The Video Detection in the Dynamic Background Based on the Sensitive Areas

Qi Liu and Kaiyue Li

ABSTRACT

During the procedure of video analysis with complex dynamic backgrounds, a new method of object detection based on the sensitive areas was presented. First, each frame in a video was converted into an effective information map by using the Harris corner detection method. Second, the sensitive areas in the frame were extracted by using the context information and the effective information maps of the consecutive video frames. The sensitive areas in the video frame were the candidate areas where the target objects would appear at high probabilities. Third, the information entropy features of each sensitive area were extracted to form the feature matrix, based on which, an SVM model was trained for selecting the target areas from the sensitive areas. Finally, the locations of the objects were detected based on the target areas in the video with a complex dynamic background. The experimental results showed that this method could achieve good results against the benchmark of CDnet 2014 on the premise of saving computing resources.

INTRODUCTION

In the last decade, the video processing technology of target detection and tracking has made great progress. The end-to-end deep convolutional models can have good performance in speed and accuracy [1]. In video processing, these models are not optimized based on the contextual information. When some interfering factors like illumination change, motion blur, or video jitter appear in the video,

Qi Liu, Kaiyue Li
Shanghai University, Shangda Road, Shanghai, 200444, China.
recognition accuracy declines sharply. The speed of video processing rests with the positioning time of suspected target areas in the video frame [2]. The computational cost of video processing depends on the complexity of the deep network model [3]. According to the regularity of the background change and the finiteness of the target category, the video processing model, like No Scope model [4], can be designed and modified specially for some certain scenes.

The video processing requires avoiding the use of complex deep neural network to reduce hardware cost. At the same time, it also requires exploiting the data redundancy between consecutive frames to reduce software (computation) costs. Zhu [5] used sparse feature propagation to save these costs. These propagated features were only calculated on the key frames in the video. However, there is not a proper method to dynamically select the key frames from all the video frames.

In this paper, a method based on the sensitive areas was proposed to detect the video object in the dynamic background. The flow chart of the method is shown in Figure 1. In videos with a complex and changeable background, it is very difficult to distinguish the target from the noise by only the motion trajectory. However, the change of local information entropy between consecutive frames is effective in identifying the object. In addition, the sensitive areas in the video frame are more likely to contain foreground targets, so it can avoid a lot of time-consuming operations by doing intensive calculation for sensitive areas rather than for the whole frame image. The objective of the experiments in this paper was to detect objects in the video with dynamic background. Comprehensive experiments showed that the method based on the sensitive areas achieved high accuracy and significant performance.

PREVIOUS WORK

Many experiments have shown that combining traditional feature extraction methods (like the corner detection operators [6,7]) and machine learning methods can solve the problem of object detection and classification well in the field of image processing [8-10]. Besides, current evidence suggests that the deep learning algorithm has the ability of approximate human beings in the field of image classification and object detection [11,12]. When all of the frames in the video are processed in the same way, the applications of deep network in image processing can be directly converted to algorithms for solving video processing tasks. Without considering the cost and speed, the task of object recognition and segmentation in video processing can be solved well by improving the deep network algorithms (like CNN [13]) in image processing [14,15]. However, as the requirements for speed and accuracy have gradually improved, more models have been specially designed as end-to-end deep network models [16,17].
The end-to-end network model is considered as the key point to improving the speed of recognition and positioning. In the object recognition task of computer vision, the YOLO model abandons the thought of regional pre-processing and truly realizes the application of the end-to-end network model [18,19]. Now, the target object in video processing tasks can be identified and tracked in real time such as through the SSD model [20].

The information entropy of one image is actually the expected value of all information saved in this image. Therefore, the change of information entropy in continuous frames is one special form of information gain. This information gain is often used in decision tree algorithms to select characteristics [21]. In fact, the random forest algorithm proposed based on decision trees is now still popular in many fields [22]. These works prove that the information theory is useful for finding the effective characteristics of different classes.

