iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IX **Month of publication:** September 2021

DOI: https://doi.org/10.22214/ijraset.2021.37972

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IX Sep 2021- Available at www.ijraset.com

Automation in Surveillance: A Review of the Developments

Sai Shashank Mukkera
Independent Researcher

Abstract: The purpose of this paper is to conduct a literature survey on the developments of surveillance systems while laying strong emphasis on AI-related developments. With the rise in availability of video data and advancements in computer vision techniques, the role of AI in surveillance has become more prominent. Most surveillance footage is used for investigative purposes after an event has occurred instead of providing real-time alerts. The systems discussed in this paper help achieve this goal of providing real-time alerts. Limitations and strengths of the systems are also discussed.

I. INTRODUCTION

The decreasing costs of video surveillance equipment have resulted in large volumes of video data. This excessive amount of information has not been met with adequate human operators. Not to mention the non-stop human attention required to monitor these footages (1,2,3). Surveillance cameras record large amounts of footage each day which consumes much space. Searching in these record files consumes much time as well (4). The video generated in current systems is being used mainly for investigative purposes after an event has happened, in contrast to using it as a mechanism for real-time alerts during an event (5). Perceiving meaningful activities in a long video sequence is challenging due to ambiguous definitions of 'meaningfulness' and clutters in the scene. The chance of an anomaly occurring is frustratingly low, making over 99.9% of the effort to watch videos wasted (6,2). Often the cost of installing such systems is very high, requiring a lot of additional hardware like new surveillance cameras. Considering how many infrastructure projects already have surveillance cameras, owners prefer systems that can work with their existing cameras (7). Recent computer vision algorithms can be exploited to avoid deploying many expensive sensors (8).

Given the video surveillance industry's phenomenal expected growth, less than 1% of video footage is likely to be analysed. (9).

This paper surveys the development of surveillance systems starting from Hardware optimization to Deep Learning based approaches while presenting the strengths and limitations of the systems.

II. HARDWARE OPTIMIZATION

Smart homes are moving towards wireless remote control, multi-media control, and high-speed data transmission. The authors of (10) design a system based on ZigBee technology that can send abnormal images and warning messages through MMS and SMS. In the proposed architecture, the ZigBee module connects household appliances and temperature and smoke sensors to a motherboard controlled by a -In the proposed architecture, the ZigBee module connects household appliances and temperature and smoke sensors to a motherboard controlled by an S3C44BOX-32 microcontroller. This system uses a CMOS camera to monitor positions and can be remotely controlled by SMS. As for the software, the system can adapt to the background environment changes through self-learning. For accurate monitoring, the system uses a combination of static and dynamic thresholds based on the difference in the background and the pictures taken for motion detection, specific object tracking, etc. These abnormal detections are also sent to the police. It can also make a judgment based on a control image.

On final testing, it was found that the system could transmit data (detections in this case) with very little packet loss (0% for distances less than 60m), making this system very optimal for family use, all of this while reducing power consumption.

As the software of this system can only perform intrusion detection, we can conclude that this system wouldn't suffice for use cases outside home surveillance, especially in more crowded scenes like parks, malls, etc.

III. OBJECT DETECTION

One of the significant issues to be resolved before developing a fully automated surveillance system is object detection. A surveillance system must be able to detect objects, classify them and track some of their activities. To tackle this issue, authors Omar Javed and Mubarak Shah proposed a method to classify objects based upon detecting recurrent motion for each tracked object. They develop a specific feature vector called a "Recurrent Motion Image" (RMI) to calculate the repeated movement of objects. These RMIs are then used to classify things as different objects yield very different RMIs (11).

Authors Francesco Turchini, Lorenzo Seidenari, Tiberio Uricchio, and Alberto Del Bimbo in (8) presented an integrated, flexible system that can tackle two primary surveillance problems: object counting and anomaly detection with localization.
The system proposed in this paper (Figure 1) was also able to perform tasks that weren't tackled in previous works like, Abnormal car behaviour detection, Parking duration understanding, Generic anomaly detection, and even accurately estimating the number of objects in the area of interest.

