January 2020

CHANGING AN ONLINE VIDEO MEETING BACKGROUND IN REAL-TIME USING DEEP LEARNING

Barrie Chen

Iris Qian

Richard Cai

CHANGING AN ONLINE VIDEO MEETING BACKGROUND IN REAL-TIME USING DEEP LEARNING

AUTHORS:

Barrie Chen
Iris Qian
Richard Cai

ABSTRACT

This proposal provides a technique to change an online video meeting background in real-time using deep learning. By changing the background video for participants of a meeting, the participants can protect their accuracy. The background-changing technique of this proposal may include a deep learning model for video semantic segmentation. One or more Spatio-Temporal Transformer Gated Recurrent Units (STGRUs) may be utilized to enhance classification accuracy and to add an optical flow into the model; thereby increasing calculation speed.

DETAILED DESCRIPTION

There are many online conference tools that are useful and portable for people that may work at home or may travel on business trips. However, people sometimes feel embarrassed to open their cameras when they are in private zones, such as their bedrooms, or when their family members are around them. In an online conference, attendees hope to see others' faces but may not want to show their own background due to privacy concerns.

This proposal provides a technique that involves utilizing a video semantic segmentation model that can be configured to classify the background in a video and cover the background with other pictures in real-time to protect the privacy of meeting attendees. Figure 1, below, illustrates example details that may be associated with the video semantic segmentation model of this proposal.

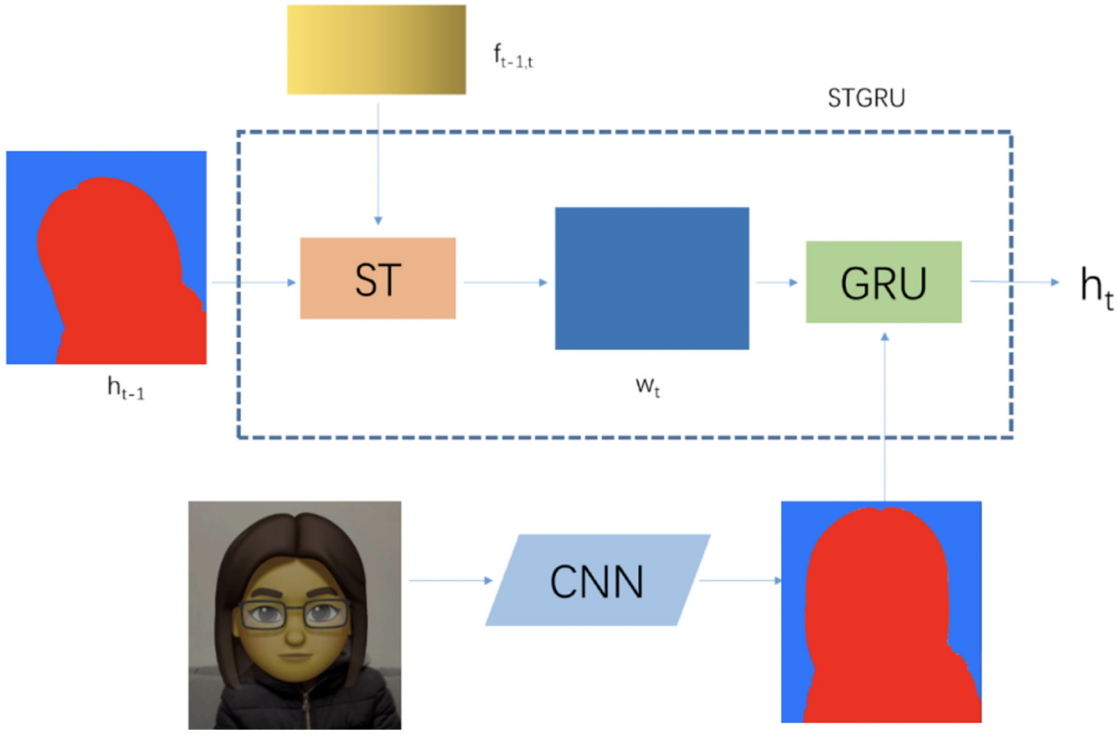

Figure 1 - Video Semantic Segmentation Model

Because people in video meetings do not move frequently and the difference between each selected frame may not be very large, every frame in a video may not need to be processed. Rather, it may be suitable to focus on the difference between time-sequence images.

There are typically two classes in a video, people and background. These classes may represent the basic unit for the model of this technique as illustrated in Figure 1. For the model, inputs can be listed by time such that at a moment in time (t), a raw image (I_t) can be sent to Convolution Neural Networks (CNN) for feature extraction. Meanwhile, an output (h_{t-1}) of a last moment in time is also sent to the unit to warp the output with an optical flow (f_{t-1}), which is the difference between two adjacent moments in time.

The warping operation is referred to as Spatio-Temporal Transformer (ST) in which the output of ST (W_t) is fed to Gated Recurrent Units (GRU) that have the capacity to process current information according to past and future information. Thus, two feature maps of different moments in time can be fed into the GRU such that the output (h_t) of the GRU is a prediction of the current moment in time.

The CNN can be implemented using a Pyramid Scene Parsing Network (PSPNet) to extract features from input images, as shown in Figure 2, below.

