

# Suspicious Behavior Detection Using Computer Vision

Dexter Sing Fong Leong, Hao Feng Chan, Shakir Hussain Naushad Mohamed, Wui Chung Alton Chau  
*School of Engineering, Deakin University, Australia*

Andi Prademon Yunus  
*Telkom University, Indonesia*

Takao Ito  
*Hiroshima University, Japan*

Zheng Cai, Xinjie Deng, Yit Hong Choo\*  
*Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Australia*  
*\*Email: s223026243@deakin.edu.au, y.choo@deakin.edu.au*

## Abstract

Detecting suspicious activity is a crucial task for public safety. The determination of class suspicious behavior is based on the facial cues of a person. Research has been conducted in this field using computer vision tools. However, accuracy still has room for improvement. Hence, this paper aims to use a novel approach in using other deep learning models to classify behavior as either suspicious or normal based on facial cues. By enhancing the detection process, this paper contributes to improving the reliable and effective surveillance system.

*Keywords:* Facial Cue, Computer Vision, Deep Learning Models, Suspicious, Surveillance

## 1. Introduction

Detecting suspicious behavior accurately and efficiently is critical to public safety and security. In the context of increasing global security concerns, surveillance systems equipped with computer vision technology provide an essential system for monitoring public and private spaces. However, while traditional surveillance systems are adept at capturing and recording vast amounts of footage, these systems often fall short of actively analyzing and interpreting behavioral cues that may indicate suspicious activities [1]. Enhancing these systems with the capability to detect subtle facial expressions and cues can significantly improve threat identification and response times, thereby bolstering overall security protocols [2].

The current surveillance system incorporated with computer vision predominantly utilizes object detection and motion tracking technologies [3]. Object detection involves identifying and classifying objects of interest within video feeds, such as people, vehicles, or specific items like weapons. This method uses deep neural networks (DNNs) [4] to automatically classify objects in real time, helping to quickly pinpoint elements that may be relevant to security concerns. Motion tracking, on the other hand, refers to the process of following the movement of identified objects across multiple frames of a video, allowing the system to track individuals or

objects as they move through different camera views. This technology is crucial for monitoring activities that occur over time, such as detecting the movement of individuals in restricted areas or tracking the behavior of a person across a public space. While both object detection and motion tracking are foundational for surveillance systems, they primarily focus on recognizing objects and their movement, which is insufficient for identifying subtle behavioral cues that may indicate suspicious activities or criminal intent. Hence, this study focuses on incorporating facial expression cues analysis. There is a substantial gap in the deployment of facial cue analysis within these systems, which could otherwise provide insights into the intent of an individual or emotional state [5] [6]. This gap highlights the need for an integrated approach that combines robust object detection with advanced facial expression analysis.

DenseNet [7, 8] is characterized by the unique structure, which connects each layer to every other layer in a feed-forward manner. This dense connectivity ensures that all layers directly receive feature maps from preceding layers, which promotes feature reuse and substantially reduces the vanishing gradient problem. This architecture arranges the network into several densely connected blocks separated by transition layers that perform convolution and pooling to reduce dimensionality and manage parameter growth.

ResNet18 [9] is designed to facilitate the training of deeper neural networks through residual blocks, which incorporate skip connections. These skip connections enable the network to bypass certain layers by adding the input of the block directly to the output, allowing the network to learn identity functions more effectively. This mechanism mitigates the risk of performance degradation as the network depth increases. By enabling the training of deeper networks without a proportional increase in training difficulty, the ResNet architecture improves both the efficiency and scalability of neural network training.

A further enhancement in the performance of ResNet was done by incorporating Squeeze-and-Excitation (SE) block [10]. The SE block performs dynamic channel wise feature recalibration, significantly improving the ability of the network to emphasize important features while suppressing less useful ones. This process is achieved by first squeezing global spatial information into a channel descriptor through global average pooling, followed by excitation, which recalibrates the feature maps based on the descriptor. This approach helps the network focus on the most relevant features for the task at hand. This mechanism is particularly beneficial for tasks requiring high sensitivity in feature detection, such as facial expression analysis in security systems, where accurately detecting subtle variations in facial expressions is crucial for identifying suspicious behavior.

