Advancing Video Object Segmentation with Spatio-Temporal Markov Random Fields

Wang Ganyi
School of Computing
National University of Singapore
Singapore
e1101663@u.nus.edu

Li Xingchen
School of Computing
National University of Singapore
Singapore
e0945876@u.nus.edu

Zhang Rongqi
School of Computing
National University of Singapore
Singapore
e1132299@u.nus.edu

Yang Boran
School of Computing
National University of Singapore
Singapore
boranyang@u.nus.edu

1 Introduction

Video Object Segmentation is a critical challenge in the field of computer vision, which aims to accurately separate foreground object(s) from the background in consecutive frames of a video. This task requires not only the recognition of objects in a single frame just like the typical image segmentation task, but also the continuous tracking of these object(s). In addition, in real scenarios, the video object segmentation task is affected by a number of factors, such as lighting conditions, occlusions, and the variety of object appearances and backgrounds in video streams, which significantly increase the complexity of the task.

There are multiple reasons that drive us to address the video object segmentation challenge. First, the video object segmentation task is the basis for a large number of advanced applications that often require a high-level understanding of the video content. For example, in autonomous driving, which has become very popular in recent years, the ability to apply accurate segmentation and track objects is important for the safety and reliability of navigation systems. Similarly, in video content AI creation and editing, accurate recognition and processing objects in a scene is fundamental to enabling further complex editing.

On the other hand, the nature of video object segmentation involves several other challenging domains in computer vision, including object recognition, motion analysis, and scene understanding. Attempting to solve the video object segmentation problem often requires researchers to have a broad understanding of how models interpret dynamic visual information. Any developments in these domains are likely to provide new ideas for developing better approaches to solving video object segmentation tasks.

Furthermore, the video object segmentation task is very interesting due to its complexity and the creative solutions required. There are many related works that exist in this area, and we will introduce some of interesting ideas and studies in Section 2.

2 Related Work

Historically, video object segmentation research began with simpler, rule-based algorithms that relied heavily on colour and texture features to distinguish objects from the background. As artificial intelligence evolved, the advent of machine learning and subsequently deep learning opened up many

Preprint. Under review.

new possibilities. In this section, we review some interesting related work in the domain of video object segmentation.

2.1 Handcrafted Feature-based Methods

One of the typical handcrafted feature-based approaches is based on the histogram statistics. Lam and Lee [1] introduced a video segmentation algorithm utilizing a color difference histogram (CDH), designed to be robust against changes in lighting, object movements, and camera motions.

Lei, et al. [2] introduced a framework for video shot detection and sequence segmentation, grounded in statistical hypothesis testing. This method employs critical statistical characteristics that incorporate both local and global spatial and temporal data from video sequences.

Optical flow, depicting object motion due to object or camera movement between frames, is a pivotal feature in video segmentation. By effectively capturing the temporal and spatial relationships of moving objects, it estimates motion vectors for pixel position changes across frames, aiding in understanding scene motion patterns and facilitating object segmentation. FlowNet [3] significantly enhances optical flow estimation with Convolutional Neural Networks (CNNs), addressing traditional method limitations through deep learning. FlowNet includes two main architectures: FlowNetS, stacking two consecutive frames for CNN processing, and FlowNetC, using a correlation layer for refined motion analysis between frames.

Tsai, et al. [4] proposed a method that considers the optical flow estimation and video segmentation at the same time. Specifically, they compute the flow independently in the segmented regions and recompose the results to estimate the optical flow. And they also design a multi-scale spatio-temporal objective function to transfer information between frames by using the optical flow.

2.2 Classical Machine Learning Methods

Yu, et al. [5] introduced a video segmentation framework utilizing parametric graph partitioning (PGP), a method with minimal parameter requirements that works by identifying and eliminating edges between clusters to create clusters of nodes. PGP is known for its computational efficiency and its ability to cluster the spatio-temporal volume without the need for pre-determining the number of clusters.

Grundmann, et al. [6] proposed an approach for the spatiotemporal segmentation of long video sequences through a hierarchical graph-based method. The process is initiated by over-segmenting a volumetric video graph into space-time regions that are categorized based on their appearance. Subsequently, a "region graph" is formed from the segmentation, and this procedure is iteratively applied across several layers to generate a hierarchical tree of spatio-temporal segmentations.

Jang and Kim [7] proposed a method that employs short-term hierarchical segmentation (STHS) and couples it with frame-by-frame optimization using a Markov random field (MRF). STHS initially segments by moving through a window of frames, applying spatial agglomerative clustering to each frame, followed by inter-frame bipartite graph matching to create initial segments. Subsequently, by minimizing an MRF energy function that includes both unary and pairwise costs, they obtained the final segments for each frame.

