Ohio State University, Dept. of Computer Science and Engineering Technical Report OSU-CISRC-5/13-TR13
(original work completed December 2012)

Localizing and Removing Moving Objects in Aerial Video

James W. Davis & Kyle Salberg
Dept. Computer Science and Engineering
Ohio State University

1 Introduction

Given a sequence of images form aerial video, between frames, objects moving on the ground, typically cars, can be observed. Additionally, since the plane is circling the area, as the perspective changes, the appearance of tall structures in the scene will change relative to the ground. The goal of this project is to find the moving cars in the scene, while also distinguishing these from the structural movement from the changing viewpoint.

2 Methods Used

The project involves first pre-processing the input image sequence to register and normalize the brightness of the sequence. After pre-processed, three different methods were examined: using the median, background subtraction, and intrinsic analysis. Additionally, for the median and intrinsic methods, two different approaches were tested. The full pass approach utilized the entire sequence of data, while a temporal window approach only used a small window of frames around the frame of interest, such as the pre-processed examples frames in Figure 1, Figure 2, and Figure 3.

Figure 1: Time step 103114.

Figure 2: Time step 103115.

Figure 3: Time step 103116.

2.1 Pre-Processing

2.1.1 Registration

In order to perform any of the analytical methods applied for this project, the images in the sequence must be properly aligned. While the input data is already aligned to some extent, finer registration is required. To accomplish this, a single image from the sequence was selected, and good feature points, located throughout the image on the ground plane, were then identified within that image, shown in Figure 4.

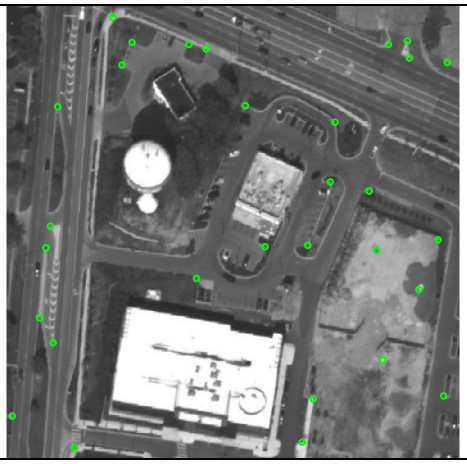

Figure 4: Base image for registration

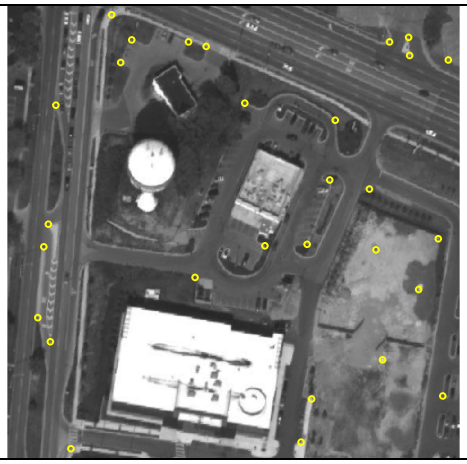

Figure 5: Input image to be registered

Given the base image and pre-selected feature points, for any input image to be registered, template matching is used to identify the corresponding features in the input image, resulting in the points displayed in Figure 5. This is accomplished using a small window around the feature point in the base image as the template, and searching for the best Normalized Cross-Correlation (NCC) match in a local region in the input image. After the matching points have been found, a projective transform is used to warp the input image to the base image, shown in Figure 6.

Figure 6: Registered input image result.

However, this can result in an uneven black border, depending on the images being registered, so in order to only work on the valid portion of the image, each image is also then cropped slightly to remove several pixels from the border, removing possible issues that would otherwise arise.

2.1.2 Brightness Normalization

To further ensure every image in the sequence is on the same scale, the brightness of every image is normalized to that of the base image. To accomplish this, the histogram of each input image is approximately equalized to the histogram of the base image.