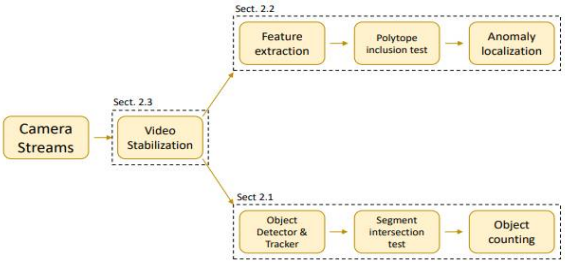


Figure 1: source: Deep Learning Based Surveillance System for Open Critical Areas

The method proposed uses YOLOv2 for object detection trained on the COCO dataset capable of detecting 80 classes.
Tracking, however, comes with a few challenges of maintaining consistency, especially amidst a few inconsistent classifications. To solve this problem, the authors use a greedy association multi-tracker algorithm. To associate a set of detections $\mathbf{D_t}$ to tracks $\mathbf{T_{t-1}}$ (possibly empty) from the previous frame, an association matrix $\mathbf{M}$ is computed such that $\mathbf{M_{ij}} = \frac{d_i \cap t_j}{d_i \cup t_j}$ (intersection over union measure). Then the function associate described in Algorithm1 (Figure 2:source: Deep Learning Based Surveillance System for Open Critical Areas) is applied.

Algorithm 1 Data association algorithm. We associate tracks and unassociated detection if IoU $> \tau$ and remove a track if it is "dead" for ω frames. Matrix **A** keeps track of associations and vector **l** counts the number of frames in which a track i is not associated with any detection.

FUNCTION associate $(\mathcal{T}_{t-1}, \mathcal{D}_t)$
Data: $\mathcal{T}_{t-1} : \{t_1 \dots t_n\}, \mathcal{D}_t : \{d_1 \dots d_m\}, \mathbf{M}_{ij} = \frac{d_i \cap t_j}{d_i \cup t_j}$
Result: $\mathcal{T}_t$
while $\max_{ij} \mathbf{M}_{ij} > \tau$ **do**
  if not $\mathbf{A}_{ij} \wedge \mathbf{M}_{ij} > \tau$ **then**
    $\langle \hat{i}, \hat{j} \rangle \leftarrow \arg\max_{ij} \mathbf{M}_{ij}$;
    $t_{\hat{j}} \leftarrow d_{\hat{j}}$ $\mathbf{A}_{\hat{i}} \leftarrow$ TRUE;
    $\mathbf{A}_{\hat{i}\hat{j}} \leftarrow$ TRUE;
/* Unassigned detections initialize new tracks. */
$\mathcal{T}_t \leftarrow \mathcal{T}_{t-1} \cup \{d | \mathbf{A}_{ij} = \text{TRUE}\}$;
/* Remove tracks not assigned for ω frames. */
$\mathcal{T}_t \leftarrow \mathcal{T}_{t-1} \setminus \{t_i | \mathbf{l}_i > \omega\}$

Figure 2:source: Deep Learning Based Surveillance System for Open Critical Areas

Finally, counting is performed by testing the segment intersection between the gate and the sequence of points on each track, making this a segment intersection problem. To detect people under a whole variety of camera viewpoints, human poses, lighting conditions, and occlusions, authors Tuan-Hung Vu, Anton Osokin, and Ivan Laptev in (12) put forth a way to leverage person-scene relations and proposed a Global CNN model trained to predict positions and scales of heads directly from the full image. They explicitly model pairwise relations between objects and train a Pairwise CNN model using a structured-output surrogate loss. The Local, Global and Pairwise models are then combined into a joint CNN framework.

IV. EVENT DETECTION IN VIDEOS

Tracking the actions of people in videos is another challenge. The authors of (13) proposed a hybrid feature extraction technique to recognize human activity. The extracted features are fused based on serial-based fusion, and later on, fused features are utilized for classification. Support vector machines (SVMs) were used for classification with a 90% accuracy. A weight-based segmentation approach is implemented to detect frame differences using cumulative mean and update the background by using weights and identifying the foreground regions. A hybrid feature extraction technique is then used to recognize human action. The extracted features are fused based on serial-based fusion, and later on, the fused feature is utilized for classification.
Authors Piotr Dollar, Vincent Rabaud, Garrison Cottrell, and Serge Belongieshow that the direct 3D counterparts to commonly used 2D interest point detectors are inadequate and propose an alternative. They devised a recognition algorithm based on spatio-temporally windowed data. Cuboids extracted from many sample behaviours from a given domain are clustered to form a dictionary of cuboid prototypes. The only information kept from all subsequent video data is the location and type of the cuboid prototypes present. They argue that such a representation is sufficient for recognition and robust with respect to the variations in the data (14).

