Anomalous Activity Detection for Surveillance Robots
Extended Abstract

Mritunjoy Halder[1], **Snehasis Banerjee**[1,2], **Balamuralidhar Purushothaman**[1]
[1] TCS Research, Tata Consultancy Services, India
[2] Robotics Research Center (RRC), IIIT-Hyderabad, India

Abstract

Can a surveillance robot autonomously detect anomalous activity from its ego view camera perception? This is a challenging task as it requires identifying what is normal and what is an abnormal pattern – given the variations of possible anomalies and abnormalities. This paper presents an architecture and method based on a spatio-temporal convolution neural network to detect and classify anomalies. This work is inspired by the 'Konio-Magno-Parvocellular' cells of the human brain, which is claimed to aid humans in organizing changes in perceived scenes. The model is trained and tested on a benchmark video dataset (Kaushik 2020) of human activity. We have obtained 74% testing F1-score on this dataset. Experiments in simulation as well as deployment on a real robot shows that the proposed methodology can identify anomalous activities effectively. We also list down the observations from practical deployment of the model.

Introduction

With the recent advances in robotics and deep learning, surveillance robots have increasingly become relevant as a solution offering for large facilities like airports, parks, campus, factories, office, etc. The cost of installing huge number of CCTV cameras for securing large spaces is high – both in terms of installation and maintenance; and there is always a chance of missing some important view angles, and scope of manual camera connectivity tampering. In contrast, surveillance robots, similar to a human security, can serve as trustworthy autonomous agents, with more dynamic range of area coverage. A surveillance robot's task is to patrol a pre-specified area to ensure the space's security, and report to an administration stakeholder or an authority if any anomaly (unusual pattern) is perceived by its onboard sensors (like ego view camera). This anomaly detection task helps prevent varied crimes and enables authorities to react quickly to unpleasant situations(Sahay and et. al. 2022; Bozcan and et. al. 2021; Al-amri and et. al. 2021). Among various types of anomalies possible, unusual human activity detection is a key sub-task of a surveillance robot. Prior work has focused on detecting anomalies using auto-encoders, unsupervised learning-based neural nets, one-class classification (Bozcan and et. al. 2020; 2021) etc. However, in prior work, the datasets used are synthetically generated, having low train-test accuracy, prone to overfitting, lacks real world testing, and are mostly heuristic-based. This paper's contributions are enlisted below:

(1) As explained in the Fig. 1, the proposed model architecture tries to mimic the 'Konio-Magno-Parvocellular' cells of the human brain, which are responsible for detecting slow, normal and swift changes in perceived scenes. To the best of our knowledge, this is a new way to look at this problem.

(2) To detect anomalous human activity, the static (non-moving) portion of the scene need to be treated as redundant. So using SSIM (Wang and et. al. 2004) method, only the region of interest is used to train the model. To detect the abnormal change in a scene, we have used Farneback (Farneback 2000) algorithm to identify the change in the flow of the intensities in different frames. We have presented the experimental results obtained on a human activity video dataset and validated the functionality with a real life robot deployment.

Related Work

Here we discuss some of the relevant prior work on anomalous activity detection in robotics. (Wellhausen, Ranftl, and Hutter 2020) presented the idea of an auto-encoder being intentionally made to overfit the normal activities class, but had a drawback of classifying even slightly unseen normal activities as abnormal. (Shin and Na 2020) has used thermal image processing and point cloud analysis for anomaly detection, however they cater to specific anomaly types and need relevant and often costly sensors for their method. (Bera and et. al. 2016) used a Kalman filter and clustering to detect outliers, but their approach did not generalize well to crowds. (Fadjrimiratno and et. al. 2021) used a combination of (a) fast modules to detect co-ordinate pairs in image (b) slow modules to detect neighboring region anomalies, however its performance was reported as low. Another work (Park and et. al. 2021) used an auto-encoder to minimize the reconstruction error; and while testing if that error was above a threshold value, an anomaly was said to be present; however, that work heavily depended on the threshold parameter and is not generalizable to a varied class of anomalies. To overcome the aforementioned limitations, we have developed a method to effectively process video data obtained from mobile robot's on-board camera to raise an

