International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 10 | Oct - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Object Detection Algorithms for Video Surveillance Applications

A. Swapna[1], C.R.K. Gayatri [2], Sai Kishan Kranthi Kumar Vuriti [3]

[1&2]*Assitant Professor, Department of Electronics and Communication Engineering*

[3] *Postgraduate student, Department of Electronic and Communication Engineering*

Abstract—Object Detection algorithms find applications in various fields, such as defense, security, and healthcare. In this paper, various Object Detection Algorithms, such as face detection, skin detection, color detection, shape detection, and target detection, are simulated and implemented using MATLAB 2017b to detect various types of objects with improved accuracy for video surveillance applications. Further, various challenges and applications of object detection methods are elaborated.

1. INTRODUCTION

The paper focuses on improving object detection techniques used in video surveillance systems. It addresses various challenges related to detecting and tracking moving objects in real-time video feeds. The thesis simulates and implements object detection algorithms such as face detection, color detection, and target detection using MATLAB, offering practical improvements in accuracy.

2. OBJECT DETECTION AND TRACKING

Video surveillance is an active research topic in computer vision that tries to detect, recognize and track objects over a sequence of images, and it also makes an attempt to understand and describe object behavior by replacing the aging old traditional method of monitoring cameras by human operators. Object detection and tracking are important and challenging tasks in many computer vision applications such as surveillance, vehicle navigation and autonomous robot navigation. Object detection involves locating objects in the frame of a video sequence. Every tracking method requires an object detection mechanism either in every frame or when the object first appears in the video. Object tracking is the process of locating an object or multiple objects over time using a camera. The high-powered computers, the availability of high quality and inexpensive video cameras and the increasing need for automated video analysis has generated a great deal of interest in object tracking algorithms. There are three key steps in video analysis, detection interesting moving objects, tracking of such objects from each and every frame to frame, and analysis of object tracks to recognize their behavior. Therefore, the use of object tracking is pertinent in the tasks of, motion-based recognition. Automatic detection, tracking, and counting of a variable number of objects are crucial tasks for a wide range of home, business, and industrial applications such as security, surveillance, management of access points, urban planning, traffic control, etc. However, these applications were not still playing an important part in consumer electronics. The main reason is that they need strong requirements to achieve satisfactory working conditions, specialized and expensive hardware, complex installations and setup procedures, and supervision of qualified workers. Some works have focused on developing automatic detection and tracking algorithms that minimize the necessity of supervision. They typically use a moving object function that evaluates each hypothetical object configuration with the set of available detections without to explicitly computing their data association. Thus, a considerable saving in computational cost is achieved. In addition, the likelihood function has been designed to account for noisy, false and missing detections. The field of machine (computer) vision is concerned with problems that involve interfacing computers with their surrounding environment. One such problem, surveillance, has an objective to monitor a given environment and report the information about the observed activity that is of significant interest. In this respect, video surveillance usually utilizes electro-optical sensors (video cameras) to collect information from the environment. In a typical surveillance system, these video cameras are mounted in fixed positions or on pan-tilt devices and transmit video streams to a certain location, called monitoring room. Then, the received video streams are monitored on displays and traced by human operators. However, the human operators might face many issues, while they are monitoring these sensors. One problem is due to the fact that the operator must navigate through the cameras, as the suspicious object moves between the limited field of view of cameras and should not miss any other object while taking it. Thus, monitoring becomes more and more challenging, as the number of sensors in such a surveillance network increases. Therefore, surveillance systems must be automated to improve the performance and eliminate such operator errors. Ideally, an automated surveillance system should only require the objectives of an application, in which real time interpretation and robustness is needed. Then, the challenge is to provide robust and realtime performing surveillance systems at an affordable price. With the decrease in costs of hardware for sensing and computing, and the increase in the processor speeds, surveillance systems have become commercially available, and they are now applied to a number of different applications, such as traffic monitoring, airport and bank security, etc. However, machine vision algorithms (especially for single camera) are still severely affected by many shortcomings, like occlusions, shadows, weather conditions, etc. As these costs decrease almost on a daily basis, multi-camera networks that utilize 3D information are becoming more available. Although, the use of multiple cameras leads to better handling of these problems, compared to a single camera, unfortunately, multi-camera surveillance is still not the ultimate solution yet. There are some challenging problems within the surveillance algorithms, such as