2.2 Finding Cars

2.2.1 Median

The simplest approach attempted for finding cars is simply using the median. For the set of input images, whether that includes the entire pass of data or only several frames in a sliding window, the median is computed. On any road, if a given patch of road is empty the majority of the time, then the median result for that patch will be a road patch, even if there is a car in that space some of the time. If this assumption holds, then given a scene where cars are constantly moving, the median image will contain only the road, without any non-static cars. Therefore, we can obtain a weighting of likely moving car objects in a given input image, by taking the absolute difference of that image with the median.

2.2.2 Background Subtraction

As an alternative method, since the large sequence of images, after registration, all contain the same background scene, with only some noise in the form of cars and moving buildings, background subtraction can be applied to construct a model for the scene. The model for the scene is represented by the per-pixel mean and variance across the image sequence. The Mahalanobis distance can then be computed for each input image. If the variance of the roads is not too large, then cars on the roads should have a large Mahalanobis distance, so this resulting per-pixel distance can be interpreted as a weighting of likely car objects.

2.2.3 Intrinsic

The third approach attempted involves obtaining the intrinsic image, outlined in (Weiss, 2001). The essential idea is that an image can be decomposed into a reflectance image and an illumination image. The reflectance image contains only the underlying scene structure, common to any image of that scene. Thus, for a sequence of images of the same scene, there are an equal number of illumination images, each when applied to the core reflectance image, results in the original image.

For the problem of finding cars, within a small window of frames, the images all contain essentially the same underlying scene. The only differences between frames are the cars that have moved, and small shifts in the tall structures of the scene. Since the reflectance will only contain the parts of the scene common to the sequence, the cars and the shifts in the structures will be evident in the illumination image instead, appearing as either a shadow or a bright spot. From this illumination image, the global illumination of the image is removed by subtracting the

median of the illumination from it, and then taking the absolute value of that difference gives a weighting of where the moving cars are visible in the scene.

3 Problems Encountered

One noticeable problem encountered is that the tall structures of a scene move, even within a small window. The most noticeable issues arise around the edges of the structures appearing in the results of the windowed approach. However, this effect can be reduced by weighting against the edges of the underlying structure. For both the windowed median and intrinsic approaches, an estimation of the scene, which includes these edge structures but not moving cars, is obtained; the median and the reflectance images. The gradient magnitude of the median or reflectance image will give a weighting of edges in the underlying scene, which can then be used to suppress the edges in the result, while preserving the moving cars.

An additional problem encountered was that if the input images are not properly aligned, all of the above analyses will have bad results around the misaligned edges. This was solved by using template matching to automatically register all images prior to analysis, which is significantly more accurate than attempting to register images by hand.

4 Final Results

4.1 Median

Using the full pass approach, the median of the entire image sequence, Figure 7, does indeed not contain any cars visible on the roads. However, the tall structures of the scene are distorted as a result of significant movement over the course of the entire image sequence. Most noticeably, only the base of the water tower in the center of the image is present in the median. From this median, Figure 8 is obtained by taking the absolute difference from the median and the input image, Figure 2.

Figure 7: Median of full pass.

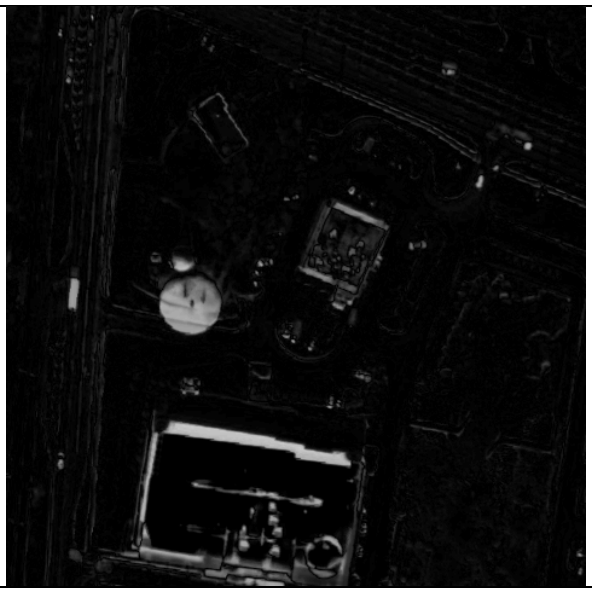

Figure 8: Full pass median result.

