Video Stabilization with respect to Object

Project by: Lakshay Tyagi

Under Guidance of: Prof. K S Venkatesh

[Indian Institute of Technology, Kanpur]

Abstract

The goal of the project was to develop a Computer Vision Algorithm which can track particular objects and stabilize the video with respect to the object. This problem is a variation of the Video Stabilization Algorithm with the constraint that the video of the main object should remain stable

Introduction

Video Stabilization is a video enhancement technique which aims at removing shakiness from videos. The focus of this project is a variant of this problem which may be considered as Video Stabilization with respect to an object. To understand this problem better first a definition of an object is needed. We define an object as an instance of some semantic object which may be judged to be an individual entity in the video. Thus, our goal is to stabilize the video with respect to this entity which may move with respect to the background of the video. Hence, the overall goal of the algorithm is to transform the video in such a way that this object which is of concern to us remains stable or does not move.

Video used for experiments

The video used for the experiments has been taken from [3] and can be viewed on this Link . The object we wish to track in this video is the person's face. Which moves across the video due to camera motion.

Proposed Algorithm

The overall Problem can be split into multiple steps as follows:

Object Detection and Tracking:

YOLO Based [4]:

- This involves first detecting the object of concern and tracking it throughout the video. A naïve and easy way to do this is by running deep learning-based object detection algorithms like YOLO on the entire video. The result of running YOLOv5 on the entire video can be seen in this Link .
- This results in bounding boxes which can be tracked throughout the video. A simple way to convert these bounding boxes into the object's trajectory is to take the centre of these bounding boxes as the trajectory of the object through the video.
- However, we find that for the purpose of tracking these bounding boxes fluctuate a lot and the resultant tracking is not accurate. This is demonstrated in this Video where the centres of the bounding boxes predicted by YOLO are shown by a white dot.
- For a good tracking these dots should remain somewhat stable with respect to the face. However, this is far from true as the white dot which indicates the tracked centre moves all

over the face indicating a bad tracking. This along with the higher resource consumption in terms of needing GPU support makes this a poor method.

Filter Based Tracking [6]:

- Filter Based Tracking Algorithms offer an alternative to YOLO based tracking. They are faster than YOLO and do not require GPUs to run. The following three filter-based tracking algorithms were tried on the problem:
    - MOSSE (Minimum Output Sum of Squared Error): Very Fast, Low Accuracy
    - KCF (Kernelized Correlation Filters): Relatively fast, Mid Accuracy
    - CSRT (Discriminative Correlation Filter): Relatively slow, High Accuracy
- Out of these CSRT was chosen as the final alternative for YOLO because our application requires decent amount of accuracy otherwise the object still appears to be unstable in the video.
- Even though CSRT is slower than KCF and MOSSE, it is still considerably faster than running YOLO on the whole Video.
- Filter Based Algorithms need an initial bounding box, we can get this by running YOLO on the first frame of the video. This entire tracking algorithm now is as shown in figure 1.1.

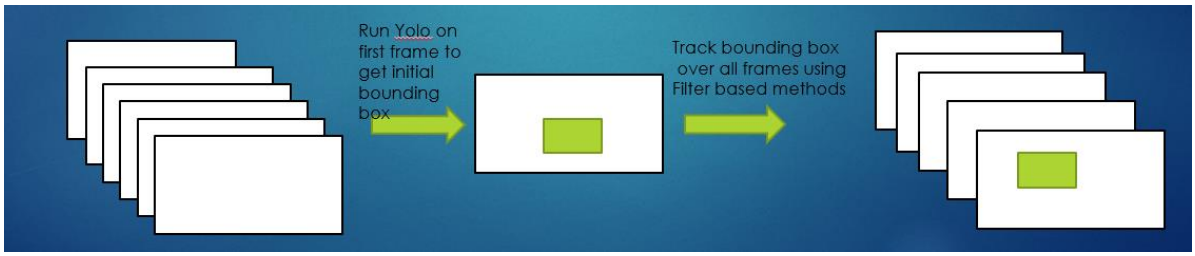

Fig 1.1 Modified Filter Based Tracking

- As we can see in this Video , Modified Filter based tracking is much more accurate than YOLO. The White dot which denotes the tracked centre is more or less stable with respect to the face which is the indication of a good tracking.

Stabilization with respect to object[1]:

The motion of the object in the video is assumed to be affine in nature or that it lies in a 2D plane. After this to stabilize the video with respect to the object we simply perform a affine transformation to put the Object at the centre of the video. The result of this using CSRT based tracking can be seen in this Video .

The missing area problem[1], [2]:

- Since when we perform an affine transform to put the object in the centre of the video, the part of the frame which is not known to us getting shifted into the video leads to a black missing area.
- This is seen in classical Video Stabilization problems as well. However, in this case due to large amount of object motion, sometimes around 50 percent of the frame is missing which means that the problem is much more pronounced here.

- There are two main ways to handle this problem:
    - Crop out the missing part:
      However, in our case sometimes around 50% of the frame is missing which will lead to a scenario in which we have cropped out the entire background other than the face!
    - Use Video Completion Techniques to fill in the missing Area. However, this only works when the part to be filled has been revealed in other parts of the video.

Video Completion Techniques:

Median Based Filling [1]:

- This is naïve approach in which we simply replace the missing areas by median intensity of transformed images taken over time. Median of intensity is taken over all possible transformed frames after removing the missing pixels and sorting them according to intensity. This median operation is over time and through this we obtain a single median frame. This static frame is the used for filling the missing areas leading to a static filling of these areas. This is shown in the figure below in Figure 1.2. The result of this can be seen in this Video .

Figure 1.2 Showing the method used in Median Frame Filling

- Median completion does not give good completion due to high variance in the pixels over time. This problem is more complex and better more complex methods need to be used

Deep Learning Based Completion [5]:

- The Deep learning-based approach specified in Learning Joint Spatial-Temporal Transformation Networks for Video Inpainting by Zeng et al is applied for this problem of Video Completion. The results are considerably better than Median Completion. The completed Video can be seen Here .
- However, the completion at the fringes tends to be poor. This is because we try to complete a very large percent of the frame for which the algorithm does not work well. These methods work well when we have to complete small portions like 10% of the frame.

Limitation of Video Completion Approaches:

- Video filling approaches can only work when the rest of the scene is revealed in the other frames i.e., if the other frames don't have the information to fill the frame it will not be possible to do so
- Hence, in videos where we have a moving object filling may not be possible. Since the camera being stable does not tell us about other frames.
- Hence, video completion works best for no object motion and only camera motion

Future Work

- Video Completion performance can be improved using Mosaicking Techniques which store the part of the video revealed in other frames which can be used later for completion.
- Iterative Completion approaches where we stabilize the video and complete alternatively can be experimented with to improve completion performance.

References

1 Full-frame Video Stabilization by Matsushita et al
2 Video stabilization: Overview, challenges and perspectives by Guilluy et all
3 Full-Reference Stability Assessment of Digital Video Stabilization Based on Riemannian Metric by Zhang et all
4 YOLOv3: An Incremental Improvement
5 Learning Joint Spatial-Temporal Transformations for Video Inpainting
6 https://www.pyimagesearch.com/2018/07/30/opencv-object-tracking/