Team 12 - Moving Object Removal from Images: Final Report

Nandita Lakshmi Tunuguntla
cs19btech11051@iith.ac.in
Arsalan Ahmed Sheikh
es19btech11025@iith.ac.in
Sharan Basav Patil
ma19btech11012@iith.ac.in
Spandan More
es19btech11009@iith.ac.in

December 6, 2021

Abstract

Knowledge regarding the background of an image has many applications in computer vision tasks. Removing moving objects that come in the way of these backgrounds is the problem we will aim to solve in this project. We will explore the conventional method of using a Median Stack Filter for this purpose and improve the existing methodologies using self-attention Vision Transformers. The code we have used is made available at https://github.com/cybars69/ivp-project

1 Introduction

Moving objects can often be seen as outliers in consecutive images with the same frame of reference, when compared to other objects from a scene. Removing these outliers is an image processing problem that is gaining importance these days because understanding scenes and backgrounds is essential for various computer vision tasks. If any foreground object is captured by accident or is decided to be unnecessary for the task, it is essential to erase it from the picture/sequence. One application of this is that with knowledge of the background precisely, background subtraction techniques can be applied for tasks like surveillance, human pose/gesture recognition and so on.

Although various methodologies have been proposed, state of the art techniques using deep learning and transformers can be used for improving this problem. In this project, we study the limitation of the conventional approach for this problem, the Median Stack Filter(MSF). We then develop techniques that utilize machine learning algorithms and models for improved results. We visually show the the resulting images from our experimentation and their difference with images from the Median Stack Filter approach.

2 Problem Statement

Moving objects in the foreground may interfere with the stationary background, the subject of importance, in consecutive images taken within a short period of time. Traditionally, this has been achieved by methods which do not employ machine learning techniques, like the Median Stack Filter approach where all images are stacked on top of each other and each pixel of the output image is obtained by taking the median value of the pixel at that corresponding location by iterating over all stacked images.

Speed of the moving objects in the scene affects the result of the filter, MSF does not perform well on images with slow moving objects. Thus, the aim of this project is to use MSF as a baseline method and develop our own pipeline which aims to develop techniques for removing moving objects in a sequence of images with the same frame of reference and a stationary camera to output a single image with the background extracted from the input images.

3 Literature Review

One problem with the median filter for images is their inconsistency in lighting due to reflections and changes in shadows. As Maiya et. al. [4] mentioned, this can be overcome by linear color transformations. For moving objects, LMedS can be utilized assuming the moving objects are comparable to the iterating window.

The main motivation behind this project is from Moving Object Removal[2]. FgSegNet, K-means clustering and LGTSM Deep Video Inpainting was the main pipeline followed which inspired us to use state of the art techniques to improve each component of the pipeline.

In general scenes with multiple objects, it is difficult for the generator to remove only the desired object while re-synthesizing the rest of the image exactly. Shetty et. al. [3] propose a two-stage generator method to tackle this: the first stage involving a mask generator which identifies the target class and the next stage being an in-painter, which takes the generated mask and the masked-out image as input and learns to in-paint to produce a realistic output. The second part of their solution is a GAN based framework to impose shape priors on the mask generator to encourage it to produce compact and coherent shapes. This motivated the idea of using neural networks and deep learning in our model.

Park et. al. [5] discuss the applications of Object Removal from a moving background in Film Post Processing. They first track all the background transformations, from which they create a mosaic image in which only the foreground appears to be moving. They achieve this using the RANSAC Algorithm for background feature matching, which is followed by extrapolating information from multiple frames for background completion.

The FgSegNet[9] is a encoder-decoder structure Deep Neural Network which does foreground segmentation by background subtraction. The salient Feature of this implementation is that it uses feature fusion for extracting mulitscale features.

Recently transformers are being explored in the place of neural networks. DINO[6], a self-supervised model inspired from natural language processing does both semantic segmentation (foreground and background separation) and k-NN classification with great results, both of which are a motivation to our solution methodology.

4 Proposed Approach

We will approach the problem statement of this project by further splitting it into 3 sub problems as seen below.

Our goal for optimization of the existing methodologies is to use machine learning techniques like deep neural networks and attention models for these sub problems and observe the results.

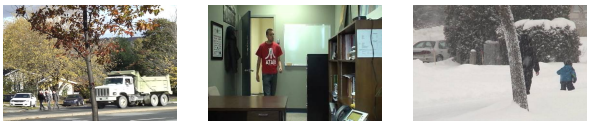
4.1 Image Segmentation

In the pipeline presented above, we mainly focused on improving the segmentation block as it influences the final output the most. In the methods that have been explored before[2] segmentation was done by using FgSegNet v2 which uses a lot of outdated libraries, hence we aimed to use state of the models to do the same task.

Image Segmentation has been evolving since new models have improved the precision of object detection and foreground identification. DINO is a one such vision transformer (ViT) that requires no supervision. DINO is a new state of the art self supervised method that we have planned to employ in our pipeline due to its efficiency and high accuracy of segmentation.