Using a windowed approach with Figure 1, Figure 2, and Figure 3, the median is shown in Figure 9. Because these three images are close together in time, and also because the cars on the ground move relatively more than the tall structures, the moving cars from the road are also removed, while the building structure is largely retained. However, there are two noticeable cases to be observed here. First, the cars in the top left corner are still visible in the median, because they are temporarily stopped over the duration of these input frames. Second, the there are two white patches on the left side on the road, where the truck was in the input images. These patches are caused by an overlap in the trucks location in this small windowed sequence. Even with these defects, the result, show in Figure 10, still gives a stronger weighting for the cars that were moving during this window, while also containing significantly less building structure than the full pass.

Figure 9: Median of 3-frame window.

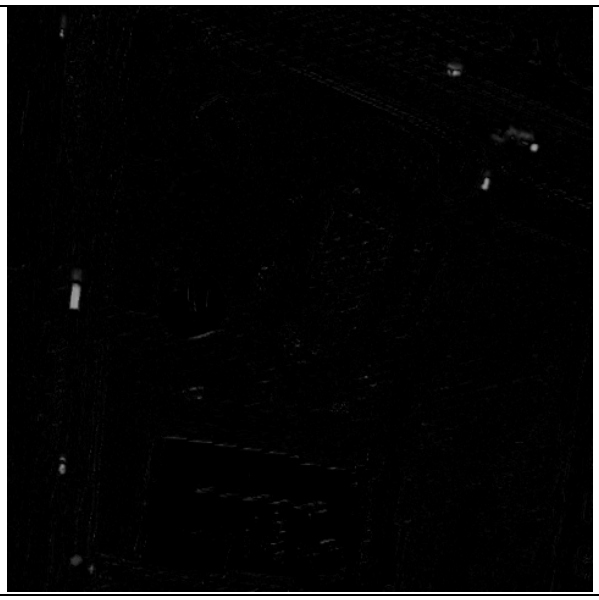

Figure 10: Absolute difference from median.

This result can then be further improved by weighting the result against the gradient magnitude of the median image, Figure 11, to suppress the response along the scene structure edges, shown in Figure 12.

