


Single Object Tracking In A Video

Anal Bera,
Computer Science(II),
Roll - B1930068

A decorative blue wavy bar at the bottom of the slide, consisting of a solid blue area with a white wavy line separating it from the white background above.

Introduction:

This project is focused on detecting and tracking an object in a particular video. Visual tracking can be defined as the process of detecting the position of an object in each frame of a video. Visual tracking has many real life applications in security, surveillance, traffic control, etc. Once an object can be tracked it can be labelled and more information can be extracted.

Tracking Algorithm(Brief)

The tracking algorithm is briefly presented in this section. There are two main components to the algorithm, initialization and tracking. For initialization purposes, an object is selected using either the first few frames for a simple version of the tracker or the first frame for a more complex version of the tracker. The object is cropped and centered to initialize the filter. The initialized filter is then correlated with a tracking window in the video to find the new location of the object. In a complex version, the tracker updates the filter as it tracks.

An object is selected by a user by clicking on its center for initializing the filter. A bounding box is drawn around the object after selecting an object for labeling. The bounding box represents the template during initialization and the tracking window during tracking. The template is cropped to initialize and update the filter during tracking.

Challenges in Visual Tracking

Visual tracking is difficult due to variations in the appearance of an object and its surroundings in a video. Object appearance changes due to pose, lighting, natural variations or non rigid transformation. Objects may be occluded or move out of the frame of the video. These variations make tracking difficult.

When an object being tracked is occluded or moves out of the frame, the tracker might lose the object and stop tracking, or the tracker might start tracking an object that resembles the original object. The tracker may stop updating a filter once the object is no longer in the frame or is occluded, but it may resume tracking and updating the filter once the object reenters the field of view.

Terminologies

Object/Target: An object is a person or thing which has definable characteristics. An object in this tutorial is referred to a physical entity to be tracked. Sometimes, the object is also referred to as a target.

Template: A template is a sub-image cropped from a larger image. The template is used as an input image to initialize the filter. An image of an object is present in the template. Also it may be referred to as correlation template.

Correlation: Correlation is a measure of similarity between two images. It is the sum of pairwise products of corresponding pixels of two images.

Template matching uses correlation of a template and an image to find the location of an object in the image. The template is moved to different locations in the image and the position of the best match, highest correlation is found.

Filter : Typically a filter is just a sub-image and the act of filtering is the process of placing the filter over the image at different locations and computing the pairwise-sum of the filter and corresponding image pixels. When the filter is applied across an entire image the result is itself a new image, i.e. a filtered image.

Tracking window : A tracking window is a sub-image in which the tracker looks for the new location of the object. The tracking window is retrieved from a frame of the video. The tracking window correlates with the filter to give the new location of the object being Tracked.

Synthetic target: A synthetic target is a synthetically generated image with a Gaussian peak at the location of the object to be tracked. A synthetic target is used to map the input image to its corresponding correlation output to generate a filter.

Occlusion: An occlusion is caused by an object which blocks the view of another object in a video.

Tracking: Tracking is an action where the algorithm finds the new location of an object in all the frames of a video over time.

Initialization: Here initialization is a process where the filter used for tracking an object in a video is generated. In this tutorial, initialization is also referred to as training.

Updating: In this tutorial, updating is a process where a filter is updated with the new information about the object, for example change in the pose of a person or change in the scale of an object is a new information.

Comparing Image To a Template

A simple technique known as template matching can be used to find similarity between a template and an image by comparing their pixels. A dot product compares the pixels between a template and an image.

In template matching, correlation is used to slide a template over an image to find every possible alignment of the template over an image. The pixels for every possible alignment are compared by computing a dot product to find a similarity score.

Starting at the top left corner, the template is moved by one pixel at a time from left to right and top to bottom. Correlation produces a new image with the dot product for every alignment. The peak in the correlation image is used to find the location of the template in the image.

Preprocessing

The tracking algorithm has two main components, initializing the filter and tracking the object. A preprocessing step is performed on every frame of the video before initialization or tracking.

- (1) A template centered on the object is cropped from the frame of a video.
- (2) The template obtained in the previous step is converted from its color to a gray scale image.
- 3) A log transformation is performed on the template obtained from the previous step by using the following equation,

$$x = \ln(y + 1)$$

In Equation, x and y are the pixel values of the output and the input images respectively. The log function is applied to reduce lighting effects and enhance contrast, making high contrast features available for the filter to initialize on.

- (4) The pixel values of the template from the previous step are normalized to get a mean of zero and a normal of one. The normalization helps in reducing the effects of change in illumination and maintaining a consistency in illumination between the different frames of the video.

