Expose:

Improving the Motion Guided Attention Model with long-term motion information

Anna Lepak

2lepak@informatik.uni-hamburg.de
Studiengang Informatik
Matr.-Nr. 6423845

Contents

1 Introduction **1**

2 Recent Work **1**

2.1 Image saliency models 1

2.2 Optical flow estimation 2

2.3 Video saliency models 2

3 Proposed Approach **4**

4 Experiment **4**

5 Timeline **5**

1 Introduction

Object detection is a big topic in Computer Vision with many different possible applications, such as surveillance alerts, image segmentation or medical applications [1]. But it is computationally expensive task in a real world cluttered scenario. Because of this it is important to have algorithms that find objects quickly and reliably. Salient object detection, which means finding objects of interest that draw the attention while looking at an image, has therefore become a good researched topic. Many of such salient object detection algorithms are based on the feature integration theory(FIT) [2], which was one of the first psychological theories of the human visual attention system. This theory states that the human brain processes several features in parallel and then fuses these features in one "master map of locations", from which regions of interest are selected. In recent years fully convolutional neural network (FCNs) have been widely employed to detect salient objects, since they learn the features themselves, instead of relying on handcrafted features which might not be optimal. With big datasets becoming more available, deep neural networks have improved the accuracy for many saliency algorithms and became the new state-of-the-art. While static image salient object detection works well on spatial information alone, temporal information becomes available in videos and needs to be taken into account. Few works are available for video salient object detection. Most rely on motion information from optical flow estimation, which estimates the apparent motion of each pixel in a sequence of images [3]. This is computationally expensive and slows down the model considerably. While such models achieve good accuracy, they are very slow and not applicable to real-world applications. Moreover, since optical flow estimation is calculated on two consecutive frames, the gained motion information only represents a short-term context. However, videos are usually longer than a few frames, so the motion information should also be considered in a long-term context. To address the above issues, I will modify an existing video salient object detection model to not rely on optical flow images and incorporate long-term motion information.

2 Recent Work

2.1 Image saliency models

One of the first image salient object detection models was developed in 1998 by Itti et al and it heavily relied on the FIT theory. In their model they compute feature maps for color, intensity and orientation channels using dyadic Gaussian pyramids and center-surround contrast [4]. The obtained feature maps are then fused together in a bottom-up manner to one saliency map. Since this first computational saliency model, the interest in saliency computation has increased strongly and many traditional saliency methods based on center-surround contrast have been developed.

But in recent years the popularity of deep neural networks increased and various

deep learning models were proposed. They use different approaches, like *Wang et al.* [5] proposed an encoder-decoder architecture, which was trained over multiple scales. The encoder consists of convolutional layers, which capture high-level features and down-sample the low-feature maps. The decoder upsamples these feature maps so that they get the original resolution of the input image and then fuses them to a saliency map.

2.2 Optical flow estimation

FlowNet2.0 [6] is a popular optical flow estimation method. It is based on the Flownet by *Dosovitskiy et al.* which was one of the first optical flow estimation methods to use convolutional neural networks. FlowNet 2.0 employs two parallel streams to estimate small and large displacements. The stream to compute the large displacements combines multiple FlowNets while the stream for small displacements uses only one FlowNet. At the end both streams are fused and an optical flow is estimated.

2.3 Video saliency models

Luo et al [7] developed a model which first divides each frame into superpixels using the SLIC algorithm and then computes a motion vector field for each superpixel. They used this vector field to compute two measures: the motion uniqueness – which represents the difference of each superpixel element from all other elements of the frame – and motion distribution - which represents the spatial variance of the motion vector. Both measures were fused together to derive motion saliency for that element. To get the saliency map, they combined all element motion values linearly.

In [8], *Le and Sugimoto* developed a video salient object detection using spatiotemporal deep (STD) features. They used conditional random fields (CRF) enhanced with the STD features to compute a saliency map based on region-wise features.