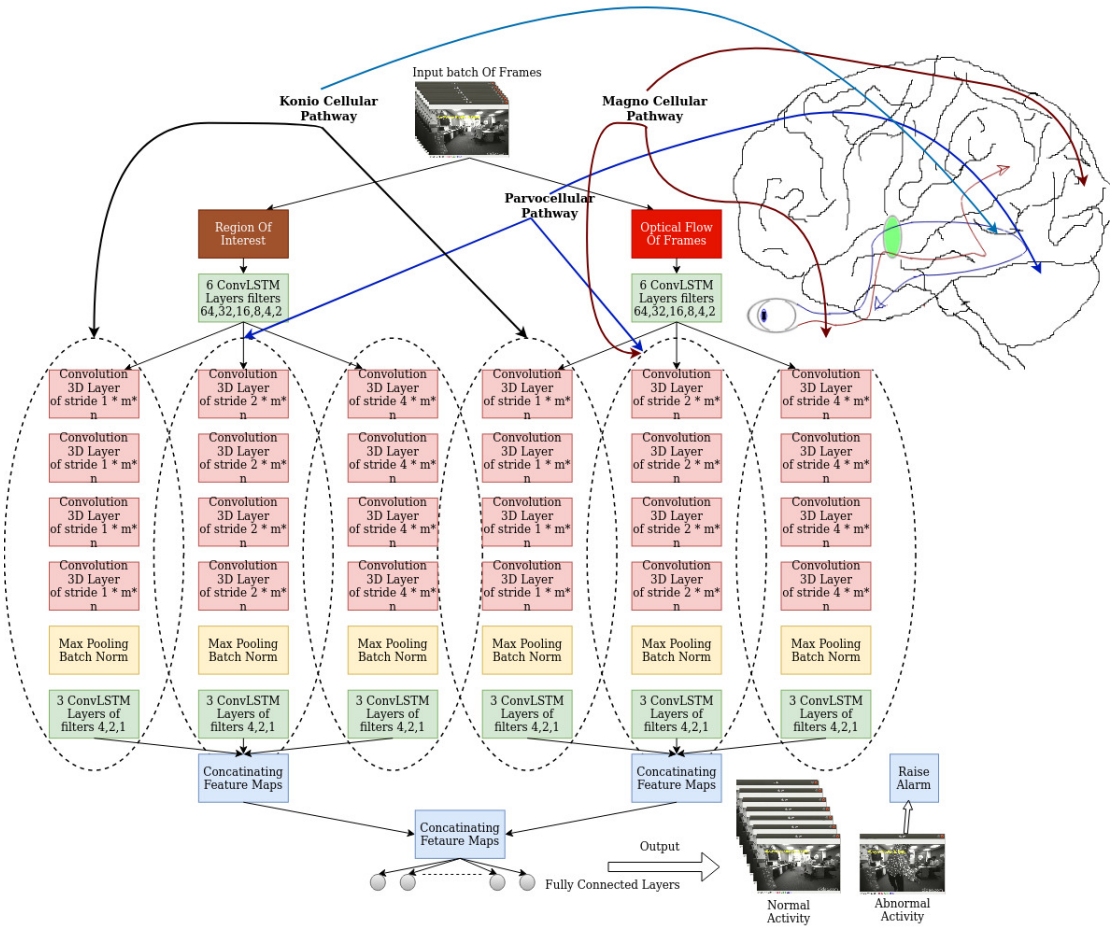

Figure 1: Architecture of anomaly detection for mobile robot, inspired by human brain

alarm if a specific type of anomaly (here violent human activity) is detected.

