

Article

Deep-Learning-Based Action and Trajectory Analysis for Museum Security Videos

Christian Di Maio ^{1,2,†} , Giacomo Nunziati ^{1,3,*,†}  and Alessandro Mecocci ¹ 

¹ Department of Information Engineering, University of Siena, 53100 Siena, Italy; c.dimaio1@student.unisi.it or christian.dimaio@phd.unipi.it (C.D.M.); mecocci@unisi.it (A.M.)

² Department of Computer Science, University of Pisa, 56127 Pisa, Italy

³ Department of Information Engineering, University of Florence, 50139 Firenze, Italy

* Correspondence: giacomo.nunziati@student.unisi.it or giacomo.nunziati@unifi.it

† These authors contributed equally to this work

Abstract: Recent advancements in deep learning and video analysis, combined with the efficiency of contemporary computational resources, have catalyzed the development of advanced real-time computational systems, significantly impacting various fields. This paper introduces a cutting-edge video analysis framework that was specifically designed to bolster security in museum environments. We elaborate on the proposed framework, which was evaluated and integrated into a real-time video analysis pipeline. Our research primarily focused on two innovative approaches: action recognition for identifying potential threats at the individual level and trajectory extraction for monitoring museum visitor movements, serving the dual purposes of security and visitor flow analysis. These approaches leverage a synergistic blend of deep learning models, particularly CNNs, and traditional computer vision techniques. Our experimental findings affirmed the high efficacy of our action recognition model in accurately distinguishing between normal and suspicious behaviors within video feeds. Moreover, our trajectory extraction method demonstrated commendable precision in tracking and analyzing visitor movements. The integration of deep learning techniques not only enhances the capability for automatic detection of malevolent actions but also establishes the trajectory extraction process as a robust and adaptable tool for various analytical endeavors beyond mere security applications.



Citation: Di Maio, C.; Nunziati, G.; Mecocci, A. Deep-Learning-Based Action and Trajectory Analysis for Museum Security Videos. *Electronics* **2024**, *13*, 1194. <https://doi.org/10.3390/electronics13071194>

Academic Editor: Gemma Piella

Received: 16 February 2024

Revised: 17 March 2024

Accepted: 21 March 2024

Published: 25 March 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: video; museum; action; trajectory; deep learning

1. Introduction

Security in museums stands as a paramount concern, transcending the mere safeguarding of physical assets to preserving invaluable cultural heritage. The vulnerability of museums to various threats, ranging from theft to vandalism and even terrorism, underscores the necessity for robust security measures. Over the years, traditional security methods, which rely heavily on manual surveillance and intermittent patrols, have shown themselves to be insufficient in adequately addressing the evolving nature of security challenges within museum spaces [1]. These methods often suffer from limitations such as human error, limited coverage, and delayed response times, leaving museum collections susceptible to potential harm.

Moreover, the increasing value and visibility of cultural artifacts have made museums attractive targets for illicit activities, exacerbating the urgency for enhanced security protocols. The repercussions of security breaches extend far beyond mere financial losses [2], encompassing irreparable damage to irreplaceable artifacts and the erosion of public trust in the institution's ability to preserve cultural treasures. In light of these challenges, there exists a pressing need for innovative security solutions that can effectively mitigate risks while minimizing disruption to the visitor experience.

Recognizing the inadequacy of traditional security measures, museums have increasingly turned to technological advancements to bolster their security infrastructure. Among the most common innovations is the integration of video surveillance systems, which offer comprehensive monitoring capabilities across museum premises. Video surveillance not only provides a means of continuous monitoring but can also serve as a valuable tool for post-incident analysis and forensic investigations. The proliferation of cameras across museum premises has significantly increased the volume of video data generated by surveillance systems. This abundance of footage presents a formidable challenge that requires the development of intelligent video analysis techniques capable of processing and extracting actionable insights in real time.

In recent years, significant strides have been made in the field of video analysis, which were largely driven by the rapid advancement of deep learning techniques [3,4]. Deep learning, which is a subset of artificial intelligence inspired by the structure and function of the human brain, has revolutionized the field of computer vision by enabling machines to learn complex patterns and features directly from raw data [5–7]. This paradigm shift has led to remarkable progress in tasks such as image classification, object detection, and semantic segmentation, laying the groundwork for innovative applications in surveillance and security.

One of the most notable advancements enabled by deep learning is the development of convolutional neural networks (CNNs) [8,9], which were shown to be highly effective in analyzing visual data, including video streams [10].