(5) The template obtained from the previous step is converted from the spatial domain to the Fourier domain. The Fourier transform is used to decompose a signal into its sine and cosine components. An image is not an analogue signal therefore, discrete Fourier transform (DFT) is used for an image. The output of this transformation represents the image in the Fourier or frequency domain. The following equation is used for the DFT in two dimensional image,

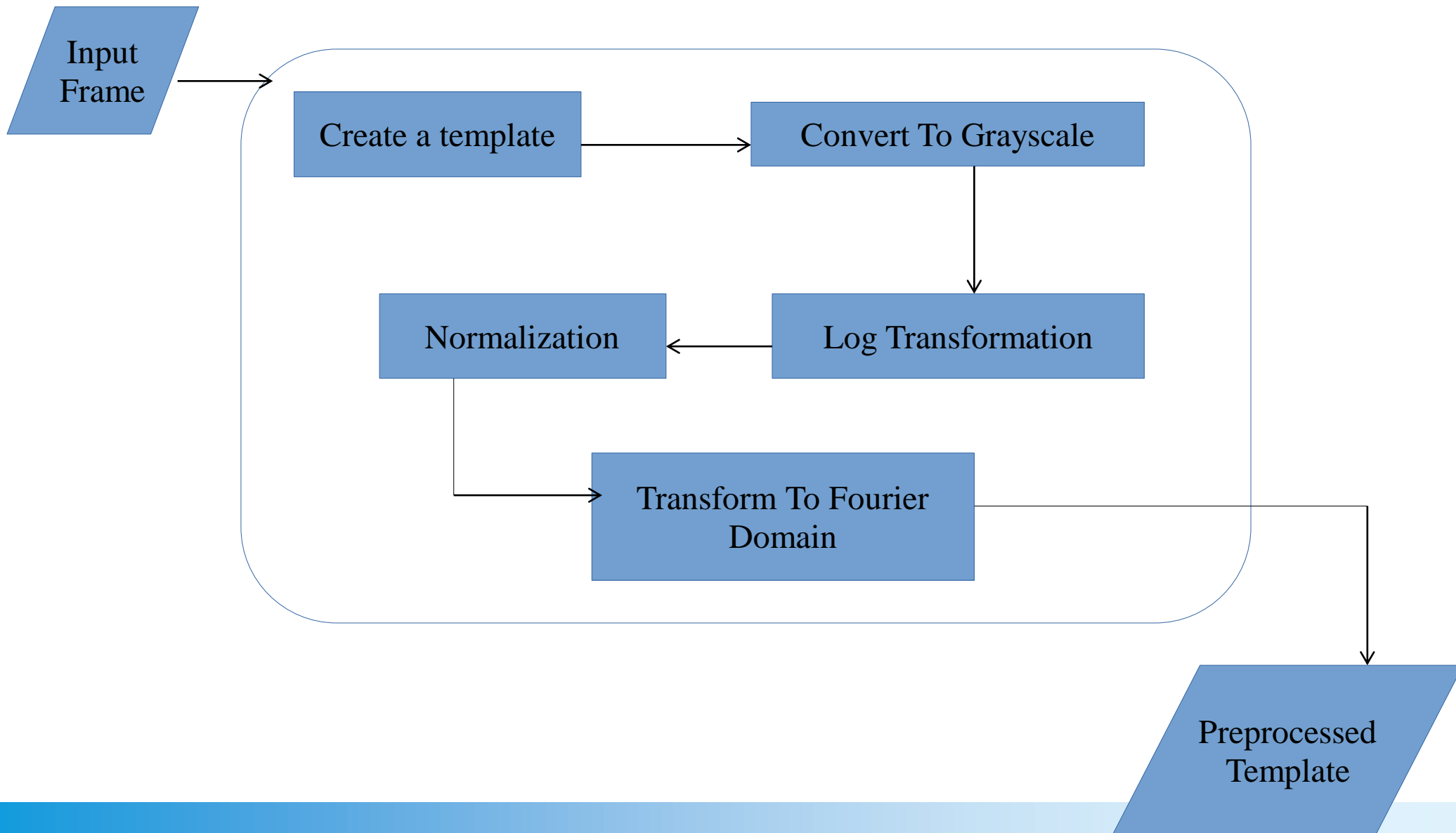
$$F(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) e^{-2\pi i \left(\frac{xu}{M} + \frac{yv}{N} \right)} \quad u = 0, \dots, M-1; \quad v = 0, \dots, N-1$$

In the above equation, $F(u, v)$ represents the image in frequency domain and $f(x, y)$ represents the image in the spatial domain. The following equation is used to convert from the frequency domain to the spatial domain known as inverse DFT.

$$f(x, y) = \frac{1}{NM} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} F(u, v) e^{2\pi i \left(\frac{xu}{M} + \frac{yv}{N} \right)} \quad x = 0, \dots, M-1; \quad y = 0, \dots, N-1$$

In the above equation, $F(u, v)$ represents the image in frequency domain and $f(a, b)$ represents the image in the spatial domain.