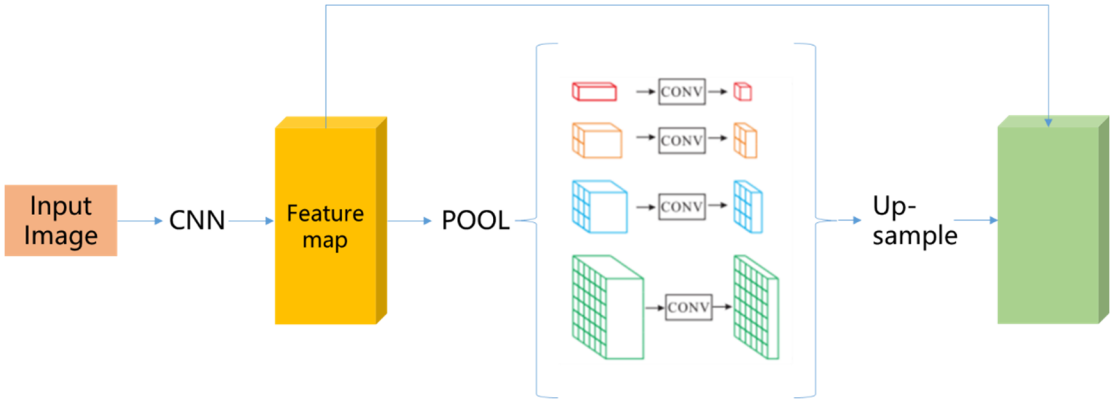

Figure 2

The PSPNet shown in Figure 2 utilizes a structure referred to as a pyramid module, which provides additional contextual information for the network, thereby providing relatively high accuracy for the PSPNet. Figure 3, shown below, illustrates a structure of the model for the technique of this proposal.

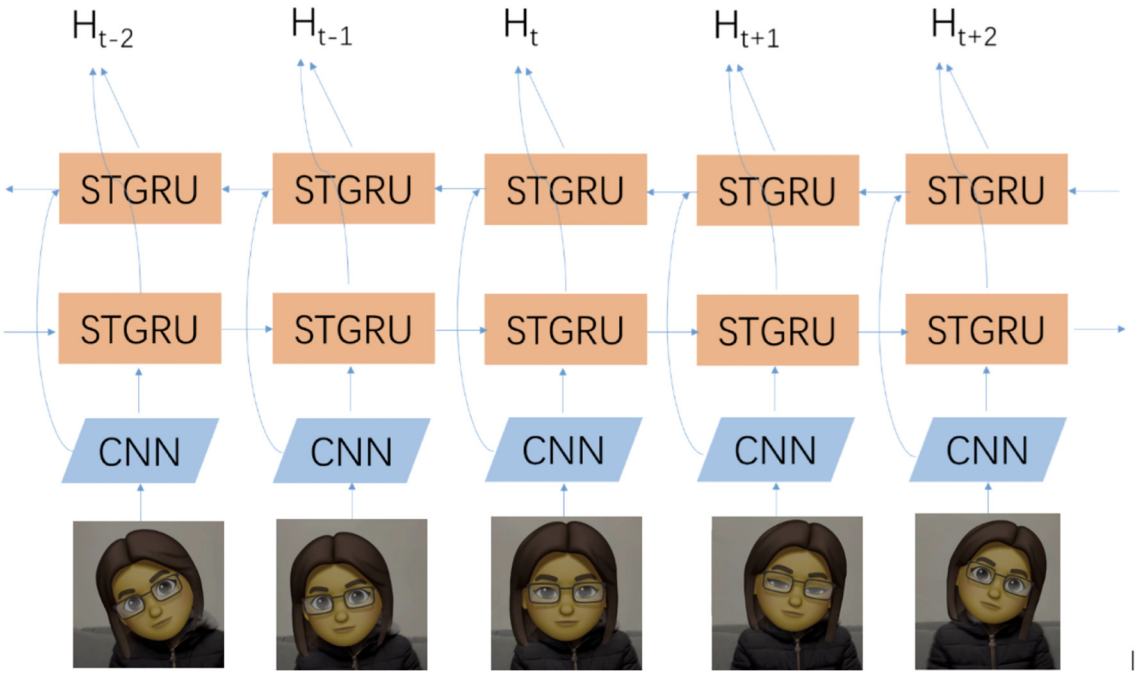

Figure 3

As illustrated in Figure 3, images within a time-sequence can are fed to the CNN. Forward and backward STGRU is applied so that the network can justify current outputs not only according to past information, but also to future information. The output of each moment is concatenated by the outputs of two STGRUs. After the online meeting tools

classify backgrounds, meeting participants can change the private background to any picture they may desire.

The model can be trained and tested using video from one or more actual meeting videos. For each video, one image may be intercepted every 500 milliseconds. The proportion of training images and testing images may be a 4:1 ratio. When making labels for a video, the background in the dataset may be set to be red (including people that may not be attending the meeting but may be shown in the video) and meeting attendees may be set to another color.

In at least one implementation, the model may be trained with a Mean Square Error (MSE) loss, as shown in Equation 1 (Eq. 1) as follows:

$$Loss = \sum_{i=j=1}^{a,b} (X_{i,j} - Y_{i,j})^2 \qquad \text{Eq. 1}$$

For Eq. 1, variables 'a' may be the length of an image, 'b' may be the width of the image, 'X' may be the serial number of a predicted class, and 'Y' may be the serial number of the true class. In at least one implementation, a Gradient Descent Optimizer may be selected as the optimizer in which the learning rate can be set to 2*10e-11 and the momentum can be set to 0.95.

Once trained, a checkpoint (.ckpt) file can be obtained that contains optimal parameters for the model and the model may be deployed (e.g., in a cloud, etc.). Because the difference between each frame may not be very large, online meeting tools may be capable of classifying a video background in a short period of time such that a background classified area can be covered with a desired picture in real-time.

In summary, this proposal provides a technique to change an online video meeting background in real-time using deep learning. Various advantages may be provided by the technique of this proposal. For example, the technique may provide for the ability to protect the privacy of meeting attendees. Further, a higher classification accuracy may be provided for the technique through the use of STGRU. Moreover, the model provided by

this technique may have a faster calculation speed through the utilization of past and future information.