Object Detection of Animals from Camera Trap Images

ENGR484: Justin Kim

Advisor: Professor Cheng

ABSTRACT

In this paper, Eigenbackground method of object detection is explored as a means of object detection of animals from camera trap images in a pre-compiled dataset. Finding a way to model and subtract a common background from a sequence of images or video is becoming increasingly popular and important, because it is the first step in computer vision and object identification of non-static objects. An automated process that can differentiate between the foreground and background of images greatly reduces the human input required in processing these images.

The data set that this paper is based on is from Trinity College's //tbos/Projects/Smedley database, formerly a biology/ecology project concerned with compiling biodiversity data. The images from this data set were obtained through camera traps, which is a motion triggered stationary camera source. The methods outlined in this paper are generalized, and they can be applied to any sequence of images from other stationary image data sources such as traffic cameras and security cameras. By using a combination method as detailed, the suggested Eigenbackground model creates a robust probability density function that makes for a forgiving object detection scheme in even the shortest data sets with only 10 images.

INTRODUCTION

Camera Trap images generate tens of thousands of images. Trying to process these by hand can take hours of human labor, and is unnecessarily repetitive. Find if there are non-static portions in an image is important in identifying that there are foreground objects. This is the first step in detection, and can provide a training data set for later identification.

While there was no previous personal knowledge on the subject, object detection, background modelling, and object identification is an extremely popular field due to the increasing demand of computer vision and machine learning engineers, and there are multitudes of journals covering various methods of detection.

The objectives of this study is to develop an Eigenbackground based method of object detection that will work robustly in a pre-compiled data sets that have short range (10 to 40 images) and are generally based in a dynamic natural setting. The difference between this project and the numerous other publications is that the data was pre-assembled and gathered without the intention of being used for testing image processing methods. Other studies typically have over 200 carefully controlled image frames to use as a training set build their ground truth background model.

An acceptable outcome would be detection accuracy over 90%. This will help compile a neatly processed data for an identification training set.

METHODS

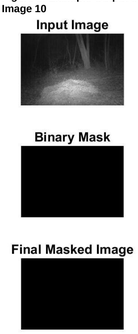
Eigenbackgrounds

- **All images come from a 2015 database of Trinity's tbos/Projects/Smedley**
- **Every image processing step was done on a personal laptop with Matlab R2016a**

Figure 1: 10 Image Dataset

Figure 2: Example Output of Image 1

Figure 3: Example Output of Image 10



DESIGN

The process for obtaining the Eigenbackground images are as follows[1]:

For each image i of size $[w*h]$ (width by height), the images are transformed into a $[wh*1]$ column vector x_i. From this follows that each proceeding image is also reshaped into a column vector, and placed in the second column, the third, and so forth.

The model is taken from the data set consisting of N images. The mean image, m, of the data set is then calculated to be:

$$m = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

The mean normalized image vectors are then formed into the matrix X which has the dimension $[wh * N]$:

$$X = [x_1 - m \qquad x_2 - m \qquad \cdots \qquad x_N - m] \tag{2}$$

The columns of X all lie on a wh-dimensional space. The assumption here is that the frames are related and similar, since they are taken from a single, stationary camera source. Since the frames are similar, it is likely that these columns can be represented in a lower dimensional subspace. Following that logic, the singular value decomposition (SVD) of X is then calculated:

$$X = U\Sigma V^T \tag{3}$$

The first r columns of U that is used will now be referred to as U_r, which is also known as the "Eigenbackgrounds". Any new image y can be projected onto the reduced subspace represented by U_r through the following equation:

$$\tilde{y} = U_r p + m \tag{4}$$

and since U_r is orthogonal (per the definition of SVD), p (the principal component of the image data set of mean normalized vector X) is obtained by performing the following computation:

$$p = U_r^T (y - m) \tag{5}$$

Then, by computing and thresholding the absolute difference between the input image and the projected image, the moving objects present in the input image can be represented as follows:

$$|y_i - \tilde{y}_i| > T \tag{6}$$

Figure 4: Example of Singular Values

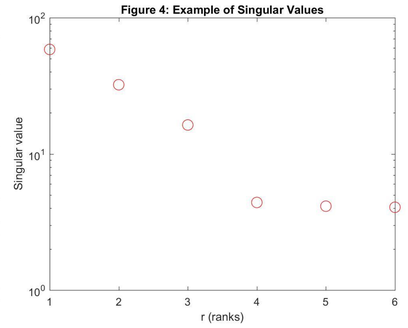
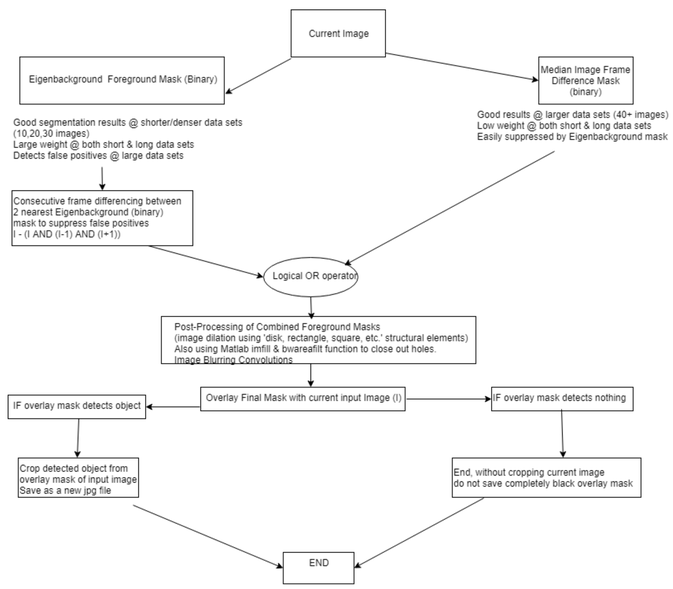


