Object Detection in Wide Area Aerial Surveillance Imagery with Deep Convolutional Networks

Gregoire Robinson
University of Massachusetts Amherst
Amherst, MA
gregoirerobi@umass.edu

Dong Zhang
University of Central Florida
Orlando, FL
dzhang@eecs.ucf.edu

Introduction

Wide Area Aerial Surveillance (WAAS) technology, alternatively called Wide Area Motion Imagery (WAMI), and Wide Area Persistent Surveillance (WAPS), is capable of monitoring many square kilometers of land for extended periods of time, particularly if used in conjunction with Unmanned Aerial Vehicle (UAV) platforms. It has applications in security, traffic monitoring, and emergency management, where it is important to detect and track the movements of large numbers of vehicles on roads, highways, and parking lots. However, WAAS imagery comes with multiple trade offs, including low image resolution, low image capture frequency, thousands of moving objects in every image, low object saliency from background, small object size in the range of 50 pixels, lack of color information, and significant camera motion, which leads to registration and parallax errors. Human monitoring of this data is prohibitive, necessitating an automated detection and tracking solution.

We use two WAAS datasets, Columbus Large Image Format (CLIF) 2007 from the Air Force Research Laboratory (AFRL) found at (https://www.sdms.afrl.af.mil/) and another from PV Labs (http://www.pv-labs.com/). We attempt to use deep convolutional neural networks (CNN) to detect moving vehicles. CNNs have the potential to learn both appearance features and motion cues (differences between consecutive frames) as well as contextual information, since vehicles are mostly found on pavement. The output of the network is a detection heat-map, which can then be used for object localization and tracking.

Related Work

Object detection is not a new problem in computer science. However, the unique challenges of WAAS imagery have spawned a number of different approaches. Previous methods used on the CLIF dataset have primarily focused on exploiting frame differences. Reilly, Idrees & Shah used median background modeling [1], while Saleemi & Shah developed a two-frame differencing method [2]. These use sophisticated algorithms and classical computer vision techniques to achieve results.

Many other attempts have been made to create a robust detection system in WAAS imagery or WAMI. Shi et al. use image context to reduce false positives [11]. Liang et al. used HOG and Haar features along with Generalized Multiple Kernel Learning [12]. The same authors further improved on these results using background subtraction [13]. Prokaj et al. make use of motion patterns to improve both tracking and detection [14]. The same authors later expanded on this and ran two trackers in parallel, one using background subtraction, and the other using a target state regressor [15]. Progress has also been made on real-time tracking of vehicles in WAMI [16]. Pelapur et al. use a Likelihood Of Features Tracking system [17]. These are just some of the approaches that have been taken to solve the object detection and tracking problem in WAAS, WAPS, and WAMI, but the challenges arising from this data are numerous, making robust solutions elusive.
Interest in deep learning has grown significantly in recent years due to a number of successes in the field. Previous uses of deep learning have focused on detecting and localizing fairly large objects in the PASCAL VOC [3,4,9] and ILSVRC [5,6,8] datasets for example. Additionally, Yi et al. used deep networks for vehicle classification in WAAS imagery, but not localization [7]. Chen et al. used a hybrid deep convolutional neural network to detect vehicles in satellite imagery [10]. Ours is the first attempt to use deep learning for both detection and localization of thousands of very small objects within the same image.

**Dataset**

Both CLIF and PV Labs images are produced by an array of cameras mounted on an electro-optic platform flying at ~7000 ft. Each camera in the array records its own images, and these must first be stitched together into larger images. The CLIF dataset consisted of 261 large stitched images, which were then sliced into ~150,000 smaller sub-images. For the PV Labs dataset, a smaller section of the area under surveillance was selected and subsequently registered to compensate for camera motion. There were 359 images, which were then divided into 4 smaller sub-images to yield 1436 training images.

**Method**

We chose to train a network from scratch. First we created output heat-maps for each of the images in the dataset. This was accomplished using the ground truth annotations; the CLIF dataset indicates a single x, y coordinate at the center of every vehicle, while PV Labs indicates a bounding box around each vehicle. At each annotation, we drew a two-dimensional Gaussian, large enough to capture most vehicles found in the image. An example of the heat-map can be found in Figure 1 along with the original image.