In (15), the authors presented an efficient action recognition system that combines three state-of-the-art low-level descriptors like MBH(motion boundary histogram), SIFT(scale-invariant feature transform), and MFCC (mel-frequency cepstral coefficients) along with the recent Fisher vector representation. Their experimental evaluation is among the most extensive and diverse ones to date, including five of the most challenging action recognition benchmarks, action localization in feature-length movies, and large-scale event recognition with a test set of more than 1,000 hours of video.

V. ANOMALY DETECTION

Event recognition techniques have their limitations. Most of the videos recorded on surveillance are of normalcy. Beyond that, it is unrealistic to predict every single anomaly in the first place (16). The examples of anomalous events are relatively low. Due to the sheer number of variations and types of "anomalous" events, it becomes practically impossible to train a model to recognize each of them (17). This problem can be addressed by modelling the temporal regularity of videos with limited supervision rather than modelling the sparse irregular or meaningful moments in a supervised manner (6).

In (8), Anomaly detection is performed on the trajectories detected using a Polytope Ensemble technique. The method proposed involves only training on videos depicting "regular circumstances" as it is easy to acquire this data. A sample is defined as abnormal if it doesn't belong to the convex hull of X (It is not "regular"). In fact, the convex hull used is a Robust Convex Hull to avoid overfitting. For the frame-level evaluations, a frame was considered anomalous if at least an abnormal detection occurs for that frame, regardless of its location. However, the accuracy of trajectory analysis relies heavily on tracking, and precise tracking remains a significant challenge in computer vision. It becomes even more challenging in crowded situations (1). Hence a deep-learning based approach would be more suitable.

One such model was proposed by authors Viorica Patraucean, Ankur Handa, and Roberto Cipolla in (18). They used a classic convolutional image encoder-decoder with a nested memory module composed of convolutional LSTM cells, acting as a temporal encoder. Since they focused on learning features for motion prediction, they used a robust optical flow prediction module and an image sampler as a temporal decoder, which provided immediate feedback on the predicted flow map. At each time step, the system receives as input a video frame, predicts the optical flow based on the current frame and the LSTM content as a dense transformation map, and applies it to the current frame to predict the next frame. By minimizing the reconstruction error between the predicted next frame and the ground truth next frame, the authors could train the whole system for motion prediction without any supervision effort.

M. Dahmane and J. Meunier in (19) develop an innovative way of using self-organizing maps for anomaly detection. They consider local motion properties (flow vector) and more global ones expressed by elliptic Fourier descriptors. From these temporal trajectory characterizations, two Kohonen maps are used to distinguish normal behaviour from abnormal or suspicious ones.

The authors (20) of this paper propose a spatiotemporal architecture for anomaly detection in videos which even works with crowded scenes. The model uses a spatial feature extractor along with an encoder-decoder which together learn the temporal patterns of the input volume frames. This model is trained on videos containing only "normal" scenes with the objective of minimizing reconstruction loss. The idea is that when the model encounters "normal" scenarios, the reconstruction loss would be significantly lesser than when the model encounters "abnormal" scenarios.

The spatial encoder and decoder parts have two convolution and deconvolution layers, respectively, while the temporal encoders a three-layered convolutional LSTM model (Figure 3).