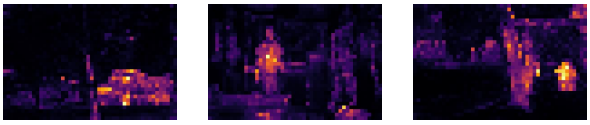
To demonstrate the attention maps obtained from DINO, we have used the CDNet2014 dataset [8] Figure 1 shows images corresponding to 3 categories of the dataset anf Figure 2 the corresponding DINO attention maps.

The above images are attention maps of individual images in the sequence. We have observed that all the objects, even those which are not moving are precisely being highlighted in the maps. To remove these



(a) Fall (b) Office (c) Skating

Figure 1: Images in Dataset



(a) Fall (b) Office (c) Skating

Figure 2: DINO Output

we plan on applying filters to the attention maps so that they highlight only the required moving objects.

One advantage that DINO give us over other methods of segmentation is the coloured attention maps. This way we can adjust the threshold while binarizing the attention map to a black and white mask to our convenience on which threshold gives a more precise final mask.

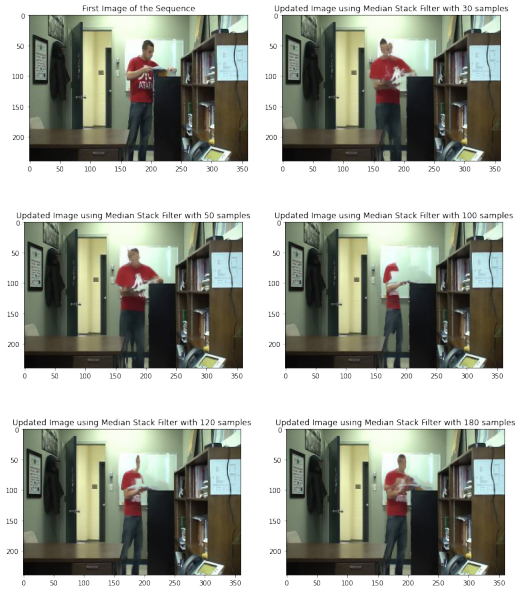
The DINO output masks are first converted to grayscale, after which the Otsu Algorithm is used to find a correct threshold, and the mask is then binarized. Due to the nature of DINO's energy maps, there are many minor blobs in the masks. We remove these by counting the contours in the image and removing those contours with area less than a specified number, which can be tweaked according to the mask. This mask is used in the remaining parts of the pipeline.

4.2 Object Removal using Filters

To begin with the core object removal method, we start with the implementation of the classic Median Stack Filter without any segmentation. The following are some resultant images obtained:

(a) Fall (b) Office (c) Skating

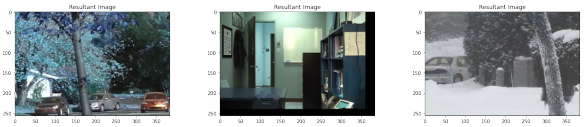
Figure 3: Resulting Images after MSF



Linear filters, like the mean filter, tend to blur sharp edges, destroy lines and other fine details, performs poorly on heavily tailed noise and signal dependent noise. Nonlinear filters, like the median filter, have edge preserving properties and are robust to impulsive noise, but still face issues in terms of preserving fine details, like lines and corners.

From experimental results in [12], the median stack filter has proven to show more satisfactory results in these areas. It is also proven that the stack filter algorithm has a much more efficient run time while maintaining better image detail.

However, it lacks in ability to remove slow moving objects from the scene. As seen in the examples above, a slow moving object (one that remains in the scene for a large number of images with respect to the total number of images in the sequence) does not get removed fully just by using the Median Stack Filter.



(a) Fall (b) Office (c) Skating

Figure 4: Resulting Images after Inpainting

We see that as the sample size of images increases, the object removal is better. However, this cannot be demanded in practical applications where video feed would be limited. Furthermore, as the number of images in the sequence increases, the time taken to run the function increases exponentially, since the time complexity of the algorithm is of the order O(ijkc), where i,j are the dimensions of the image, k is the number of images in the sequence, and c is the number of colour channels in the image.

As this is not sufficient, the next step would be to use the masks obtained from the DINO method and decrease the number of images used to get a better quality image when compared to using MSF with a large number of images.

4.3 Image In-painting

Image inpainting started with restoration of artifacts by hand. Now as we progress into computers, there are 2 main methods of inpainting that we use with pure image processing excluding machine learning.

Navier Stokes method was derived by Bertalmio et. al. [10] as they drew an analogy between image intensity function for the image in-painting problem and the stream function in a two-dimensional in-compressible fluid. This function can be approximated by numerically approximating the steady state solution of the 2D Navier-Stokes Equations.