Maczyta et al. developed a motion saliency map estimation method which relies on optical flow cues only [9]. It is based on the notion that motion itself can create attention, especially when an object motion is very different from its surrounding motion. An example for this would be a crowd moving in one direction while one person in it moves in the other direction. In their method they compare the optical flow in a given area - which most likely is a salient moving object - with the optical flow field that would have been induced in the same area with the surrounding motion. Because the later is not observable, they used an inpainting method to obtain this optical flow field. A discrepancy between this two flows is an indication for motion saliency.

But this method only uses motion cues to detect saliency, which is not a good way to detect objects, because only the parts of an object that are actually in motion would be detected. In a video with a person waving their hand only the hand would be recognized instead of the whole person. It is better to use this motion information to attend appearance features. In 2019, *Haofeng Li et al.* developed an system which combines both

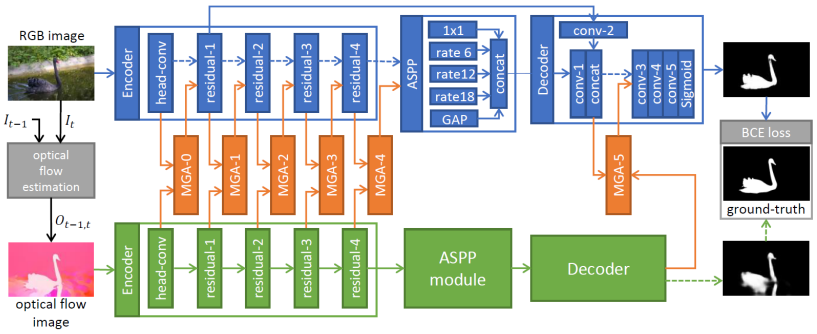

Figure 1: The Network Architecture for Motion Guided Attention [10]

saliency problems. For their model they developed two sub-networks - the appearance branch and the motion branch - and motion attention modules, as seen in Figure 1.

Both the appearance and motion branch consist of an encoder, an ASPP (Atrous Spatial Pyramid Pooling) module and a decoder. The encoder consists of five layers. Each layer outputs either appearance features or motion information. The appearance branch - the blue one in Figure 1 - takes a static image as input while the green motion branch takes optical flow image as input. The Motion Guided Attention Modules (orange) take motion information from the motion branch and attend them with the appearance features from the appearance branch. There are four different Motion Guided Attention Modules as seen in Figure 2. The MGA-m module takes a saliency map as the motion input, while the other three MGA modules take a motion feature tensor as the motion input. MGA-tm converts the motion feature tensor into a spatial weights map before attending it with the appearance feature. The MGA-tmc module first does the same as the MGA-tm module, but afterwards uses a global poooling average algorithm to harvest a global representation and based on this predicts a vector of C scalar weights for channels. With these channel-wise attention weights the module aims at selecting attributes such as edges, colors, etc which generate more attention. The decoder in the architecture in Figure 1 outputs a saliency map, which is why the linked MGA-5 module has to be MGA-m. The MGA-0,1,2,3,4 modules could be MGA-t, MGA-tm or MGA-tmc. Li et al. propose to use MGA-tmc for the positions MGA-0-4 and MGA-m for MGA-5.

Yi Tang et al. [11] used a ConvLSTM in their architecture. They still used two subnetworks to extract feature maps, similar to the work in [10]. But theses feature maps are then fed to the ConvLSTM unit, which refines the motion information and outputs a feature map. All feature maps are then fused together through a 1x1 convolution to produce the saliency map.