This study aims to bridge the existing technology gap by employing advanced deep learning models ResNet 18 [9], DenseNet [7, 8], and SEResNet18 [10] to analyze facial cues for suspicious behavior detection. By training these models with custom datasets of micro facial expressions, this research seeks to improve the accuracy and reliability of automated threat detection, thus contributing to safer public environments [11].

## 2. Methodology

In this study, a robust methodology was employed to enhance the predictive accuracy of neural networks in detecting suspicious behaviors, utilizing facial expressions as primary indicators. A dataset was compiled, encompassing a wide range of facial expressions, categorized into two classes: suspicious (Class 1) and non-suspicious (Class 0). Advanced data management ensured uniformity and quality. Specialized deep learning architectures such as deep learning models ResNet 18 [9], DenseNet [7, 8], and the proposed SEResNet18 [10] were chosen for their effectiveness in image analysis. These models were trained and validated against specific metrics to measure accuracy and generalization across unseen data, contributing to advancements in automated surveillance technologies. The performance of the models [12] will be assessed using Mean Average Precision, Accuracy, Recall, Precision, and confusion matrix metrics [13]. These metrics are essential for evaluating the ability of the model to correctly identify suspicious behaviors while

minimizing false positives, thus ensuring the reliability of surveillance interventions.

### 2.1. Dataset Preparation

The preparation of the dataset for this study involved a meticulous and structured process. The initial step comprised the acquisition of images from external datasets, each representing a diverse array of facial expressions. This selection was guided by a review focused on identifying facial cues commonly associated with suspicious behavior.

After acquisition, the images underwent a labeling process, wherein bounding box annotations were meticulously applied to accurately delineate facial features. The enhancement of efficiency in data handling and preparation was done in Roboflow platform [14]. The capability of this platform is known to streamline dataset organization, annotation, class customization, and format conversion, facilitating swift and precise data processing. Data preprocessing techniques such as cropping, resizing, and normalization were conducted on this platform to ensure uniformity across the dataset.

The processed images were then formatted into the Common Objects in Context (COCO) format [15], ensuring integration with the deep learning architectures utilized in this research. A quality assurance protocol was implemented, involving a review and refinement of annotations to ensure accuracy and consistency within the dataset.

The custom dataset [16] comprises 1599 RGB images with a nearly balanced distribution across each class. This dataset was methodically partitioned into training (73%), validation (18%), and testing sets (9%). This distribution aimed to support robust model training, facilitate effective parameter tuning, and enable comprehensive performance evaluation.

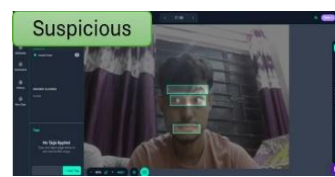


Fig. 1 Example of Suspicious Behavior.



Fig. 2 Example of Non-Suspicious Behavior.

### 2.2. Models For Suspicious Behavior Detection

In this study, three specialized convolutional neural network (CNN) architectures were employed to enhance the predictive accuracy in detecting suspicious behaviors via facial expressions (Fig.1, Fig.2). The models selected for this purpose include DenseNet [7, 8], ResNet 18 [9],

and. SEResNet18 [10], each optimized for efficacy in processing complex image data.

DenseNet models, accessible through the 'torchvision.models.densenet' module in PyTorch [17], offer flexibility in configuring network depth and growth rates to meet various computational and application needs. ResNet 18, part of the standard offerings within the 'torchvision.models.resnet' module in PyTorch, is readily available for both research and production use. Additionally, this study introduces SEResNet, an enhanced version of ResNet 18, incorporating SE blocks into each residual unit [10]. The 'torchvision.models.resnet' module was modified to integrate these SE blocks, thus improving model performance.

