A New Approach to Moving Object Detection and Segmentation: The XY-shift Frame Differencing

N. Wondimu[1,4][a], U. Visser[3][b] and C. Buche[1,2][c]
[1]*Lab-STICC, Brest National School of Engineering, 29280, Plouzané, France*
[2]*IRL CROSSING, CNRS, Adelaide, Australia*
[3]*University of Miami, Florida, U.S.A.*
[4]*School of Information Technology and Engineering, Addis Ababa University, Addis Ababa, Ethiopia*

Keywords: Moving Object Detection, Frame Differencing, Object Segmentation, XY-shift Frame Differencing, Three-Frame Differencing.

Abstract: Motion out-weights other low-level saliency features in attracting human attention and defining region of interests. The ability to effectively identify moving objects in a sequence of frames help to solve important computer vision problems, such as moving object detection and segmentation. In this paper, we propose a novel frame differencing technique along with a simple three-stream encoder-decoder architecture to effectively and efficiently detect and segment moving objects in a sequence of frames. Our frame differencing component incorporates a novel self-differencing technique, which we call XY-shift frame differencing, and an improved three-frame differencing technique. We fuse the feature maps from the raw frame and the two outputs of our frame differencing component, and fed them to our transfer-learning based convolutional base, VGG-16. The result from this sub-component is further deconvolved and the desired segmentation map is produced. The effectiveness of our model is evaluated using the re-labeled multi-spectral CDNet-2014 dataset for motion segmentation. The qualitative and quantitative results show that our technique achieves effective and efficient moving object detection and segmentation results relative to the state-of-the-art methods.

1 INTRODUCTION

Motion is a key social stimulus that engages visual attention and induces autonomic arousal in the viewer (Williams et al., 2019; Mahapatra et al., 2008). Motion detection is extensively used in computer vision applications to facilitate the analysis of real world video scenes. Video surveillance (Sehairi et al., 2017) being a significant one, monitoring traffic and pedestrian (Zhao et al., 2016), robot control (Motlagh et al., 2014), target detection and counting (García et al., 2012), and detection of human activity (Zhang et al., 2012) are some of its applications in computer vision.

The process of relating moving region of interests (ROI) with object/s is called moving object detection. Moving object detection concerns how to take out moving objects from video frames and remove the background region and noise. Robust moving object detection enable computationally optimal foreground object detection and saliency prediction (Sengar and Mukhopadhyay, 2017).

The optical flow method use sequence of ordered images to estimate motion of objects as either instantaneous image velocities or discrete image displacements (Beauchemin and Barron, 1995). It is based on the properties of flow vector of the object over time to detect moving object regions. Optical flow method is highly complex and susceptible to noise corruption, fake motion and illumination variation.

On the other hand, background subtraction comprises of two steps: (i) background modeling and (ii) computation of difference between the current background model and the current video frame. The performance of background subtraction method is highly dependent on the accuracy of the background model. This method yield outstanding performance whenever the background model is accurate; otherwise wrong moving object detection will be undertaken (Garcia-Garcia et al., 2020). The performance of background subtraction technique degrades when it face videos

[a] https://orcid.org/0000-0002-0726-9892
[b] https://orcid.org/0000-0002-1254-2566
[c] https://orcid.org/0000-0003-0264-2683

Wondimu, N., Visser, U. and Buche, C.
A New Approach to Moving Object Detection and Segmentation: The XY-shift Frame Differencing.
DOI: 10.5220/0011664500003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 309-318
ISBN: 978-989-758-623-1; ISSN: 2184-433X

with smaller frame rate, camera jitter, and significant illumination change.

The other widely used moving object detection technique is frame differencing. It detects moving objects by taking pixel-by-pixel difference of consecutive frames in a video sequence. Frame differencing is the most common and computationally less complex method for moving object detection in scenarios where the scene is dynamic due to camera movement and mobility of objects in a video sequence. However, this method fails to detect whole relevant pixels of some types of moving objects (foreground aperture problem) (Xu et al., 2017). It also wrongly detects a trailing regions as moving object (known as ghost region) when there is an object that is moving fast in the frames (Fei et al., 2015). Most significantly, frame differencing fails to detect objects that preserve uniform regions.

To this end, we propose a novel XY-shift frame differencing technique along with three-stream encoder-decoder architecture to address the setbacks of frame differencing based moving object detection and segmentation techniques. The effectiveness of our frame diffrencing technique is analyzed both as a standalone high-pass filter algorithm and as an input for our improved three-frame differencing and encoder-decoder network. Furthermore, we implement three-stream encoder-decoder architecture to build simple but robust moving object detection model.

The rest of the paper is organized as follows. The second part briefly introduce related research works, the third part introduce the proposed method in detail, the fourth part of this paper show experimental analysis of the research work, and finally, a summary of this paper is presented.

2 RELATED WORKS

In the early days of moving object detection, researchers formulated a well established background subtraction techniques for stationary camera setting. These techniques have been extended for many years and are able to successfully detect moving objects as long as the camera is stationary (Stauffer and Grimson, 1999; Barnich and Van Droogenbroeck, 2010; Stauffer and Grimson, 1999; Dou et al., 2017; Cucchiara et al., 2003; Bouwmans, 2012; Sengar and Mukhopadhyay, 2017). However, a relatively long initialization time to model the background and residual image alignment error on non-stationary camera setting are the main setbacks of this technique (Zhang et al., 2006). These problems are incontestably addressed by optical flow based methods (Narayana et al., 2013; Ochs et al., 2013; Horn and Schunck, 1981). However, optical flow is highly dependent on optical flow vectors. The quality of the these vectors is crucial for the motion segmentation performance. Besides, optical flow is highly complex and due to high sensibility of noise corruption, it cannot meet the need of real time object detection. Optical flow exceptionally works well on large moving objects and fails to detect smaller objects due to blurry edges and low resolution (Rozantsev et al., 2016).