background modeling, feature extraction, tracking, occlusion handling and event recognition. Moreover, machine vision algorithms are still not robust enough to handle fully automated systems and many research studies on such improvements are still being done. This work focuses on developing a framework to detect moving objects and generate reliable tracks from surveillance video. The problem is most of the existing algorithms works on the gray scale video. But after converting the RGB video frames to gray at the time of conversion, information loss occurs. The main problem comes when background and the foreground both have approximately same gray values. Then it is difficult for the algorithm to find out which pixel is the foreground pixel and which one background pixel. Sometimes two different colors such as dark blue and dark violet, color when converted to grayscale, their gray values will come very near to each other, it can't be differentiated which value comes from dark blue and which comes from dark violet. However, if color images are taken then the background and foreground color can be easily differentiated. So without losing the color information, this modified background model will work directly on the color frames of the video.

3. LITERATURE REVIEW

The research conducted so far for object detection and tracking objects in video surveillance system are discussed in this chapter. The set of challenges outlined above span several domains of research and the majority of relevant work will be reviewed in the upcoming chapters. In this section, only the representative video surveillance systems are discussed for better understanding of the fundamental concept. Tracking is the process of the object of interest within a sequence of frames, from its first appearance to its last. The type of object and its description within the system depend on the application. During the time that it is present in the scene, it may be occluded by other objects of interest or fixed obstacles within the scene. A tracking system should be able to predict the position of any occluded objects. Object tracking systems are typically geared towards surveillance applications where it is desired to monitor people or vehicles moving about an area. There are two district approaches to the tracking problem, top-down and another one is bottom-up. Top-down methods are goal-oriented and the bulk of tracking systems are designed in this manner. These typically involve some sort of segmentation to locate regions of interest, from which objects and features can be extracted for the tracking system. Bottom-up responds to stimulus and has according to observed changes. The top-down approach is the most popular method for developing surveillance systems. The system has a common structure consisting of a segmentation step, a detection step, and a tracking step.

As per the description in Chapter 1, object tracking has a lot of application in the real world. But it has many technological lacuna still exist in the methods of background subtraction. In this section, some previous works is discussed for frame difference that use of the pixel-wise differences between two frame images to extract the moving regions, Gaussian mixture model based on background model to detect the object and finally background subtraction to detect moving regions in an image by taking the difference between current and reference background image in a pixel-by-pixel, and previous works done for the background modelling. After the detection scenario is over, tracking part is done. Once the interesting objects have been detected it is useful to have a record of their movement over time. So tracking can be defined as the problem of estimating the trajectory of an object as the object moves around a scene. It is necessary to know where the object is in the image at each instant in time. If the objects are continuous observable and their sizes or motion does not vary over time, then tracking is not a hard problem. In general surveillance systems are required to observe large area like airports, shopping malls. In these scenarios, it is not possible for a single camera to observe the complete area of interest because sensor resolution is finite and structures in the scene limit the visible area. Therefore, surveillance of wide areas requires a system with the ability to track objects while observing them through multiple cameras. But here no discussion about multiple camera network is done. Lipton et al. [5] proposed frame difference that use of the pixel-wise differences between two frame images to extract the moving regions. In another work, Stauffer & Grimson et al. [6] proposed a Gaussian mixture model based on background model to detect the object. Liu et al. [7] ,proposed background subtraction to detect moving regions in an image by taking the difference between current and reference background image in a pixel-by-pixel. Collins et al. [8], developed a hybrid method that combines three-frame differencing with an adaptive background subtraction model for their VSAM (Video Surveillance and Monitoring) project. Desa & Salih et al [9], proposed a combination of background subtraction and frame difference that improved the previous results of background subtraction and frame difference. Sugandi et al. [10], proposed a new technique for object detection employing frame difference on low resolution image. Julio cezar et al. [3] has proposed a background model, and incorporate a novel technique for shadow detection in gray scale video sequences. Satoh et al. [11], proposed a new technique for object tracking employing block matching algorithm based on PISC image. Sugandi et al. [12], proposed tracking technique of moving persons using camera peripheral increment sign correlation image. Beymer & konolige et al. [2],1999 proposed in stereo camera based object tracking, use kalman filter for predicting the objects position and speed in x-2 dimension. Rosals & sclaroff et al.,1999 proposed use of extended kalman filter to estimate 3D trajectory of an object from 2D motion.