Our Approach

In this work, we have proposed an optical flow frame(OFF) (Shah and Xuezhi 2021) and SSIM based Spatio-temporal neural network (Fig. 1) that can detect if any human activity (like fighting – violence) is happening in indoor scenes. This architecture can be used to detect different types of anomalies like arson, vandalism, shoplifting, etc. At a timestep, the network takes 20 frames (20 fps being the frame rate for the robot ego view camera) as an input; and predicts presence or absence of anomaly by analyzing the Spatio-temporal features of those frames.

Optical Flow between Frames: It has been observed that many of the previous state-of-the-art methods used the complete frame as an input to detect an anomaly (Bozcan and et. al. 2020; Zaheer and et. al. 2021). However, if we are detecting some abnormal movement in humans only, then most of the non-moving object in the frame becomes redundant, leading to more computation and probable misclassification. To deal with this problem, we have used the optical flow of frames. Instead of taking all the frames as input, we fix the first frame, and then with respect to the first frame, we have calculated the optical flow of the other frames within a timestep of a second (20 frames). For example, if the first frame has an intensity of $I(x, y, t)$, where (x, y) are the pixel co-ordinates at timestep t, and the i^{th} frame has an intensity of $I(x+dx, y+dy, t+dt)$ then from the Taylor approximation, we can say that:
$\frac{dI}{dx}\delta x + \frac{dI}{dy}\delta y + \frac{dI}{dx}\delta t = 0$,
as the change in intensity is infinitesimal. Using this formula, we have calculated the optical flow between frames and fed that as a part of the input to the neural network architecture described later.

Removal of Non moving objects: As the non-moving objects are redundant features (that do not change over a prolonged time span), we can safely remove them from the images. To achieve our goal, we have fixated one frame, then using equation 1, which gives the Structural Similarity Index Measure (SSIM) between the initial and i^{th} frame, we have calculated the net change in the frames.

$$SSIM(F_0, F_i) = \frac{(2\mu_{F_0}\mu_{F_i} + K_1)(2\sigma_{F_0F_i} + K_2)}{(\mu^2{}_{F_0} + \mu^2{}_{F_i} + K_1)(\sigma^2{}_{F_0} + \sigma^2{}_{F_0} + K_2)} \quad (1)$$

where F_0 is the initial frame, F_i is the i^{th}. frame, μ_x stands

for mean, σ_x is the variance and $\sigma_x y$ is the covariance. By this equation, we can get the changes in the image at the pixel level. So the other parts which are not changing are made white colored (or any consistent pattern over frames – to be ignored by model). This processed image is given as an input to the neural network architecture.

Network Architecture As shown in Fig. 1, the network takes optical flow between frames and only the moving object view as an input. The network has two symmetric parts. The first part takes the optical flow between frames as input, and the second part takes the moving object as input. Each piece of the network initially has six convolution LSTM layers having filters 64,32,16,8,4,2,1; used to memorize the changes in the moving object. Then the feature map obtained from the LSTM layer is fed into the convolution 3D network. The convolution 3D network has three parts, 'the normal pathway', 'the fast pathway', and 'the faster pathway', where the instances of the video are analyzed in different frame rates. The model is inspired by and resembles the magnocellular cells, parvocellular cells, and koniocellular cells of the lateral geniculate nucleus of the human brain. The video clip (frame of robot camera) is inputted as a spatio-temporal volume in the convolutional model. Finally, all the sub-architectures (pathways – mentioned below) are concatenated, and ultimately the last two components of the neural network are concatenated with a fully connected layer to give the output as either normal or anomaly. Grouping of temporally close anomalies is done to avoid repeating alarms.

The Normal Pathway (Koniocellular Pathway): This part of the neural network can be considered as monocellular cells of the human brain. This part has a stride, half of the fast pathway and one-fourth of the faster path. So this part of the neural network catches the minimal change in the scene like the koniocellular cells. After getting τ frames, this part of the neural network processes ρ frames at a time. Here, ρ (stride) equals one. So it records the minimal change in the frame-to-frame transition.

