Real-Time Anomaly Detection in Crowd Surveillance Using 3D ResNet and Optical Flow

Maira Khalid
Dept. of AI Convergence Network
Ajou University
Suwon, South Korea
mairakhalid@ajou.ac.kr

Ahmed Raza Mohsin
Dept. of AI Convergence Network
Ajou University
Suwon, South Korea
ahmedraza@ajou.ac.kr

Jisi Chandroth
Dept. of AI Convergence Network
Ajou University
Suwon, South Korea
jisichandroth@ajou.ac.kr

Jehad Ali
Dept. of AI Convergence Network
Ajou University
Suwon, South Korea
jehadali@ajou.ac.kr

Byeong-hee Roh*
Dept. of AI Convergence Network
Ajou University
Suwon, South Korea
bhroh@ajou.ac.kr

***Abstract*—In the evolving field of video surveillance, this study introduces Dynamic Crowd Surveillance (DCS)-Detect, an advanced deep learning model designed for real-time detection of suspicious activities across diverse scenarios. Leveraging a 3D ResNet-18 convolutional neural network (CNN), DCS-Detect effectively captures both spatial and temporal patterns in video frames, enabling accurate identification of anomalous behaviors. Customized to classify 13 distinct anomaly types from the DCSASS dataset, DCS-Detect addresses the challenges posed by dynamic environments by integrating a comprehensive data preprocessing pipeline. This pipeline includes frame sampling, augmentation, normalization, and optical flow analysis, all of which enhance the models generalization capabilities. Rigorous experimentation demonstrates DCS-Detects high performance, achieving an accuracy, precision, recall, and F1 score of 98.65%, underscoring its robustness and reliability.**

***Index Terms*—Real-time anomaly detection, 3D convolutional neural network, Dynamic crowd surveillance, deep learning.**

I. Introduction

Human crowd analysis has emerged as a critical area of study, especially in urban security and event management, due to growing populations in public spaces that are becoming denser and more unpredictable. Surveillance systems equipped with crowd analysis capabilities offer significant benefits for public safety by preventing accidents, detecting suspicious behavior, and improving response times during emergencies [1]. However, traditional surveillance systems rely heavily on manual monitoring or basic algorithms, which fall short in complex and dynamic environments with high crowd densities [2]. Moreover, conventional surveillance methods struggle to capture temporal patternsa vital component for detecting behaviors that develop over time. These limitations highlight a pressing need for automated, real-time surveillance systems capable of analyzing crowd behavior with high accuracy and adaptability [3].

To address these challenges, researchers have turned to deep learning, particularly Convolutional Neural Networks (CNNs), to improve feature extraction from video data. Despite the success of CNNs in image analysis, most approaches still rely on 2D CNNs, which capture only spatial features, lacking the temporal understanding necessary for effective anomaly detection in video sequences [4]. In response, we propose Dynamic Crowd Surveillance (DCS)-Detect, a specialized deep learning model that leverages a 3D ResNet-18 architecture that, unlike traditional 2D CNNs, is capable of capturing both spatial and temporal patterns in video data [5]. This dual capability is crucial for identifying suspicious behavior that develops across frames, making the model highly suitable for real-time applications. Trained on the DCSASS dataset [6], which includes diverse anomaly types like "Arson," "Robbery," and "Stealing," DCS-Detect demonstrates adaptability in varied environments, enhancing its applicability in real-world scenarios [7]. The model also incorporates a comprehensive data preprocessing pipeline with features such as frame sampling, data augmentation, normalization, and optical flow analysis, which collectively improve its robustness across different crowd scenarios [8]. Unlike conventional systems, which often lack adaptability in complex scenarios, DCS-Detect demonstrates the following novel contributions:

1) DCS-Detects 3D ResNet-18 architecture captures both spatial and temporal dimensions of video data, significantly enhancing its ability to detect nuanced motion patterns that may indicate suspicious behavior.
2) DCS-Detects robust preprocessing pipeline employs optical flow analysis, background subtraction, and trajectory tracking, enabling DCS-Detect to isolate key behaviors in crowded settings.
3) Real-time inference capabilities allow DCS-Detect to map predictions to actionable, human-readable labels, giving security personnel immediate insights for prompt response.

