MSAI 495 MP7: Template-matching based Target Tracking

Marthinus Johannes Nel

Template-matching based Target Tracking is a computer vision technique used for object tracking in video streams. A template image is defined to represent the object to be tracked. The template is then compared with each frame of the video stream to find the best match, using a correlation-based approach. Once the best match is found, the location of the candidate within the frame is determined, and the object is tracked in subsequent frames. The template is updated after each frame with the new candidate.

Template-matching based Target Tracking is widely used in computer vision applications such as surveillance, autonomous driving, and robotics. However, it has limitations such as being sensitive to variations in lighting and background, and requiring a good initial template for accurate tracking.

The initial template/target (region of interest) is selected in the first frame using ***'cv2.selectROI()'***.

Three algorithms were programmed namely:

1) **SSD (Sum of Squared Difference):** The SSD metric is computed as the sum of the squared differences between the pixels of a template region, T in the previous frame and a candidate region in the current frame, both defined within a bounding box of fixed size (Figure 1). The algorithm minimizes the SSD by exhaustively searching the candidate regions within a search window with a step size, updating the bounding box with the region that provides the minimum SSD.

1. SSD: sum of squared difference

$$D = \sum_{u,v} [I(u,v) - T(u,v)]^2$$

Figure 1: Sum of Squared Difference equation

2) **CC (Cross-Correlation):** The CC metric is computed as the sum of the product between the pixels of a template region in the previous frame and a candidate region in the current frame, both defined within a bounding box of fixed size (Figure 2).. The algorithm maximizes the CC by exhaustively searching the candidate regions within a search window with a step size, updating the bounding box with the region that provides the maximum CC.

2. CC: cross-correlation

$$C = \sum_{u,v} I(u,v)T(u,v)$$

Figure 2: Cross-Correlation equation

3) N**CC (Normalized Cross-Correlation):** The NCC metric is computed with the equation on Figure 3. The algorithm maximizes the NCC by exhaustively searching the candidate regions within a search window with a step size, updating the bounding box with the region that provides the maximum NCC.

NCC: normalized cross-correlation

$$\hat{I}(u,v) = I(u,v) - \bar{I},\ \hat{T}(u,v) = T(u,v) - \bar{T},$$

where $\bar{I}$ and $\bar{T}$ are the average intensity of I and T.

$$N = \frac{\sum_{u,v} \hat{I}(u,v)\hat{T}(u,v)}{\sqrt{\left[\sum_{u,v} \hat{I}(u,v)^2\right]\left[\sum_{u,v} \hat{T}(u,v)^2\right]}}$$

Figure 3: Normalized Cross-Correlation equation

All three functions takes a video, a HSV video, the size of the bounding box, the center of the object to track in the first frame, and other parameters. They return a list of BGR images with the bounding box on the tracked object.

Results:

The result video for SDD displays good initial tracking with the person moving side ways, closer to the camera and further away. SSD is unable to track the person after she fully rotates, with the back of their head facing the camera, and starts moving side ways. Later in the video the person's face passes the bounding box and SSD latches on to the person's face and follows it. When the man's face passes the bounding box SSD starts following the man's face.

The result video for CC displays the bounding box rapidly jumping around and almost never stays 100% on the person's face. This can be due to the face that CC only relies on maximizing the sum of the template with a candidate and thus any are with HSV values greater than the face/template will result in a larger value.

The result video for NCC is similar to that of the SDD, but NCC is more accurate in the case where the woman rotates her head and moves the NCC algorithm is still able to somewhat track the woman's head for a short while.

The algorithms are thus not robust to illumination changes, occlusion, large motions, large changes in scale, and other factors, but the NCC algorithm is more robust than the SSD and CC algorithms.

From the result videos it is evident that NCC is the most accurate, but also the must computational intensive taking on average 30 times longer to process the video than SDD and CC.