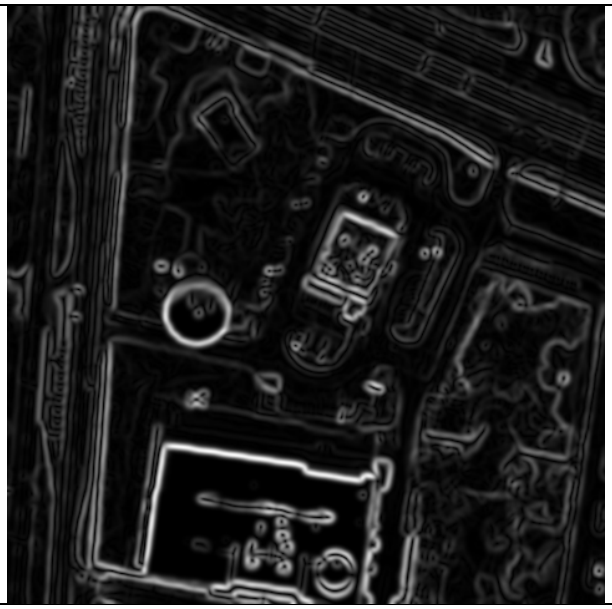

Figure 11: Gradient Magnitude

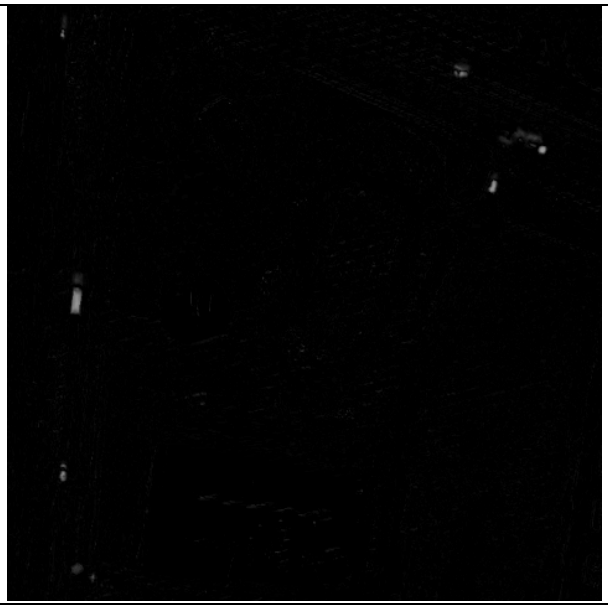

Figure 12: Result after edge suppression.

4.2 Background Subtraction

Since it is a statistical approach, background subtraction can only reasonably be applied to a large sequence of images. The mean and variance of the entire image sequence are shown in Figure 13 and Figure 14 respectively. The Mahalanobis distance, using this mean and variance image, gives a weighting of likely car objects.

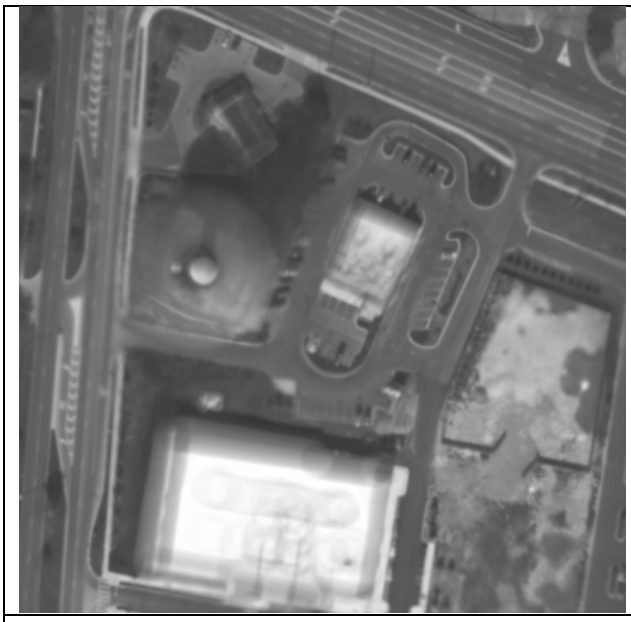

Figure 13: Full sequence mean.

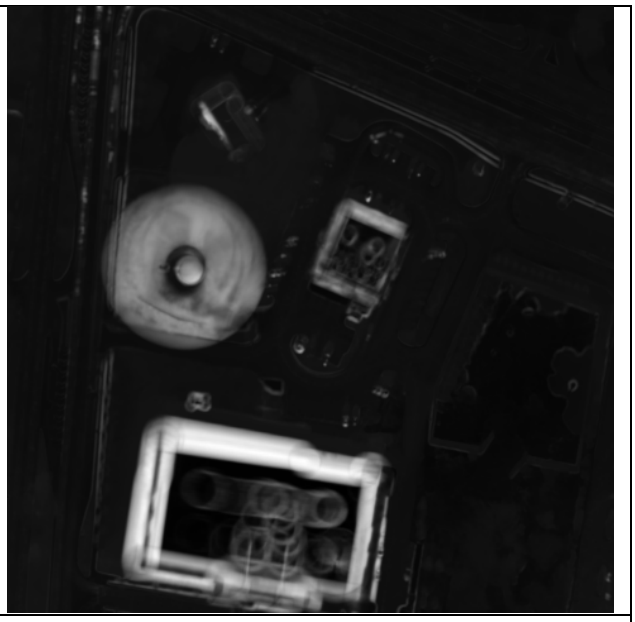

Figure 14: Full sequence variance.