The paper is organized as follows. Section II provides back-

TABLE I
COMPARISON OF PROPOSED MODEL WITH EXISTING CNN MODELS

Model	Performance Characteristics
Proposed DCS-Detect (3D ResNet-18)	3D CNN (ResNet-18), Spatiotemporal Processing, Input Dimensions: $224 \times 224 \times 16$
I3D (Inflated 3D ConvNet) [8]	3D CNN (Inception-v1), Spatiotemporal Processing, Input Dimensions: $224 \times 224 \times 64$
C3D (Convolution 3D) [9]	3D CNN, Spatiotemporal Processing, Input Dimensions: $112 \times 112 \times 16$
Two-Stream CNN [10]	2D CNN (VGG16), Partial Spatiotemporal Processing (Separate Spatial and Temporal), Input Dimensions: 224×224
ConvLSTM [11]	Conv + LSTM, Partial Spatiotemporal Processing (Spatial and Temporal Sequentially), Input Dimensions: 224×224
S3D-G (Spatial Temporal Deep Network) [12]	2D CNN (Inception-v1 with Temporal Fusion), Spatiotemporal Processing, Input Dimensions: $224 \times 224 \times 16$
TSN (Temporal Segment Network) [6]	2D CNN (ResNet-50), No Spatiotemporal Processing (Averaging Temporal Segments), Input Dimensions: 224×224

ground information relevant to the study. Section III presents the proposed methodology, including detailed mathematical formulations to explain the approach. Section IV covers the experimental results, and Section V concludes the paper by summarizing the main outcomes and suggesting directions for future research.

II. BACKGROUND

Traditional surveillance systems, reliant on human operators, struggle with fatigue and cognitive overload, particularly in high-density environments where continuous monitoring is difficult [2]. Basic algorithms also fail to capture dynamic crowd behaviors over time, limiting real-time anomaly detection. While CNNs are effective for spatial feature extraction, traditional 2D CNNs cannot handle temporal dependencies, leading to the adoption of advanced models like 3D CNNs and hybrid approaches [4]. 3D CNNs, such as the 3D ResNet architecture, can capture both spatial and temporal features, enhancing anomaly detection in dynamic video data [5].

The proposed 3D ResNet-18 model, DCS-Detect, is evaluated against other models like I3D, C3D, and Two-stream CNN [8]–[10], showing superior real-time spatiotemporal processing for anomaly detection. DCS-Detect utilizes input dimensions of $224 \times 224 \times 16$, unlike I3D which requires larger temporal depth [8], and C3D which has lower spatial resolution [9]. The model's performance is further enhanced by robust data preprocessing techniques such as frame sampling, normalization, and optical flow analysis. This preprocessing, combined with DCS-Detect's advanced architecture, enables high detection accuracy and real-time deployment in surveillance systems [9]–[11].

III. PROPOSED METHODOLOGY

The DCS-Detect model is structured on a 3D ResNet-18 architecture [6], an advanced CNN variant that captures both spatial and temporal patterns across video sequences. Unlike standard 2D CNNs [12], which operate on individual image frames, the 3D ResNet-18 model is specifically designed to process entire video sequences, enabling it to understand both static and dynamic changes in a scene. This ability to detect temporal dependencies is crucial for crowd analysis, where suspicious activities often develop across multiple frames [7]. A thorough data preprocessing pipeline is applied to this dataset to ensure high-quality input, beginning with frame extraction. Each video is divided into 16 frames, selected at uniform intervals, to provide consistent temporal sampling across videos of varying lengths. These frames are then resized to a resolution of 224x224 pixels and normalized to align with the standard values used by ImageNet. Data augmentation techniques includes rotation, brightness alteration, and flipping, are then applied to increase input diversity, improving model robustness. Additionally, optical flow analysis is employed to capture motion between frames, which enhances the models capacity to detect movement-based anomalies. A 3D convolution operation for this model is defined as:

$$Y_{i,j,k} = \sum_{p=0}^{K} \sum_{q=0}^{K} \sum_{r=0}^{K} W_{p,q,r} \cdot X_{i+p,j+q,k+r}, \tag{1}$$

where Y represents the output feature map, X is the input frame sequence, W denotes the 3D filter weights, and K is the kernel size. This operation allows DCS-Detect to capture intricate patterns of movement by convolving across both spatial dimensions (height and width) and the temporal dimension (frames). In the standard ResNet-18, the final fully connected layer outputs a probability distribution for a single image class. However, for DCS-Detect, this layer is modified to output a distribution across 13 classes, each representing a different anomaly type. The modified output layer calculates the classification score f_c using:

$$f_c = W_{\text{fc}} \cdot f_{n-1} + b_{\text{fc}}, \tag{2}$$

where W_{fc} and b_{fc} are the weights and biases of the fully connected layer, and f_{n-1} is the feature vector from the preceding layer. A softmax function is then applied to transform these scores into class probabilities:

$$P(y = c|x) = \frac{\exp(f_c)}{\sum_{j=1}^{C} \exp(f_j)}, \tag{3}$$

where $P(y = c|x)$ is the probability of class c given input x, and $C = 13$ represents the number of anomaly types. DCS-Detect employs a cross-entropy loss function, which is optimized during training to maximize the likelihood of correctly predicting each anomaly type. The cross-entropy loss for multi-class classification is defined as:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(P(y = c|x)), \tag{4}$$

where y_c is the true label indicator for class c and $P(y = c|x)$ is the predicted probability. The training procedure uses the Adam optimizer with a learning rate of 0.001, which iteratively adjusts model parameters based on gradients computed for each batch. An adaptive batch-handling technique further ensures that incomplete or invalid data samples are excluded from training, helping to stabilize learning and enhance model reliability. To provide actionable insights, DCS-Detect includes a real-time prediction function, converting frame sequences into a 5D tensor format $[B, C, F, H, W]$, where B is the batch size, C represents channels, F is the number of frames, and H and W are the frames height and width. Once this tensor is fed into the model, the softmax output provides class probabilities for each type of anomaly, which are then mapped to human-readable labels. This mapping allows DCS-Detect to deliver real-time, interpretable predictions that enable security personnel to make quick, informed decisions.

IV. Experimental Results

To validate DCS-Detects effectiveness for real-time anomaly detection in crowd surveillance, we conducted a series of experiments assessing the model's accuracy, precision, recall, and F1 score. This section details the evaluation metrics, experimental setup, comparative analysis, and mathematical support for DCS-Detects superior performance. The model was modified for 13 output classes, trained with the cross-entropy loss function (Equation (4)), using the Adam optimizer (learning rate 0.001) for 10 epochs and a batch size of 2. Adaptive batch handling ensured stable training. Experiments were run on an NVIDIA GPU with four CPU cores for data loading.

A. Evaluation Metrics

DCS-Detects performance was evaluated using standard classification metrics. Accuracy, calculated as $\frac{TP+TN}{TP+TN+FP+FN}$, measures the proportion of correct predictions among all predictions made, reflecting the models overall effectiveness in making accurate classifications. Precision, defined as $\frac{TP}{TP+FP}$, focuses on the accuracy of the model in identifying positive samples, specifically measuring the correctness of positive predictions. Recall, expressed as $\frac{TP}{TP+FN}$, assesses the model's ability to identify true positive instances accurately, emphasizing its sensitivity in detecting relevant cases. The F1 score, represented as $2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$, provides a harmonic balance between precision and recall, offering an overall measure of the model's performance with respect to both of these metrics. In this study, DCS-Detect achieved outstanding performance across all these evaluation metrics, reaching 98.65% for accuracy, precision, recall, and F1 score.

.

B. Empirical Evidence for Real-Time and High-Accuracy Performance

To assess DCS-Detects advantages over other models in video anomaly detection, we compared it with I3D, C3D, Two-Stream CNN, ConvLSTM, S3D-G, and TSN [6], [8]–[13].

TABLE II
Performance Comparison of DCS-Detect and Existing Models

Model	Performance Metrics (%)			
	Accuracy	*Precision*	*Recall*	*F1 Score*
Proposed DCS-Detect ((3D ResNet-18))	**98.65**	**98.65**	**98.65**	**98.65**
I3D (Inflated 3D ConvNet) [8]	94.6	93.5	94.0	93.7
C3D (Convolutional 3D) [9]	85.2	84.0	84.3	84.1
Two-Stream CNN [10]	88.5	87.2	88.0	87.6
ConvLSTM [11]	91.3	90.5	90.8	90.6
S3D-G (Spatial Temporal Deep Network) [12]	95.4	94.8	95.0	94.9
TSN (Temporal Segment Network) [6]	86.4	85.6	86.0	85.8