The models used in this study are trained with key hyperparameters that significantly impact their performance. The learning rate is set to 0.001 and optimized using Adam. The perks of using Adam optimizer are the adaptive learning rate properties, which aids in efficient training. The batch size is set to 32, balancing computational efficiency and memory usage, while 20 epochs provide enough iterations for the model to learn without overfitting. Cross Entropy Loss is used as the loss function for classification tasks, optimizing model performance by comparing predicted probabilities with true labels. Additionally, data augmentation techniques like random horizontal flips and rotations are applied to improve model robustness and prevent overfitting.

### 2.3. Metrics For Evaluation

The evaluation of convolutional neural network models SEResNet 18, DenseNet, and ResNet 18 in this study involves several key metrics including Accuracy, Precision, Recall, F1-Score, and Mean Average Precision (mAP). Each metric contributes to understanding the effectiveness of these models in detecting suspicious behaviors based on facial cues.

This Accuracy quantifies the overall correctness of the models in classifying facial expressions either as suspicious or non-suspicious. This metric involves calculating the ratio of correct predictions (true positives and true negatives) to the total number of cases tested.

Precision [12] reflects the reliability of the model in identifying an expression as suspicious, determined by the ratio of true positives to the combined total of true positives and false positives. High Precision reduces unnecessary alarms in surveillance scenarios. Eq. (1) describes the formula for calculating Precision in a classification context where TP is the correctly predicted positive instances and FP is the incorrectly predicted as positive.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \quad (1)$$

Recall [12], known as sensitivity, measures the capability of model to identify all relevant instances of suspicious expressions. This metric calculates the ratio of true positives to the sum of true positives and false negatives, ensuring comprehensive detection of suspicious activities. Eq. (2) presents the formula for calculating Recall in classification tasks where FN is the actual positive instances incorrectly predicted as negative.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \quad (2)$$

F1-Score [12] represents the harmonic mean of Precision and Recall. This metric evaluates the balance between Precision and Recall, critical for informed decision-making in security contexts. Eq. (3) ensures that the F1-Score considers both the Precision (and the Recall of the test. By balancing these two metrics, the F1-Score serves as a useful measure when you need a single metric to evaluate the overall accuracy of a model, especially in scenarios where uneven class distribution might make Precision or Recall alone misleading. The use of the harmonic mean punishes extreme values, making the F1-Score a more robust measure that requires both Precision and Recall to be relatively high to achieve a high score.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Mean Average Precision (mAP) [12] evaluates model consistency and reliability by measuring Precision at varying Recall levels, averaged over multiple thresholds. Eq. (4) presents the formulas for calculating Average Precision (AP) and mAP, where N is the number of classes, p(r) is Recall r, Ap indicates Average Precision for each class and mAP indicates Mean Average Precision. Eq. (4) assess object detection performance across classes.

$$\text{AP} = \int_0^1 p(r) dr$$

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^N \text{AP}_i \quad (4)$$

A confusion matrix [13] describes the performance of a classification model with known true values. It details how predictions are distributed across categories. In this study, the confusion matrix has two classes: Class 1 for suspicious behavior and Class 2 for nonsuspicious behavior. An ideal model would show most detections in

TP (True Positives) and TN (True Negatives), while poor results indicate lower model performance.

### 3. Results And Discussion

#### 3.1. Results

This study evaluates the performance of three convolutional neural network models. The primary model SEResNet 18 will be fairly compared with ResNet 18 and DenseNet for model evaluation. The task of these models is to detect suspicious behavior through facial cues based on the custom dataset. The following results will present Confusion Matrix, Precision, Recall, F1-Score, Accuracy and mAP. These results will reflect the competence of each model and provide a clear conclusion for the study.

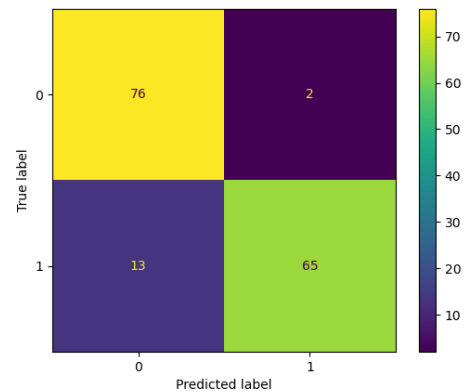


Fig. 5 DenseNet Confusion Matrix.

