IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155234

# Violence Detection in Video Surveillance Using Frame Extraction

**[1]Priyanka Sunil Kognole, [2]Shruti Mahadev Pisal, [3]Suparshwa Sudhir Kognole**

[1,2,3]Department of computer Science DBATU University Lonere, India

**Abstract:** This study presents a novel approach for enhancing the automation and effectiveness of real-time threat detection in video surveillance systems. Traditional surveillance methods require continuous human monitoring, are resource-intensive, and of ten fail to consistently identify suspicious activities with precision. Addressing these challenges, we propose the Mono-Scale CNN-LSTM Fusion Network, an advanced deep learning model designed for automated, sustainable, and high-accuracy CCTV systems. The model utilizes Convolutional Neural Networks(CNN) in combination with Long Short Term Memory (LSTM) networks to improve recognition capabilities by capturing temporal and spatial features. For feature extraction, the Oriented FAST and Rotated BRIEF (ORB) techniques are employed to enhance detection efficiency. The model was tested using the UCF crime image dataset and achieved an accuracy rate of approximately 99%, surpassing traditional models like CNN, VGG-16, VGG-19, ResNet-50, and Dense Net. This study highlights the contributions of our approach, which offers a significant reduction in the need for human oversight and sets new standards in the field of automatic threat detection. Furthermore, it emphasizes the model's capability to support contemporary security systems with high precision, reliability, and scalability, making it a valuable tool for the next generation of intelligent surveillance systems.

**Keywords:** Real-time threat detection, video surveillance, deep learning, CNN-LSTM, ORB feature extraction, UCF crime dataset, automated security, intelligent surveillance, high-accuracy CCTV.

## I. INTRODUCTION

Violence detection techniques using computer vision, analyze the surveillance camera videos. Over the last few years, these cameras and other surveillance equipment are installed on different places for the public safety e.g. Educational institutions, hospitals, banks, markets, streets etc. Surveillance cameras are widely used for monitoring public and private spaces, but most systems depend on human supervision. Manual monitoring is time-consuming, inefficient, and prone to human error.

With the rapid growth of surveillance cameras to monitor the human activity demands such system which recognize the violence and suspicious events automatically. Abnormal and violence action detection has become an active research area of computer vision and image processing to attract new researchers. The relevant literature presents different techniques for detection of such activities from the video proposed in the recent years. This research study reviews various state-of-the-art techniques of violence detection.

Detecting violent or suspicious activities in real time can prevent crimes, ensure public safety, and enhance response times. This project aims to develop an AI-based system capable of detecting suspicious or violent activities from surveillance videos using video frame extraction and deep learning techniques. By analyzing both spatial (image-level) and temporal (motion-level) patterns, the system can automatically classify actions as normal or violent.

Surveillance cameras are ubiquitous and appear in environments such as hospitals, schools or banks. These cameras record an entire day's length of activities resulting in very long video footages that makes the process of browsing video content a laborious and time consuming task for a human observer. Fully automated video analysis methods eliminate the need for a human observer and rely on computer vision and machine learning to detect events of interest. However, these methods are not fully reliable particularly when the search criteria are subjective or vaguely defined.

These advancements are critical for improving real-time monitoring and response sy

'stems, which are fundamental for public safety. However, despite the progress made, there are still several challenges in current crime detection technologies. One main issue is the need for real-time processing, which is crucial for timely identification of suspicious activities. Many models still struggle to achieve high accuracy in real-time applications, where latency and quick responses are critical. Furthermore, the scalability of these models remains a concern, as they often perform well on small datasets but face difficulties when deployed in large-scale systems or with diverse video sources. Additionally, the accuracy of existing models is still a work in progress, as they may struggle with complex or overlapping activities, requiring continuous refinement.