**METHOD**

The method proposed in this paper was to achieve the object detection task in a complex environment. Video processing in a complex environment is usually more difficult. Thus, a new method was proposed in this paper on the basis of information theory. In this method, the information entropy was utilized to quantify the information change process of the local area. The quantify result could well indicate the different nature between the object and the noise. As shown in Figure 1, the focus of the method was the extraction of sensitive areas in the video frame, then the feature matrix was extracted by information entropy calculation to classify sensitive areas.

**Effective Information Map**

The effective information represents the valuable information used for identifying foreground objects in the frame of a video. There is a greater probability of having targets in the area where effective information is assembled in the image. In the selective search model, the candidate area set is composed of local regions that are merged after image segmentation [2]. Therefore, the angular points and
edges that are important during the extracting processing of candidate areas often represent effective information. Since the Harris corner detection algorithm [23] can distinguish the flat regions, the edge regions, and the corner regions in the image, this algorithm can be used as the extraction method of effective information.

The Harris corner detection algorithm can obtain the response value of the corner information through the transformation of the corner response function. Assume that the non-flat region in the video frame is the area containing effective information, so the transformation formula of the effective information map is

\[ I_{\text{info}} = f(x,y) = \begin{cases} 255, & |\text{dst}| > (\alpha \cdot \text{Max}(|\text{dst}|)) \\ 0, & |\text{dst}| \leq (\alpha \cdot \text{Max}(|\text{dst}|)) \end{cases} \]

where \( \text{dst} \) represents the corner detection result of the video frame. \( \alpha \) times of the maximum value of the corner detection result is the threshold to divide the flat region and the non-flat region, i.e., the ineffective information region and the effective information region. Finally, the effective information map can be obtained by Gaussian filtering and simple morphological processing.

**Sensitive Area Extraction**

The dynamic background often contains complex information. So, the video with dynamic background is generally difficult to process by using the background difference method. The frame difference method is characterized by low complexity, fast running speed, and strong adaptive ability of the dynamic environment. Some noise in the dynamic background can sometimes be mistaken for the foreground object during the processing of the frame difference method.

According to the effective information map rather than the original video frame as the processing unit, the frame difference method can obtain better results. The method used in this paper was to calculate on the basis of effective information maps of three consecutive frames. The calculation formula is

\[ D_n(x,y) = [f_n(x,y) - f_{n+1}(x,y) \land f_n(x,y)] \lor [f_n(x,y) - f_n(x,y) \land f_{n-1}(x,y)] \]

where \( f_n(x,y) \) represents the effective information map of the \( n \)th video frame. The \( \lor \) operation in the formula calculates the mean value of the corresponding pixels in two effective information maps. The \( \land \) operation in the formula reserves the corresponding pixel value of the effective information map of \( n \)th video frame. Threshold processing is required for \( D_n(x,y) \) after the difference operation. The threshold value \( T_{\text{Otsu}} \) is obtained by the Otsu [24] method automatically. Then, the influence of light fluctuation is added on this basis to get the optimal threshold value \( T_{\text{optimal}} \). The optimal threshold calculation formula is

\[ L_{\text{Diff}} = \frac{1}{N} \sum_{(x,y) \in A} \frac{|f_{n+1}(x,y) - f_n(x,y)| + |f_n(x,y) - f_{n-1}(x,y)|}{2} \]

\[ T_{\text{optimal}} = T_{\text{Otsu}} + \lambda \cdot L_{\text{Diff}} \]
where $\lambda$ represents the influence factor of light fluctuation in the current environment. Through the threshold processing, the difference image can be finally obtained.

The difference image obtained after threshold processing is divided into many small regions. These regions have the same size and are non-overlapping. The number of effective information pixels in each region represents the occurrence possibility of the foreground object. This paper assumed that the region was judged as a sensitive area as long as its occurrence possibility was greater than zero.

