# Video Propagation Networks

Xuewen Yang

June 12 2018

## Abstract

*We propose a technique that propagates information for- ward through video data. The method is conceptually sim- ple and can be applied to tasks that require the propagation of structured information, such as semantic labels, based on video content. We propose a Video Propagation Network that processes video frames in an adaptive manner. The model is applied online: it propagates information for- ward without the need to access future frames. In particular we combine two components, a temporal bilateral network for dense and video adaptive filtering, followed by a spatial network to refine features and increased flex- ibility. We present experiments on video object segmentation and semantic video segmentation and show increased performance comparing to the best previous task-specific methods, while having favorable runtime. Additionally we demonstrate our approach on an example regression task of color propagation in a grayscale video.*

## 1. Introduction

In this work, we focus on the problem of propagating structured information across video frames. This problem appears in many forms and is a pre-requisite for many applications. An example instance is shown in Figure 1. Given an object mask for the first frame, the problem is to propagate this mask forward through the entire video sequence. Propagation of semantic information through time and video color propagation are other problem instances.

Videos pose both technical and representational challenges. The presence of scene and camera motion lead to the difficult pixel association problem of optical flow. Video data is computationally more demanding than static images. A naive per-frame approach would scale at least linear with frames. These challenges complicate the use of standard convolutional neural networks (CNNs) for video processing. As a result, many previous works for video propagation use slow optimization based techniques.

Our architecture is composed of two components (see Figure 1. A temporal bilateral network that performs imageadaptive spatio-temporal dense filtering. The bilateral network allows to connect densely all pixels from current and previous frames and to propagate associated pixel information to the current frame. The bilateral network allows the specification of a metric between video pixels and allows a straight-forward integration of temporal information. This is followed by a standard spatial CNN on the bilateral network output to refine and predict for the present video frame. We call this combination a Video Propagation Network (VPN). In effect, we are combining video-adaptive filtering with rather small spatial CNNs which leads to a favorable runtime compared to many previous approaches.

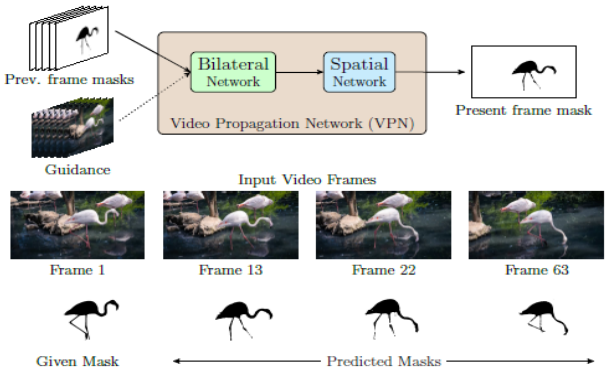


Figure 1. Video Propagation with VPNs. The end-to-end trained VPN network is composed of a bilateral network followed by a standard spatial network and can be used for propagating information across frames. Shown here is an example propagation of foreground mask from the $1^{st}$ frame to other video frames.

## 2. Related Work

### 2.1. General propagation techniques

Techniques for propagating content across image/video pixels are predominantly optimization based or filtering techniques. Optimization based techniques typically formulate the propagation as an energy minimization problem on a graph constructed across video pixels or frames. A classic example is the color propagation technique. Although efficient closedform solutions [3] exists

for some scenarios, optimization tends to be slow due to either large graph structures for videos and/or the use of complex connectivity. Fullyconnected conditional random fields (CRFs)open a way for incorporating dense and long-range pixel connections while retaining fast inference.

Filtering techniques [1] [2] aim to propagate information with the use of image/video filters resulting in fast runtimes compared to optimization techniques. Bilateral filtering is one of the popular filters for long-range information propagation. A popular application is joint bilateral upsampling that upsamples a low-resolution signal with the use of a high-resolution guidance image. The works showed that one can backpropagate through the bilateral filtering operation for learning filter parameters or doing optimization in the bilateral space. Recently, several works proposed to do upsampling in images by learning CNNs that mimic edge-aware filtering [6] or that directly learn to upsample. Most of these works are confined to images and are either not extendable or computationally too expensive for videos. We leverage some of these previous works and propose a scalable yet robust neural network approach for video propagation.

### 2.2. Video object segmentation

Prior work on video object segmentation can be broadly categorized into two types: Semi-supervised methods that require manual annotation to define what is foreground object and unsupervised methods that does segmentation completely automatically. Unsupervised techniques such as [5] [4] use some prior information about the foreground objects such as distinctive motion, saliency etc.

### 2.3. Semantic video segmentation

Earlier methods such as use structure from motion on video frames to compute geometrical and/or motion features. More recent works construct large graphical models on videos and enforce temporal consistency across frames.It used dynamic temporal links in their CRF energy formulation. It proposes to use Perturb-and-MAP random field model with spatial-temporal energy terms and propagate predictions across time by learning a similarity function between pixels of consecutive frames

## References


[1] J. H. R. Chang and Y. C. F. Wang. Propagated image filtering. In *CVPR*, 2015. 2

[2] K. He, J. Sun, and X. Tang. Guided image filtering. *IEEE TPAMI*, 2013. 2

[3] A. Levin, D. Lischinski, and Y. Weiss. A closed form solution to natural image matting. *IEEE TPAMI*, 2007. 1

[4] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2014. 2

[5] B. Taylor, V. Karasev, and S. Soattoc. Causal video object segmentation from persistence of occlusions. In *CVPR*, 2015. 2

[6] L. Xu, J. S. J. Ren, Q. Yan, R. Liao, and J. Jia. Deep edge-aware filters. In *ICML*, 2015. 2