2.3 Deep Neural Network Models

Perazzi, et al. [8] presented the idea of using a convolutional network-based method in video object segmentation. The proposed model considers the results of the previous frame and guides the recognition of the interest object in the next frame.

Chandra, et al. [9] proposed a method that integrates neuronal decisions across both spatial and temporal dimensions. By leveraging the progress in deep Gaussian Conditional Random Fields (GCRFs), they demonstrated the capability for exact and efficient inference on a densely connected spatio-temporal graph.

Ren, et al. [10] proposed a reciprocal mechanism that integrates the intra-frame contrast, motion cues, and temporal coherence of recurring objects to filter out ambiguous distractions from videos.

One-Shot Video Object Segmentation (OSVOS) [11] is a technique for semi-supervised video object segmentation. The goal of OSVOS is to separate an object from the background in a video sequence. By transferring this knowledge to the task of foreground segmentation and fine-tuning on a single annotated object in the test sequence, OSVOS achieves impressive results. Notably, OSVOS processes each frame of the video independently, leading to temporally coherent and stable segmentations without the need for explicit temporal constraints.

3 Methodology

The following ideas are basically based on [12].

3.1 Energy Function Design

A state $\mathbf{x}$ is defined as the collection of pixel values $\mathcal{V} = \{1, 2, ..., N\}$ in the video, where each pixel's value belongs to the range $\mathcal{L} = \{0, 1\}$. The state $\mathbf{x}_c$ for a specific frame c denotes the pixel values in that frame, given by $\mathcal{V}_c = \{1, 2, ..., N_c\}$. The overall energy of state $\mathbf{x}$ is represented by the Eq. 1.

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} E_u(x_i) + \sum_{(i,j) \in \mathcal{N}_T} E_t(x_i, x_j) + \sum_{c \in \mathcal{S}} E_s(\mathbf{x}_c), \tag{1}$$

which encompasses three distinct energy sources: unary energy $E_u(x_i)$, temporal energy $E_t(x_i, x_j)$, and spatial energy $E_s(\mathbf{x}_c)$. Here, $\mathcal{N}_T$ denotes the temporal relationships between different frames, and $\mathcal{S} = \{1, 2, ..., C\}$ represents the set of frames within the video.

The unary energy $E_u(x_i)$ is acquired through the OSVOS model. The temporal energy $E_t(x_i, x_j)$ is obtained using the FlowNet2 model, providing us with $\mathcal{N}_T$, representing the relationships between the same pixel in different frames. Finally, the spatial energy $E_s(\mathbf{x}_c) \propto \|\mathbf{x_c} - g_{CNN}(\mathbf{x_c})\|_2^2$ is derived from the CNN model, specially trained to produce refined image masks.

The rationale behind the design of the energy function is rooted in the following intuitions: Firstly, the OSVOS model provides an initial coarse mask, and the objective is for the MRF output to closely align with it, resulting in a low one-dimensional energy state. Secondly, pixels at different positions in different frames should correspond to the same mask state, leading to a low temporal energy state. Lastly, considering that the image segmentation boundaries produced by the CNN are more refined compared to those of the MRF, the hope is for the MRF output to be less divergent from the refined CNN boundaries, resulting in a low spatial energy state.

3.2 Training

The training process consists of three key models: OSVOS for object segmentation, FlowNet2 for optical flow computation, and a two-stage CNN training strategy. OSVOS is initially trained on the DAVIS2016 dataset [13] to acquire general object segmentation skills. Simultaneously, the FlowNet2 model undergoes pre-training using a dedicated dataset to augment its optical flow computation capabilities.

The CNN training unfolds in two phases: initially, pre-training occurs on the DAVIS2016 training set to establish fundamental segmentation capabilities. In the subsequent phase, the model undergoes fine-tuning using the first frame's ground truth masks from the DAVIS2016 test set, coupled with the integration of data augmentation techniques to enhance generalizability. This adaptive training process empowers the CNN to adapt its segmentation capabilities for specific objects in each video sequence, consequently improving its performance in predicting subsequent frames.