Surveillance systems can be categorized into two types: traditional and autonomous. Traditional systems rely heavily on human monitoring, whereas autonomous systems use DL, AI, and ML algorithms to independently detect human

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

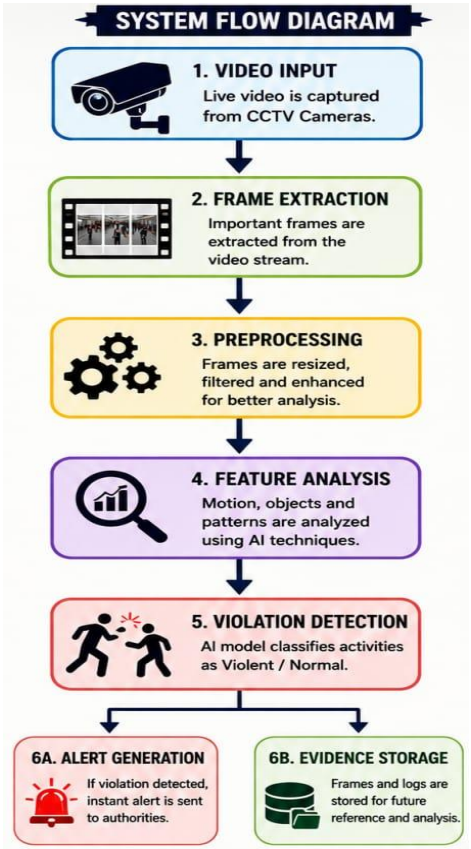
DOI: 10.17148/IJARCCE.2026.155234

activities. These advanced algorithms excel in feature extraction and pattern recognition, making them particularly adept at identifying complex activities. By training on large datasets, deep learning models can learn to recognize intricate human actions and adapt to unseen data, improving their ability to detect suspicious behavior across various environmental conditions.

## II. LITERATURE REVIEW

Video surveillance systems are widely used in public and private areas to improve security and monitor activities. Traditional surveillance systems mainly depend on human observation, which can be inefficient because continuous monitoring of multiple camera feeds is difficult. Human operators may miss important events due to fatigue, lack of attention, or heavy workload. To overcome these limitations, researchers have introduced Artificial Intelligence, Machine Learning, and Computer Vision techniques into surveillance systems. Many research studies focus on automated violence and abnormal activity detection using deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). These models help in identifying suspicious activities by analyzing human movement, object behavior, and motion patterns from surveillance videos. Some systems also use Optical Flow methods to detect sudden movements and aggressive actions in crowded environments. Frame extraction plays an important role in modern surveillance systems. Instead of analyzing the complete video continuously, only important frames are selected and processed. This method reduces computational complexity, decreases storage requirements, and improves processing speed while maintaining detection accuracy. Researchers have found that key-frame extraction helps improve overall system performance and makes real-time surveillance more efficient .Several existing systems are capable of generating real-time alerts when suspicious activities are detected. These systems send notifications, alarms, or warning messages to security personnel for immediate action. However, many existing approaches still face challenges such as false detections, poor performance in low-light conditions, high hardware requirements, and difficulties in recognizing complex human activities.The proposed system aims to overcome these challenges by combining frame extraction with AI-based violence detection techniques. The system focuses on improving accuracy, reducing unnecessary processing, and providing faster response time. Overall, the literature shows that intelligent surveillance systems can significantly enhance public safety and reduce the dependency on manual monitoring.

## lll. SYSTEM FLOW



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155234

**The System Architecture & Workflow**

The pipeline begins with the **Video Input** phase, where the system continuously captures live footage from strategically placed CCTV cameras or surveillance devices to provide the main input feed required for real-time observation. To optimize resource consumption, this dense stream is immediately passed to the **Frame Extraction** module, which divides the continuous video into multiple discrete image frames. Rather than processing the entire video file, the system intelligently selects key frames for analysis, a technique that drastically reduces storage overhead and accelerates overall processing speeds.

Once isolated, these key frames enter the **Preprocessing** stage, where they are cleaned and prepared for algorithmic evaluation through critical operations like resizing, filtering, noise removal, and image enhancement. This phase is vital for maximizing image quality, ensuring that subsequent computer vision models receive clean data to optimize detection accuracy.