In object detection method, many researchers have developed their methods. Liu et al., 2001 proposed background subtraction to detect moving regions inan image by taking the difference between current and reference background image in a pixel-by-pixel. It is extremely sensitive to change in dynamic scenes derived from lighting and extraneous events etc. In another work, Stauffer & Grimson, 1997 proposed a Gaussian mixture model based on background model to detect the object. Lipton et al., 1998 proposed frame difference that use of the pixel-wise differences between two frame images to extract the moving regions. This method is very adaptive to dynamic environments, but generally does a poor job of extracting all the relevant pixels, e.g., there may be holes left inside moving entities. In order to overcome disadvantage of two-frames differencing, in some cases three-frames differencing is used. For instance, Collins et al., 2000 developed a hybrid method that combines three-frame differencing with an adaptive background subtraction model for their VSAM (Video Surveillance and Monitoring) project. The hybrid algorithm

successfully segments moving regions in video without the defects of temporal differencing and background subtraction. Desa & Salih, 2004 proposed a combination of background subtraction and frame difference that improved the previous results of background subtraction and frame difference. In object tracking methodology, this article will describe more about the region based tracking. Region-based tracking algorithms track objects according to variations of the image regions corresponding to the moving objects. For these algorithms, the background image is maintained dynamically and motion regions are usually detected by subtracting the background from the current image. Wren et al., 1997 explored the use of small blob features to track a single human in an indoor environment. In their work, a human body is considered as a combination of some blobs respectively representing various body parts such as head, torso and the four limbs. The pixels belonging to the human body are assigned to the different body part's blobs. By tracking each small blob, the moving human is successfully tracked. McKenna et al., 2000 proposed an adaptive background subtraction method in which color and gradient information are combined to cope with shadows and unreliable color cues in motion segmentation. Tracking is then performed at three levels of abstraction: regions, people, and groups. Each region has a bounding box and regions can merge and split. A human is composed of one or more regions grouped together under the condition of geometric structure constraints on the human body, and a human group consists of one or more people grouped.

4. DIGITAL IMAGE PROCESSING

Object Detection is the process of finding and recognizing real-world object instances such as car, bike, TV, flowers, and humans out of an images or videos. An object detection technique lets you understand the details of an image or a video as it allows for the recognition, localization, and detection of multiple objects within an image. It is usually utilized in applications like image retrieval, security, surveillance, and advanced driver assistance systems (ADAS). Object Detection is done through many ways:

- Feature Based Object Detection
- Viola Jones Object Detection
- SVM Classifications with HOG Features
- Deep Learning Object Detection

Object detection from a video in video surveillance applications is the major task these days. Object detection technique is used to identify required objects in video sequences and to cluster pixels of these objects. The detection of an object in video sequence plays a major role in several applications specifically as video surveillance applications. Object detection in a video stream can be done by processes like pre-processing, segmentation, foreground and background extraction, and feature extraction.

Humans can easily detect and identify objects present in an image. The human visual system is fast and accurate and can perform complex tasks like identifying multiple objects with little conscious thought. With the availability of large amounts of data, faster GPUs, and better algorithms, we can now easily train computers to detect and classify multiple objects within an image with high accuracy.

5. METHODOLOGY

Frame differencing is a pixel-wise differencing between two or three consecutive frames in an image sequence to detect regions corresponding to moving object such as human and vehicles. The threshold function determine's change and it depends on the speed of object motion. It's hard to maintain the quality of segmentation if the speed of the object changes significantly. Frame differencing is very adaptive to dynamic environments, but very often holes are developed inside moving entities.Videos are actually consists of sequences of images, each of which called as a frame. For detecting moving objects in video surveillance system, use of frame difference technique from the difference between the current frame and a reference frame called as 'background image' is shown. That method is known as frame difference method. Frame differencing is the simplest moving object detection method which is based on determining the difference between input frame intensities and background model by using pixel per pixel subtraction.Grad. Sch. of Eng. et al. [5] have proposed frame difference method to detect the moving objects. In this case, frame difference method is performed on the three successive frames, which are between frame Fk and Fk−1 and also the frame between Fk & Fk+1 and the output image as frame difference image is two difference images dk−1 and dk+1 is expressed as

$$d_{k-1} = |f_k - f_{k-1}| \tag{3.1}$$

$$d_{k+1} = |f_k - f_{k+1}| \tag{3.2}$$

$$d_{k'}(x,y) = \begin{cases} 1, \, if \, d_{k'}(x,y) > T \\ 0, \, otherwise \end{cases} \tag{3.3}$$

Where $k' = k-1$ *and* $k+1$

The process is followed by applying and operator to dk−1 and dk−1 This method is already discussed in details in chapter (2). Here original frames are shown and after pre-processing segmented results frames are also shown.

