Limited-Anchor Deep Neural Network for Moving Object Detection

Chih-Yang Lin¹, Han-Yi Huang², Wei-Yang Lin³, Chuan-Yu Chang¹, and Yih-Kuen Jan⁴
¹Dept. of Electrical Engineering, Yuan-Ze University, Taoyuan, Taiwan
²Dept. of Computer Science & Information Engineering, National Chung Cheng University, Chiayi, Taiwan
³Dept. of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Taiwan
⁴Dept. of Kinesiology and Community Health & Computational Science and Engineering Program, University of Illinois at Urbana-Champaign

Abstract—This paper proposes a new method that integrates a deep learning-based object detection network into traditional background modeling to detect moving objects. The proposed method allows us to efficiently identify candidates that contain moving objects while only setting a small number of anchors in the moving area of the image through guidance from the traditional background modeling method. This paper overcomes the disadvantages of conventional background modeling methods and conventional deep learning based object detection methods in terms of dynamic backgrounds and objects’ motion states.

I. INTRODUCTION

Moving object detection is a fundamental and important research field in computer vision, particularly in video surveillance. A robust and real-time moving object detection algorithm can easily solve technical challenges such as object tracking and abnormal event detection.

Most traditional moving object detection methods establish a background model, and then compare the original image against the model to obtain the foreground object. Thus, it becomes very important to determine how to initialize and maintain the model to fit the real background over time. GMM [1] and Vibe [2] are well-known methods offering ways to improve and maintain a background model. GMM and Vibe both output a binary mask, in which white pixels represent moving objects and black pixels represent foregrounds. However, these traditional methods cannot easily recognize types of objects; therefore, they often misclassify moving objects such as swaying trees, shadows, and noise.

Deep learning has been widely used in classification and object detection. However, when well-known object detection CNN-based methods such as Fast R-CNN, Faster R-CNN, SSD, or even YOLO are applied, they cannot distinguish whether detected objects are static or moving. Some CNN-based methods [3] use LSTM to detect moving objects, but these methods are quite complicated and time-consuming to implement.

In this paper, we propose a new CNN-based method that combines the background model established by Vibe with Fast R-CNN[4] to detect moving objects. Unlike Faster R-CNN, which requires setting anchors everywhere on the feature map, the proposed method sets anchors only on the candidate moving areas produced by Vibe. This step can greatly improve the accuracy of moving object detection while speeding up the overall process. Fig. 1 shows the outline of the proposed method and is detailed in the sections below.

II. RELATED WORK

Object Detection: Fast R-CNN uses selective search to obtain candidate proposals in an image that may contain objects. The candidate proposals are mapped onto the feature map to obtain the regions of interest (ROIs). Finally, the ROIs are sent to the second stage in which classification and regression are performed to obtain the final prediction results. We apply Fast R-CNN instead of Faster R-CNN, SSD, or YOLO, to the proposed method because doing so enables us to easily set anchors only on the possible moving object areas. The definition of the anchor in the proposed method is the same as that of Faster R-CNN.

III. METHOD

The details of the proposed method in Fig. 1 are introduced in the following section.

A. Candidate Regions for Anchors

We apply the connected-component algorithm to the output of Vibe to demarcate moving objects, and set anchors on these areas. Note that the output of Vibe usually contains not only potential moving objects, but false “moving” objects as well. If anchors are set directly within the bounding rectangles, called minimum bounding rectangles (MBRs), some of the anchors may be set outside the moving object itself. That will increase false alarm rates. Therefore, instead of using the MBR, we use the minimum volume enclosing ellipsoid (MVEE) of the connected element as shown in Fig. 2. Then, we set the anchors on the major axis of the ellipse to prevent erroneous detection.

Fig 2. Different settings of candidate regions. (a) MBR, (b) Major axis of MVEE.

