MP7 Template-matching based Target Tracking

In this homework, sum of squared difference (SSD), cross-correlation (CC), and normalized cross-correlation (NCC) are implemented to track a girl's head in a sequence of images(video). To implements all three functions, a structure is required before implementation.

The function needs to initialize the area of interest at the beginning to indicate the girl's face in the first frame of the video. This step is done manually. The bounding box in the image indicates the area of interest to show the girl's face location. Then three functions will take this frame, the next frame, and the bounding box information to search for the area in the next image which is similar to the area of interest in the current frame. In this homework, the search method is the local exhaustive search, which depends on the previous bounding box. The search region in the next image has the same centroid as the bounding box, but the area is doubled. Then the function goes through all the possible regions in the search area and compares them with the area of interest in the previous image. Three different image matching methods are used in this step:

Sum of squared difference (SSD):

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

where T (u, v) is the template (i.e., the target region in the previous frame), and I(u, v) is the matching candidate. All the possible areas are scored and the function will select the area with the lowest score as the next image's bounding box location.

cross-correlation (CC):

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

where T (u, v) is the template (i.e., the target region in the previous frame), and I(u, v) is the matching candidate. All the possible areas are scored and the function will select the area with the highest score as the next image's bounding box location.

normalized cross-correlation (NCC):

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

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

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

All the possible areas are scored and the function will select the area with the highest score as the next image's bounding box location.

Then the function will take updated bounding box information to predict the bounding box for the next (third) image until loops through all the images. In the end, the function will output all the images with corresponding bounding boxes to a video output.

The result should be able to locate the girl's face in each frame. However, all the methods cannot perfectly track the face from beginning to end. When the girl starts to turn around, none of the functions can still keep on track perfectly. The NCC has the best performance and it can track half of the girl's face for the most of time, but it still lost track for about 24 frames. The CC can generally mark the face location, but the bounding box it generated always jumps around. It may be caused by the multiplication in the function which could return a high value when some area time with other area than it times itself. It has around 238 frames that do not correctly indicate the girl's face. The SSD is smoother than CC, but it has around 250 frames that are tracing the background objects.