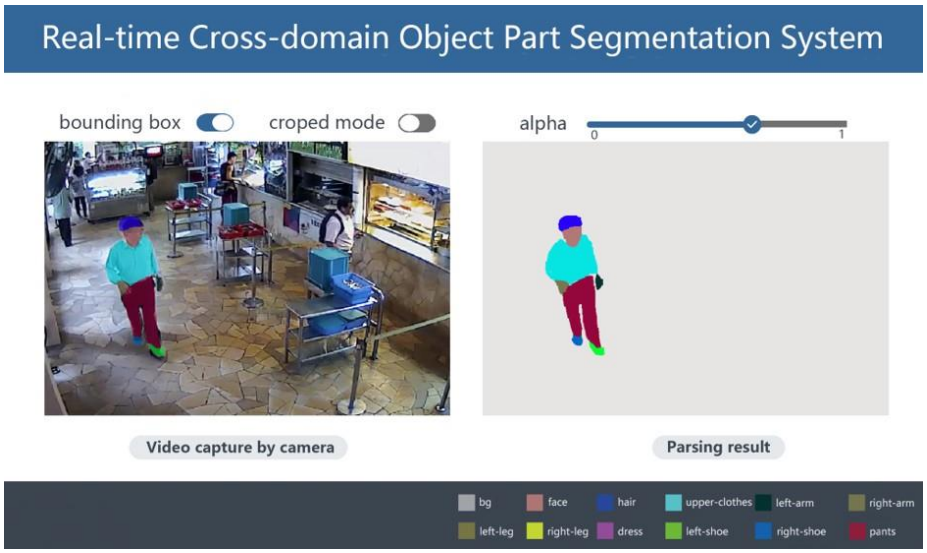

Fig. 2. The user interface of COPS system.

4 Demonstration

In order to attest the performance of the real-time cross-domain object part segmentation system, the demonstration needs a laptop, a video camera as well as a large screen. The laptop processes the frames captured by the video camera or loaded from disk and then the segmentation results are shown on the screen.

When selecting a local video file or the live video filmed by the real-time camera whose user interface is displayed in Figure 2, this video is going to be shown on screen in the left side, while the outcome of object part segmentation module will also be presented on screen in the right side during system execution. Meanwhile, users can choose whether to present bounding box of the object or crop the object from video. Alpha denotes the degree of mergence between the video and the segmentation result and it can be adjusted as well.

We present the comparisons of experimental results human parsing both with domain adaptation and without adaptation respectively in Figure 3. The cross-domain adaptation human parsing model can predict the details of the pictures so that these running results are more robust and adapted to the target domain. Overall, the real-time cross-domain object part segmentation system shows good performance and can be more advantageous for practical applications.

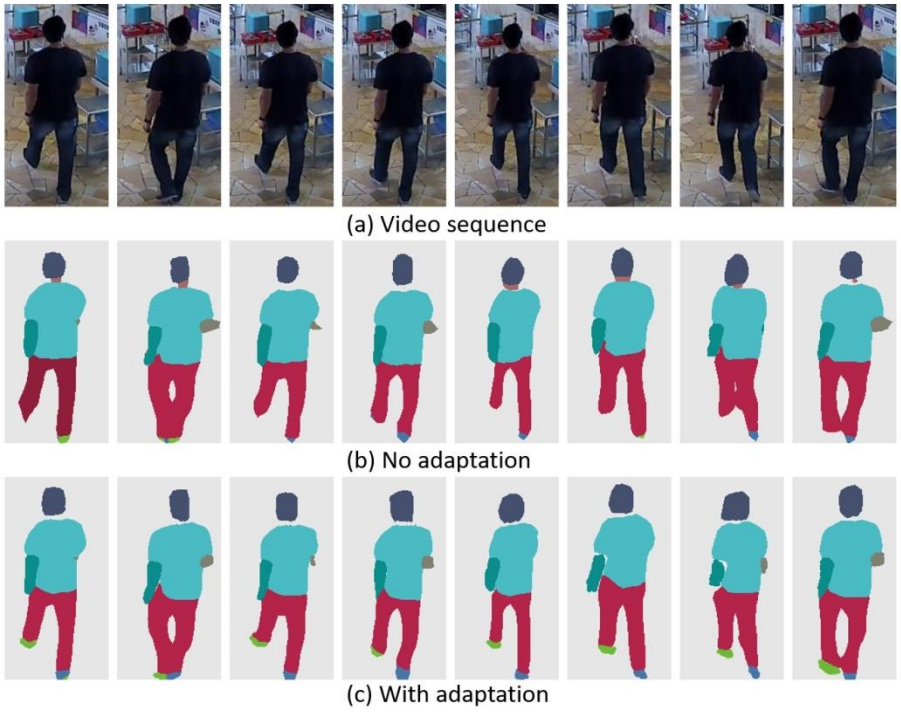

Fig. 3. Qualitative results on video sequence from target domain.

5 Discussion

Substantial benefit is obvious to applying our system over existing approaches. Using the cross-domain human parsing model without any additional manual labeling in real-time gives rise to excellent performance in practice. Additionally, our system tries to combine the human parsing and cross-domain feature transformation and a future direction is to learn these two jointly. Nowadays, our work focuses on challenging the issue, how to acquire an ideal parser that can annotate automatically by the system itself for the target domain. Not only the human parsing field, we also intend to extend our approach to scene parsing field. Scene semantic segmentation can be used as the core technology in various fields such as autonomous vehicles and smart home， and the cross-domain adaptive techniques tend to be extremely essential for these applications.

6 Conclusion

As mentioned above, we address the problem of taking an unconstrained benchmark dataset with rich pixel-wise labeling as the source domain and the real-time cross–domain object segmentation system is put forward to solve the challenging issue. The system is demonstrated to be an excellent one with higher practical advantages and more suitable applications. And according to the demonstration, an adversarial feature and structured label adaptation method are adopted to the object part segmentation module of the real-time cross-domain object segmentation system and they can jointly guarantee a precise and stable parser. Apart from the work we have done, we plan to continue researching on this approach and updating the domain adaptation technology to improve performance of our real-time cross-domain object part segmentation system of various fields in the network.

7 Acknowledgments

This work was performed and supported by the China Agricultural University. And we also would like to thank Defa Zhu for his technical and theoretical help.

References

1. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv preprint arXiv:1606.00915, (2016).
2. Liu, S., Sun, Y., Zhu, D. and Ren, G., Chen, Y, Feng, J., Han, J.: Cross-domain human parsing via adversarial feature and label adaptation. In AAAI (2018).
3. Liu, S., Zhang, T.Z., Cao, X.C., Xu, C.S.: Structural correlation filter for robust visual tracking. In CVPR (2016).
4. Liang, X., Shen, X., Feng, J., Lin, L., Yan, S.: Semantic object parsing with graph lstm. In ECCV (2016).

5. McClosky, D.E.C., Mark J.: Automatic domain adaptation for parsing. In ACL (2010).
6. Mao, X.D., Li, Q, Xie, H.R., Lau, R.Y., Wang, Z., Smolley, S.P.: The Grid: Least squares generative adversarial networks. arXiv:1611.04076, (2016).
7. Ren, S.Q., He, K.M., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS (2015).
8. Sasank, C.: http://blog.qure.ai/notes/semantic-segmentation-deep-learning-review.
9. Zhang, Y.: Fully Convolutional Adaptation Networks for Semantic Segmentation. In CVPR (2018).