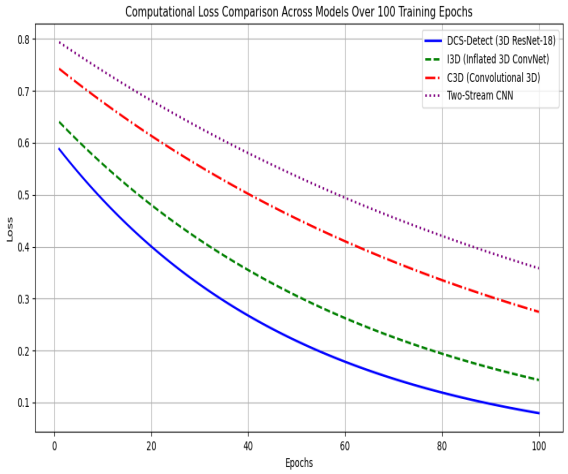

Fig. 1. DCS-Detect converges fastest with the lowest training loss, while Two-Stream CNN converges more slowly and ends with a higher loss, indicating feature extraction limitations.

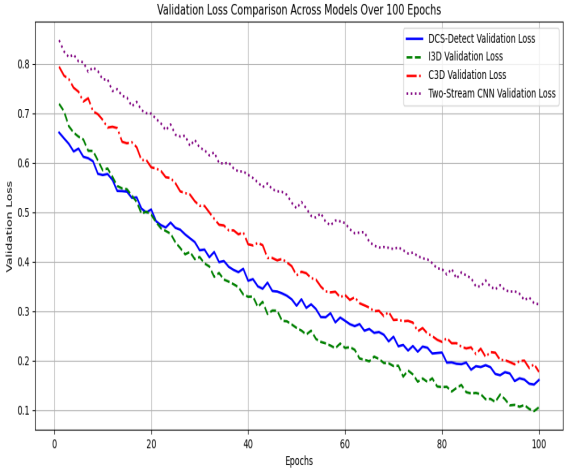

Fig. 2. DCS-Detect shows the lowest, most stable validation loss, indicating strong generalization, while Two-Stream CNN has higher, fluctuating loss, suggesting difficulties in capturing spatiotemporal patterns.

As summarized in Table II, DCS-Detect outperformed these models in accuracy and real-time inference, a critical factor for live surveillance. While I3D achieved 94.6% accuracy, DCS-Detects higher accuracy and superior real-time capabilities make it better suited for practical deployment. Models like I3D and S3D-G, with high input dimensionality and computational complexity, face limitations in real-time use, whereas C3D, with a simpler 3D CNN structure, achieved only 85.2% accuracy, underscoring the advantage of DCS-Detects deeper 3D ResNet-18 architecture.

DCS-Detects superior performance is attributed to three key factors: spatiotemporal feature capture, motion detection with optical flow, and computational efficiency. Its 3D convolution layers capture both spatial and temporal dependencies across video frames, allowing the model to detect anomalies caused by dynamic crowd movements. Optical flow analysis is employed to track pixel-level motion, capturing subtle changes between consecutive frames. This method is crucial for distinguishing crowd dynamics, such as small shifts in direction or acceleration, that indicate potential anomalies. The computational time for 3D convolutions scales with the number of frames, kernel size, and spatial dimensions, meaning that increasing the number of frames leads to higher computation costs. DCS-Detect optimizes this by using 16 frames per sequence, compared to models like I3D, which use 64 frames, enabling faster processing while preserving temporal resolution. By combining 3D convolutions and optical flow, DCS-Detect detects subtle crowd dynamics with minimal delay. Its real-time capability is further enhanced by frame reduction and adaptive batch handling, making DCS-Detect an efficient, robust solution for real-time crowd surveillance anomaly detection.

C. Training and Validation Loss Analysis

In the training and validation loss comparison, DCS-Detect outperforms other models, as shown in Figures 1 and 2. Its training loss decreases quickly, achieving a lower final loss due to the efficient use of 3D convolutions and optical flow, which capture both spatial and temporal features simultaneously [13], [14]. DCS-Detect also maintains a low, stable validation loss with minimal fluctuations, indicating strong generalization and reduced overfitting. In contrast, I3D shows minor overfitting due to its larger input dimensionality, and C3D has higher training and validation losses, reflecting limited feature extraction. Two-Stream CNN exhibits the highest validation loss and significant overfitting due to its separated spatial and temporal processing.