Flowchart Of The Preprocessing



Synthetic Target

A synthetic target is a synthetically generated desired correlation output. A synthetic target image contains a Gaussian peak centered on the object. The synthetic target is used in the initialization of the filter and for updating the filter during tracking by mapping the template to its desired output. The following equation is used to synthetically generate the synthetic target image,

$$g_i = \sum e^{-\frac{(x-x_j)^2+(y-y_j)^2}{\sigma^2}}$$

In above equation, g_i is the synthetically generated target. x and y represent the location of pixels in the image. x_j and y_j represent the location of the center of the object to be initialized on. The radius of the peak is specified by σ .

The Filter Used

The filter minimizes the sum of squared error between the actual output of the correlation and the desired output of the correlation . The filter uses a set of image pairs for initialization. First a template is cropped, then it is preprocessed and converted to the Fourier domain. After that a synthetic target image is created and converted to the Fourier domain.

The filter is initialized using the following formula,

$$H^* = \frac{\sum_i G_i \odot F_i^*}{\sum_i F_i \odot F_i^* + \epsilon}$$

Here, H^* is a complex conjugate of the filter. F_i represents the preprocessed cropped template in the Fourier domain for the i th frame of the video. G_i is the synthetic target image in the Fourier domain for the i th frame of the video. F_i^* is the complex conjugate of F_i . An element-wise multiplication is denoted by the symbol \odot . ϵ is the regularization parameter.

Mathematics Of The Tracker

A filter initialized using one template can track a simple object for a small number of frames. The filter has to be initialized on a set of templates for efficient and continuous tracking. The tracker uses the first few consecutive frames of the video to initialize a filter. The preprocessed cropped template in the Fourier domain, F , is obtained previously. The synthetic target in the Fourier domain, G , is also obtained previously. Now, F and G are used to initialize the filter,

$$N_i = \eta(G_i \odot F_i^*) + (1 - \eta)N_{i-1}$$

$$D_i = \eta(F_i \odot F_i^* + \epsilon) + (1 - \eta)D_{i-1}$$

In above equations, F_i is the i th preprocessed template in the Fourier domain for the i th frame in the video. G_i is the i th synthetic output image in the Fourier domain for the i th frame in the video. The learning rate is represented by η . The learning rate lies between 0 and 1. The symbol $*$ represents the complex conjugate. The symbol \odot represents element-wise multiplication. The regularization parameter is represented by ϵ .

The cumulative values of N_i and D_i over a set of initial frames is used in the following equation to initialize the filter:

$$H_i^* = \frac{N_i}{D_i}$$

In above equation, H_i^* denotes the complex conjugate of the filter. Once the filter is initialized, the next frame of the video, becomes the current frame for the tracker. A tracking window is cropped from the current frame of the video with the tracking window centered on the position of the object in the previous frame. The tracking window is preprocessed and transformed to the Fourier domain. The tracking window is then multiplied with the filter to get the new position of the object in the current frame using the following equation:

$$G = H^* \odot F$$

In above equation, F is the cropped and preprocessed tracking window. The size of the template, the tracking window, and the filter is the same so that element-wise multiplication can be performed.

Tracking Window Update

Tracking without updating the tracking window gives a basic tracker that is limited. If the object moves out of the tracking window, the tracker loses the object and begins tracking the most similar object inside the tracking window. This is not the desired result, therefore the position of the tracking window should be updated to continue tracking the object. When the tracker starts tracking after initializing the filter, a new position of the object is obtained for every frame. For a moving tracking window, the center of the tracking window is updated with the new position of the object obtained for every new frame. The new $x - y$ coordinates of the object becomes the new center of the tracking window.

Affine Transformations

An option for initializing the filter is to use affine transformations of a single template to create more training samples. Affine transformations such as scale and rotation can be applied to the template obtained from the first frame of a video. The affine transformations are used to initialize the filter.

Affine Transformation Mathematics

An affine transformation has 6 degrees of freedom. It includes scale, rotation, translation and shear. For tracker, scale, rotation and translation are used. The general affine transformations can be applied using the following equation:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

In above equation, x and y represent the coordinates of the pixels on an image, u and v represent the new coordinates of the pixels after transformation. The values of a through f are used to perform different affine transformations.

Translation

The coordinates of an image can be translated using the following equation. In the equation, t_x and t_y represent the number of pixels by which the x and y coordinates of an image are translated.

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

Scaling

The following equation is used to change the scale of an image. S_x and S_y represent the magnitude by which the x and y coordinates of an image are scaled respectively. The origin for changing the scale of an image is the top left corner of the image.

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

Rotation

The following equation is used to rotate an image around with Theta angle.

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

Problem Faced

The tracker is detecting and tracking a single object. For frames at beginning it is working good but for the intermediate frames it is facing problems.

Further work , is to work on this problem and optimize the tracker more to track two or more Objects.

Thank You