Figure 1: Original video frames

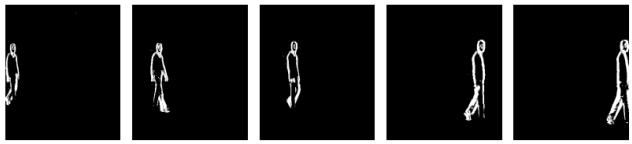


Figure 2: Output after frame difference

The background subtraction [10] is the most popular and common approach for motion detection. The idea is to subtract the current image from a reference background image, which is updated during a period of time. It works well only in the presence of stationary cameras. The subtraction leaves only non-stationary or new objects, which include entire silhouette region of an object. This approach is simple and computationally affordable for real-time systems, but are

extremely sensitive to dynamic scene changes from lightning and extraneous event etc. Therefore it is highly dependent on a good background maintenance model. Here in this chapter simulation of different background subtraction techniques available in the literature, for motion segmentation of object is performed. Background subtraction detects moving regions in an image by taking the difference between the current image and the reference background image captured from a static background during a period of time. The subtraction leaves only non-stationary or new objects, which include entire silhouette region of an object. The problem with background subtrac-tion [14], [8] is to automatically update the background from the incoming video frame and it should be able to overcome the following problems:

- **Motion in the background:** Non-stationary background regions, such as branches and leaves of trees, a flag waving in the wind, or flowing water, should be identified as part of the background
- **Illumination changes:** The background model should be able to adapt, to gradual changes in illumination over a period of time.
- **Memory:** The background module should not use much resource, in terms of computing power and memory
- **Shadows**: Shadows cast by moving object should be identified as part of the background and not foreground.
- **Camouflage:** Moving object should be detected even if pixel characteristics are similar to those of the background
- **Bootstrapping**: The background model should be able to maintain background even in the absence of training background (absence of foreground object)

In simple background subtraction, a absolute difference is taken between every current image It(x, y) and the reference background image B(x, y) to find out the motion detection mask D(x, y). The reference background image is generally the first frame of a video, without containing the foreground object.

$$D(x,y) = \begin{cases} 1, & if|I_t(x,y) - B(x,y)| \geq \tau \\ 0, & otherwise \end{cases} \quad (3.4)$$

where τ is a threshold, which decides whether the pixel is foreground or background. If the absolute difference is greater than or equal to τ, the pixel is classified as foreground, otherwise, the pixel is classified as background.

To implement an existing Gaussian mixture model based on background model to detect the moving objects. For detecting moving objects in video surveillance system the use the Gaussian mixture model, is essential this model has the color values of a particular pixel as a mixture of Gaussians. But the pixel values that don't fit the background distributions are considered as foreground. Nowak 2003 showed how the parameters of a mixture of Gaussians for which each node of a sensor network had different mixing coefficients could be estimated using a distributed version of the well-known expectation-maximization (EM) algorithm. This message-passing algorithm involves the transmission of sufficient statistics between neighboring nodes in a specific order and was experimentally shown to converge to the same results as centralized EM. Kowalczyk and Vlas-sis Kowalczyk and Vlassis, 2004 proposed a related gossip-based distributed algorithm called Newscast EM for estimating the parameters of a Gaussian mixture. Random pairs of nodes repeatedly exchange their parameter estimates and combine them by weighted averaging.

In this section, another technique that is commonly used for performing background segmentation. Stauffer and Grimson et al. [5] have proposed, suggest a probabilistic approach using a mixture of Gaussian for identifying the background and foreground objects. The probability of observing a given pixel value Pt at time t is given by:

$$P(p_t) = \sum_{i=1}^{k} \omega_{i,t}\eta(p_t, \mu_{i,t}, \sum i, t) \quad (3.9)$$