**Analysis, Decision Making, & System Outputs**

With the data fully optimized, the system advances to **Feature Analysis**, leveraging advanced AI and computer vision techniques to analyze critical elements within the frames. By identifying and tracking motion dynamics, object shapes, human movement patterns, and sudden kinetic anomalies, the machine learning model can accurately comprehend the context of the behavior being observed.

This leads directly into **Violation Detection**, the intelligent decision-making core of the architecture. Here, the deep learning model classifies the analyzed actions as either normal or abnormal, automatically isolating high-risk behaviors such as physical fighting, aggressive posturing, trespassing, or other prohibited activities.

**Incident Response & Forensic Archiving**

The final phase of the workflow executes a dual-pronged response strategy the moment a violation is confirmed, splitting into immediate threat mitigation and long-term logging.

- **6A. Alert Generation:** The system instantly triggers its notification engine to send immediate alerts—ranging from localized alarms and SMS text messages to automated email warnings—directly to security personnel to facilitate rapid intervention.
- **6B. Evidence Storage:** Simultaneously, the system executes an archiving protocol that securely commits the corresponding evidence frames, detailed activity logs, and incident metadata to a centralized database, establishing a tamper-evident repository for future forensic investigations, legal reporting, and iterative system optimization.

## lV. REQUIREMENT ANALYSIS

The system specifications for the automated surveillance project encompass hardware, software, functional, and non-functional requirements to ensure successful deployment and operation.

**1. Hardware & Software Infrastructure**

The hardware architecture requires a baseline processing unit, such as an Intel Core i5 or AMD Ryzen 5 processor (or higher), paired with a minimum of 8 GB RAM (though 16 GB is strongly recommended to handle deep learning model training efficiently). Storage demands dictate at least a 500 GB HDD or a faster 256 GB SSD, while real-time testing necessitates an integrated webcam or a connected CCTV camera feed, displayed on a monitor supporting a minimum resolution of $1366 \times 768$ pixels.

On the software front, the environment is built upon a Windows 10/11 or Ubuntu 20.04+ operating system running Python 3.8+ within an IDE like Jupyter Notebook, PyCharm, or VS Code. The core library ecosystem leverages OpenCV for video processing and automated frame extraction, alongside TensorFlow or Keras for constructing and executing the deep learning models. Data handling and mathematical operations are managed via NumPy and Pandas, while Matplotlib handles data visualization, and Scikit-learn provides the necessary evaluation metrics. For extended features, an optional backend database like MySQL or SQLite can be integrated for system logging, and a web framework such as Flask or Django can be utilized to deploy a user-facing dashboard interface.

**2. Functional & Non-Functional Deliverables**

The system's functional requirements define its operational capabilities, dictating that the pipeline must seamlessly capture either live or pre-recorded video feeds and automatically execute frame extraction. The core AI model must then analyze these extracted frames to classify human behavior into normal or violent categories, instantly triggering automated alert notifications the moment suspicious activity is detected. All resulting classifications and system outputs must either be displayed dynamically on-screen or securely committed to the database.

To ensure this functionality translates effectively into real-world use, the system must adhere to strict non-functional constraints. It must deliver high-speed, low-latency processing to guarantee real-time performance and maintain a strict detection accuracy benchmark of at least 85%. Furthermore, the final system must feature a user-friendly, highly intuitive

DOI: 10.17148/IJARCCE.2026.155234

interface, possess the architectural scalability to manage multiple simultaneous camera feeds, and remain consistently reliable across varying environmental conditions, such as poor lighting or shifting camera angles.

## V. METHODOLOGY

**Dataset and Pre-Processing :**

For our study, a widely known UCF Crime Dataset is used on the 3 most common crime categories: Robbery, Shoplifting, and Fighting. These images are divided into Train and Test subsets. The images within the dataset possess dimensions of 64x64 pixels and are stored in .png format. These images were extracted from video footage, with a systematic sampling approach where every 10th frame was chosen from each video. As part of the data preprocessing pipeline, first, convert the images to grayscale. By conversion, complexity is reduced while preserving intensity information. In our study, we suggest using a tool called Oriented FAST and Rotated BRIEF (ORB) to work with images. ORB is like a detective for pictures. It spots special points in an image and figures out what's unique about them. This helps us describe the important features of an image simply.