**Sensitive Area Screening**

Another key of the method in this paper was how to select real target areas from all of the sensitive areas. In videos with complex backgrounds, the location recognition of the foreground object is more important than the classification of the foreground object. This paper proposed a method that uses information entropy to quantify the process of information change of the area. An N dimensional feature vector can be obtained by calculating the information entropy of the local areas of the same location in consecutive N frames. However, the local areas of the video frame are first preprocessed by M kinds of algorithms before the calculation of the feature vector. Thus, one area in the video frame can be finally represented by a $M \times N$ dimensional feature matrix. There are many options for the image processing algorithm in the preprocessing operation. Two preprocessing algorithms used in this paper were the image gray algorithm and the image gradient algorithm. During the training of the machine learning model, one feature matrix is used as the sample data of one sensitive area. These samples are used for training a supervised binary classification model.

**Target Object Locating**

The target object is calibrated on the basis of target areas in the video frame. Since the previous operation has divided the whole image into multiple regions, the resulting image that contains all the target areas is resized for the convenience of subsequent processing. The adjustment rule of image size is that each target area corresponds to one pixel in the adjusted image. Through the morphological analysis of the adjusted image, target areas that belong to the same object are merged into a large target block. After the image is adjusted back to its previous size, the four boundaries of this irregular target block can be utilized to form a complete rectangle. Finally, these rectangular boxes can be used as the calibration results of the target objects in the video frame.
EXP:ERIMENTS

The experimental data in this paper is from a video dataset in CDnet 2014 [25]. The main test set used in this paper was the video with dynamic background. The experimental flow chart of the method proposed in this paper is shown in Figure 2. During the transform process of an effective information map, the threshold of the Harris corner detection algorithm was 0.01. If the shot environment of the video was not very complex, extreme, or uncommon, then the influence of this parameter was little. During the extraction and screening of sensitive areas, the experiments proved that the optimal pixel size of a sensitive area was 11 × 11. After the sensitive area was determined, the same location area of eight consecutive frames were selected as the data sample source for subsequent classification work. Then, each sensitive area was transformed into a 2 × 8 dimensional feature matrix.

The supervised binary classification model used in the experiments was the SVC model. Three representative series of video were selected as the primary experiment data: one contained the dynamic background of water fluctuation, one contained the dynamic background of leaves shaking, and the other one contained the heavy jitter of the camera. As shown in Table I, the experimental results proved that the object detection framework based on the sensitive areas could achieve the same ideal effect in different scenes.

<table>
<thead>
<tr>
<th>Video Category</th>
<th>Dynamic Background</th>
<th>Camera Jitter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boats</td>
<td>Highway</td>
</tr>
<tr>
<td>Accuracy</td>
<td>85.29%</td>
<td>85.61%</td>
</tr>
<tr>
<td>Speed(fps)</td>
<td>64</td>
<td>52</td>
</tr>
</tbody>
</table>

Figure 2. The experimental flow chart of method.
The different classification models based on the sensitive areas were trained in different scenes. Then, these models were tested in the corresponding scene. The test results are shown in Figure 3. In the dynamic scene where the content was a boat, approximately 1000 frames in the first half of the video were used in the training of the detection model, and around 2000 frames in the second half of the video were used in the test. In Figure 3, the first line is the processing result obtained by using the validation set of the video, and the second line is the processing result obtained by using the test set of the video. The above experiment proved that the video detection framework based on the sensitive areas had strong generalization ability in the same scene.

CONCLUSIONS

The method of this paper provides a new way, which is different from the end-to-end deep network, to solve the video processing task at low cost. The method can be treated as a framework for target location detection in the dynamic background to save as much costs as possible. A large number of non-candidate areas avoided to be considered by using the context information of the video. At the same time, a lot of hardware costs can be saved by using a lightweight machine learning model. Although the algorithm had strong generalization ability in the same scene, different models needed to be trained respectively for videos with different dynamic backgrounds as the noise had different characteristics in different dynamic backgrounds.

REFERENCES