For SE-ResNet18, Fig. 3 showed 73 true positives and 74 true negatives, with minor misclassifications evident through 5 false positives and 4 false negatives. The performance of ResNet 18 in Fig. 4 was slightly lower with 68 true positives and 75 true negatives but had a higher number of false positives at 10, and fewer false negatives at 3. DenseNet (Fig. 5) demonstrated a strong ability to identify true positives (76) and true negatives (65) but struggled with Recall, evidenced by a higher number of false negatives (13) compared to only 2 false positives.

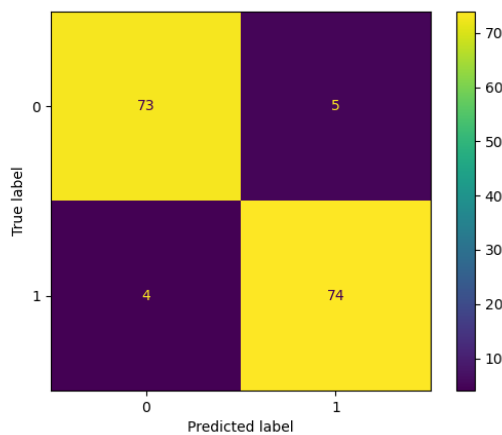


Fig. 3 SE-ResNet 18 Confusion Matrix.

Table 1. Performance Result of Models.

	Suspicious Class			Accuracy (%)
	Precision (%)	Recall (%)	F1-Score (%)	
<b>SE-ResNet 18</b>	94	95	94	97
<b>DenseNet</b>	97	83	90	90
<b>ResNet 18</b>	88	96	92	92

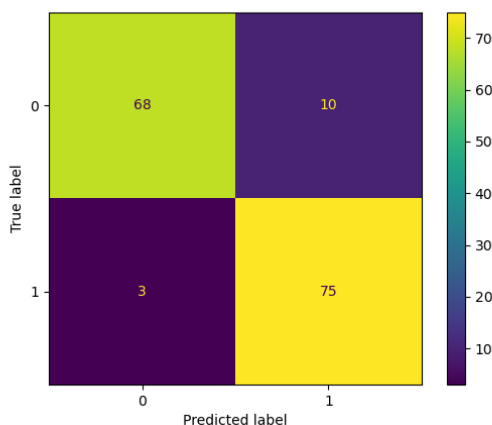


Fig. 4 ResNet 18 Confusion Matrix.

SE-ResNet18 (Table 1) demonstrated the highest overall Accuracy at 97%, supported by impressive Precision, Recall, and F1-Scores of 94%, 95%, and 94% respectively. DenseNet followed with an Accuracy of 90%, boasting the highest Precision of 97% but the lowest Recall at 83%, which affected F1-Score of DenseNet, also at 90%. ResNet18 showed robust performance with an overall Accuracy of 92%, and while it had a lower Precision at 88%, it recorded the highest Recall of 96%, culminating in an F1-Score of 92%.

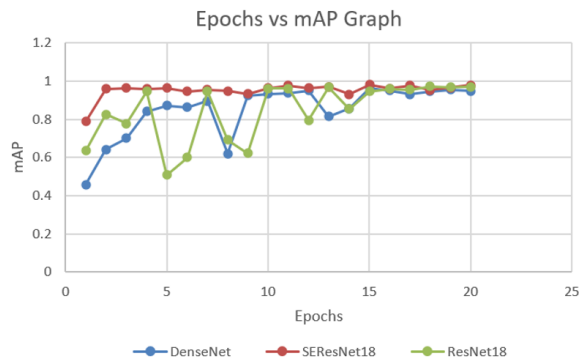


Fig. 6 mAP For Positive Detection Results.

Fig. 6 shows that DenseNet exhibited significant mAP variability, indicating performance instability. In contrast, SE-ResNet18 demonstrated stable, consistent mAP, while ResNet18, though slightly lower in mAP, was more consistent than DenseNet.