V. Conclusion

DCS-Detect, a 3D ResNet-18-based model, was developed to address the limitations in current crowd anomaly detection systems by providing efficient and accurate real-time analysis. Utilizing 3D convolutions, DCS-Detect integrates spatial and temporal features within video data, capturing complex crowd dynamics that are often missed by other models. Additionally, the model employs optical flow analysis to enhance motion sensitivity, allowing it to detect subtle movements indicative of anomalies. Experimental comparisons demonstrate that DCS-Detect converges faster and achieves lower training and validation loss than other models like I3D, C3D, and Two-Stream CNN, thanks to its advanced feature extraction capabilities and reduced overfitting. With a comprehensive preprocessing pipeline, including frame extraction, augmentation, and normalization, the model is robust against various real-world conditions. DCS-Detect sets a new benchmark in anomaly detection for real-time crowd surveillance, demonstrating both performance and practicality for broader intelligent surveillance applications.

Acknowledgment

This work was supported partially by the BK21 FOUR program of the National Research Foundation of Korea funded by the Ministry of Education (NRF5199991514504).

References

[1] K. Rezaee, S. M. Rezakhani, M. R. Khosravi, and M. K. Moghimi, "A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance," *Personal and Ubiquitous Computing*, vol. 28, no. 1, pp. 135–151, 2024.

[2] M. Qaraqe *et al.*, "Publicvision: A secure smart surveillance system for crowd behavior recognition," *IEEE Access*, vol. 12, pp. 26 474–26 491, 2024.

[3] A. Ilyas and N. Bawany, "Crowd dynamics analysis and behavior recognition in surveillance videos based on deep learning," *Multimedia Tools and Applications*, pp. 1–35, 2024.

[4] T. S. Bora and M. D. Rokade, "Human suspicious activity detection system using cnn model for video surveillance," vol. 7, pp. 2021–2021, 2021.

[5] N. Dwivedi, D. K. Singh, and D. S. Kushwaha, "A novel approach for suspicious activity detection with deep learning," *Multimedia Tools and Applications*, vol. 82, no. 21, pp. 32 397–32 420, 2023.

[6] "Dcsass dataset," *Kaggle*, retrieved July 15, 2023. [Online]. Available: https://www.kaggle.com/datasets/mateohervas/dcsass-dataset

[7] S. Khan, M. A. Khan, J. H. Shah, F. Shehzad, T. Kim, and J. Cha, "Suspicious activities recognition in video sequences using darknet-nasnet optimal deep features," *Comput. Syst. Sci. Eng.*, vol. 47, no. 2, pp. 2337–2360, 2023.

[8] B. Ganga, B. T. Lata, and K. R. Venugopal, "Object detection and crowd analysis using deep learning techniques: Comprehensive review and future directions," *Neurocomputing*, p. 127932, 2024.

[9] B. H. Narayan, A. Chowdhury, and P. Shukla, "Anomaly detection in surveillance videos using deep learning and svm based data reduction method," in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, vol. 5. IEEE, Feb 2024, pp. 1733–1738.

[10] D. Kurchaniya and S. Kumar, "Two stream deep neural network based framework to detect abnormal human activities," *Journal of Electronic Imaging*, vol. 32, no. 4, p. 043021, 2023.

[11] J. H. Park, M. Mahmoud, and H. S. Kang, "Conv3d-based video violence detection network using optical flow and rgb data," *Sensors*, vol. 24, no. 2, p. 317, 2024.

[12] A. M. Abbas, K. Siddhardha, G. S. Dattatreya, and K. S. Reddy, "Machine learning-based suspicious activity detection for surveillance application," in *AIP Conference Proceedings*, vol. 2965, no. 1, July 2024.

[13] U. Gawande, K. Hajari, and Y. Golhar, "Real-time deep learning approach for pedestrian detection and suspicious activity recognition," *Procedia Computer Science*, vol. 218, pp. 2438–2447, 2023.

[14] S. Kansal, A. K. Jain, M. Biswas, S. Bansal, N. Mahindru, and P. Kansal, "Suspact: novel suspicious activity prediction based on deep learning in the real-time environment," *Neural Computing and Applications*, pp. 1–14, 2024.