Where k is the number of Gaussian Mixture and that is used. The number of k varies depending on the memory allocated for simulations. Then the normalized Gaussian η is a function of ωi,t, μi,t, ∑i, t which represents weight, mean and co-variance matrix of the ith Gaussian at time respectively.The weight indicates the influence of the ith Gaussian and time t. In this case k=5 to maximize the distinction amongst pixel values.Since it is an iterative process that all parameters are updated, with the inclusion of every new pixel. Before update take place, then the new pixel is compared to see if it matches any of the k existing Gaussian. A match is determined if |pt − μi,t| < 2.5σ Where correspond to the standard deviation of the Gaussian. Depending on the match, the Gaussian mixture is updated in the following manner

6. PROPOSED SYSTEM

Visual detection: Visual problems can surface in many ways. Visual processing challenges present as difficulty in reading, handwriting, sports, navigating a hallway, or many other areas. Sometimes, the issue is a result of visual detection challenges. Read on to find out exactly what visual detection is and what an eye detection problem looks like in kids, including common visual detection difficulties that present in the classroom or during academic work. We've shared a few visual detection tips and soon on the site, we'll share a collection of visual detection activities, too. You've probably seen it before: The child who struggles with letter reversals. The child who has challenges in navigating obstacles when playing…the child who labors with reading and commonly skips words or lines of words when reading. These are all signs of a visual detection problem. There are many more, in fact. The thing is visual detection is a part of almost everything we do! Before we talk more about what visual detection looks like and other common signs of visual detection problems, let's discuss what exactly visual detection is. Visual detection is a visual processing skill that occurs when the eyes focus on an object as it moves across the field of vision. Visual detection occurs with movement of the eyes to follow a moving object and not movement of the head. The eyes have the ability to track an object in the vertical and horizontal, diagonal, and circular planes. There should also be an ability to track across the midline of the eyes and with smooth pursuit of the object. Visual detection requires several skills in order to efficiently occur. These include oculomotor

control abilities, including visual fixation, saccadic eye movement, smooth pursuit eye movements, along with convergence, and visual spatial attention. Here is more detailed information on saccades and their impact on learning.

These are the visual processing skills that need to occur in conjunction with visual detection. They are necessary to enable visual detection in functional tasks. Page | 86 Visual Fixation- The ability to visually attend to a target or object. Visual fixation occurs while maintaining focus on the object and typically occurs at a variety of distances and locations within the visual field. This is a skill that typically develops at about 4 weeks of age. Saccadic Eye Movement- These eye movements are those that occur very rapidly and allow us to smoothly shift vision between two objects without turning or moving the heads. Saccadic eye movement, or visual scanning is necessary for reading a sentence or paragraph as the eyes follow the line of words. This skill also allows us to rapidly shift vision between two objects without overshooting. In copying written work, this skill is very necessary. Smooth Pursuit Eye Movement- This ability allows us to steadily follow an object as it is visually tracked. When a smooth pursuit of eye movements occurs, the eyes do not lose track of the object, and occur without jerky movements or excessive head movements. Visual scanning occurs in vertical, horizontal, diagonal, and circular movements. Convergence- This ability is the simultaneous shift of both eyes together in an adducted position toward an object. The eyes work together to shift inward toward a target object, with single vision occurring with fixation on the object. Convergence is needed to focus on an object with both eyes together. Visual Spatial Attention- This skill includes awareness and attention in the body and the environment and allows us to attend to all visual fields. When visual spatial inattention occurs, visual neglect can occur.

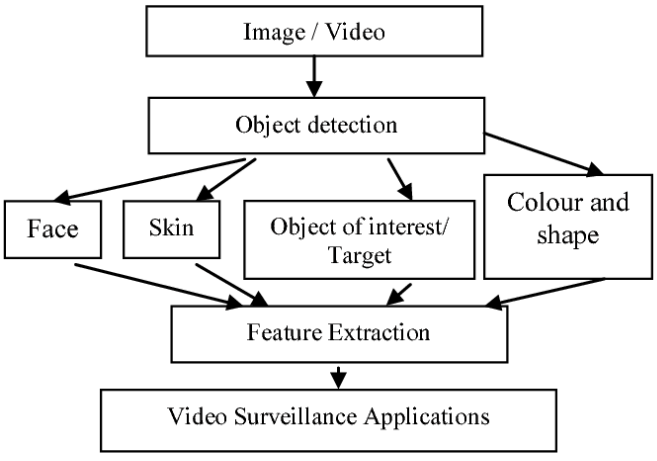
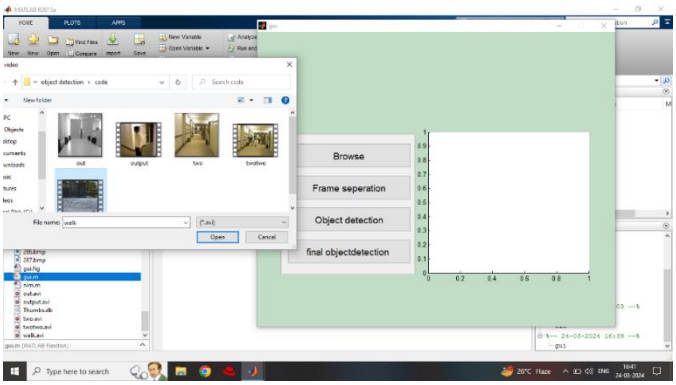