### 3.2. Discussion

The evaluation of SEResNet 18, alongside ResNet 18 and DenseNet, reveals a superior capability in detecting suspicious behaviors through the analysis of facial cues using a custom dataset. As the primary model, SEResNet 18 not only achieved the highest Accuracy, Precision, Recall, and F1-Scores but also demonstrated robustness across different testing conditions. This result suggests that the integration of Squeeze-and-Excitation (SE) blocks, which enhance feature recalibration capabilities, significantly contributes to the effectiveness of detection. Despite the standout performance of SEResNet 18, this model did not achieve 100% Accuracy, which underscores the presence of intrinsic challenges and room for improvement. Even though the custom dataset used in this study offers a substantial improvement in clarity compared to more generalized datasets, like crime UCF videos, it still presents limitations. The variation in facial cue clarity, especially under less-than-ideal conditions such as low light or partial obstructions, affects even the most advanced models like SEResNet 18. Enhancing dataset quality with more varied and challenging examples could help improve the robustness of models.

Despite SEResNet 18 overall efficacy, this model showed some fluctuations in Mean Average Precision (mAP) over training epochs, indicative of potential overfitting to the training data or an inability to generalize perfectly to new data. This suggests that while SEResNet 18 effectively learns from the dataset provided, performance of the model could be impacted by the presence of noise or non-representative data. While SEResNet 18 leads in many performance metrics, incorporating features from other architectures like dense connectivity of DenseNet could provide a balance between depth and feature

richness, potentially leading to even higher accuracy and stability

In summary, while SEResNet 18 emerges as the most capable model for detecting suspicious behavior, further enhancements to the dataset and model architecture could elevate the accuracy and generalization capabilities of the trained model. Addressing these areas could make SEResNet 18 even more effective, paving the way for the deployment of model in surveillance applications requiring high reliability and Precision.

### 4. Conclusion

This research has validated the efficacy of advanced CNN models, particularly SEResNet 18, ResNet 18, and DenseNet, in detecting suspicious behaviors through facial cues. SEResNet 18 stood out for the superior performance in Accuracy, Precision, Recall, and F1-Scores, benefiting from SE blocks that enhance feature recognition. Despite these successes, none of the models achieved perfect accuracy, highlighting the need for improved datasets and model architectures that better mirror real-world complexities. Future research should focus on refining datasets and exploring hybrid architectures to enhance the generalization capabilities of these systems. This study underscores the potential of deep learning models to significantly improve automated surveillance, advancing public safety in both theory and application.

### References

1. Jiyang Xie, Y.Z., Ruoyi Du, Weiyu Xiong, Yufei Cao, Zhanyu Ma, Dongpu Cao, and Jun Guo, *Deep Learning-Based Computer Vision for Surveillance in ITS: Evaluation of State-of-the-Art Methods*. IEEE Transactions on Vehicular Technology, 2021. 70(4): p. 3027–3040.
2. Harikrishnan, J., Sudarsan, Arya., Ajai, Remya A. S., Sadashiv, Aravind., *Vision-Face Recognition Attendance Monitoring System for Surveillance using Deep Learning Technology and Computer Vision*, in *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*. 2019, IEEE.
3. Haroon Idrees, M.S., and Ray Surette, *Enhancing camera surveillance using computer vision: a research note*. arXiv, 2018. 1808.03998.
4. Szegedy, C., Toshev, A., & Erhan, D, *Deep neural networks for object detection*, in *Advances in Neural Information Processing Systems (NIPS) 2013*. 2013, 1-9. p. 26.
5. Marco Leo, P.C., Pier Luigi Mazzeo, Paolo Spagnolo, Dario Cazzato, Cosimo Distanto, *Analysis of Facial Information for Healthcare Applications: A Survey on Computer Vision-Based Approaches*. Information, 2020. 11(128).
6. Zhe Li, T.Z., Xiao Jing, Youning Wang, *Facial expression-based analysis on emotion correlations, hotspots, and potential occurrence of urban crimes*. Alexandria Engineering Journal, 2021. 60: p. 1411–1420.