Telea [11] described a method which is as effective as Navier Stokes method that does not require intensive calculations and complex methods called as the "Fast Marching Method" (FFM). This method derived a way to paint the points inside the boundary, according to the increasing distance from the boundary of the region to be in-painted.

5 Results and Discussions

From the resulting images seen above, we can say that our pipeline improves the baseline MSF process drastically. Here, we follow a subjective quality assessment. For 80 images, the baseline MSF takes 20 seconds to execute, whereas the pipeline takes only an additional 10 seconds but provides much better results. To achieve the same results using only MSF, we would need to use more images, which would increase the computation time exponentially.

A few improvements can also be made on the current pipeline. Currently, no modifications have been made to any of the pretrained models. However, if extra layers are added to the DINO model, the resolution and sharpness of the masks can be manipulated according to requirement. Moreover, the image binarization can also be done by using Deep Neural Networks modelled over the Otsu Algorithm which would provide a small increase in performance.

One part of the pipeline which would increase the final resultant quality drastically would be the inpainting block. Research has been made to replace the traditional Navier-Stokes method with Deep Neural Networks. One such CNN based approach involved the model being made to learn inpainting features rather than image content features [13]. Another paper includes a semantic image inpainting model, able to fill in large missing regions upto surprising accuracy given a dataset and a trained generative model, achieving pixel level photorealism [14].

Finally, we look at the core of the pipeline, namely the filtering process. The Median Stack Filter is limited by a high time complexity. This has the most potential to be replaced by a Machine Learning Model, as this would drastically improve the performance of the pipeline. The MSF Algorithm cannot be optimised any further. Hence, any improvements in this part of the pipleline involve a change of approach.

Overall, most of the possible improvements that can be made to the pipeline are improvements in the final image quality. The nature of object removal is specific, in the sense that any adjustments to improve quality will be made on a case specific basis. However, employing Machine Learning Models which have trained on diverse datasets will remove any need

of human intervention for different scenes that need object removal. Thus, a better mask generator paired with an inpainter which employ Machine Learning are the way forward.

6 Contributions

- Nandita Lakshmi Tunuguntla - Image Segmentation
- Sharan Basav Patil, Spandan More - Mask Postprocessing and Image Filtering
- Arsalan Ahmed Sheikh - Image Inpainting

7 References

[1] Simple algorithm to remove moving objects from pictures

[2] Victor Chen, Daniel Mendoza, Yechao Zhang: Moving Object Removal in Unlabeled Image Databases

[3] R. Shetty, M. Fritz, and B. Schiele, "Adversarial Scene Editing: Automatic Object Removal from Weak Supervision." Computer Vision and Pattern Recognition, Jun. 2018. arXiv:1806.01911

[4]M. Hori, H. Takahashi, M. Kanbara, and N. Yokoya. (2010). Removalof Moving Objects and Inconsistencies in Color Tone for an Omnidirectional Image Database. 6469. 62-71. 10.1007/978-3-642-22819-3 7.

[5] Soon-Yong Park, Chang-Joon Park, Inho Lee: Moving Object Removal and Background Completion in a Video Sequence

[6] Caron, Mathilde and Touvron, Hugo and Misra, Ishan and Jégou, Hervé and Mairal, Julien and Bojanowski, Piotr and Joulin,Armand:Emerging Properties in Self-Supervised Vision Transformers, Proceedings of the International Conference on Computer Vision (ICCV), 2021 arxiv:2104.14294

[7] Pritika Patel, Ankit Prajapati, Shailendra Mishra: Review of Different Inpainting Algorithms, International Journal of Computer Applications (0975 – 8887) Volume 59– No.18, December 2012

[8] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "Cdnet 2014: an expanded change detection benchmark dataset," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 387–394, 2014.

[9] L. A. Lim, and H. Y. Keles. "Learning Multi-Scale Features for Foreground Segmentation." Pattern Analysis and Applications, Aug. 2019. arXiv.org, doi:10.1007/s10044-019-00845-9.

[10] M. Bertalmio, A. L. Bertozzi, G. Sapiro, Navier-Stokes, Fluid Dynamics, and Image and Video Inpainting, IEEE CVPR, 2001.

[11] A Telea, An image inpainting technique based on the fast marching method , J. GRAPHICS TOOLS, 2004.

[12] Zhu, Li. "Application of Stack Filter on Image Processing." Applied Mechanics and Materials. Trans Tech Publications, Ltd., March 2014. https://doi.org/10.4028/www.scientific.net/amm.543-547.2163.

[13] Xinshan Zhu, Yongjun Qian, Xianfeng Zhao, Biao Sun, Ya Sun, A deep learning approach to patch-based image inpainting forensics, Signal Processing: Image Communication, Volume 67, 2018, Pages 90-99, ISSN 0923-5965, https://doi.org/10.1016/j.image.2018.05.015.

[14] Raymond A. Yeh, Chen Chen, Teck Yian Lim, Alexander G. Schwing, Mark Hasegawa-Johnson, Minh N. Do; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5485-5493