The Fast Pathway (Parvocellular Pathway): The fast pathway acts like the parvocellular cells of the human brain. The parvocellular cells of the human brain can detect slow and sustained changes. The stride of the fast pathway is $(\alpha * \rho)$ (here it is two), which is α times the normal pathway. So if there is a regular change in human behavior, like moving from one place to another, it detects that. As it processes $(\alpha * \rho)$ frames at a time, it can see the changes, which is fast but not rapid.

The Faster Pathway (Magnocellular Pathway): The faster pathway imitates the magnocellular cells of the human brain, that are responsible for detecting swift changes in perception. This part of the neural network processes the spatio-temporal features of $(\beta * \alpha * \rho)$ frames at a time (β being the multiplier, here stride is four). So if there are any swift changes in a scene, it can handle that change.

Experimental Results and Deployment

Experiments are done on Video Fight Detection Dataset of Kaggle (Kaushik 2020), containing videos of 100 fights and 101 normal situations (2 classes). We have used 190 videos from training and ten videos for testing. On training the model for 10 epochs, we obtained a training accuracy of 95.06% and a testing accuracy of 75% percent. The hyper-parameter settings are: Learning Rate: 0.0001; Loss Function: Sparse Categorical Cross-entropy; Optimizer: Adam; Dropout: 0.25. Table 5 describes the training and testing accuracies of the model per epoch, averaged over cross-validation. Table 2 shows the precision,recall and accuracy scores. Table 1 represents the confusion matrix.

	Predicted Yes	Predicted No
Actual Yes	49	20
Actual No	17	54

Table 1: Confusion Matrix

Class	Precision	Recall	F1 Score
0	0.74	0.71	0.73
1	0.73	0.76	0.74

Table 2: Classification Report

To prove that SSIM and Optical flow frames produces superior result, we shows results on the entire image frame. Table 3 shows the accuracy and loss for this case which is lower. We have compared our method with ViViT and CNN-RNN and the results of the methods are as enlisted in Table 4. We can see that the we have achieved better results in terms of precision and F1 Score and also have obtained better accuracy than ViViT and almost similar accuracy as CNN-RNN's best reported accuracy. Furthermore, the methods has not been tested to work in real world.

Deployment in Robot

We have deployed our model in a Double3[1] robot over a custom built software layer in office campus settings, with onboard camera having frame capture rate of 20 fps. By creating situations of conflict (based on hand gestures and body movements), we have verified the anomaly prediction to be correct most of the time. The major observations are: (i) if the input frame rate is low, the model does not work well for rapid activities; (ii) frame losses can be expected due to (a) compute time of the incoming frames and (b) network connectivity glitches; (iii) frames captured in dim light conditions leads to more false alarms; (iv) robot's positioning at an optimal camera view angle leads to better accuracy. This makes way for research on finding the best camera view for surveillance in the world frame. Fig. 2a demonstrates how the model has classified normal situation using robot ego view camera and in Fig. 2b the same is tested to classify abnormal situations.

Conclusion and Future Work

In this work, we have build a model for a surveillance robot to detect anomalous human activities. By evaluating the

[1]https://www.doublerobotics.com/

Epoch	1	2	3	4	5	6	7	8	9	10
Train Loss	0.69	0.69	0.68	0.67	0.65	0.63	0.62	0.60	0.58	0.57
Train Accuracy	0.53	0.54	0.54	0.59	0.66	0.64	0.66	0.68	0.69	0.70
Test Loss	0.70	0.70	0.70	0.65	0.65	0.62	0.65	0.65	0.65	0.65
Test Accuracy	0.47	0.47	0.47	0.68	0.60	0.66	059	0.66	0.64	0.64

Table 3: Table For Average Loss and Accuracy during Experimentation for 10 epochs (Entire Image Frame)

Method	Accuracy	Recall	Precision	F1 Score
ViViT(Aritra Roy Gosthipaty 2022)	0.69	0.95	0.61	0.73
CNN-RNN(Paul 2021)	0.78	1.0	0.60	0.75