A variety of frame differencing techniques address the aforementioned problems. For instance, Inter-frame differencing technique generate the difference between two consecutive frames over a period of time for identifying background and foreground pixels. A research in (Liang et al., 2010) use inter-frame differencing algorithm to detect moving target in aviation video. Their experiment indicate that the algorithm has few computations and high accuracy to extract moving-target in aviation videos. In (Nakashima and Yabuta, 2018), interframe differencing and dynamic binarization using discriminant analysis is applied. The positions of the moving object in the image are determined by observing the histograms of each frame.

A slightly different method with comparable result with that of inter-frame differencing is three-frame differencing. This method put three adjacent frames as a group, subtracts both adjacent frames and lets two differential results do the logical AND operation. This has been the most widely used and traditional three-frame differencing method. In (Yin et al., 2016), traditional three-frame differencing technique and W4 algorithms are used to detect foreground objects in the infrared video datasets. Another research, (Sengar and Mukhopadhyay, 2016), propose a moving object detection method under static background. The algorithm use a non-overlapping blocks of the difference frames and calculate the intensity sum and mean of each block.

The inter-frame differencing and three-frame differencing techniques suffer from the foreground aperture and ghosting problems due to slow-moving as well as fast-moving foreground objects. Besides, these methods are known for partial or splitted detection of objects (Tsai and Yeh, 2013). As a result, frame differencing techniques are prone to false positives and sometimes false negatives.

To this end, combined moving object detection methods improved the performance of frame differencing techniques. Relatively robust moving object detection methods combine background subtraction with frame differencing (Weng et al., 2010; Xiao

et al., 2010; Cheng and Wang, 2014; Fei et al., 2015) or optical flow (Halidou et al., 2014; Zhang et al., 2006; Fernández-Caballero et al., 2010).

A static background based on three-frame differencing method in combination with background subtraction method is proposed in (Weng et al., 2010) and (Cheng and Wang, 2014). A combination of optical flow and three-frame differencing based moving object detection method is employed in (Halidou et al., 2014). It use region of interest (ROI) and multi-block local binary pattern descriptors. Another frame differencing and optical flow based moving object detection technique is proposed in (Fernández-Caballero et al., 2010). Here, a thermal infrared camera mounted on autonomous mobile robot is used as a feed to the detection module.

A method for detecting moving people in the indoor environment is proposed with the help of frame differencing and neural network based classification techniques (Foresti et al., 2005). This method reduce the false alarm and provides a robust classification with the help of a finite state automation. Similarly, a new approach based on fuzzy adaptive resonance theory, neural network with forgetting method for foreground detection and background establishment in natural scenes is proposed in (Dou and Li, 2014). On the other hand, the frame differencing and the non-pyramidal Lucas-Kanade approaches (Bouguet et al., 2001) are used to detect human candidates based on thermal signatures when the robot stops and moves.

In (Xu et al., 2017), an efficient foreground detection method is proposed by combining three-frame differencing and Gaussian mixture model. Another research work, (Lee et al., 2013), present a moving object detection method by combining background subtraction, separable morphological edge detector, and optical flow.

Recently, a more sophisticated and efficient moving object detection methods have been proposed by intriguing improved frame differencing techniques with deep neural network technologies. For instance, (Ellenfeld et al., 2021) propose a deep learning based moving object detection method. It use a Deep Convolutional Neural Network (DCNN) for multi-modal motion segmentation. Improved three-frame differencing and current RGB frame is used to capture temporal information and appearance of the current scene respectively. These inputs are later fused in the DCNN component for effective, efficient and robust motion segmentation. This model improved the performance of three-frame differencing techniques in detecting tiny moving objects.

A research work in (Yang et al., 2017) applied a frame differencing technique with Faster Region-Convolutional Neural Network (R-CNN) for highly precise detection and tracking characteristics. Similarly, (Mohtavipour et al., 2022) propose a multi-stream CNN and frame differencing based moving object detection method for deep violence detection. It use a handcrafted features related to appearance, speed of movement, and representative image and fed to a convolutional neural network (CNN) as spatial, temporal, and spatiotemporal streams.

Furthermore, (Siam et al., 2018) and (Wang et al., 2018) show promising results using CNN for moving object detection. They use a two-stream convolutional network to jointly model motion and appearance cues in a single convolutional network. In (Wang et al., 2018) a new framework named moving-object proposals generation and prediction framework (MPGP) is proposed to reduce the searching space and generate some accurate proposals which can reduce computational cost. In addition,they explored the relation of moving regions in feature map of different layers. This method utilize spatial-temporal information to strengthen the detection score and further adjust the location of the bounding boxes.

3 OUR APPROACH

We propose a novel moving object detection and segmentation method using XY-shift and improved three-frame differencing. Furthermore, we have extended our method by feeding it to a three-stream encoder-decoder network. In this section, we discuss details of our proposed technique.