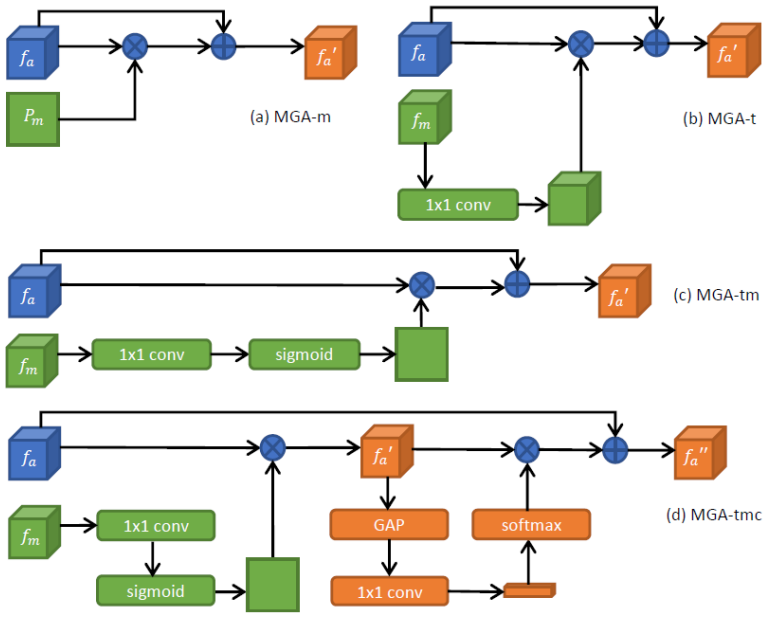

Figure 2: Motion Guided Attention Modules [10]

3 Proposed Approach

In this work, I will analyze these various motion saliency approaches. In the evaluation data provided in [10] their MGA model reaches the lowest mean absolute error (MAE) and highest max F-measure. With the division of the architecture into two sub-networks, it is possible to modify them independently. The motion sub-network uses optical flow images as input. Such calculations are expensive and time-consuming [12]. Furthermore, with optical flow as input, only short term context in a video is considered. But videos usually are longer and as such longer time spans should be taken into consideration.

Tang et al. [11] used a ConvLSTM in their model to consider long-term motion information, but they still relied on optical flow images as input to their temporal sub network.

For my work I will keep the appearance sub-network of the MGA model as it is, but completely disregard the motion sub-network and replace it with a ConvLSTM. The ConvLSTM will take the feature maps generated from the appearance sub-network as input. For each frame the LSTM units output feature maps, which then are fed into a 1x1 convolutional layer and a sigmoid function to produce the final saliency map for the input frame.

4 Experiment

Since my model will be mainly compared with the model from [10], I will use the same datasets they used. The DUTS dataset [13] is a static-image salient object detection dataset