B. Anchor Setting

With guidance from Vibe, the candidate moving object areas can be identified, so there is no need to waste computational resources by setting dense anchors across the entire image. The idea in anchor setting is that a large moving

Fig. 1. Proposed two-step model. First, we use Vibe to obtain areas that may contain moving objects, and select the candidate proposals from these areas. Next, the original image and the candidate proposals are used as the deep neural network’s input to perform object detection. Finally, the output containing the category and the bounding boxes is obtained.
object requires longer interval between anchors; in contrast, if a connected component contains multiple small moving objects, it should have a shorter interval between anchors. The interval in the proposed method is determined by the semi-minor axis of the MVEE.

In video surveillance, human and vehicles are the most important categories to monitor. A semi-minor axis can help distinguish these objects from the MVEE. In general, the semi-minor axis of a single car is quite different from the semi-minor axis of people. This approach allows us to set up anchors without using fixed intervals, and elastically adjust them for different types of foreground objects. If the moving object belongs to a vehicle type, the anchors are set only on the major axis of the ellipse; otherwise, each anchor on the major axis is regarded as a base point, and then additional anchors are set on the neighbors of the base point. In this paper, we use different approaches to set neighbors around the base point; Fig. 3 shows settings with two neighbors, four neighbors, and eight neighbors. In addition, we use the halfway point of the semi-minor axis as the interval between the neighbor and the base point. Our experimental results show that the setting with four neighbors starting from 0 degrees achieves the best performance.

![Fig 3. Different neighbor settings. (a) Two neighbors, (b) Four neighbors (0 degrees), (c) Four neighbors (45 degrees), and (d) Eight neighbors.](image)

C. Object detection

The object detection backbone in the proposed method is based on Resnet with the feature pyramid network (FPN). The original image is fed into the backbone network to extract features. Then, the candidate proposals can be generated through the anchors mentioned above. These proposals next pass through the second stage for classification and regression to get the final results.

IV. EXPERIMENTAL RESULTS

The dataset used in our experiments was St-Marc[5], which contains approximately 2,000 video surveillance images. These videos include moving, static, and intermittent moving objects.

Our experimental results show that the proposed method is able to discern an object’s motion status. As shown in Fig. 4, we use different colors to describe an object. Since Vibe cannot distinguish the object’s type, only one color is used in the Vibe result. In Fig. 4, the car that is temporarily stopped on the road is not a foreground object in the Vibe result; thus, we do not set any anchors on this area, and no moving object will be detected. Furthermore, although the anchors are set on the noise caused by the dynamic background (swaying trees, shadows, noise, or illumination changes), these anchors will be invalidated by our detection backbone, which can filter out these false proposals. These two capabilities, motion status detection and object type detection, are difficult to achieve by either the traditional Vibe or conventional CNN-based methods alone. Faster R-CNN labels all vehicles and pedestrians, no matter whether they are stationary or moving. A more detailed demo video is available at https://youtu.be/fjOndtXQm7o.

![Fig 4. Experimental results. (a) Vibe, (b) The proposed method, (c) Faster R-CNN. The red bounding box identifies the person while the blue box identifies the vehicle.](image)

Table 1 shows the mAP for different neighbor settings. In the two-neighbor setting, some objects in scenes of multiple small moving objects are not detected. In the eight-neighbor setting, too many anchors result in the generation of redundant bounding boxes. The four-neighbor setting can balance the two cases, and starting from 0-degrees is better than beginning from 45-degree in terms of mAP.

<table>
<thead>
<tr>
<th>Number of neighbors</th>
<th>Two (0 degrees)</th>
<th>Four (45 degrees)</th>
<th>Eight</th>
</tr>
</thead>
<tbody>
<tr>
<td>mAP</td>
<td>0.8948</td>
<td>0.9008</td>
<td>0.9001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.8564</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS

In this paper, we propose a new method that integrates a deep learning based object detection network and a traditional background modeling method to detect moving objects. We perform object detection only on the areas of moving object candidates. This new method overcomes the inability of conventional background modeling methods to remove dynamic backgrounds, and also overcomes the inability of conventional CNN-based methods to distinguish the motion status of objects, which is critical for video surveillance.

REFERENCES