**Long Short-Term Memory :**

In the proposed architecture the LSTM model is created Keras-based LSTM layers in a linear stack to learn temporal dependencies from input data. It starts with two LSTM layers (8 units each), where the f irst retains sequences to capture temporal info from 2500 time steps with a single feature per step, using tanh activation. After LSTM, a dense layer (4 neurons) enhances complex relationship learning, followed by a 20% dropout layer for generalization. Finally, a flattened layer reshapes the 3D output from the preceding LSTM layers into a 1D vector, preparing the data for the classification task. The LSTM model, with its sequential data processing capabilities and the ability to capture temporal patterns, complements the CNN model effectively when combined with a given overall approach for detecting and classifying criminal activities in surveillance footage. The integration of both CNN and LSTM components contributes to the models' exceptional accuracy in identifying criminal activities, enhancing security measures, and ensuring public safety.

**Convolution Neural Network :**

The CNN model for activity detection processes 50x50 grayscale images through 3 successive convolutional layers (64, 128, and 256 filters, 3x3 filter size), each using Leaky Re LU. Activation to capture essential features. Max pooling layers (2x2) down sample feature maps, while dropout layers (25% and 40%) prevent overfitting. After convolution and pooling, the output is flattened into a 1D vector. A fully connected layer with 256 neurons, Leaky Re LU activation, and a 50% dropout finetunes features and adds regularization. The output layer uses softmax to classify criminal activities based on the highest probability, ensuring effective eature extraction and classification.

The technical implementation of the proposed automated surveillance system is structured into seven sequential phases:

1. **Data Collection:** Comprehensive video datasets depicting normal and violent activities are gathered from open-source repositories like the **Hockey Fight**, **Violent-Flows**, and **UCF Crime** datasets. These videos are pre-processed to ensure uniform framerates, resolutions, and durations, eliminating data inconsistencies before training.
2. **Video Frame Extraction:** The system samples individual frames from the input videos at fixed intervals (e.g., every **10–15 frames** per second). These isolated sequences serve as the primary data for training and testing, allowing the model to bridge the gap between static spatial elements and temporal motion patterns.
3. **Preprocessing:** Extracted frames are uniformly resized to a standard dimension (such as **$128 \times 128$ pixels**) to reduce computational load. Pixel values are normalized, noise-reduction filters are applied to improve image quality, and data augmentation (rotation, flipping, brightness adjustments) is performed to increase dataset diversity and prevent overfitting.
4. **Feature Extraction:** A hybrid deep learning approach maps the data. A **Convolutional Neural Network (CNN)** extracts spatial features—such as objects, environmental details, and body movements—from individual frames. These features are then passed to a **Long Short-Term Memory (LSTM)** network, which captures temporal features by tracking motion sequences and behavioral changes over time.
5. **Model Training:** The unified CNN-LSTM architecture is trained on the labeled dataset. The CNN processes each frame while the LSTM analyzes the sequential frame progression to detect violence patterns. The dataset is split into training, validation, and testing sets to rigorously evaluate model performance.
6. **Real-Time Detection:** The trained model is integrated directly into live CCTV or video streams. The system continuously extracts and processes frames in real time, classifying behavior as normal or violent. The moment a violation is identified, an automated alert notification is instantly generated for immediate security response.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155234

7. **System Testing and Validation:** The integrated system undergoes rigorous validation under diverse real-world conditions, including poor lighting, varying camera angles, and high crowd densities. Testing the model on entirely unseen footage ensures its generalization, reliability, and low false-alarm rates before field deployment.