Table 4: Classification Results of ViVit and CNN-RNN

change in structural similarity between several frames, we created a neural network design that exploits optical flow between frames and only takes the moving portion of the image into account. The network also uses varied strides to change the frame processing mode. We were able to achieve a training accuracy of 95% and a testing accuracy of roughly 70%. We tested our model using intentionally generated anomalous scenes after deploying it in a real robot. In future work, we plan to test this model in a variety of anomalous activities that can be detected from camera sensor. Furthermore, by attaching the robot with additional sensors (like infrared, thermal, microphone array), we will test the architecture's generality to various modes of input and varied surveillance use cases.

References

Al-amri, R., and et. al. 2021. A review of machine learning and deep learning techniques for anomaly detection in iot data. *Applied Sciences* 11(12):5320.

Aritra Roy Gosthipaty, A. T. 2022. Video vision transformer.

Bera, A., and et. al. 2016. Realtime anomaly detection using trajectory-level crowd behavior learning. In *IEEE CVPR workshops*, 50–57.

Bozcan, I., and et. al. 2020. Uav-adnet: Unsupervised anomaly detection using deep neural networks for aerial surveillance. In *IROS*, 1158–1164. IEEE.

Bozcan, I., and et. al. 2021. Gridnet: Image-agnostic conditional anomaly detection for indoor surveillance. *RAL* 6(2):1638–1645.

Fadjrimiratno, M. F., and et. al. 2021. Detecting anomalies from human activities by an autonomous mobile robot based on" fast and slow" thinking. In *VISIGRAPP (5: VISAPP)*, 943–953.

Farneback, G. 2000. Fast and accurate motion estimation using orientation tensors and parametric motion models. In *ICPR*, volume 1, 135–139. IEEE.

Kaushik, N. 2020. Video fight detection dataset.

Park, K. M., and et. al. 2021. Collision detection for robot manipulators using unsupervised anomaly detection algorithms. *IEEE/ASME Transactions on Mechatronics*.

Paul, S. 2021. Video classification with a cnn-rnn architecture.

Epoch	1	2	3	4	5	6	7	8	9	10
Train Loss	0.68	0.61	0.57	0.54	0.52	0.50	0.45	0.42	0.36	0.31
Train Accuracy	0.55	0.69	0.71	0.72	0.74	0.74	0.79	0.81	0.83	0.86
Test Loss	0.63	0.54	0.58	0.50	0.47	0.51	0.51	0.52	0.49	0.53
Test Accuracy	0.71	0.73	0.71	0.75	0.76	0.75	0.76	0.79	0.74	0.74

Table 5: Table For Average Loss and Accuracy during Experimentation for 10 epochs (SSIM and Optical flow)

(a) Some instances where Normal situation is detected with robot ego view camera

(b) Some instances where Abnormal situation is detected with robot ego view camera

Figure 2: Detecting Normal and Anomalous Scenes in camera enabled Double3 wheeled Robot

Sahay, K. B., and et. al. 2022. A real time crime scene intelligent video surveillance systems in violence detection framework using deep learning techniques. *Computers and Electrical Engineering* 103.

Shah, S. T. H., and Xuezhi, X. 2021. Traditional and modern strategies for optical flow: an investigation. *SN Applied Sciences* 3(3):1–14.

Shin, H.-c., and Na, K. 2020. Anomaly detection using elevation and thermal map for security robot. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, 1760–1762. IEEE.

Wang, Z., and et. al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13(4):600–612.

Wellhausen, L.; Ranftl, R.; and Hutter, M. 2020. Safe robot navigation via multi-modal anomaly detection. *IEEE Robotics and Automation Letters* 5(2):1326–1333.

Zaheer, M. Z., and et. al. 2021. An anomaly detection system via moving surveillance robots with human collaboration. In *IEEE ICCV*, 2595–2601.