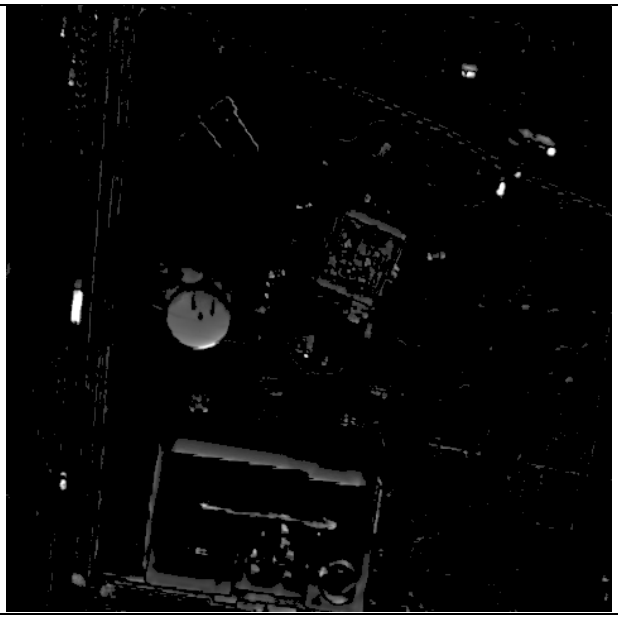

Figure 15: Background subtraction result.

For the image in Figure 2, Figure 15 is the resulting Mahalanobis distance, showing only values greater than one standard deviation and capped at five standard deviations away from the mean. Although the tower and building edges are still visible in the result, they are not weighted as heavily as the moving cars in the scene, which is due to taking into account the high variance around these buildings.

4.3 Intrinsic

Using the images in Figure 1, Figure 2, and Figure 3, the reflectance image in Figure 16 is obtained. From this reflectance image, and the input image in Figure 2, the illumination image, Figure 17, is also obtained.

Figure 16: Reflectance image.

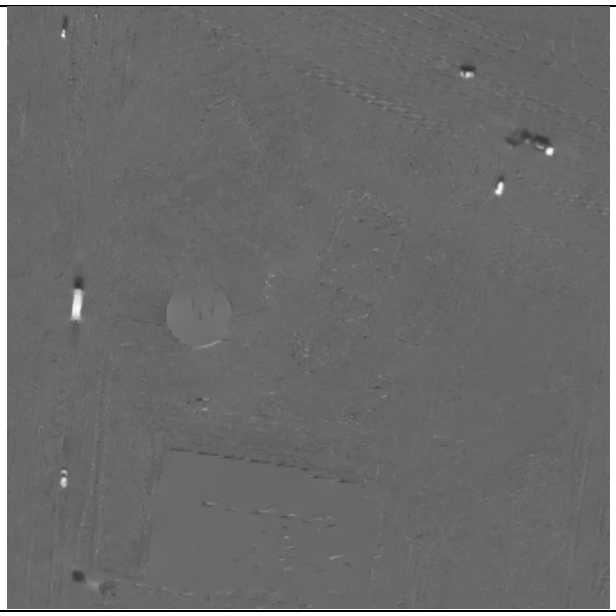

Figure 17: Illumination image.

The illumination image contains a global illumination of the image, shown by the gray level over most of the image, plus additional bright and dark spots, representing areas in the scene which are either darker or lighter than the global scene illumination. Taking the absolute difference of the illumination with its median gives the result shown in Figure 18. The result contains the car regions of interest, as well as some noise. Some of the noise can be removed through edge suppression using the gradient magnitude of the reflectance image, the result of which is shown in Figure 19.