3.1 The Proposed Framework

Fig 2 show our proposed framework, consisting of two major components namely: frame differencing and three-stream encoder-decoder network. The frame differencing component consists of two frame differencing methods. The first method is a novel XY-shift frame differencing technique and the second one is an improved three-frame differencing technique. The XY-shift frame differencing differs a frame from its shifted replica. The effect of this operation is equivalent to the result of a high-pass filter method, but with significantly smaller computational resource. We have mainly used this method to reduce the visibility of irrelevant background objects and expose foreground object even if they are in a temporarily static position. The mathematical formalization of XY-shift frame differencing is depicted as follows in equation 1. Let a be the first channel of image A with a shape of (h,w,3). Then, the XY-shift frame differ-

encing of a is calculated as:

$$g(a) = \begin{cases} a(x_i, y_j) - a(x_{i+f}+, y_{j+f}), & \text{if } i <= h - f \\ & j <= w - f \\ a(x_i, y_j) - a(x_{i-f}+, y_{j-f}), & \text{if i = h or j = w} \end{cases} \quad (1)$$

where h and w stands for the height and width of the channel and f is a shift-factor.

Minuned

00	01	02	03	04	05
10	11	12	13	14	15
20	21	22	23	24	25
30	31	32	33	34	35
40	41	42	43	44	45
50	51	52	53	54	55

Minuned ↦ Subtrahend

11	12	13	14	15	15
21	22	23	24	25	25
31	32	33	34	35	35
41	42	43	44	45	45
51	52	53	54	55	55
51	52	53	54	55	55

Figure 1: Tabular representation of XY-shift operands.

Figure 1 depicts a notational representation of XY-shift frame differencing. Assuming a 6X6 pixel raw image as a minuend, the subtrahend of XY-shift frame differencing technique with 1 shift-factor is constructed starting from the second row and column of the minuend.

What follows the XY-shift frame differencing is an improved three-frame differencing technique. This technique use the output of XY-shift differencing. It takes three consecutive frames, compute the difference between the current frame and the previous frame, the current frame and the next frame separately, and extract a pixel-wise max between these two resulting frames. This technique is adapted and enhanced to improve the extraction of temporal features from datasets in spatio-temporal domain. The improved three-frame differencing method is formalized as follows in equation 2. Let A,B, and C be the first channel of three consecutive XY-shift frame differenced frames with a shape of (h,w). Let B be the first channel of the current frame. Then the improved three-frame differencing, f(A,B,C), is calculated as:

$$f(A,B,C)_{i,j} = max_{i,j}(|B_{i,j} - A_{i,j}|, |B_{i,j} - C_{i,j}|) \quad (2)$$

where for i,j >= 0 and i<= h and j <= w. Furthermore, the pixel-wise maximum of two images is computed as shown in 3. Let Q1 be the absolute difference of the current frame B and its predecessor frame A. Let Q2 be the absolute difference of the current frame B and its successor frame C. Let's say both differenced images have a size of (h,w). Then, the pixel-wise maximum, P_max, of these two frames is calculated as:

$$max(Q1,Q2)_{i,j} = \begin{cases} Q1_{i,j}, & \text{if } Q_{i,j} > Q2_{i,j} \\ Q2_{i,j}, & \text{if otherwise} \end{cases} \quad (3)$$

where for i,j >= 0 and i<= h and j <= w.

The second component of our model constitutes a VGG-16 based encoder and a decoder network. The top five convolutional layers of VGG-16 along with the weights of ImageNet (Simonyan and Zisserman, 2014) are used as an encoder with the intention of avoiding excessive sparsity of hidden units.

The implementation detail is as follows. Three consecutive raw image frames are passed to the frame differencing component. The XY-shift frame differencing component takes each frame separately and perform XY-shift frame differencing. The result of XY-shift frame differencing on random images is presented in figure 3 Concurrently, the improved three-frame differencing takes all XY-Shift frame differenced frames and perform three-frame differencing for each consecutive frames using a pixel-wise maximum function, shown in equation 3.

The purpose of the XY-shift frame differencing as depicted in column two of Fig. 3 is to clear irrelevant background objects. This contributed a lot in reducing textures that affects the three-frame differencing negatively. The third column of Fig. 3 clearly shows the contribution of the XY-shift frame differencing in enhancing three-frame differencing techniques.

Furthermore, the result of the improved three-frame differencing, the XY-shift frame differencing, and the raw RGB image frames is fused into one future space using a residual layer and fed to the last encoder-decoder network. The VGG-16 based encoder extracts further features and the decoder segment produces the desired segmentation map as shown in Fig. 2.

3.2 Loss Function

Moving object detection and segmentation task is a binary pixel-wise classification task. Consequently, binary cross-entropy is chosen as the loss function to evaluate the model's performance during training.

The binary crossentropy loss function calculates the loss as shown in equation 4:

$$L_{BCI} = -\frac{1}{S}\sum_{i=1}^{S} y_i.\log\hat{y}_i + (1-y_i).(\log(1-y_i) \quad (4)$$

where $\hat{y}_i$ is the i^{th} scalar value in the model output, y_i is the corresponding target value, and S, the output size, is the number of scalar values in the model output.