Next, we sliced the images and their corresponding heat-maps into sub-images and sub-heat-maps of equal size. For the CLIF 2007 dataset, the patches were 512 x 512, and for PV Labs, 600 x 700 pixels. An example of this is shown in Figure 2. For the CLIF dataset, we ran only one type of experiment; the input to the network was a single frame, and the output was a detection heat-map that was then regressed over the ground-truth heat-map. A sample of the input can be seen in Figure 3.

For the PV Labs dataset, we ran three types of experiments, single frame, multiple frames, and single frame combined with the background subtracted image. In all the experiments, the output was a detection heat-map that we regressed over the ground-truth heat-map. In the first experiment, the network input was a single sub-image (Figure 4). In the second experiment, the network input was five consecutive frames, and the output was a detection heat-map for the third frame (Figure 5). In the last experiment, the input was two channels, a single frame and its corresponding background-subtracted image (Figure 6).

![Figure 1: A portion of the output heat-map is shown on the left, along with its corresponding input. The green circles indicate the size of the two dimensional Gaussian, in this case 25 pixels in diameter.](image1)

![Figure 2: Example of an image from the PV Labs dataset that has been sliced into sub-images of dimensions 600 x 700.](image2)
Figure 3: CLIF 2007 dataset experiment. Input was a single image, and output was regressed over the ground-truth heat-map on the right.

Figure 4: PV Labs single frame experiment sample input. Output heat-map is shown on the right.
The network structure for the CLIF 2007 dataset consisted of 3 convolution layers, and 2 pooling layers. Each convolution layer is followed by a ReLu activation function. This can be seen in Figure 7. The network structure for the PV Labs dataset consisted of 8 convolution layers, and 2 pooling layers. Each convolution layer is followed by a ReLu activation function. This can be seen in Figure 8.

Figure 5: Multiple frames experiment. The input was five consecutive sub-images. 2D convolution with five channels was used during network training.

Figure 6: Original image and background-subtracted image experiment. The moving vehicles can be clearly seen on the right, but much of the background is also visible due to parallax error.
The loss function used was mean-squared error, where $Y_i$ is the value of a predicted pixel on the output detection heat-map, and $X_i$ is the value of the corresponding pixel on the ground-truth heat-map. The optimizer used was stochastic gradient descent.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i)^2$$

**Results**

For the CLIF 2007 dataset, detection of vehicles was achieved, but localization was challenging due to the presence of excessive false positives. Road edges, road markings, and building edges were also strongly detected.
For the PV Labs dataset, in all three experiments, single frame, multiple frames, and single frame combined with background subtraction, the results were inferior to the CLIF 2007 experiment. Fewer vehicles were detected, and there were more false positives.

Figure 9: CLIF dataset sample results. Input is shown on the left and predicted detections are on the right.
Figure 10: PV Labs single frame experiment. Input is on the left, and predicted detection is on the right.

Figure 11: PV Labs multiple frames experiment. Input was the third frame in the sequence (left). Output (right) was the predicted detections heat-map.
For all experiments, the loss as computed by the mean-squared error converged to approximately 0.005, and did not decrease further. Given the fact that the ground-truth heat-maps were mostly blank, it is not surprising that the calculated loss was so small, but this value is nevertheless not a reliable indicator of detection accuracy, as can be seen from the results.

**Discussion**

It is important to note that the CLIF dataset had superior ground-truth annotations compared with the PV Labs dataset. Furthermore, while the latter’s images were registered and of higher quality, this did not appear to improve results, suggesting that the incompleteness of the annotations was a significant obstacle. At this point, it is unclear why the results for the PV Labs images were inferior to those using the CLIF images. One major finding is that Theano learns differently from Caffe, despite the fact that the network structure and input were roughly identical during the experiments. Future work must ascertain how these frameworks differ. Given the fact that road edges and many terrain features such as trees and buildings were strongly detected, one possible avenue for future research might be to incorporate geospatial information to focus network training exclusively on roads, highways, and parking lots, thereby helping to minimize background noise.

**Conclusion**

In this paper, we presented a novel approach to object detection and localization in WAAS imagery. CNNs were trained with ground truth heat-maps, but the results using Keras and Theano generated a great deal of false positives, making detection and localization difficult. Further research is needed to determine why Caffe and Theano give radically different results despite having a similar input.

**References**