## Vl. OBJECTIVES

The core operational methodology of the proposed system centers on a sophisticated, multi-stage pipeline designed to ingest raw surveillance video and transform it into actionable security intelligence. The process begins with an automated frame-extraction module that captures discrete video frames at strictly defined time intervals, ensuring continuous observation while eliminating redundant data points. These extracted frames are immediately routed to a preprocessing engine that normalizes the visual data, optimizing it for feature extraction by adjusting contrast, resolution, and noise levels. At the heart of the analytical framework sits a hybrid deep learning architecture that combines Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks; the CNN layers excel at mapping complex spatial features within individual frames, while the sequential LSTM layers simultaneously track temporal motion patterns over time. This dual-layered analysis enables the system to accurately classify video segments into distinct violent or non-violent categories based on subtle behavioral kinetics. Finally, upon the classification of a high-risk event, the pipeline triggers an integrated alert system that flags the suspicious activity and instantly routes the data to security personnel, closing the loop from raw visual input to real-time threat mitigation.

## Vll. LIMITATION AND RESULT

### a. Limitation

While the proposed automated surveillance system offers advanced detection capabilities, it faces several technical and operational limitations that must be addressed for real-world deployment. Environmental factors significantly impact performance, as system accuracy inherently decreases under poor lighting, low-quality video feeds, or sudden camera instability that disrupts precise frame extraction. Furthermore, complex, highly dense crowd environments present a substantial challenge, often triggering false alerts due to occlusion or normal, non-violent activities structurally resembling suspicious behavior. To sustain this deep learning-driven, real-time analysis, the infrastructure demands robust, high-performance hardware and stable, high-bandwidth network connectivity to prevent latency or total system failure in cloud-based deployment models. The core intelligence of the system is also entirely contingent upon the breadth and quality of its initial training dataset, necessitating continuous retraining cycles to effectively recognize evolving threat patterns. Finally, the system introduces long-term logistical hurdles, including rapidly expanding data storage requirements for high-resolution evidence logs, alongside critical ethical and legal challenges regarding data security and public privacy rights during continuous mass monitoring.

### b. Result

The proposed automated surveillance model delivers a highly scalable and intelligent solution for modern security by successfully identifying violent or suspicious behavior directly from raw video feeds with high precision. By integrating an innovative frame-extraction mechanism, the system effectively strips away redundant visual data, drastically reducing unnecessary video processing overhead and maximizing overall computational efficiency. This streamlined, AI-based analysis allows the platform to accurately isolate abnormal activities across diverse and demanding environments—such as crowded public areas, schools, corporate offices, and traffic monitoring zones—while concurrently alleviating the burden of manual, exhausting screen-monitoring on human personnel. The operational impact is immediate: the moment a violation is detected, the system triggers real-time alerts that dramatically slash incident response times, allowing security teams to intervene rapidly and prevent escalation, thereby elevating the benchmark for public safety. Additionally, the system doubles as a comprehensive administrative and legal asset by automatically cataloging and storing crucial evidence frames and detailed event logs, ensuring that law enforcement and security management have access to a reliable, organized repository for future forensic investigations and reporting.

## Vlll. CONCLUSION

The proposed system introduces an intelligent, automated framework that revolutionizes video surveillance by deploying advanced computer vision and deep learning algorithms to detect violent or suspicious activities in real time. By utilizing an optimized frame-extraction mechanism, the system isolates high-risk visual sequences for analysis rather than processing continuous, resource-heavy video streams, drastically reducing computational overhead and bandwidth consumption. This proactive approach minimizes human fatigue and monitoring oversight, enabling the system to instantly generate automated alerts for security personnel to facilitate rapid, targeted interventions. Furthermore, the architecture functions as a robust

DOI: 10.17148/IJARCCE.2026.155234

digital forensics tool by securely archiving evidence frames, timestamps, and metadata logs, providing law enforcement with tamper-evident, searchable data for future investigations. Ultimately, this project bridges the gap between raw hardware and actionable intelligence, establishing a highly scalable, efficient asset for modern public safety and security infrastructure.