3.3 Training Protocol

We use the (Tezcan et al., 2021)'s 4-fold cross validation strategy to split between training and test data

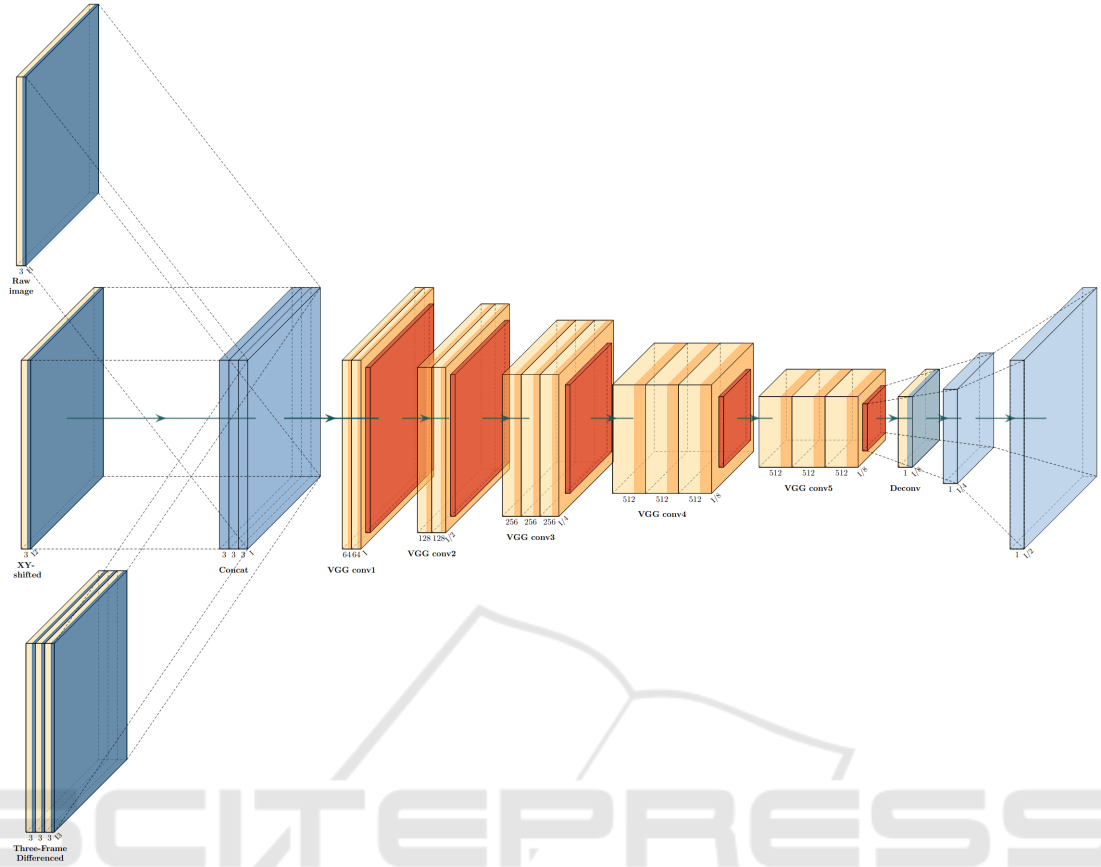

Figure 2: Moving object detection and segmentation architecture.

of CDNet-2014. All CDNet-2014 sequences are divided equally into four disjoint splits. The model is trained on three of the splits. The remaining split is used to evaluate the model's performance. We used a built-in python randomize function to select sequence of frames in each iteration from a varity of videos. A randomizer function is set to pick n number of frames in each iteration where n is the batch size. In this way we enhanced the representatives and complexity of the data; overcoming the possibility of overfitting at the same time. The designated training data is further divided into a training split (90 %) and validation split (10 %). Moreover, early stopping is used to prevent overfitting: if the validation loss does not improve in two consecutive epochs, the training process is terminated.

We used Adaptive Moment Estimation (Adam) (Kingma and Ba, 2014) to optimize the network during the training process. During training, the learning rate was set to 0.0001 and was decreased by a factor of 10 every 2 epochs. The network was trained for 10 epochs. The whole model is trained in an end-to-end manner. The entire training procedure takes about 8 hours using a single NVIDIA Quadro RTX 3000 Max-Q GPU.

4 EXPERIMENTS

4.1 Datasets

We use the relabeled version of CDNet-2014, (Wang et al., 2014) dataset, for training and evaluation. CDNet-2014 includes over 160,000 pixel-wise annotated frames in 53 video sequences subdivided in 11 categories and two spectra: VIS and thermal IR. The 53 sequences contain a large variety of different scenes with varying image quality and resolution ranging from (320 × 240) to (720 × 480) pixels. Most scenes show an urban environment with persons or cars. The dataset contains both indoor and outdoor scenes and covers many different real world challenges such as shadows, dynamic backgrounds, and camera motion. The ground truth for each image is a gray-scale image that describe the 4 motion classes: static, hard shadow, unknown motion, and motion.