Fig. 1. Basic block diagram of object detection process

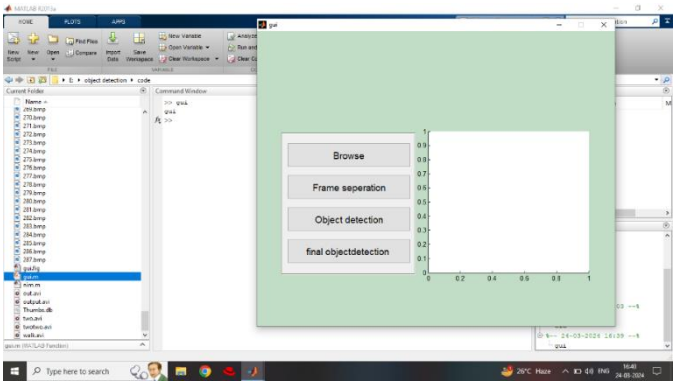
7. RESULTS AND CONCLUSION

In this section, we introduce the implementation details of our proposed model and analyze the effects of the modules in the network. We refer to our tracker as GDT (Gatedfusion Deformable object Tracker) and we compare GDT with state-of-the-art trackers in the benchmark datasets. To validate the effectiveness of each module, we separately train the following three models: network without the two modules (Baseline), network with the deformable convolution module only (Baseline+Deform), and the complete network (Baseline+Deform+Gate). We compare the performance of these models on the OTB-2013 dataset. In Fig. 6(left), we evaluate the overall detection performance by the success rate metric on the entire dataset. In Fig.(right), we measure the detection performance on the videos containing the deformation attribute. The challenges of each method – LIBS, W4, Behaviour Subtraction, Kalman Filter, Mean Shift Algorithm, Colour detection and Skin detection has been highlighted. In LIBS, the model fails to provide the most accurate results in the presence of dynamic objects in the background. If there are small changes in the background like the waving of leaves or any subtle changes that may occur in the background. In W4, only people in upright position can be detected using the cardboard model. If people are in different poses, or are crawling and climbing, it becomes challenging. In Behavior Subtraction, detection of spatial anomalies like U-turns is challenging in this method. If it is necessary in detecting outliers, only the ones that are spatially localized and temporal can be detected. Behavior camouflage takes place especially when there is a foreground object during background activity. Kalman Filter, Mean Shift Algorithm and GMM face the challenge of detecting multiple objects when there is slight occlusion. If multiple objects are present in the image, existing skin detection algorithms will fail to detect skin region. In colour Detection, existing algorithms can only detect primary colours with accuracy. If other different colours are present in an image, existing methods mis-detect the colours. Apart from these, some of the general issues are, if there is any change in illumination in the background, it could be mis-considered for a foreground object. Some methods also face challenges in detecting shadows. Similarity between the appearance of foreground object and background could create confusion of camouflage. Modelling of non-static backgrounds is another challenge. In High-traffic areas the background is frequently obstructed by many different foreground objects. Therefore, it will be difficult to classify a fixed foreground and background due to continuous change.