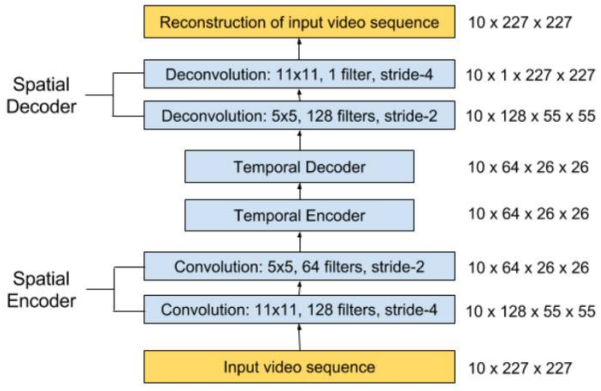


Figure 3:source: Abnormal Event Detection in Videos using Spatiotemporal Autoencoder

The reconstruction error between the input and output video sequence is taken as the Euclidean distance between them $e(t)$ using the learned weights f_W.

$$e(t) = \|x(t) - f_W(x(t))\|_2$$

This is then used to compute the abnormality score Sa(t) by scaling it between 0 and using the following formula.

$$s_a(t) = \frac{e(t) - e(t)_{\min}}{e(t)_{\max}}$$

And the regularity score can then be calculated by subtracting the abnormality score from 1.

$$s_r(t) = 1 - s_a(t)$$

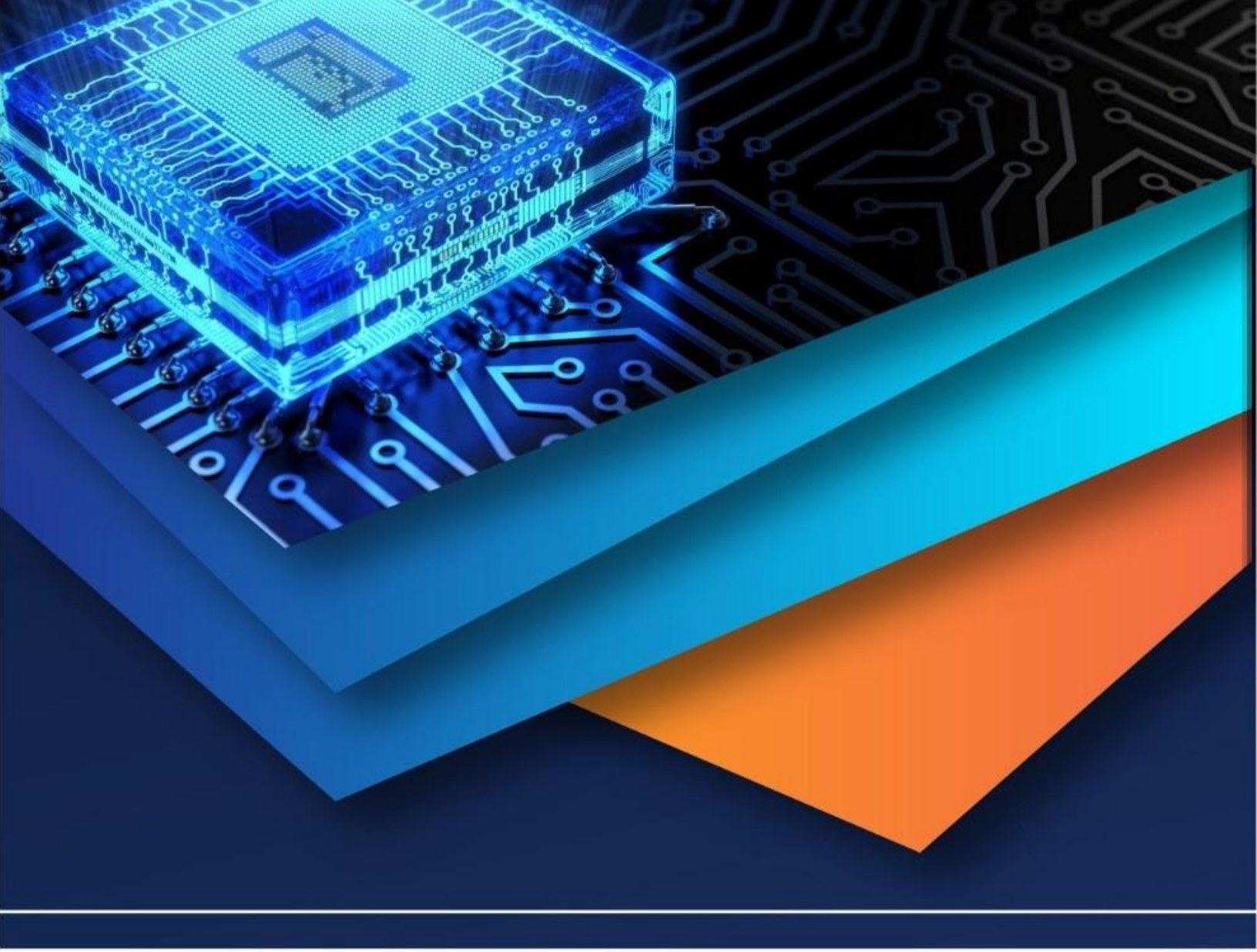
During inference, every time the reconstruction loss is above a certain threshold, the event is classified as an anomaly.