3.3 Inference

The energy function $E(\mathbf{x})$ can be minimized in MRF using an iterative method named Iterated Conditional Modes (ICM) [14]. Due to the computational cost associated with running the CNN inference in each iteration for every pixel, we maintain the CNN output unchanged to compute $E(\mathbf{x})$. Following the ICM iteration, we update the CNN output, incorporating it in the subsequent iteration

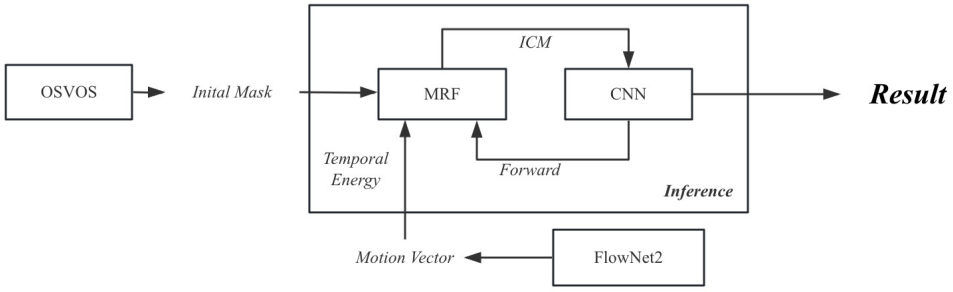

Figure 1: Computational flow

of ICM. Consequently, an iterative approach to the overall inference algorithm has been established, as illustrated in Fig. 1.

3.4 Motivation

Fusing CNN for feature extraction and MRF for dependency modelling can significantly enhance video segmentation. While CNNs excel at hierarchical feature extraction, they face challenges in capturing temporal and spatial relationships, especially in dynamic scenes. MRF models, with their superior ability to model spatial and temporal dependencies, can address these limitations. Structured probabilistic modelling in MRF models accurately represents the interactions of pixels over time, thus ensuring improved segmentation consistency and reliability. The use of complex spatio-temporal MRF models can be effectively combined with CNN output to provide a fine-grained approach to spatio-temporal relationships. This cooperative approach bridges the gap between independent CNN applications and provides a more comprehensive solution to video segmentation challenges.

4 Social Imapct

This approach makes innovations to traditional MRF by integrating it with machine learning models to advance traditional probabilistic models. We believe the exploration and advancement of video segmentation technology potentially has broad societal implications that affect many aspects of daily life. For instance, in the medical field, improved video segmentation techniques can also improve diagnostic procedures. By accurately understanding and analyzing medical images, we can detect medical conditions earlier with higher accuracy. One of the most important impacts is for public safety, where better video segmentation technology allows for more accurate and real-time analysis of surveillance footage, allowing us to identify suspicious activity and prevent crime. In addition, in the video entertainment and media industry, improved techniques have the potential to allow content creation, enabling filmmakers and video editors to produce higher quality visual effects at lower costs.

5 AI Tool Use

In writing the abstract, ChatGPT fulfills three main purposes. Firstly, it provides a broad introduction to the various applications of probabilistic modelling in statistics, machine learning and artificial intelligence. Subsequently, it played a key role in elucidating the complex methods and concepts in the paper after we had identified the topic of our choice, helping us to understand the methods in the paper. Finally, the AI tool optimised the syntax of the paper, ensuring accuracy and compliance with academic writing standards.

References

[1] C. F. Lam, & M. C. Lee. (1998, August). Video segmentation using color difference histogram. In *IAPR International Workshop on Multimedia Information Analysis and Retrieval* (pp. 159-174). Berlin, Heidelberg: Springer Berlin Heidelberg.

[2] Z. Lei, W. Chou, J. Zhong, & C. H. Lee. (2000, July). Video segmentation using spatial and temporal statistical analysis method. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)* (Vol. 3, pp. 1527-1530). IEEE.

[3] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., & Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2462-2470).

[4] Y. H. Tsai, M. H. Yang, & M. J. Black. (2016). Video segmentation via object flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3899-3908).

[5] C. P. Yu, H. Le, G. Zelinsky, & D. Samaras. (2015). Efficient video segmentation using parametric graph partitioning. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3155-3163).

[6] M. Grundmann, V. Kwatra, M. Han, & I. Essa. (2010, June). Efficient hierarchical graph-based video segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 2141-2148). IEEE.

[7] W. D. Jang, & C. S. Kim. (2016). Streaming video segmentation via short-term hierarchical segmentation and frame-by-frame Markov random field optimization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14* (pp. 599-615). Springer International Publishing.

[8] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, & A. Sorkine-Hornung. (2017). Learning video object segmentation from static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2663-2672).

[9] S. Chandra, C. Couprie, & I. Kokkinos. (2018). Deep spatio-temporal random fields for efficient video segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8915-8924).

[10] S. Ren, W. Liu, Y. Liu, H. Chen, G. Han, & S. He. (2021). Reciprocal transformations for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15455-15464).

[11] Caelles, S., Maninis, K. K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., & Van Gool, L. (2017). One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 221-230).

[12] Bao, L., Wu, B., & Liu, W. (2018). CNN in MRF: Video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5977-5986).

[13] *Benchmark Video Object Segmentation on DAVIS*, 2016. Available at https://davischallenge.org/davis2016/soa_compare.html

[14] Bishop, C. (2006). Pattern recognition and machine learning. *Springer google schola*, 2, 5-43.