We propose a new visual tracking algorithm leveraging double-level visual attention to full use of the information during tracking. In our tracking framework includes spatial attention and channel-wise attention. Considering that deep features of different levels may be suitable for different scenario, we propose to train an attention network in the off-line stage to facilitate feature selection in online tracking. Different from the image classification task, background clutter is more complicated in the tracking task. Thus, we purify the features by attention and channel-wise attention to effectively suppress the background noise and highlight the target region.

Introduction

Object tracking is a fundamental problem that is widely concerned in the field of computer vision today. It is widely used in many fields. Although the tracking field has made significant progress in recent years, there are still many problems in the tracking field that need to be solved.

2. Related works

The proposed tracking algorithm consists of four components: a feature extraction, spatial attention, channel-wise attention and a tracking module for target localization.

Spatial attention

In general, in order to distinguish the target from the background, only partial regions of the image should be focused. Therefore, applying a global image feature map may lead to tracking drifts due to the interference of superfluous regions. Instead of treating every region of the image equally, spatial attention mechanism gives higher weights for the valuable regions of the task. ….. To compute the spatial attention, we first apply average-pooling and max-pooling operations along the channel axis and concatenate them to generate an efficient feature descriptor. Applying pooling operations along the channel axis is shown to be effective in highlighting information regions. On the concatenated feature descriptor, we apply a convolution layer to generate a spatial attention map  which encodes where to emphasize or suppress. We describe the detailed operation below.

We aggregate channel information of a feature map by using two pooling operation, generating two 2D maps:  and. Each denotes average-pooled features and max-pooled features across the channel. Those are then concatenated and convolved by a standard convolution layer, producing our 2D spatial attention map. In short, the spatial attention is computed as:



where  denotes the sigmoid function and  represents a convolution operation with the filter size of .

Channel-wise attention

As mentioned in [10], CNN filters can be seen as pattern detectors. Some are sensitive to color information while others may be strongly responsive around object edge. We aim to select certain filters which fit well with the current tracking. To achieve this, we introduce a channel-wise attention mechanism to dynamically choose effective channels, since each channel of CNN features is exactly a response activation of the corresponding convolution filter. Give a visual image feature , we first apply the global average pooling on the visual image feature V to obtain the channel feature v:



Where.

Where aggregate spatial information of a feature map by using average-pooling operation, generating one spatial context description: which denote average-pooled feature. Descriptors is then forwarded to a shared network to produce our channel attention map. The shared network is composed of multi-layer perception (MLP) with one hidden layer. To reduce parameter overhead ratio, the hidden activation size is set to, where r is the reduction ratio. After the shared network is applied to descriptor. In short, the channel attention is computed as:



wheredenotes the sigmoid function,