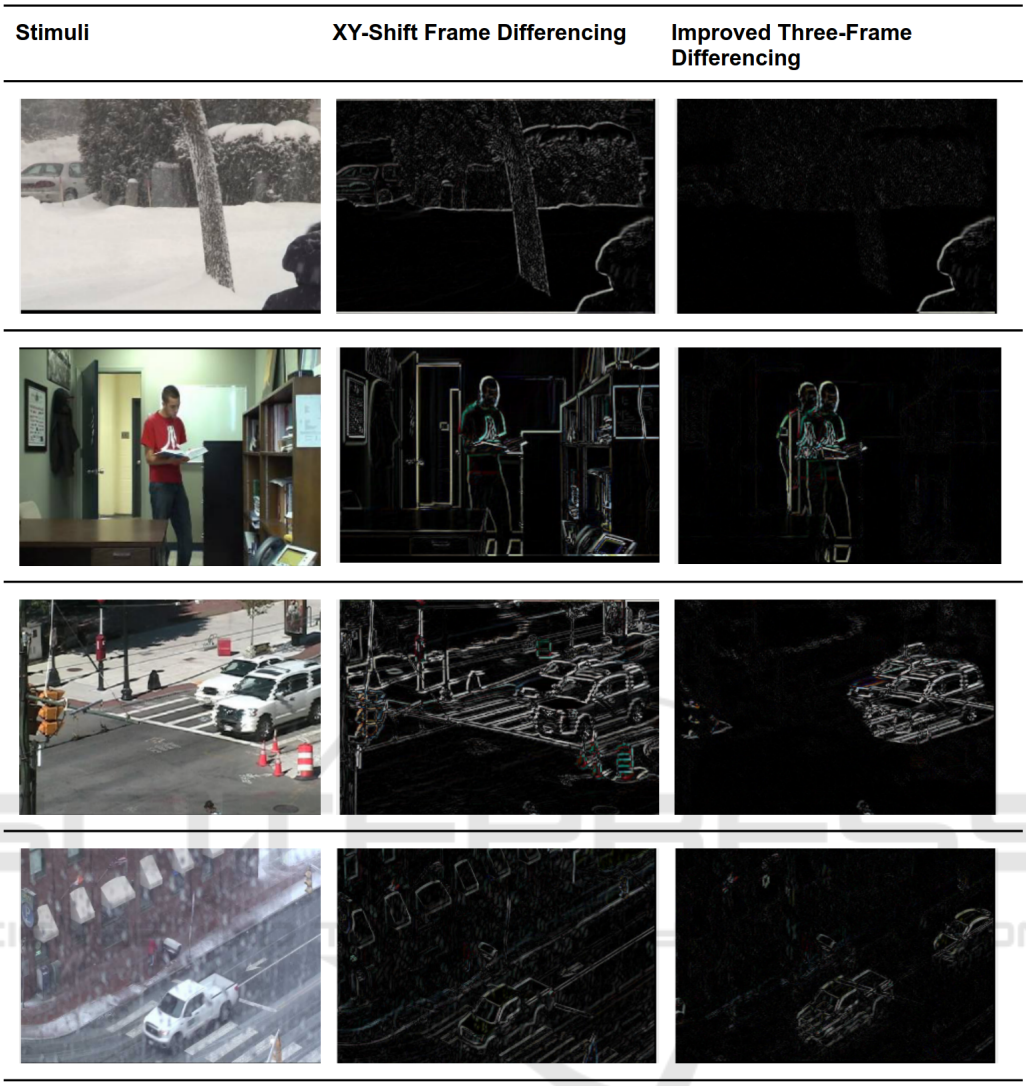

Figure 3: A random presentation of XY-shift frame differenced and Three-Frame differenced frames: frames not thresholded.

An additional class is used to mark areas that are outside the region of interest (non-ROI). Pixels annotated as non-ROI are discarded during evaluation.

4.2 Evaluation Metrics

We evaluate our model on the testing sets of CDNet-2014, in total of 11 video sequences with nearly 39,820 frames. We emphasized on the Recall (Re), Precision (Pr), and F1-score from the the standard evaluation measures (Wang et al., 2014). This is mainly due to the sufficiency of the selected metrics for the problem at hand.

4.3 Frame Differencing Experiments

We undertook an ablation analysis on the different components of our architecture. The significance of the XY-shift frame differencing and improved three-frame differencing is thoroughly analysed by eliminating and replacing each of them with inter-frame differencing (Liang et al., 2010; Nakashima and Yabuta, 2018) and traditional frame differencing (Yin et al., 2016; Sengar and Mukhopadhyay, 2016) techniques. The substitution of both of these frame differencing methods exhibit a reduced model performance. Especially, elimination of our XY-shift frame differencing technique and the use of inter-frame differencing as a source of input for our model caused a major performance degradation. The elim-

Table 1: Quantitative comparison of results.

Approach	Precision	Recall	F1-Score
(Bosch, 2021)	0.626	0.673	0.553
(Xiao et al., 2010)	0.462	0.513	0.42
STBGS (Bouwmans, 2012)	0.406	0.549	0.401
(Sengar and Mukhopadhyay, 2016)	0.375	0.58	0.389
(Ellenfeld et al., 2021)	0.774	0.751	0.745
Ours	**0.801**	**0.795**	**0.772**

ination of our frame differencing techniques significantly reduced our model's performance on most of its salient features, such as robustness against false-motion, temporarily at-rest object detection, and tiny moving object detection.

The use of our XY-shift frame differencing technique along with the improved three-frame differencing technique exhibit an outstanding performance. This is mainly due to the power of our XY-shift frame differencing technique in eliminating dynamic and noisy backgrounds. Moreover, the use of XY-shift differenced frames for three-frame differencing component further improved the performance of our model, especially in dynamic background scene videos. From the qualitative analysis point of view, the absence of XY-shift frame differencing affected our model's performance on tiny moving objects and temporarily at-rest foreground objects.

4.4 Optical Flow Experiments

Our second phase of ablation analysis is concerned with the popular optical flow technique. We converted our model into a two-stream architecture and assessed the impact of optical flow technique. Here, we used optical flow output and raw image as input source of a two-stream encoder-decoder architecture. The ability of our model to detect false-negatives was slightly compromised in this setup. We extended our model to a three-stream network by replacing the improved three-frame differencing segment by the optical flow output. This setting slightly improved its performance but with major deficiency to yield a competitive result.