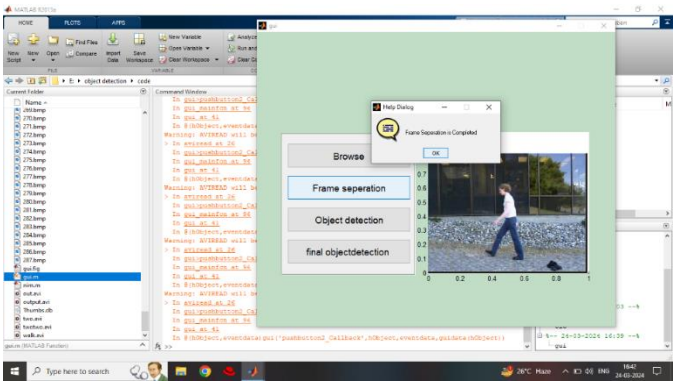


GUI Interface

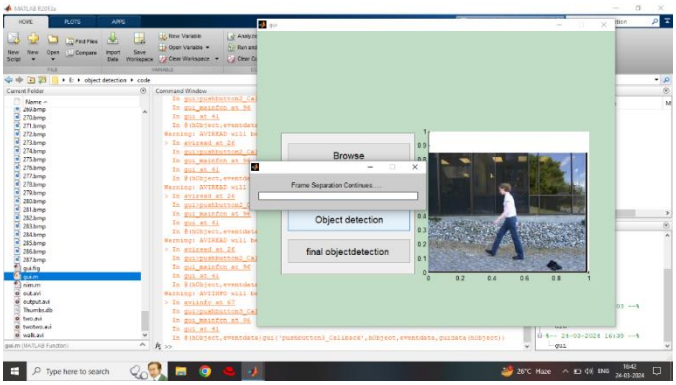
- The research successfully implemented algorithms that improved the real-time detection and tracking of objects in video surveillance.
- The thesis suggests further improvements could focus on handling more complex real-world scenarios and integrating deep learning methods for enhanced accuracy



Taking Input Video Data



Taking Frame Separation from Video



Taking Frame Separation from Video

8. CONCLUSIONS

We have proposed a deformable convolution layer to model target appearance variations in the CNN-based detection-by detection framework. We aim to capture target appearance variations via deformable convolution and supplement its normal convolution features through the online learned gating module. The gating module controls how the deformable convolutional features, and the normal features are fused. Experimental results show that the proposed tracker performs favourably against the state-of-the-art methods. There are still limitations in our proposed model. Our deformable convolution slightly degrades the robustness of the model, since its deformation estimation may fail in some extreme situations, including long-term occlusion, fast and large deformation, or significant illumination variation. For the future work, we would like to improve the robustness of the deformable convolution by enhancing the features extraction stage of our framework. This can be accomplished by collecting more data or adopting a data augmentation technique (e.g., image warping) to independently train the deformable convolution module. In addition, the online learned gating module may not be adequately adaptive to difficult videos. To alleviate this problem, we aim to improve the gating module in the offline training stage. Moreover, in the future, we will consider generalizing our approach to different sources of image data, e.g., RGB-D data and medical images.

9 REFERENCES

[1] Z. Kalal, K. Mikolajczyk, and J. Matas, “Detection learning detection,” TPAMI.
[2] H. Grabner, M. Grabner, and H. Bischof, “Real-time detection via on-line boosting.” in BMVC, 2006.
[3] B. Babenko, M.-H. Yang, and S. Belongie, “Visual detection with online multiple instance learning,” in CVPR, 2009.
[4] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S. L. Hicks, and P. H. Torr, “Struck: Structured output detection with kernels,” TPAMI.
[5] S. Hong, T. You, S. Kwak, and B. Han, “Online detection by learning discriminative saliency map with convolutional neural network,” in ICML, 2015.
[6] L. Zhang, J. Varadarajan, P. N. Suganthan, N. Ahuja, and P. Moulin, “Robust visual detection using oblique random forests,” in CVPR, 2017.
[7] H. Nam and B. Han, “Learning multi-domain convolutional neural networks for visual detection,” in CVPR, 2016.
[8] G. Ning, Z. Zhang, C. Huang, X. Ren, H. Wang, C. Cai, and Z. He, “Spatially supervised recurrent convolutional neural networks for visual object detection,” in IEEE International Symposium on Circuits and Systems, 2017, pp. 1–4.
[9] Y. Song, C. Ma, X. Wu, B. L. Gong, Lijun, W. Zuo, C. Shen, R. W. H. Lau, and M.-H. Yang, “Vital: Visual detection via adversarial learning,” in CVPR, 2018.
[10] S. Pu, Y. Song, C. Ma, H. Zhang, and M.-H. Yang, “Deep attentive detection via reciprocative learning,” in NIPS, 2018.
[11] B. Han, J. Sim, and H. Adam, “Branchout: Regularization for online ensemble detection with convolutional neural networks,” in ICCV, 2017.
[12] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, “Visual object detection using adaptive correlation filters,” in CVPR, 2010.