Available on: http://gjetr.ep-journals.org/ E-ISSN: 3051-3782, P-ISSN: 3051-3774

Global Journal of Engineering and Technology Review

Volume: 02 Issue: 11 November 2026

Pose-Based Suspicious Activity Detection in Surveillance Video: A Systematic Literature Review

Cole Rodney Migisha
Department of Computing and Technology, Faculty of Engineering Design and Technology, Uganda Christian University, Mukono,Uganda.
ORCID: 0009-0000-5311-2485.

Innocent Ndibatya
Department of Computing and Technology, Faculty of Engineering Design and Technology, Uganda Christian University, Mukono, Uganda

Peter Jamugisa
Department of Computing and Technology, Faculty of Engineering Design and Technology, Uganda Christian University, Mukono, Uganda

Published Date: November 26, 2026 **Article DOI : 10.65150/EP-gjetr/V2E11/2026-12**

ABSTRACT:

objective: Security cameras record more video than staff can review, so suspicious behaviour can go unnoticed and later be hard to locate. This review began with face and pose recognition in video, found that face-based identification is limited by privacy concerns and inconsistent footage quality, and narrowed to pose-based detection of suspicious activity, including examination malpractice.

Methods: Scopus and ScienceDirect were searched. The selection process, reported with a PRISMA 2020 flow diagram, screened 19,043 records and included nine studies. Six further sources from outside the search, three on examination surveillance, gave context. The evidence was synthesised narratively around three questions: tracking in crowds, changing viewing angle, and low resolution with limited computing.

Results: Face-based pipelines reported high accuracy in simple settings but 29.67% face recognition accuracy on complex footage in one study. Pose-based work included a retail benchmark with a best unsupervised AUC-ROC of 67.46% and an examination study using YOLOv3 and OpenPose that reported 99.58% test accuracy on its own dataset but 26.3% precision on a simulated video. The examination studies used small, self-collected data. None of the examined sources combined pose-based suspicious action classification with tests for crowds, viewing angle, resolution, explanation and retrieval from long footage.

Conclusions: Pose is a promising, privacy-conscious input for examination halls and similar settings, but the evidence is small, mostly indirect or self-collected, and calls for a system tested against the conditions identified here.

KEYWORDS: pose estimation, suspicious activity detection, video surveillance, multi-person tracking, anomaly detection, low-resolution video, examination malpractice, systematic literature review

INTRODUCTION

Security cameras record more video than the people responsible for them can watch. Hodgetts et al. (2017) review the cognitive demands placed on CCTV operators, including visual search, perceptual overload, attentional failures and vulnerability to distraction, and argue that intelligent technologies can support operators but do not replace them entirely. Martínez-Mascorro et al. (2020) make a related observation: a person can attend to only a small part of a screen at a time, so an event on another screen, or in another part of the same screen, may go unnoticed. Footage that is not watched as it is recorded can only be examined afterwards, when finding one short event may mean searching through hours of video. The same difficulty appears in examination rooms, where Ding et al. (2016) report that it is hard for censors to detect transitory abnormal behaviours in time when many monitoring images are shown at once, and that relevant departments must review all the recordings afterwards if an examinee is suspected of cheating.

Computer vision supplies building blocks for helping with this problem. Voulodimos et al. (2018) review deep learning schemes, namely convolutional neural networks, deep Boltzmann machines and deep belief networks, and stacked denoising autoencoders, and describe their use in tasks that include object detection, face recognition, action and activity recognition, and human pose estimation. These tasks are the ingredients of an automated surveillance assistant. Whether they work together on real surveillance footage is a separate question, and it is the one this review pursues.

The review began with a broad question: how do computer vision methods for face and pose recognition and tracking perform on video footage? Two difficulties with face-based identification shaped what followed. The first is privacy, because recognising a face requires identifiable imagery. The second is the quality of the footage, which varies with camera distance, viewing angle and resolution, and which bears directly on face recognition. Pose-based methods describe a person by the positions of body joints instead of facial appearance, which reduces the identifiable information that is kept (Rashvand et al., 2025). Pose-based recognition of suspicious behaviour also suited a practical need: one of the intended

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

applications is the detection of examination malpractice, a concern regarded as highly relevant at the first author's university and one for which automated analysis of examination video has been studied for some years (Cai et al., 2016; Ding et al., 2016; Wan et al., 2021). The review therefore narrows from the wide field of face and pose analysis in video to a specific target, the detection of suspicious human activity from body pose in surveillance video.

Three research questions guide the synthesis. **RQ1:** Can a system handle one target person or action in a crowd of several people, including similar-looking or similar-motion cases? **RQ2:** How can a model adapt to viewing people and actions from different angles without storing every possible view? **RQ3:** How can a system run efficiently on low-resolution video while keeping useful detection results? The remainder of the paper gives the background literature, the methods used to find and select studies, the results of the review, and a discussion that traces how the evidence points towards pose-based detection and what remains unresolved.