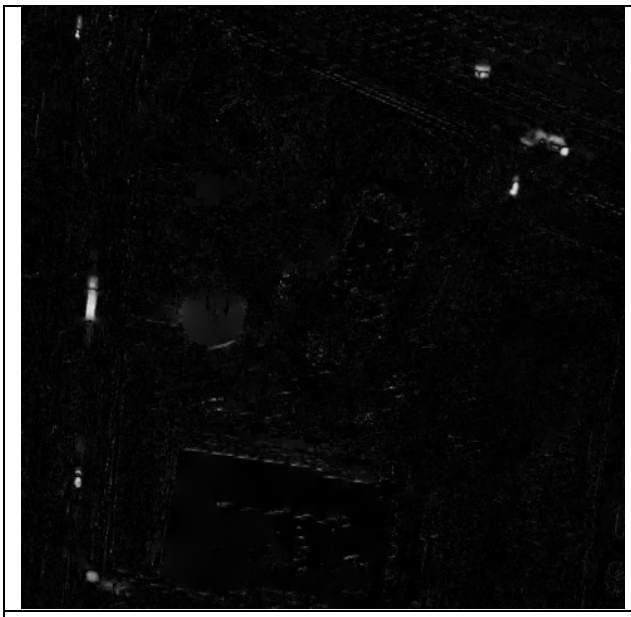

Figure 18: Median subtracted illumination.

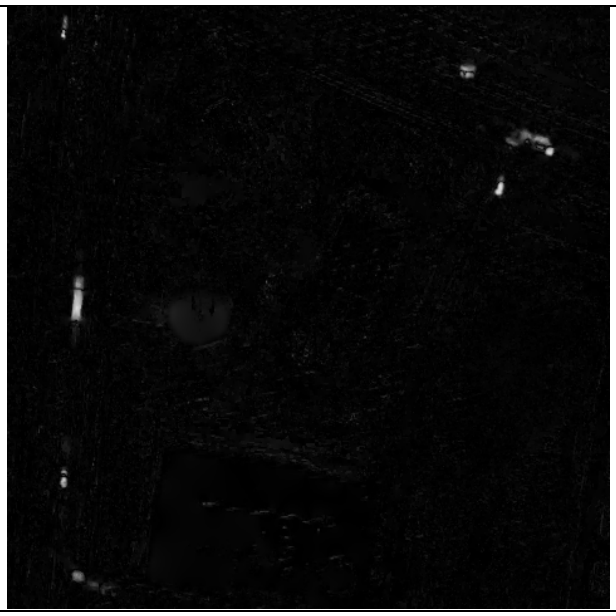

Figure 19: Result after edge suppression.

5 Conclusions

5.1 Full Pass Approach

Between the median and background subtraction for the full pass, background subtraction gives clearly better results, due to the added information provided by the variance of the sequence. In the results from the median approach (Figure 8), the cars and the building edges and water tower appear as roughly the same weighting, giving little information as to which is a car, and which is simply caused by the view change. In the background subtraction approach, Figure 15 clearly shows the known car objects, such as the truck in the left side of the image, as have a greater weight than the buildings.

5.2 Windowed Approach

Comparing the median to the intrinsic approaches when using a temporal window, the intrinsic seems to perform slightly better. In general, the strength of the car results is much stronger than the strength of the noise in the intrinsic results, when compared to the median results. This can easily be observed by comparing Figure 20 and Figure 21, which show the median and intrinsic results after a threshold has been applied. In the below figures, the yellow circle highlights the truck, which is a desired result, while the red circle shows the edge of a building, which is considered noise. In the median result, after this particular threshold, portions of the truck are already being lost, while there is still a reasonable amount of noise left. However, in the intrinsic result, the entire truck can still be retained, while at the same time, containing significantly less noise than the median result.