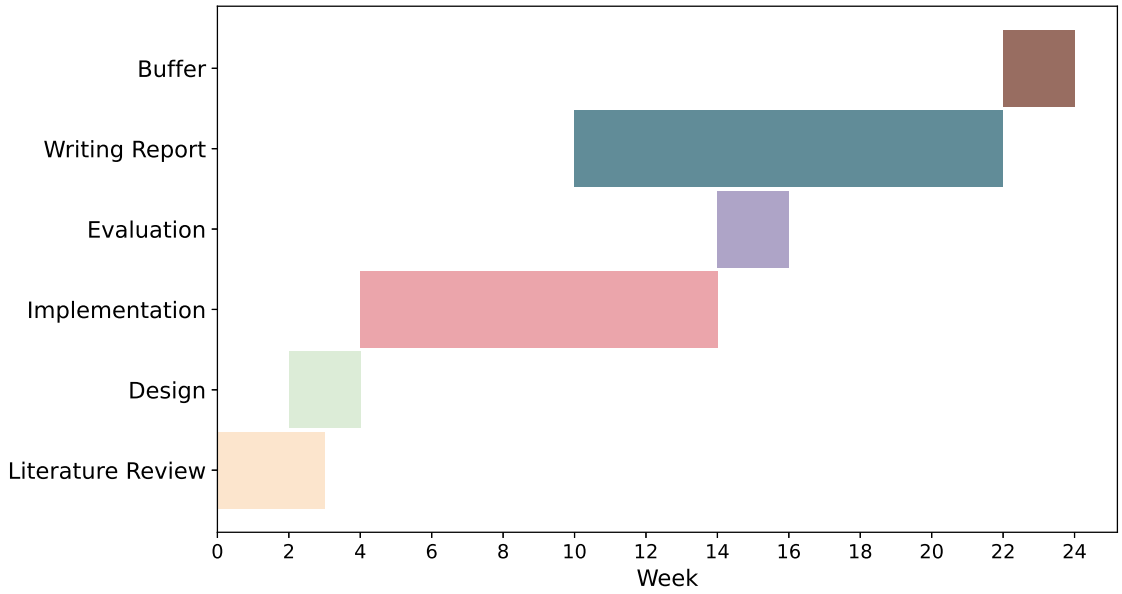

Figure 3: Timeline for the thesis

and will be used to train the appearance sub-network. DAVIS [14] and ViSal [15] are both video salient object detection datasets with pixel-precise annotations. DAVIS has different sets where either a single object is annotated or multiple object are annotated. I will use the single object annotation dataset. This DAVIS train dataset will be used to train the whole model. To evaluate the model I will use mean absolute error (MAE), structure-measure (S-m), maxF-measure (maxF) and Precision-Recall curves. Furthermore I will measure the speed of both models. I expect my model to be faster while reaching at least the same accuracy as the MGA model.

5 Timeline

The Figure 3 shows my estimated timeline for my Thesis. First I will do a deep literature review. This involves the papers mentioned in this expose, as well as papers referenced in these, which could prove useful for my thesis. During the "Design" phase I will plan the architecture of my model. Then I will implement the model, for which I plan the most time. After the model has been fully implemented and tested I will test it against the MGA model, which is planned under the "Evaluation" section. Lastly I will finish my report. I have included a two week buffer for any potential unforeseen problems

References

[1] A. Vahab, M. S. Naik, P. G. Raikar, and P. S. R, "Applications of object detection system," in *International Research Journal of Engineering and Technology (IRJET)*, 2019.

[2] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, Jan. 1980.

[3] P. Turaga, R. Chellappa, and A. Veeraraghavan, "Advances in video-based human activity analysis: Challenges and approaches," in *Advances in Computers* (M. V. Zelkowitz, ed.), vol. 80 of *Advances in Computers*, pp. 237–290, Elsevier, 2010.

[4] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[5] W. Wang and J. Shen, "Deep visaul attention prediction," in *IEEE Transactions on Image Processing*, March 2018.

[6] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] L. Luo, R. Jiang, X. Tian, and Y. Chen, "Video saliency detection using motion saliency filter," in *Proceedings of 2013 3rd International Conference on Computer Science and Network Technology*, IEEE, 2013.

[8] T.-N. Le and A. Sugimoto, "Video salient object detection using spatiotemporal deep features," *IEEE Transactions on Image Processing*, vol. 27, pp. 5002–5015, Oct 2018.

[9] L. Maczyta, P. Bouthemy, and O. L. Meur, "Unsupervised motion saliency map estimation based on optical flow inpainting," in *2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019*, pp. 4469–4473, IEEE, 2019.

[10] H. Li, G. Chen, G. Li, and Y. Yu, "Motion guided attention for video salient object detection," 2019.

[11] Y. Tang, W. Zou, Z. Jin, and X. Li, "Multi-scale spatiotemporal conv-lstm network for video saliency detection," in *ICMR '18: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, June 2018.

[12] W. Wang, T. Zhou, F. Porikli, D. Crandall, and L. V. Gool, "A survey on deep learning technique for video segmentation," 2021.

[13] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *CVPR*, 2017.

[14] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Computer Vision and Pattern Recognition*, 2016.

[15] W. Wang, J.Shen, and L.Shao, "Consistent video saliency using local gradient flow optimization and global refinement," in *IEEE Trans. on Image Processing, 24(11):4185-4196*, 2015.