4.5 Comparisons with the State-of-the-Art

Comparing ones model with the state-of-the-art requires ready-made repositories of research works that constitute the state-of-the-art code base. It has been difficult to find code bases of moving object detection and segmentation research works. However, we were able to compare our proposed model with five other state-of-the-art methods for motion detection and segmentation namely, (Bosch, 2021), (Xiao et al., 2010), (Sengar and Mukhopadhyay, 2016), and (Ellenfeld et al., 2021). Table X shows the quantitative results regarding Precision, Recall, and F1-score. Our approach outperforms most of these methods by a large margin and score a competitive result with the recent research work that combines three-frame differencing with DCNN (Ellenfeld et al., 2021).

The qualitative result is visualized in figure 4. What is depicted in the first row is the ground truth set by the CDNet-2014 dataset. The second raw show the raw appearance images for the corresponding ground truth. We used columns to showcase the performance of the state-of-the-art papers and our model over these images. Images are selected from different scene videos to show the generalization capacity of models in different environment. The qualitative analysis show the poor object detection and segmentation performance of background subtraction and frame differencing based algorithms in complex scene environment like, images with water in the background and dynamic scenes. As it can be seen in figure 4, these algorithms are highly prone to false positives and negatives. Compared to these algorithms, our model was able to detect both moving objects and temporarily at-rest objects effectively and efficiently.

The most robust model that we analysed in this section is (Ellenfeld et al., 2021). As it is discussed in the previous sections, this model combine improved three-frame differencing technique with appearance frame in a two stream encoder-decoder architecture. Compared to other deep learning based algorithms, these model show relatively better moving object detection and segmentation performance. Finally, our qualitative and quantitative results show the efficiency and effectiveness of our model. It has also scored a competitive result compared to the state-of-the-art methods. However, given the big error gap shown in Table 1, there is still a need to further enhance methods. Exploiting robust digital image processing and deep neural network technologies should be the focus of our next phase of research.

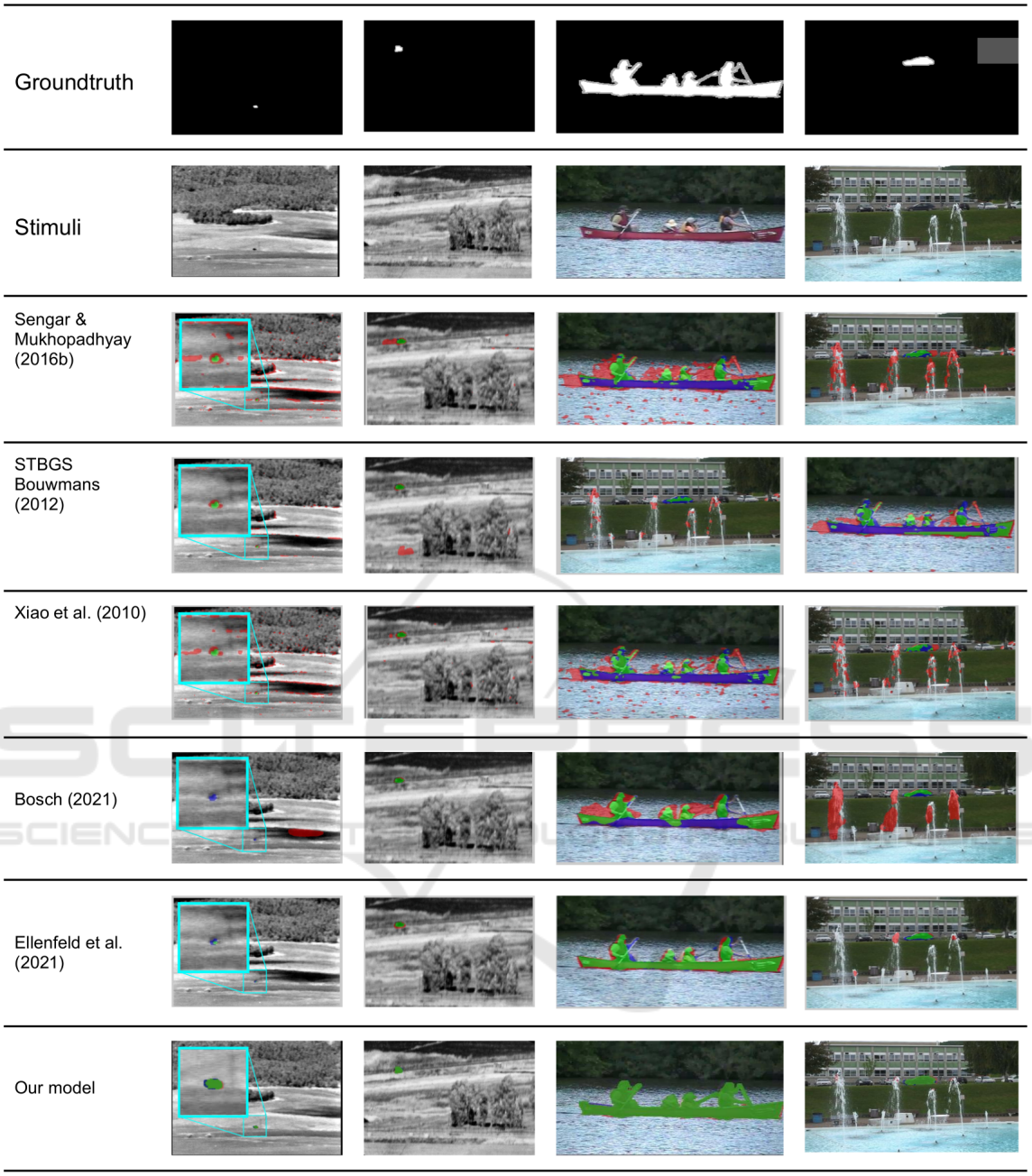

Figure 4: Qualitative evaluation of different models against our model: The green, blue, and red shades indicate the correct, false negative, and false positive classifications.