LITERATURE REVIEW

This section sets out the wider field with the help of four reviews, so that the studies examined later can be placed in context. It moves from deep learning in general, through action recognition and the detection and tracking of people, to image segmentation.

Deep learning for computer vision

Voulodimos et al. (2018) describe how deep learning methods have been shown to outperform earlier machine learning techniques in several fields, with computer vision among the most prominent. Their review covers convolutional neural networks, deep Boltzmann machines and deep belief networks, and stacked denoising autoencoders, giving a brief account of each family's history and structure before turning to applications. Those applications include the tasks on which this review depends: face recognition, action and activity recognition, and human pose estimation.

Human action recognition

Poppe (2010) surveys vision-based human action recognition and adopts a hierarchy of three levels from earlier work by Moeslund and colleagues: the action primitive, the action and the activity. An action primitive is an atomic movement described at the limb level, such as "left leg forward". An action consists of action primitives and describes a whole-body movement, such as "running". An activity contains a number of subsequent actions and gives an interpretation of the movement being performed, as "jumping hurdles" does. Poppe states that his survey focuses on actions and does not explicitly consider context such as the environment or interactions between persons. This review treats suspicious behaviour as an activity-level idea whose meaning depends on the setting. That reading is the present authors' interpretation and not a claim made by Poppe.

Poppe also draws attention to practical difficulties. Occlusion can be alleviated when observations from several views are combined, dynamic backgrounds make it harder to locate the person and observe the motion, and the rate at which an action is performed varies. The publicly available datasets give a sound basis for comparison, but they often lack these variations. More realistic datasets built from movies or web videos have been introduced, yet they are still limited in the number of training and test sequences (Poppe, 2010).

Detecting and tracking people

Brunetti et al. (2018) survey pedestrian detection and tracking, analysing systems by field of application, acquisition technology, computer vision techniques and classification strategies. They identify video surveillance, human-machine interaction and analysis as the three main application fields, discuss 2D and 3D vision systems and indoor and outdoor systems, and give a dedicated section to deep learning. Their reported results highlight the importance of testing pedestrian detection systems on different datasets to evaluate how robust the features used are. They also note that background subtraction, which separates moving objects from a scene, is not robust to illumination variability, dynamic background, shadows or noise.

Image segmentation

Minaee et al. (2022) review deep learning for image segmentation, covering more than 100 methods grouped into 10 categories by model architecture, together with the datasets commonly used and comparisons of performance. Segmentation assigns a label to image pixels, for instance separating the pixels of people from the background. It can support region extraction in a vision pipeline, but it does not by itself describe how a person moves, so it is treated here as background and not as a core component.

These reviews describe the tools. They do not establish how well pose-based detection of suspicious activity works on real surveillance footage. The next sections explain how the primary studies were identified and what they report.

MATERIALS AND METHODS

Review design and reporting

The review followed a documented sequence of identification, duplicate checking, screening and full-text eligibility assessment, and its selection process is reported with a PRISMA 2020 flow diagram. The search and the eligibility criteria below were written for the original, broader question on face and pose recognition and tracking in surveillance video. The narrowing to suspicious pose detection is described at the end of this section.

Information sources and search strategy

Two bibliographic databases were searched: Scopus and ScienceDirect, on Wednesday, 30 September, 2026. Four Boolean search strings were used (Table 1). The first was the anchor query, and the other three were applied across the title, abstract and keyword fields to capture tracking, pose estimation and video resolution constraints.

Table 1. Search strings used in the database search.

No.	Search string
1	"computer vision" AND "face-pose recognition"
2	("face recognition" OR "pose estimation") AND ("tracking" OR "multi-target tracking") AND "video footage"
3	("computer vision") AND ("face-pose tracking" OR "action recognition") AND ("surveillance" OR "unmonitored feed")
4	("pose-based action recognition") AND ("low-resolution video" OR "video resolution") AND "security"

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Eligibility criteria

Studies were eligible if they were peer-reviewed journal articles or primary conference papers written in English; introduced or evaluated computer vision architectures designed for face recognition, pose estimation, multi-target tracking or combined face-pose models; used real-time or offline video footage, such as surveillance, ambient or multi-camera security feeds, as the main input; addressed performance under visual constraints such as resolution, extreme pose angles or distance; and presented empirical evaluations with quantifiable performance metrics such as accuracy, mean average precision, multiple object tracking accuracy or frames per second.

Studies were excluded if they were non-peer-reviewed publications, editorial commentaries, extended abstracts, book reviews or patents; focused only on single static images without temporal or sequential frame evaluation; evaluated spatial tracking or biometric verification without facial or pose features; concerned specialised domains outside general surveillance, such as medical or microscopic imaging; or lacked key operational metrics, validation data or reproducible methods.