7. Gao Huang, Z.L., Laurens van der Maaten, Kilian Q. Weinberger. *Densely Connected Convolutional Networks*. in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017. Honolulu, HI, USA: IEEE.
8. Yi Zhu, S.N. *DenseNet for Dense Flow*. in *IEEE International Conference on Image Processing (ICIP)*. 2017. Beijing, China: IEEE.
9. Kaiming He, X.Z., Shaoqing Ren, Jian Sun, *Deep Residual Learning for Image Recognition*, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, IEEE: Boston, MA, USA. p. 770–778.
10. Jie Hu, L.S., Gang Sun, *Squeeze-and-Excitation Networks*, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)*. 2018, IEEE. p. 7132–7141.
11. Fiza Abdul Razzaq, W.T., Muhammad Abbas Chaudary, Muhammad Waqas, Sumaria Fareed, Shoaib Javaid. *Enhancing Public Safety: Detection of Weapons and Violence in CCTV Videos with Deep*. in *25th International Multitopic Conference (INMIC)*. 2023. IEEE.
12. Rafael Padilla, S.L.N., Eduardo A. B. da Silva. *A Survey on Performance Metrics for Object-Detection Algorithms*. in *Proceedings of the International Workshop on Signal Processing (IWSSIP)*. 2020. Niterói, Brazil: IEEE.
13. Mohammadreza Heydarian, T.E.D., Reza Samavi, *MLCM: Multi-Label Confusion Matrix*. *IEEE Access*, 2022. 10: p. 19083–19095.
14. Team, R. *Roboflow: End-to-End Computer Vision Workflow*. 2023; Available .
15. Bhadani, R. *How to Work with Object Detection Datasets in COCO Format*. 2022.
16. *Sus\_or\_non\_v1*, D.S.F. Leong, Editor. 2025, Roboflow.
17. Team, P. *PyTorch Vision: Models*. 2023.

---



---

### Authors Introduction

Mr. Dexter Sing Fong Leong



He is currently an undergraduate pursuing Bachelors Of Mechatronic Engineering in Deakin University, Australia.

Mr. Hao Feng Chan



He is currently an undergraduate student majoring in Bachelor of Mechatronics Engineering (Honours) at Deakin

Mr. Shakir Hussain Naushad Mohamed



He is currently an undergraduate student majoring in Bachelor of Electrical and Electronics Engineering (Honours) at Deakin University, Australia.

Mr. Chau Wui Chung Alton



He is currently an undergraduate student in Bachelor of Mechanical Engineering (Honours) at Deakin University, Australia.

Mr. Zheng Cai



He is currently a PhD candidate at Deakin University's Institute for Intelligent Systems Research and Innovation (IISRI). His research interests include multiobjective optimisation algorithms such as metaheuristic algorithms and evolutionary algorithms for scheduling problems. He is also exploring the integration of machine learning with optimisation algorithms.

Ms Deng Xinjie



She is pursuing a PhD in Information Technology at the Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, with her research centered on creating lightweight deep learning algorithms for computer vision applications.

Andi Prademon Yunus, Ph.D



He is an Assistant Professor at Telkom University and he received his PhD in Engineering from Mie University, Japan. His research focuses on applied and fundamental machine learning for motion and behavior computing. He also collaborates with industry partners to develop AI-based tools for language modeling and image analytics.

Dr. Takao Ito



He is Professor of Management of Technology (MOT) in Graduate School of Advanced Science and Engineering at Hiroshima University. His current research interests include automata theory, artificial intelligence, systems control, quantitative analysis of interfirm relationships using graph theory, and engineering approach of organizational structures using complex systems theory.

Dr. Yit Hong Choo



He has completed his PhD and is now a Research Fellow in Operations Analytics at Deakin University's Institute for Intelligent Systems Research and Innovation (IISR). His research focuses on advanced multi-objective optimisation algorithms for complex maintenance scheduling in rolling stock. He also collaborates with transportation industry partners to develop AI-based tools for video and image analytics.