5 CONCLUSION

We proposed a novel moving object detection and segmentation technique. Our contribution in terms of architectural component is two fold: we propose a novel frame differencing technique, XY-shift frame differencing, and enhanced the traditional three-frame differencing technique. The XY-shift frame differencing is mainly used to sharpen the image and transform it to a high-pass filtered frame in relatively smaller computational resource. Image passed through this component eliminate most types of noises and irrelevant background objects. Feeding the output of the XY-shift frame differencing to the traditional three-

frame differencing technique happen to overcome the most common defects of three-frame differencing. The performance of the the three-frame differencing technique is improved by feeding XY-shift frame differenced frames instead of raw image and computing the pixel wise maximum of differences.

We have also introduced an efficient way of fusing raw appearance images with frame differenced images resulting in a significant improvement on moving object detection and segmentation models. We used a three-stream encoder-decoder deep neural network architecture. The raw appearance image, XY-shift frame differenced image, and three-frame differenced images are fed to their corresponding stream and later a feature space which is a deep fusion of the three stream data is produced. This intermediate feature space is fed to the encoder, which the top 5 layers of VGG-16, and the resulting feature has been deconvolved to get the desired segmentation map. The CDNet-2014 change detection dataset were used to build and analyse our model.

Our experimental analysis covered multiple perspectives of moving object detection and segmentation. We undertook an exhaustive ablation analysis by replacing our proposed frame differencing component with background subtraction, three-frame differencing, and optical flow based moving object detection and segmentation techniques. For the evaluation, we used precision, recall and F1-score metrics. Both qualitative and quantitative results show that the proposed approach outperform the state-of-the-art moving object detection and segmentation methods.

ACKNOWLEDGEMENTS

This work would not have been possible without the financial support of the French Embassy in Ethiopia, Brittany region administration and the Ethiopia Ministry of Education (MoE). We are also indebted to Brest National School of Engineering (ENIB) and specifically LAB-STICC for creating such a conducive research environment.

REFERENCES

Barnich, O. and Van Droogenbroeck, M. (2010). Vibe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image processing*, 20(6):1709–1724.

Beauchemin, S. S. and Barron, J. L. (1995). The computation of optical flow. *ACM computing surveys (CSUR)*, 27(3):433–466.

Bosch, M. (2021). Deep learning for robust motion segmentation with non-static cameras. *arXiv preprint arXiv:2102.10929*.

Bouguet, J.-Y. et al. (2001). Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel corporation*, 5(1-10):4.

Bouwmans, T. (2012). Background subtraction for visual surveillance: A fuzzy approach. *Handbook on soft computing for video surveillance*, 5:103–138.

Cheng, Y. H. and Wang, J. (2014). A motion image detection method based on the inter-frame difference method. In *Applied Mechanics and Materials*, volume 490, pages 1283–1286. Trans Tech Publ.

Cucchiara, R., Grana, C., Piccardi, M., and Prati, A. (2003). Detecting moving objects, ghosts, and shadows in video streams. *IEEE transactions on pattern analysis and machine intelligence*, 25(10):1337–1342.

Dou, J. and Li, J. (2014). Modeling the background and detecting moving objects based on sift flow. *Optik*, 125(1):435–440.

Dou, J., Qin, Q., and Tu, Z. (2017). Background subtraction based on circulant matrix. *Signal, Image and Video Processing*, 11(3):407–414.

Ellenfeld, M., Moosbauer, S., Cardenes, R., Klauck, U., and Teutsch, M. (2021). Deep fusion of appearance and frame differencing for motion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4339–4349.

Fei, M., Li, J., and Liu, H. (2015). Visual tracking based on improved foreground detection and perceptual hashing. *Neurocomputing*, 152:413–428.

Fernández-Caballero, A., Castillo, J. C., Martínez-Cantos, J., and Martínez-Tomás, R. (2010). Optical flow or image subtraction in human detection from infrared camera on mobile robot. *Robotics and Autonomous Systems*, 58(12):1273–1281.

Foresti, G. L., Micheloni, C., and Piciarelli, C. (2005). Detecting moving people in video streams. *Pattern Recognition Letters*, 26(14):2232–2243.

García, J., Gardel, A., Bravo, I., Lázaro, J. L., Martínez, M., and Rodríguez, D. (2012). Directional people counter based on head tracking. *IEEE Transactions on Industrial Electronics*, 60(9):3991–4000.

Garcia-Garcia, B., Bouwmans, T., and Silva, A. J. R. (2020). Background subtraction in real applications: Challenges, current models and future directions. *Computer Science Review*, 35:100204.

Halidou, A., You, X., Hamidine, M., Etoundi, R. A., Diakite, L. H., et al. (2014). Fast pedestrian detection based on region of interest and multi-block local binary pattern descriptors. *Computers & Electrical Engineering*, 40(8):375–389.