Figure 4 above shows the singular values resulting from SVD of the image matrix of the dataset shown in Figure 1. It can be seen that the first few r values are non zero and decrease rapidly, thereby giving an appropriate lower dimensional subspace of the background of the images in the dataset

Two methods of thresholding were considered: Online adaptive thresholding, and the Otsu's method. After reviewing the results, using a generalized threshold using the Otsu's method was chosen. (Otsu's method, in short, looks at the image histogram to determine a threshold value that would minimize the intra-class variance between the pixels[2]).

Suppressing false positives and ghost images that occur became the next challenge. In order to create a forgiving foreground detection method, the following combination was used.

Figure 5: Combination Method

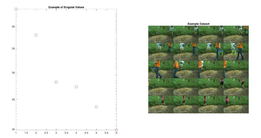


The key assumption here is that the foreground object will have moved significantly enough in pixel location over the course of 3 frames, such that it will not be removed by the logical masks designed to eliminate the false positive detection of the various changes in the dynamic background.

RESULTS

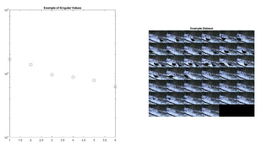
For the selected data ranging from 10 to 40 frames, the results show high accuracy for foreground object detection across various background contexts, time of day, foreground object sizes, and camera sources.

Figure 6: 20 Image dataset, SVD r values, Detection Accuracy Confusion Matrix



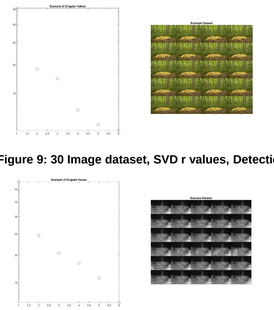
Detection vs. Truth	0 (Truth)	1 (Truth)
0 (Detected)	1	0
1 (Detected)	0	19
Accuracy (%)	100	
False Positive Rate (%)	0	
False Negative Rate (%)	0	

Figure 7: 40 Image dataset, SVD r values, Detection Accuracy Confusion Matrix



Detection vs. Truth	0 (Truth)	1 (Truth)
0 (Detected)	7	0
1 (Detected)	0	33
Accuracy (%)	100	
False Positive Rate (%)	0	
False Negative Rate (%)	0	

Figure 8: 20 Image dataset, SVD r values, Detection Accuracy Confusion Matrix



Detection vs. Truth	0 (Truth)	1 (Truth)
0 (Detected)	11	0
1 (Detected)	0	9
Accuracy (%)	100	
False Positive Rate (%)	0	
False Negative Rate (%)	0	

Figure 9: 30 Image dataset, SVD r values, Detection Accuracy Confusion Matrix

Detection vs. Truth	0 (Truth)	1 (Truth)
0 (Detected)	8	2
1 (Detected)	0	20
Accuracy (%)	86.66667	
False Positive Rate (%)	0	
False Negative Rate (%)	20	

As shown by the figures above, most of the datasets within the N range 10 to 40 fit the conditions of PCA fairly well. The combination method also works rapidly to decrease false positive rate effectively. The cropped results can then be fed into a supervised classification network as a training set, such that the images only contain the features that are pertinent to the corresponding classification label. As for the dataset shown in Figure 1, the combination method achieves 100% accuracy, and correspondingly 0% false positive and false negative rates.

An example of a convolutional neural network performing binary classification using the datasets mentioned here, with the combination method to gather the training set can be found at the following google colabs notebook[3]:
https://colab.research.google.com/drive/13pHC50V5ietsx0x2DosXU-xHjHvAkKtR

CONCLUSIONS

Conclusively, the Eigenbackground method can form a robust probability density function for the static portions of the dynamic background scene of camera trap images in as few as 10 sequential image frames. Combining the logical mask and using consecutive frame differencing can also help lower false positive detection ate due to changes in the dynamic background from sunlight, shadows, wind, and animal disturbances. These techniques can also be generalized to other stationary camera sources, such as traffic and security cameras.

ACKNOWLEDGEMENTS

I'd like to thank Professor Cheng for his continual advise over the entire project.

REFERENCES

1. **Xu, Z., Shi, P., & Gu, I. Y. (2006). An Eigenbackground Subtraction Method Using Recursive Error Compensation. *Advances in Multimedia Information Processing - PCM 2006 Lecture Notes in Computer Science*, 779-787. doi:10.1007/11922162_89**
2. **Makkar, H. (n.d.). Image Analysis Using Improved Otsu's Thresholding Method. Retrieved from https://pdfs.semanticscholar.org/c309/154d5fb9fd9b8273abb73cffd2347f3c4872.pdf**
3. **Moroney, L. (n.d.). Introduction to TensorFlow for Artificial Intelligence, Machine Learning, and Deep Learning. Retrieved April 28, 2019, from https://www.coursera.org/learn/introduction-tensorflowdeeplearning.ai**