## REFERENCES

1] **Soheil Vosta and Kin-Choong Yow** ,"ACNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras", 2022

2] **Souvik Kumar Parui** ,"An Efficient Violence Detection System from Video Clips using Conv LSTM and Keyframe Extraction", 28 June ,2023

3] **Hamza Naveed , Junaid Asghar,** "Efficient and Sustainable Video Surveillance Using CNN-LSTM Model for Suspicious Activity Detection", 2 March , 2025

4] **Hikmat Ullah Khan** ,"A Review on state-of-the-art Violence Detection Techniques", 2019

5] **Frédéric Dufaux** ,"Scrambling for Privacy Protection in Video Surveillance Systems"

6] Varadarajan, J.; Odobez, J.M. Topic models for scene analysis and abnormality detection. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, Kyoto, Japan, 27 September–4 October 2009; pp. 1338–1345

7[ Sodemann, A.A.; Ross, M.P.; Borghetti, B.J. A review of anomaly detection in automated surveillance. IEEE Trans. Syst. Man, Cybern. Part C (Appl. Rev.) 2012, 42, 1257–1272. [CrossRef]

8] Zweng, A.; Kampel, M. Unexpected human behavior recognition in image sequences using multiple features. In Proceedings of the 2010 20th International Conference on Pattern Recognition, ICPR, Istanbul, Turkey, 23–26 August 2010; pp. 368–371.

9] Jodoin, P.M.; Konrad, J.; Saligrama, V. Modeling background activity for behavior subtraction. In Proceedings of the 2008 Second ACM/IEEEInternational Conference on Distributed Smart Cameras, Trento, Italy, 9–11 September 2008; pp. 1–10.

10] Dong, Q.; Wu, Y.; Hu, Z. Pointwise motion image (PMI): A novel motion representation and its applications to abnormality detection and behavior recognition. IEEE Trans. Circuits Syst. Video Technol. 2009, 19, 407–416. [CrossRef]

11] Mecocci, A.; Pannozzo, M.; Fumarola, A. Automatic detection of anomalous behavioural events for advanced real-time video surveillance. In Proceedings of the 3rd International Workshop on Scientific Use of Submarine Cables and Related Technologies, Lugano, Switzerland, 31 July 2003; pp. 187–192.

12] Li, H.P.; Hu, Z.Y.; Wu, Y.H.; Wu, F.C. Behavior modeling and abnormality detection based on semi-supervised learning method. Ruan Jian Xue Bao (J. Softw.) 2007, 18, 527–537. [CrossRef]

13] Yao, B.; Wang, L.; Zhu, S.C. Learning a scene contextual model for tracking and abnormality detection. In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.

14] Yin, J.; Yang, Q.; Pan, J.J. Sensor-based abnormal human-activity detection. IEEE Trans. Knowl. Data Eng. 2008, 20, 1082–1090. [CrossRef]

15] Benezeth, Y.; Jodoin, P.M.; Saligrama, V.; Rosenberger, C. Abnormal events detection based on spatio-temporal co-occurences. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition CVPR, Miami, FL, USA, 20–25 June 2009; pp. 2458–2465.

16] Dong, N.; Jia, Z.; Shao, J.; Xiong, Z.; Li, Z.; Liu, F.; Zhao, J.; Peng, P. Traffic abnormality detection through directional motion behavior map. In Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, Boston, MA, USA, 29 August–1 September 2010; pp. 80–84

17] Loy, C.C.; Xiang, T.; Gong, S. Detecting and discriminating behavioural anomalies. Pattern Recognit. 2011, 44, 117–132. [CrossRef] 18] Zhang, J.; Liu, Z. Detecting abnormal motion of pedestrian in video. In Proceedings of the 2008 International Conference on Information and Automation, Changsha, China, 20–23 June 2008; pp. 81–85.

19] Ruff, L.; Vandermeulen, R.A.; Görnitz, N.; Binder, A.; Müller, E.; Müller, K.R.; Kloft, M. Deep semi-supervised anomaly detection. arXiv 2019, arXiv:1906.02694.

20] Tang, Y.P.; Wang, X.J.; Lu, H.F. Intelligent video analysis technology for elevator cage abnormality detection in computer vision. In Proceedings of the 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology, Seoul, Korea, 24–26 November 2009.