Horn, B. K. and Schunck, B. G. (1981). Determining optical flow. *Artificial intelligence*, 17(1-3):185–203.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, S., Kim, N., Paek, I., Hayes, M. H., and Paik, J. (2013). Moving object detection using unstable camera for consumer surveillance systems. In *2013 IEEE*

International Conference on Consumer Electronics (ICCE), pages 145–146. IEEE.

Liang, R., Yan, L., Gao, P., Qian, X., Zhang, Z., and Sun, H. (2010). Aviation video moving-target detection with inter-frame difference. In *2010 3rd International Congress on Image and Signal Processing*, volume 3, pages 1494–1497.

Mahapatra, D., Winkler, S., and Yen, S.-C. (2008). Motion saliency outweighs other low-level features while watching videos. In *Human Vision and Electronic Imaging XIII*, volume 6806, pages 246–255. SPIE.

Mohtavipour, S. M., Saeidi, M., and Arabsorkhi, A. (2022). A multi-stream cnn for deep violence detection in video sequences using handcrafted features. *The Visual Computer*, 38(6):2057–2072.

Motlagh, O., Nakhaeinia, D., Tang, S. H., Karasfi, B., and Khaksar, W. (2014). Automatic navigation of mobile robots in unknown environments. *Neural Computing and Applications*, 24(7):1569–1581.

Nakashima, T. and Yabuta, Y. (2018). Object detection by using interframe difference algorithm. In *2018 12th France-Japan and 10th Europe-Asia Congress on Mechatronics*, pages 98–102. IEEE.

Narayana, M., Hanson, A., and Learned-Miller, E. (2013). Coherent motion segmentation in moving camera videos using optical flow orientations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1577–1584.

Ochs, P., Malik, J., and Brox, T. (2013). Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1187–1200.

Rozantsev, A., Lepetit, V., and Fua, P. (2016). Detecting flying objects using a single moving camera. *IEEE transactions on pattern analysis and machine intelligence*, 39(5):879–892.

Sehairi, K., Chouireb, F., and Meunier, J. (2017). Comparative study of motion detection methods for video surveillance systems. *Journal of Electronic Imaging*, 26(2):023025.

Sengar, S. S. and Mukhopadhyay, S. (2016). A novel method for moving object detection based on block based frame differencing. In *2016 3rd International Conference on Recent Advances in Information Technology (RAIT)*, pages 467–472. IEEE.

Sengar, S. S. and Mukhopadhyay, S. (2017). Foreground detection via background subtraction and improved three-frame differencing. *Arabian Journal for Science and Engineering*, 42(8):3621–3633.

Siam, M., Mahgoub, H., Zahran, M., Yogamani, S., Jagersand, M., and El-Sallab, A. (2018). Modnet: Motion and appearance based moving object detection network for autonomous driving. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2859–2864.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Stauffer, C. and Grimson, W. E. L. (1999). Adaptive background mixture models for real-time tracking. In *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149)*, volume 2, pages 246–252. IEEE.

Tezcan, M. O., Ishwar, P., and Konrad, J. (2021). Bsuv-net 2.0: Spatio-temporal data augmentations for video-agnostic supervised background subtraction. *IEEE Access*, 9:53849–53860.

Tsai, C.-M. and Yeh, Z.-M. (2013). Intelligent moving objects detection via adaptive frame differencing method. In *Asian Conference on Intelligent Information and Database Systems*, pages 1–11. Springer.

Wang, H., Wang, P., and Qian, X. (2018). Mpnet: An end-to-end deep neural network for object detection in surveillance video. *IEEE Access*, 6:30296–30308.

Wang, Y., Jodoin, P.-M., Porikli, F., Konrad, J., Benezeth, Y., and Ishwar, P. (2014). Cdnet 2014: An expanded change detection benchmark dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 387–394.

Weng, M., Huang, G., and Da, X. (2010). A new interframe difference algorithm for moving target detection. In *2010 3rd international congress on image and signal processing*, volume 1, pages 285–289. IEEE.

Williams, E. H., Cristino, F., and Cross, E. S. (2019). Human body motion captures visual attention and elicits pupillary dilation. *Cognition*, 193:104029.

Xiao, J., Cheng, H., Sawhney, H., and Han, F. (2010). Vehicle detection and tracking in wide field-of-view aerial video. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 679–684. IEEE.

Xu, Z., Zhang, D., and Du, L. (2017). Moving object detection based on improved three frame difference and background subtraction. In *2017 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII)*, pages 79–82. IEEE.

Yang, Y., Gong, H., Wang, X., and Sun, P. (2017). Aerial target tracking algorithm based on faster r-cnn combined with frame differencing. *Aerospace*, 4(2):32.

Yin, J., Liu, L., Li, H., and Liu, Q. (2016). The infrared moving object detection and security detection related algorithms based on w4 and frame difference. *Infrared Physics & Technology*, 77:302–315.

Zhang, Y., Kiselewich, S. J., Bauson, W. A., and Hammoud, R. (2006). Robust moving object detection at distance in the visible spectrum and beyond using a moving camera. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 131–131. IEEE.

Zhang, Y., Liu, X., Chang, M.-C., Ge, W., and Chen, T. (2012). Spatio-temporal phrases for activity recognition. In *European Conference on Computer Vision*, pages 707–721. Springer.

Zhao, N., Xia, Y., Xu, C., Shi, X., and Liu, Y. (2016). Appos: An adaptive partial occlusion segmentation method for multiple vehicles tracking. *Journal of Visual Communication and Image Representation*, 37:25–31.