VI. CONCLUSION

This paper reviewed the development of surveillance systems from low-level background subtraction techniques to Spatio-temporal analysis. This paper can be the starting point to develop a new surveillance system using the strong points of each paper reviewed like the hardware system or the underlying algorithm. Developments relating to the ease of deployment and minimal hardware use would significantly reduce the cost of implementing these systems leading to wide spread adoption of these systems. Ultimately making the world a safer place.

REFERENCES

[1] Nishino. LKaK. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009.
[2] Lu C, Shi J, Jia J. Abnormal event detection at 150 fps in matlab. Proceedings of the International Conference on Computer Vision (ICCV), Darling Harbour, Sydney, 1–8 December 2013.
[3] Virender Singh SSPG. Real-Time Anomaly Recognition Through CCTV Using Neural Networks. Procedia Computer Science. 2020; Volume 173.
[4] E. Alajrami HTYSaM.EA. On using AI-Based Human Identification in Improving Surveillance System Efficiency. 2019 International Conference on Promising Electronic Technologies (ICPET). 2019.
[5] A. Adam ERISaDR. Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2008; 30.
[6] Hasan M&CJ&NJ&RCA&DL. Learning Temporal Regularity in Video Sequences.
[7] Denman S, Kleinschmidt T, Ryan D, Barnes P, Sridharan S, Fookes C. Automatic surveillance in transportation hubs: No longer just about catching the bad guy. 2015;: p. 42.
[8] Francesco Turchini LSTUaADB. Deep Learning Based Surveillance System for Open Critical Areas. 2018;: p. 13.
[9] [Online]. Available from: https://www.accenture.com/_acnmedia/pdf-94/accenture-value-data-seeing-what-matters.pdf.
[10] Hou J&WC&YZ&TJ&WQ&ZY. Research of Intelligent Home Security Surveillance System Based on ZigBee. 2009 Third International Symposium on Intelligent Information Technology Application Workshops. 2008.
[11] Shah OJaM. Tracking and Object Classification for Automated Surveillance.
[12] T. Vu AOaIL. Context-Aware CNNs for Person Head Detection. 2015 IEEE International Conference on Computer Vision (ICCV). 2015.
[13] Ishtiaq M&AS&AR&A,M&AH. Deep Learning based Intelligent Surveillance System. International Journal of Advanced Computer Science and Applications. 2020.
[14] Dollar P&RV&CG&BS. Behavior recognition via sparse spatio-temporal features. International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. ;: p. 65-72.
[15] D. Oneata JVaCS. Action and Event Recognition with Fisher Vectors on a Compact Feature Set. 2013 IEEE International Conference on Computer Vision. 2013.
[16] [Online]. Available from: https://www.ricoh.com/technology/institute/research/tech_anomaly_detection_in_videos#:~:text=With%20anomaly%20detection%2C%20you%20need,anomaly%20in%20the%20first%20place.
[17] Zhu Y&NN&RCA. Context-Aware Activity Recognition and Anomaly Detection in Video. IEEE Journal of Selected Topics in Signal Processing. .
[18] Patraucean V&HA&CR. Spatio-temporal video autoencoder with differentiable memory. 2016.
[19] Meunier MDaJ. Real-time video surveillance with self-organizing maps. The 2nd Canadian Conference on Computer and Robot Vision (CRV'05). .
[20] Yong Shean Chong YHT. Abnormal Event Detection in Videos using Spatiotemporal Autoencoder. 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY