Universität Osnabrück
Fachbereich Humanwissenschaften
Institute of Cognitive Science

Bachelorthesis - Expose

On the potential of recurrent structures for semantic video segmentation with deep networks

Sven Groen
970219
Bachelor's Program Cognitive Science

First supervisor: Dr. Ulf Krumnack
Institute of Cognitive Science
Osnabrück

Second supervisor: Manuel Kolmet
IMANOX GmbH
Berlin

Contents

1 Goal and background information

1.1 Cooperation with IMANOX

This project is realized in cooperation with IMANOX. IMANOX is a Berlin-based startup that developed a smart photo booth for expositions, events and promotions. The photo booth enables customers virtual product placements using augmented and mixed reality. Main features are hand-tracking, digital masks and changing virtual backgrounds.

1.2 Virtual backgrounds

Currently, the photo booth developed by IMANOX is using a built in Kinect RGB-D camera that measures the distance of objects by casting infrared illumination onto the scene and indirectly measuring the time it takes to travel back to the camera. The camera struggles with correctly predicting the depth in certain situations. Pixels are rendered as invalid and no depth information is provided. The reasons for this are numerous. Pixels might get undersaturated (signal is not strong enough) or oversaturated (signal is too strong). Other artifacts occur due to the geometry of the scene. For example, the sensors of the camera might receive signals from multiple locations in the scene, leading to an ambiguous depth. Especially around the edges and borders of objects pixels contain mixed signals from fore- and background, leading to blurred outlines. In the current version of the photo booth, alpha values (0 to 1) are calculated based on the data from the depth sensor. This is done by applying an alpha matting algorithm [1] to the raw depth data. Objects in the foreground receive high alpha values and the background is considered to have an alpha value of 0, making it transparent. In this way, the background can be virtually replaced without affecting the objects in the foreground. Due to the described inaccurate data that is given by the depth sensor, the result is of low quality. The edges and borders of the objects or people in the scene are not sharp and often misclassified. Especially for small and thin objects like hair, the camera hardly recognizes them

and parts of the objects are therefore misclassified as background and are also replaced by the virtual background. For more detailed information on this issue see `https://docs.microsoft.com/de-de/azure/Kinect-dk/depth-camera` [2].

1.3 Goal of the thesis

The goal of this bachelor thesis is to improve the quality of the semantic segmentation of the current IMANOX photo booth using machine learning techniques. The machine learning model will be one of the artificial neural network architectures (see Section 2.1 for details). This model will be altered and changed towards the needs of the project and will be trained with custom training data. It will be investigated whether additional temporal information leads to an improvement of the already existing segmentation models (see Section 3). When looking at succesful semantic segmentation models, it can be noticed that the prediction occurs on bases of individual images. It can be argued that in *real world scenarios* a semantic segmentation on single images is not as required as it is on videos, which are streams of frames. Potential applications include autonomous driving or video effects in the film industry.

When working with video data, valuable temporal information might get lost when the segmentation occurs on individual frame level. Therefore, the thesis will investigate whether the temporal information that is present in video data can be used for improvement of performance by adding recurrent structures into existing semantic segmentation models. Given the collaboration with IMANOX and their goal to replace their *virtual background change*-technique, the focus of the thesis will be on binary semantic segmentation instead of the typical multiclass segmentation.

2 Segmentation of Images

Szeliski [3] refers to image segmentation as "the task of finding groups of pixels that 'go together'" ([3], p. 237). How the pixels are *grouped together* can be achieved in different ways. In the classical image classification tasks, the task is to simply name the objects that can be seen in an image. Semantic image segmentation extends this problem further, referring to a pixel-wise classification of an image [4]. Each pixel in an image is assigned to one category label given a set of categories. However, individual instances of an object in one image are not differentiated but labeled as one class. Instance segmentation combines a pixel-wise classification (similar to semantic segmentation) with an additional object detection and localization, differentiating each individual instance of a class [4]. Given the goal of this project, only binary semantic segmentation is necessary. Detailed information about which objects are in the scene is not required.

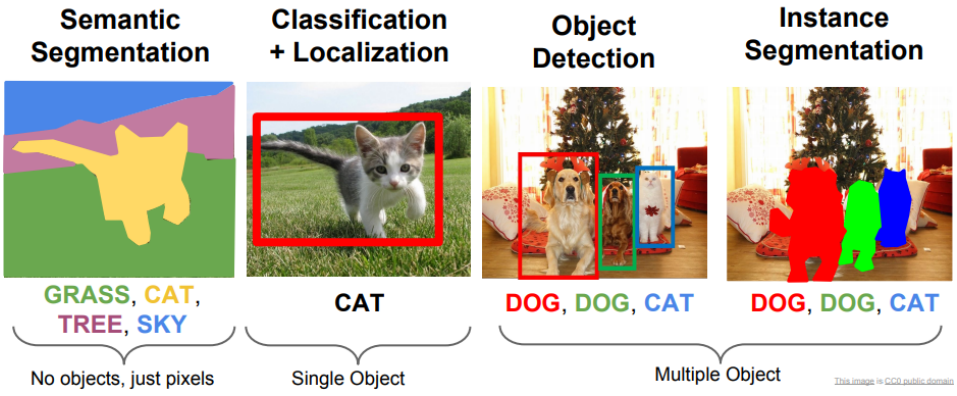

Figure 2.1: Overview of Computer Vision Tasks. (Li *et al.* [5], Slide 17).

2.1 Related work

Segmenting an image into its individual parts is a classical problem of computer vision [3]. Early approaches involve classical methods like threshold detection [6]. More modern approaches like k-means clustering [7] improved the early results.

Deep learning architectures, especially convolutional neural networks (CNNs) [8], have lead to further improvement. Long *et al.* [9] have been the first to propose a CNN architecture where a pixel-wise supervised training was achieved. This was done by upsampling the class prediction layer to the input image size, leading to an end to end pixel-wise classification, a *Fully Convolutional Network* (FCN) [9]. Following papers proposed different architectures. A *Deconvolutional Network* with special unpooling and deconvolution operations was invented by Noh *et al.* [10]. Here, the information was enconded using several convolutions and poolings and was decoded using unpooling and deconvolution. The SegNet model uses a similar Encoder-Decoder architecture by using pooling indices to upsample the image [11]. ICNet [12] was able to perform semantic segmentation not only in real-time, but also for high quality images (1024x2048 at 30 fps). This was achieved by using a cascade image input of different resolutions. The authors made use of the semantic information from the scaled down images and the details from the high resolution images. In this way, they have been able to achieve a "trade-off between efficiency and accuracy" ([12], p.2). Google's approach towards instance segmentation is called Deeplab and has evolved over the last recent years. The first DeepLab version uses a combination of Deep CNNs with fully connected conditional random fields (CRFs) that tries to grasp the semantic context of the image [13]. One of their main contributions is the use of atrous (or dilated) convolutions as an alternative to deconvolution. Originally used for wavelet transformations, a new parameter r allows to change the stride at which the samples are taken during the convolution operation [13]. This approach has been further improved and Atrous Spatial Pyramid Pooling (ASPP) (Figure 2.2) has been introduced, which was based on the idea of combining atrous convolutions with spatial pyramid pooling [14] (firstly introduced by He *et al.* [15]). In ASPP, parallel filters of different dilation rates are concatinated with the intend to cover different field-of-views [14]. The most recent approach, DeepLab V3+, has an Encoder-Decoder structure and was able to show "new state-of-the-art performance on PASCAL VOC 2012 and Cityscapes datasets." ([16], p.14). Moreover, the Google research team has shown that they are able to create an Encoder-Decoder architecture that is very light and fast. This architecture is light enough to run at real-time on modern smartphones [17].

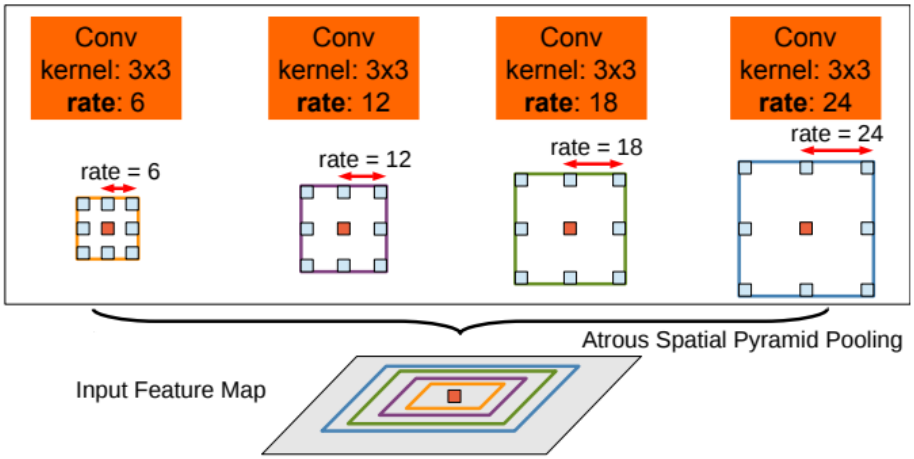

Figure 2.2: Atrous Spatial Pyramid Pooling. ([14], Fig. 4).

Most of the models are usually evaluated on individual images. However, in real world scenarios segmentation is often necessary for videos and not just single images. Naturally, the segmentation works on each individual frame but the context of the videos gets lost. The model is not able to know that the previous frame is related to the current one, it treats each frame independently. Extending existing models with a recurrent unit could enable the model to not only use spatial but also spatiotemporal information. Recurrent neural networks (RNNs), a special kind artificial neural networks, are usually the answer to time dependent problems [18] or used for sequential data such as in time series analysis or text translation. They are known to suffer from the *vanishing- and exploding-gradient-effect* [19]. To account for this problem, Long-Short-Term-Memory cells have been invented by Hochreiter and Schmidhuber [20] introducing a set of gates that enable the model to selectively forget, learn or keep specific information. Gated Recurrent Networks (GRUs) work in a similar way while being more efficient since less gates are available [21]. Research has been shown (e. g. [22–24]) that using Convolutional Long-Short-Term-Memory ("Conv-LSTM" [25]) layers help to improve the performance of already existing architectures. However, introducing a recurrent unit comes with the downside of increased computational complexity [24].

2.2 Challenges

There are several problems that have to be considered during the project:

1. achieving invariance towards environmental changes (moving, occluding/overlapping objects)
2. achieving a good trade-off between high-quality results and low-inference time
3. sucessfully integrating recurrent units (conv-LSTMs and conv-GRUs) into an already existing architecture

In order to achieve the first challenge the recurrent units in form of LSTM and GRU cells will be integrated. The assumption here is that this enables the model to recognize how the objects move throughout time and learn how to cope with uncertainty in situations where objects occlude each other. Pfeuffer *et al.* [23] and Pfeuffer & Dietmayer [24] have already shown that adding LSTMs has a massive influence on the number of parameters and inference time of the model. Therefore, the thesis will investigate whether the potential performance increase justifies the increased complexity (second challenge). For this reason, it is important to carefully consider where and how the recurrent units could be integrated in the model (third challenge).

3 Method

Since the main task is to separate humans (and objects they might carry) from the background, training data will be generated using green screen footage. A green screen setup allows to separate the background in a high quality manner using an alpha matting algorithm (e. g. [1]). The background in the training data can be virtually replaced with backgrounds that will be used in the photo booth in the final product.

It is very likely that the model will not be able to generalize very well beyond the given backgrounds, since the training data environment is highly controlled and fixed. Usual steps to cope with overfitting, e. g. data augmenting, will be explored during the training. However, the final model does not have to generalize beyond the given scenario anyway.

The project will use the most recent DeepLab architecture (DeepLabV3+) with different backbones. It will be investigated whether the complexity of the backbone has an influence on the performance change. Conv-LSTM and Conv-GRU cells will be integrated to the base models. The base model will be compared with their LSTM/GRU counterpart.

Depending on the time and scope that is left after the major questions have been evaluated, additional aspects will be explored. These include testing the model on a different dataset (probably Cityscapes dataset) and testing different models besides the DeepLabV3.

4 Preliminary structure

- Declaration of Authorship
- Abstract
- Contents
- List of figures
- List of algorithms
- Introduction
  - Motivation
  - Goal of the thesis
  - Semantic Segmentation
  - Recurrent Neural Networks
  - Related work
- Methods
  - The Dataset
  - Preprocessing
  - Network architecture
  - Network training
- Results
- Evaluation and discussion
- Conclusion
- Acknowledgements
- Bibliography

5 Time frame

Table 5.1: Previous time frame

Week	**Date**	**What will be done**
1-2	~~01.04.2020 - 14.04.2020~~	~~fixing the topics that will be investigated,~~ ~~fixing the model that will be used~~
2-4	~~15.04.2020 - 28.04.2020~~	~~preparing and recording training data,~~ ~~preprocess training data~~
4-8	~~29.04.2020 - 26.05.2020~~	training the model, evaluating and analysing the results, start writing
8-12	27.05.2020 - 22.06.2020	fix potential issues, finish implementation, finish writing,
12	23.06.2020	finish thesis

Table 5.2: Updated time frame

Week	**Date**	**What will be done**
1-4	01.06.2020 - 01.07.2020	finish integration of LSTM/GRU into DeepLab finish "Introduction"
4-8	02.07.2020 - 31.07.2020	finish training and evaluation end writing of thesis
8-12	01.08.2020 - 30.08.2020	proofreading (extend evaluation by additional models/datasets)
12	30.08.2020	finish thesis

6 Bibliography

1. Gastal, E. S. & Oliveira, M. M. Shared sampling for real-time alpha matting. *Computer Graphics Forum* **29,** 575–584. doi:10.1111/j.1467-8659.2009.01627.x (2010).
2. Sych, T., Brent, A., Phil, M. & Microsoft. *depth-camera @ docs.microsoft.com* 2019.
3. Szeliski, R. *Computer Vision: Algorithms and Applications* 185–186. doi:10.1017/cbo9780511974076.010 (Springer, 2011).
4. Minaee, S. *et al.* Image Segmentation Using Deep Learning: A Survey, 1–23 (2020).
5. Li, F.-F., Johnson, J. & Yeung, S. *Lecture 11: Detection and Segmentation* 2017.
6. Nobuyuki, O. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics* **9,** 62–66. doi:10.1109/TSMC.1979.4310076 (1979).
7. Dhanachandra, N., Manglem, K. & Chanu, Y. J. Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm. *Procedia Computer Science* **54,** 764–771. doi:10.1016/j.procs.2015.06.090 (2015).
8. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* **36,** 193–202. doi:10.1007/BF00344251 (1980).
9. Long, J., Shelhamer, E. & Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39,** 640–651. doi:10.1109/TPAMI.2016.2572683 (2014).
10. Noh, H., Hong, S. & Han, B. Learning Deconvolution Network for Semantic Segmentation. *Proceedings of the IEEE International Conference on Computer Vision* **2015 Inter,** 1520–1528. doi:10.1109/ICCV.2015.178 (2015).

11. Badrinarayanan, V., Kendall, A. & Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39,** 2481–2495. doi:10.1109/TPAMI.2016.2644615 (2015).

12. Zhao, H., Qi, X., Shen, X., Shi, J. & Jia, J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11207 LNCS,** 418–434. doi:10.1007/978-3-030-01219-9_25 (2017).

13. 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40,** 834–848. doi:10.1109/TPAMI.2017.2699184 (2016).

14. 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40,** 834–848. doi:10.1109/TPAMI.2017.2699184 (2016).

15. He, K., Zhang, X., Ren, S. & Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **8691 LNCS,** 346–361. doi:10.1007/978-3-319-10578-9_23 (2014).

16. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 833–851 (2018). doi:10.1007/978-3-030-01234-2_49.

17. Bazarevsky, V. & Tkachenka, A. *Mobile Real-time Video Segmentation* 2018.

18. Hoffmann, J., Navarro, O., Florian, K, Janßen, B. & Michael, H. A Survey on CNN and RNN Implementations. *Pesaro 2017,* 33–39 (2017).

19. Sherstinsky, A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena* **404,** 132306. doi:10.1016/j.physd.2019.132306 (2020).

20. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **9,** 1735–1780. doi:10.1162/neco.1997.9.8.1735 (1997).

21. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, 1–9 (2014).

22. Shahabeddin Nabavi, S., Rochan, M., Yang & Wang. Future Semantic Segmentation with Convolutional LSTM. *British Machine Vision Conference 2018, BMVC 2018,* 1–12 (2018).

23. Pfeuffer, A., Schulz, K. & Dietmayer, K. *Semantic Segmentation of Video Sequences with Convolutional LSTMs* in (2019). doi:10.1109/IVS.2019.8813852.

24. Pfeuffer, A. & Dietmayer, K. Separable Convolutional LSTMs for Faster Video Segmentation (2019).

25. Shi, X., Chen, Z. & Wang, H. Convolutional LSTM Network. *Nips,* 2–3. doi:[] (2015).