Selection process

The searches returned 19,043 records, of which 105 came from Scopus and 18,938 from ScienceDirect. Duplicate detection identified no duplicates. Eighteen reports were sought for full-text assessment. The three surveys among the included studies (Poppe, 2010; Brunetti et al., 2018; Minaee et al., 2022) do not report new empirical evaluations, and the record of how the criteria were applied to them is not available.

Data extraction and synthesis

A standardised extraction matrix was used for the included studies. For each study it captured publication details, the core algorithm, the datasets and conditions, the tracking and recognition approaches, the quantitative results, and the limitations and deployment constraints identified by the authors. A formal quality appraisal or risk-of-bias assessment is not documented. Results were synthesised narratively around the three research questions. Because the studies use different datasets, evaluation protocols and metrics, no meta-analysis was attempted, and reported figures are interpreted in their original contexts and are not ranked against one another.

Additional sources

Six sources identified by the author outside the database search were added to give context for the narrowed question. Three give general context: Hodgetts et al. (2017) on the cognitive limits of CCTV operators, Martínez-Mascorro et al. (2020) on suspicious behaviour before shoplifting, and Voulodimos et al. (2018) on deep learning for computer vision. Three concern examination surveillance (Cai et al., 2016; Ding et al., 2016; Wan et al., 2021). These came from a targeted search of IEEE Xplore for the phrase "abnormal behaviour in examination surveillance", and the papers were downloaded on 08 October 2026. The six sources were not screened against the eligibility criteria above and are marked in Table 2 and Figure 1. In total, fifteen sources are discussed in this review: nine from the database search and six additional.

Narrowing the scope

As the review progressed, the focus moved from face and pose recognition in general to pose-based detection of suspicious activity. Two reasons drove this change. Face recognition faced privacy concerns and depended on footage quality that was not consistent. At the same time, recognising suspicious poses suited a use case of particular interest, examination malpractice, which was regarded as highly relevant at the author's university. The nine included studies were therefore re-read against the narrowed question and the three research questions above, and the targeted examination search described above was added. The database search was originally built for faces, so the review does not claim to have searched exhaustively for pose-based suspicious activity work, and the examination search used one phrase in one database.

RESULTS

Study selection

Of the 19,043 records identified, 19,025 were excluded at title screening and 18 reports were sought for retrieval. All 18 were retrieved and assessed for eligibility, and nine were excluded because they addressed the wrong target application. Nine studies were included (Figure 1). The six additional sources supplied by the author are shown separately in the figure.

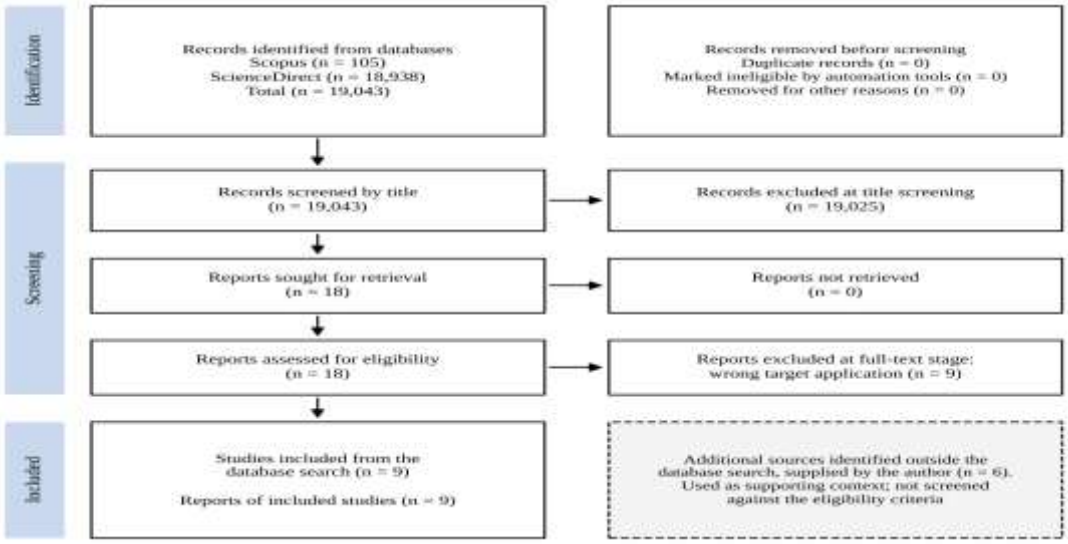

Figure 1. Flow diagram of the study selection process, showing the numbers recorded for the database search and the six additional sources.