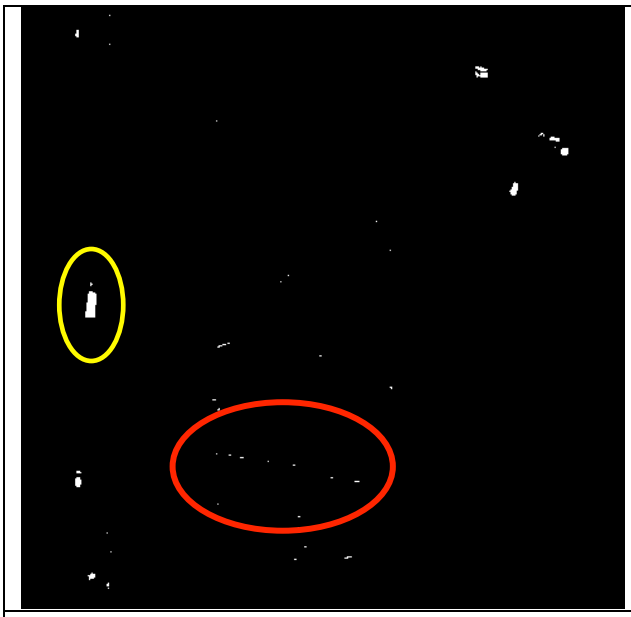

Figure 20: Median result thresholded.

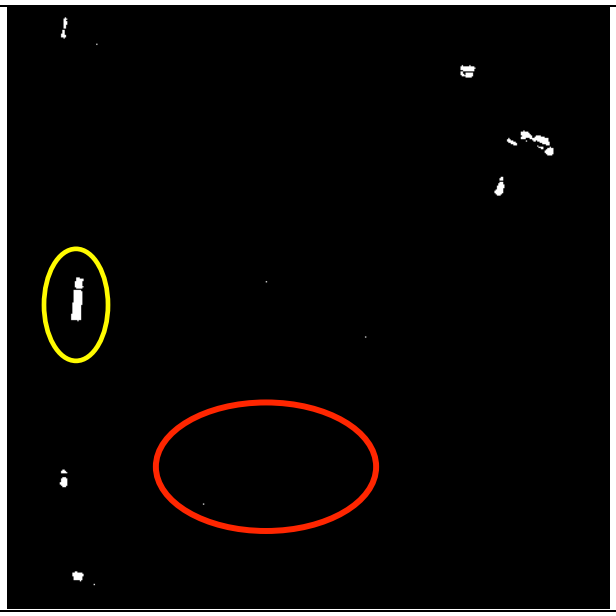

Figure 21: Intrinsic result thresholded.

5.3 Full Pass vs. Windowed

Whether the full pass or the windowed approach is better than the other largely determines on what the problem is. The windowed approach requires less computation, and can be computed quickly and online, whereas the full pass approach requires significantly more computation, and the entire dataset to be available before any results can be produced. The windowed approach also allows for more assumptions to be made about the structure of the scene. Specifically, if the tall structure objects do not move much through the duration of the window, they can be assumed to be roughly constant, and will have little impact on the final results. However, determining the size of the window can be a problem. If the window is too big, then the constant scene assumption fails, and the tall structures will interfere with the results. If the window is too small, the cars in the image may not move enough to be able to obtain the underlying scene. If a car does not move for the duration of the window, even if it moves at a later time, the windowed approach will not find it, while the full pass approach may have.

6 Lessons Learned

The most important lesson learned from this project is to understand why something behaves as it does. As a result of bad registration, the results of the analyses were all off only slightly, due to edges not lining up properly. Being able to determine what exactly causes the results to behave as they do can allow for issues such as that to be more easily corrected, and will also give a better understanding of the results.

Another lesson learned is to always explore and try every option, even the simple ones, such as median and background subtraction. One cannot know what will really work out the best until every option has been tried and exhausted. Also, even if one method is better for one situation, there may be a need where another method is better suited.

We also note that the intrinsic reflectance image, when used in conjunction with another reflectance image from nearby sequentially collected frames, can be used to provide a "clean" stereo view pair without the visual discontinuities of the movers in the scene.

7 Future Work

The output of this project is a sequence of images giving a weighting of moving objects in the scene. Applied to aerial video, this gives a sequence of images with only brightly displayed cars moving along a black background, with some scattered noise. Given more time, the next step to continue this project would be to explore the possibilities available using the output of this project is its input. For one likely example, tracking could be applied, and could potentially provide better results using the results of this project, than attempting to apply tracking algorithms directly to input image sequence.

Bibliography

Weiss, Y. (2001). Deriving intrinsic images from image sequences. *ICCV*, (pp. 68-75 vol.2).