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Characteristics of the sources

Table 2 lists the nine included studies and the six additional sources. Five of the fifteen sources are reviews or surveys that report no new results of their own on the review questions (Poppe, 2010; Brunetti et al., 2018; Minaee et al., 2022; and, among the additional sources, Hodgetts et al., 2017 and Voulodimos et al., 2018). Of the remaining ten sources, three concern faces (Kim et al., 2008; Gorodnichy, 2005; Sugiharto et al., 2025, which also includes activity recognition), one concerns person tracking (Benfold & Reid, 2011), one concerns action recognition from video (Ji et al., 2013), and five concern suspicious behaviour or anomalies from video or pose (Rashvand et al., 2025; Martínez-Mascorro et al., 2020; and the three examination studies, Cai et al., 2016; Ding et al., 2016; Wan et al., 2021).

Table 2. Sources discussed in the review. Reported results are as stated by the original authors, on different datasets and metrics, and are not directly comparable. † = additional source identified outside the database search.

Source	Type	Focus and setting	Result reported by the authors
Poppe (2010)	Survey	Vision-based human action recognition; action primitive, action and activity hierarchy; common datasets	No single performance figure. Datasets often lack variations; realistic datasets are limited in size
Ji et al. (2013)	Method	3D convolutional networks for action recognition; surveillance video from London Gatwick Airport (49 hours) and KTH	90.2% accuracy on KTH, against 91.7% for the HMAX baseline
Benfold & Reid (2011)	Method	Multi-target tracking with HOG head detection, KLT tracking and MCMCDA; Town Centre video, 1920×1080 at 25 fps	Head regions: MOTA 45.4%, MOTP 50.8%. Body regions: MOTA 61.3%, MOTP 80.3%
Brunetti et al. (2018)	Survey	Pedestrian detection and tracking; surveillance, human-machine interaction and analysis; deep learning	No single performance figure. Stresses testing on different datasets
Rashvand et al. (2025)	Dataset and benchmark	PoseLift: pose sequences from real retail CCTV, 1920×1080 at 15 fps; HRNet poses with ByteTrack; unsupervised anomaly detection	Best model STG-NF: AUC-ROC 67.46%, AUC-PR 84.06%. Lowest AUC-ROC (GEPC): 60.61%
Sugiharto et al. (2025)	Method	YOLOv8, InsightFace and CLIP on two 30-minute 720p office videos	Simple video: detection 95.73%, faces 96.58% (96.56% in the abstract), activities 53.25%. Complex video: 48.15%, 29.67%, 33.44%. About 4 FPS on a CPU
Kim et al. (2008)	Method	Face tracking and recognition with visual constraints; Honda/UCSD and YouTube videos	LDA features 97.62%; PCA (50 dimensions) 92.86%; LDA with landmarks 100%; over 70% on 1,500 YouTube sequences of 35 celebrities
Gorodnichy (2005)	Method	Associative neural memory accumulating evidence over frames; faces about 12 pixels between the eyes	Reported on low-resolution video; no summary figure is quoted in this review
Minaee et al. (2022)	Survey	Deep learning image segmentation; over 100 methods in 10 categories; datasets such as PASCAL VOC, Cityscapes and MS COCO	No single performance figure. Compares methods on benchmarks
Martínez-Mascorro et al. (2020) †	Method	3D convolutional network on pre-crime behaviour before shoplifting; samples from UCF-Crime	75% accuracy reported; above 85% in the best runs at 80×60 pixels, 75% on average over 30 runs
Ding et al. (2016) †	Method	Examinee detection (HOG, SVM and AdaBoost) and cheating detection (frame difference and sparse combination) in examination video	62 of 72 annotated behaviours detected (86.1%); about 200 FPS on a CPU in MATLAB
Cai et al. (2016) †	Method	Optical flow and a two-layer 3D convolutional network for abnormal behaviour in classroom examination video; own dataset	Accuracy 87.6% and 89.8% for two flow-image variants; precision up to 86.5%; 82 of 86 events detected
Wan et al. (2021) †	Method	YOLOv3 and OpenPose key points with a CNN classifier for peeping and note passing; own simulated dataset of over 5,000 skeleton images	99.58% test accuracy on own dataset; on a simulated video accuracy 96.2%, recall 100%, precision 26.3%
Hodgetts et al. (2017) †	Review	Cognitive challenges of CCTV operators	Narrative review; no performance metrics
Voulodimos et al. (2018) †	Review	Deep learning for computer vision	Narrative review; no performance metrics

RQ1: one target in a crowd

Benfold and Reid (2011) track pedestrians in surveillance video by detecting heads with a histogram-of-oriented-gradients detector, following feature points with Kanade-Lucas-Tomasi tracking, and associating observations over time with Markov-Chain Monte-Carlo Data Association.

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

On a busy town centre sequence they report head-region MOTA of 45.4% and MOTP of 50.8%, and body-region MOTA of 61.3% and MOTP of 80.3%. In their ablation, switching off their model of repeated false-positive detections reduced head-region MOTA from 45.4% to 22.4%, and switching off the KLT failure model reduced it to 38.0%. The method keeps track of who is where. It does not classify what each person is doing. In a different setting, Sugiharto et al. (2025) tested the same pipeline on two office videos. On the complex video, where people were farther away, more faces were sideways and some people were obstructed, YOLOv8 detected 48.15% of the people, face recognition accuracy was 29.67% and activity recognition accuracy was 33.44%. The authors recommend placing cameras so that the person detector does not capture several overlapping people in one cropped image. Rashvand et al. (2025) used ByteTrack to keep track of individuals across frames when building PoseLift. In examination rooms, Wan et al. (2021) rely on a multi-person pose estimator that is not limited by the number of people, and Cai et al. (2016) divide each frame into left, middle and right sub-regions to detect behaviour more precisely; neither they nor Ding et al. (2016) report tracking individuals across frames. A keyword search of the full texts of the ten method and benchmark papers found no report of how identity switches in tracking affect pose sequences or the classification built on them; Benfold and Reid (2011) mention identity switches only in defining the MOTA measure.

RQ2: changing viewing angle

Kim et al. (2008) address changing head pose in face tracking and recognition. Instead of storing a template for each view, they use generic pose constraints and combine pose and identity evidence over time. On the Honda/UCSD database they report 97.62% accuracy with linear discriminant analysis features, 92.86% with a 50-dimensional principal component analysis representation, and 100% when LDA and landmark features were combined, and over 70% on a set of 1,500 YouTube sequences of 35 celebrities. Gorodnichy (2005) detects the rotation of the face within the image plane and aligns the eyes before recognition. Sugiharto et al. (2025) report that their face database contained only frontal views of each person, and they attribute the lower recognition accuracy on sideways angles to this. Wan et al. (2021) included normal behaviour shot from different angles in their examination dataset and added some images with blocked key points for training, and they describe an earlier OpenPose approach that worked only from specific angles; Cai et al. (2016) list more camera angles as future work.

For pose-based work, Rashvand et al. (2025) built PoseLift from several camera views in a real store, with high-angle cameras at 1920×1080 and 15 frames per second, and describe the dataset as addressing the single camera view of earlier datasets. Poppe (2010) notes that occlusion can be alleviated when observations from multiple views are combined, and that common action datasets often lack variation in viewpoint and similar factors.

RQ3: low resolution and limited computing

Gorodnichy (2005) studies faces with as few as about 12 pixels between the eyes, where the approach accumulates evidence over consecutive frames in an associative neural memory trained by incremental projection learning. Benfold and Reid (2011) report that running their HOG detector on every video frame was roughly ten times too slow for real time. Their design runs detections asynchronously while feature tracking continues between them, which allows real-time (25 frames per second) tracking of high-definition video; their detector used a GPU implementation. Sugiharto et al. (2025) report about 4 frames per second for their pipeline on a laptop with a Ryzen 7 4800H processor and 16 GB of memory without GPU acceleration.

Martínez-Mascorro et al. (2020) varied the resolution of grayscale frames among 160×120, 80×60, 40×30 and 32×24 pixels. They report that 80×60 gave the best results, with accuracy above 85% in the best runs, and that 30 repeated runs of the best configuration gave an average accuracy of 75%. They also report that depths of 10 or 30 consecutive frames gave most of the best results. In examination video, Ding et al. (2016) report about 200 frames per second on a 2.8 GHz CPU with 4 GB of memory in MATLAB with the test data held in memory beforehand, Cai et al. (2016) resized frames to 195×100 pixels, and Wan et al. (2021) report 0.33 frames per second when every frame is processed and 10.65 frames per second when only key frames chosen by inter-frame difference are processed.

Pose-based and behaviour-based detection of suspicious activity

Rashvand et al. (2025) present PoseLift, which they describe as a privacy-preserving dataset of anonymised human pose data from a retail store, built to support shoplifting detection as an anomaly detection problem. They contrast it with earlier datasets described as real-world but based on prearranged scenarios. They benchmark three unsupervised pose-based models. STG-NF reached the highest AUC-ROC (67.46%) and AUC-PR (84.06%), and GEPC had the lowest AUC-ROC (60.61%). They also show a false-positive case in which normal activities such as walking and using a mobile phone received low normality scores. The authors state that their results show pose-based approaches achieving high detection accuracy while addressing privacy and bias concerns.

Martínez-Mascorro et al. (2020) take a different route. They model suspicious behaviour, meaning behaviour before a crime is committed, using a 3D convolutional network on short video segments taken from the shoplifting and normal classes of the UCF-Crime dataset. They report 75% accuracy in detecting suspicious behaviour before a crime, on a small dataset of 60 suspicious and 60 normal samples in its basic form. Ji et al. (2013) state that their 3D convolutional model is trained with a supervised algorithm that requires a large number of labelled samples.

Examination surveillance

Three additional sources address examination rooms directly. Ding et al. (2016) describe the difficulty plainly: with many monitoring images on screen at once it is hard for censors to detect transitory abnormal behaviours in time, and after an examinee is suspected the relevant departments must review all the recordings. Their method has two parts. Examinee presence, which covers absence, lateness and leaving early, is checked by detecting the head-shoulder region with HOG features and a two-layer classifier made of an SVM followed by AdaBoost. Suspicious cheating behaviour is detected from frame-difference foreground features and a sparse combination model learned from 22 normal-behaviour clips of about 1,400 frames each, so that motion that does not fit the learned normal patterns is flagged. On a test video containing peeking, walking, bending and passing notes, they report that 62 of 72 annotated behaviours were detected (86.1%): 25 of 30 peeking events, 7 of 7 walking, 6 of 6 bending and 24 of 29 note-passing events. They collected their own samples because open datasets for head-shoulder detection were lacking, they note that the method also locates where in the frame the suspicious behaviour occurs, and they state that the false alarm rate needs further study.

Cai et al. (2016) use optical flow and a 3D convolutional network with two convolutional layers to detect abnormal behaviour in examination video. Finding no public examination dataset, they collected classroom surveillance videos at 25 frames per second, resized frames to 195×100

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

pixels, split each frame into three sub-regions and cut the videos into non-overlapping 16-frame clips. Training used 1,055 positive and 1,162 negative clips, and the test set held 1,080 samples (370 positive and 710 negative). Two ways of converting optical flow into images gave accuracies of 87.6% and 89.8%, and the abstract reports a precision of up to 86.5%. At the level of events, 82 of 86 abnormal events in the test videos were detected, which the authors present as a recall of 95.3%. The behaviours shown include reaching out a hand, turning around, leaving and leaning. The authors explain a missed clip by noting that a student who turned and then held still generated no motion information, and they name more camera angles as future work.

Wan et al. (2021) is the only examination study among the three that works from body pose. Key frames are selected by inter-frame difference, YOLOv3 finds each candidate's body box, OpenPose extracts body, face and hand key points, and a convolutional network (VGG16) classifies the skeleton as normal or cheating. Because examination video is generally not public and involves candidates' privacy, they built their own dataset by simulating examinations in classrooms and a laboratory with a 4K camera under different lighting and angles. It holds more than 5,000 skeleton images on a black background, half normal and half cheating, with two cheating types, peeping (forward, to the side and backward turning) and passing notes (to the left and right). With an 8:2 split of this dataset, test-set accuracy was 99.58%. On a simulated video of 1 minute 28 seconds at 25 frames per second with nine examinees, 41 key frames were selected, and they report accuracy of 96.2%, recall of 100%, precision of 26.3%, a false detection rate of 3.8% and an F1 score of 0.42. The authors note that the dataset is small and its behaviours fairly limited, so precision needs to improve. A keyword search of the three examination papers found no use of post-hoc explanation methods such as SHAP or LIME and no retrieval of candidate segments from long recordings.

DISCUSSION

What the evidence says about faces

The face-based studies show both promise and fragility. Kim et al. (2008) report accuracy of up to 100% on the Honda/UCSD benchmark, and Sugiharto et al. (2025) report 96.58% face recognition accuracy on a simple office video. The same pipeline fell to 29.67% on a more complex video with distance, sideways views and obstruction, and its authors link part of the fall to a face database that held only frontal views. Gorodnichy (2005) shows that faces of about 12 pixels between the eyes can still be used when evidence is accumulated over frames, but the approach depends on continuous video. Taken together, these results fit the experience that motivated the narrowing: face recognition depends strongly on footage that is not consistent in surveillance settings. Privacy is a separate concern. Because recognising a face needs identifiable imagery, Rashvand et al. (2025) argue that pose data keep behavioural information while anonymising identities.

Why pose looked more promising, and how far the evidence goes

Pose offers a representation that does not need faces. The evidence for suspicious activity detection built on it is, however, thin. In retail footage, the one pose-based benchmark among the included studies reached a best AUC-ROC of 67.46%, which leaves considerable room for improvement, with normal activities such as using a phone being flagged. The other behaviour-focused study outside the examination setting uses a 3D convolutional network on grayscale video and reports 75% accuracy on a small dataset (Martínez-Mascorro et al., 2020).

The examination studies bear directly on the use case that motivated the narrowing. They show that automated analysis of examination video has been tried with motion-based methods (Ding et al., 2016; Cai et al., 2016) and, in one case, with pose (Wan et al., 2021). Three features stand out. First, all three built their own data, because public examination data are not available and, as Wan et al. note, such video involves candidates' privacy. That limits comparison between studies and makes the reported figures hard to generalise. Second, the figures come from small and different test sets. Wan et al. report 99.58% accuracy on a held-out part of their own skeleton dataset, but on a simulated video the precision was 26.3% with recall of 100%, which means that many flagged candidates were not cheating. Third, the motion-based methods depend on movement: Cai et al. attribute a missed clip to a student who turned and held still. A method that classifies posture could in principle respond to a held posture, which is an inference from the design of Wan et al. and not something any examined study tests. Taken together, this supports pose as a suitable input for examination halls, and it also shows that the support comes from early, small-scale evidence.

Answers to the research questions

RQ1. Trackers can keep identities in crowded scenes (Benfold & Reid, 2011), and a pose pipeline can attach poses to tracked individuals (Rashvand et al., 2025). The examination studies handle several examinees with a multi-person pose estimator or by dividing the frame into regions, without tracking individuals (Wan et al., 2021; Cai et al., 2016). The sources examined do not report how pose-based action classifiers behave when several people interact or move in similar ways, nor how identity switches corrupt the pose sequences passed to a classifier. Sugiharto et al. (2025) show that overlapping people in one detection crop is a practical problem. **RQ2.** Generic pose modelling and the fusion of evidence over time can reduce the need for stored views of each angle (Kim et al., 2008), and a frontal-only reference set is a recognised weakness (Sugiharto et al., 2025). Wan et al. (2021) added normal behaviour from different angles to their pose dataset, and Cai et al. (2016) leave more camera angles to future work. The face findings concern faces, and no examined source tests pose-based classification on camera angles held out from training. **RQ3.** Three directions emerge: aggregating evidence over short windows (Gorodnichy, 2005), avoiding expensive work on every frame (Benfold & Reid, 2011), and testing at reduced resolutions (Martínez-Mascorro et al., 2020). Reference points exist for speed: about 4 frames per second for a multi-model face pipeline on a consumer-grade CPU (Sugiharto et al., 2025), about 200 frames per second for a motion-based examination method in MATLAB (Ding et al., 2016), and 10.65 frames per second for the pose-based examination pipeline when only key frames are processed (Wan et al., 2021). No examined source reports a systematic test of a pose-based suspicious action classifier across several input resolutions and camera angles, so both should be treated as experimental variables.

The research gap

The gap is stated here in four parts, to keep what the authors say apart from what this synthesis infers. Gaps acknowledged by original authors: Sugiharto et al. (2025) call for further work on pipeline optimisation, known faces and camera positioning; Ji et al. (2013) note their dependence on many labelled samples; Wan et al. (2021) say their dataset is small and precision needs to improve; Ding et al. (2016) say the false alarm rate needs further study; and Cai et al. (2016) name more camera angles as future work. Limitations observed within individual studies: a frontal-only

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

face database, false positives on ordinary activities such as using a phone, small or self-collected training sets, and motion-based methods that miss behaviour in which the person holds still. Gaps identified by synthesis across the examined sources: none combines, in one evaluated system, a model that separates suspicious from similar non-suspicious actions on pose sequences, tests under several people, different viewing angles and reduced resolution, explanations of why a sequence was flagged, and retrieval of candidate events from long recordings. Questions left open by insufficient evidence: whether pose-based detection holds up in examination halls beyond simulated data, how its precision can be raised, and how identity switches in tracking affect it.

Limitations of this review

Several limitations qualify these findings. The database search used two sources and English-language publications, and its strings and criteria were written for face and pose recognition, so work specific to pose-based suspicious activity may have been missed. The included set is small and contains three surveys. A quality appraisal is not documented, and some procedural details, including search dates and the number of screeners, are not available. For Poppe (2010), Brunetti et al. (2018) and Hodgetts et al. (2017), only the publisher's preview pages could be examined, so statements about them are limited to the abstract and the visible text. Six additional sources were added outside the search and were not screened against the criteria; the three examination papers came from one phrase searched in one database, so other examination studies may exist. The examination studies rely on small, self-collected datasets with different metrics. Absence claims in this review mean only that the point was not found in the sources examined. Finally, the reported figures come from different datasets and metrics and are not directly comparable, and Sugiharto et al. (2025) give two slightly different values for face recognition accuracy on the simple video (96.56% and 96.58%).

CONCLUSION

This review started with face and pose recognition in video and arrived at pose-based detection of suspicious activity. Face-based pipelines worked well on simple footage but lost accuracy on complex footage and depend on identifiable imagery. Pose is a privacy-conscious alternative. Early evidence comes from a real retail benchmark, from a 3D convolutional approach to pre-crime behaviour and, for examination rooms, from a pose-based pipeline built on a small simulated dataset, alongside motion-based examination studies that also rely on self-collected data. The evidence is thin and does not yet cover crowds, viewing angle, resolution, explanation and retrieval together. Future work should build and evaluate a pose-based system in which people are detected and tracked, pose sequences are classified by a temporal model, flags are explained, and candidate events are retrieved from long footage. Its evaluation should report results separately for one person and several people, for seen and unseen viewpoints, and for several input resolutions, with identity switches counted, processing speed measured on modest hardware, and false-positive behaviour made visible, since low precision was the main weakness reported in the examination setting.

ACKNOWLEDGMENT

Special thanks to the course lecturers for guidance on problem formulation, evaluation standards and responsible computer-vision practice.

FUNDING

This research received no external funding.

CONFLICT OF INTEREST

No known conflict, no involvement of Uganda Civil Aviation Authority in funding, data, or direction, no known ties to the journal. There could be potential benefit is the system works on the author’s side.

ETHICS

Currently looking into creating a dataset of examination hall videos, and making sure the participants consent to me including their videos in the dataset.

PLAGIARISM STATEMENT

According to Duplichecker.com, 2% plaigarism.

DECLARATION OF AI USE

Generative AI tools were used as study aids, following the university's ethical AI usage guidance. Claude (Anthropic) was used to restructure and redraft this manuscript from the author's earlier literature review, to check numbers and attributions against the source papers, and to format the text to the journal template. Earlier project work used Claude and ChatGPT for brainstorming, structuring and drafting support, and Google NotebookLM for organising literature already selected by the author. The author reviewed, edited and fact-checked the text against the original sources and takes responsibility for the final submission.

REFERENCES

1) Benfold, B., & Reid, I. (2011). Stable multi-target tracking in real-time surveillance video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
2) Brunetti, A., Buongiorno, D., Trotta, G. F., & Bevilacqua, V. (2018). Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing, 300*, 17–33. https://doi.org/10.1016/j.neucom.2018.01.092
3) Cai, X., Hu, F., & Ding, L. (2016). Detecting abnormal behavior in examination surveillance video with 3D convolutional neural networks. In *Proceedings of the 6th International Conference on Digital Home (ICDH)* (pp. 20–24). IEEE. https://doi.org/10.1109/ICDH.2016.11

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

4) Ding, M., Zhao, J., & Hu, F. (2016). Abnormal behavior analysis based on examination surveillance video. In *Proceedings of the 9th International Symposium on Computational Intelligence and Design (ISCID)* (pp. 131–134). IEEE. https://doi.org/10.1109/ISCID.2016.37
5) Gorodnichy, D. O. (2005). Associative neural networks as means for low-resolution video-based recognition. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*. Montreal, Canada.
6) Hodgetts, H. M., Vachon, F., Chamberland, C., & Tremblay, S. (2017). See no evil: Cognitive challenges of security surveillance and monitoring. *Journal of Applied Research in Memory and Cognition*. https://doi.org/10.1016/j.jarmac.2017.05.001
7) Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 221–231. https://doi.org/10.1109/TPAMI.2012.59
8) Kim, M., Kumar, S., Pavlovic, V., & Rowley, H. (2008). Face tracking and recognition with visual constraints in real-world videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
9) Martínez-Mascorro, G. A., Abreu-Pederzini, J. R., Ortiz-Bayliss, J. C., & Terashima-Marín, H. (2020). *Suspicious behavior detection on shoplifting cases for crime prevention by using 3D convolutional neural networks* (arXiv:2005.02142). arXiv. https://arxiv.org/abs/2005.02142
10) Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2022). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(7), 3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968
11) Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing, 28*(6), 976–990. https://doi.org/10.1016/j.imavis.2009.11.014
12) Rashvand, N., Alinezhad Noghre, G., Danesh Pazho, A., Yao, S., & Tabkhi, H. (2025). Exploring pose-based anomaly detection for retail security: A real-world shoplifting dataset and benchmark. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops* (pp. 1123–1131). https://arxiv.org/abs/2501.06591
13) Sugiharto, N. D., Elbert, D., & Edbert, I. S. (2025). Employee monitoring system with computer vision: Face and activity recognition. *Procedia Computer Science, 269*, 362–371.
14) Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience, 2018*, Article 7068349. https://doi.org/10.1155/2018/7068349
15) Wan, Z., Li, X., Xia, B., & Luo, Z. (2021). Recognition of cheating behavior in examination room based on deep learning. In *Proceedings of the 2021 International Conference on Computer Engineering and Application (ICCEA)* (pp. 204–208). IEEE. https://doi.org/10.1109/ICCEA53728.2021.00048

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Corresponding Author: Cole Rodney Migisha