Furthermore, deep learning approaches have shown great promise in addressing some of the longstanding challenges in video analysis, such as robustness to variations in lighting conditions, camera viewpoints, and object occlusions.

Through techniques such as transfer learning [11,12] and data augmentation [13], researchers have been able to leverage pre-trained deep learning models and synthetic data to improve the generalization capabilities of video analysis systems, making them more adaptable to real-world scenarios. Additionally, the availability of powerful computational resources, such as graphics processing units (GPUs) and cloud computing platforms, has significantly accelerated the training and deployment of deep learning models for real-time video analysis applications [14–18].

Real-time analysis plays a pivotal role in modern museum security, enabling the swift detection and response to potential threats. The integration of automated surveillance systems capable of analyzing video feeds in real-time has emerged as a critical component in bolstering museum security. However, given the computational constraints inherent in real-time applications, the efficiency of algorithms becomes paramount. Lightweight algorithms, which prioritize computational speed without sacrificing accuracy, are particularly crucial in this context. Moreover, it is essential to utilize algorithms and deep learning models that are proficient in utilizing the capabilities of modern computational resources, like GPUs.

In this work, we concentrated on two crucial aspects of video analysis within museum security: action recognition and trajectory analysis. Action recognition serves as a pivotal tool for anomaly detection, with the aim to identify malevolent actions indicative of potential threats. Examples of such actions include loitering near sensitive exhibits, suspicious movements around restricted areas, or attempts at unauthorized access. By integrating action recognition into security protocols, the burden on security personnel can be alleviated, allowing for more efficient monitoring and response procedures. The detection of anomalies in museum environments remains an underexplored area in the existing literature. By addressing this gap, our study makes significant contributions to the field. First, we present the use of action classification techniques for anomaly detection in museum surveillance videos. To the best of our knowledge, our study represents the first attempt to apply action recognition techniques for anomaly detection in a museum context. The main objective of this part of the study was to evaluate whether the action recognition framework can be used to effectively model the anomaly detection problem. To this end, we selected two deep neural networks with the characteristics of allowing a highly parallel execution,

thus being lightweight, and validated their anomaly detection capabilities. We also investigated the impact of various factors such as dataset augmentation; balancing; and the inclusion of staged examples, both anomalous and normal, on the detection performance. Finally, we propose a priority-based approach to deal with classification ambiguity in crowded scenes, providing a nuanced solution to address complex surveillance scenarios.

Anomaly detection through action recognition in museums presents unique challenges that are distinct from conventional action recognition tasks. Defining actions within the context of museum settings is challenging due to the diverse range of visitor behaviors and interactions with exhibits. Furthermore, contextual factors, such as crowded spaces and privacy considerations, further complicate dataset collection and model training. In this study, we addressed these challenges by rigorously testing and comparing two deep learning methods for action recognition, focusing particularly on their effectiveness in anomaly detection. Our approach leveraged 3D CNNs that were tailored to classify actions within museum surveillance footage. We also explored the utility of data-augmentation techniques and the inclusion of staged and out-of-context videos to enhance the model robustness.

Trajectory analysis offers dual utility for both security and behavioral analysis purposes within museum environments. From a security standpoint, trajectory analysis enables the identification of anomalous behaviors, such as loitering in restricted areas or erratic movements, serving as an additional layer of security threat detection. Additionally, trajectory analysis aids in identifying high-risk areas where valuable artifacts may be vulnerable to damage or theft, facilitating targeted security measures. For behavioral analysis, trajectory data support crowd management initiatives, visitor engagement analysis, and exhibit effectiveness evaluations. Our research addresses significant gaps in the existing literature on human trajectory estimation by camera videos. Previous studies ignored the use of pose extraction techniques, which allow for the estimation of the ground position of the tracked subject based on all the detected keypoints and their measured distance from the ground plane. In addition, the integration of trajectory analysis alongside video data for anomaly detection has been little explored. In response to these gaps, our contribution to the field is multifaceted. First, we present a novel approach that utilizes deep-learning-based pose extraction, ground position estimation, and homographic projections to obtain the trajectories. Second, we proposed a new distance metric, which we then used for trajectory clustering, allowing for more effective grouping of trajectory data for anomaly identification. Furthermore, we validated the effectiveness of our trajectory reconstruction technique and confirmed its reliability in real-world scenarios. Finally, we demonstrated the utility of the extracted trajectories, performed a clustering for an illustrative application, and highlighted the versatility and practicality of our proposed methodology in the context of museum surveillance.

Our paper presents a novel approach to trajectory extraction, integrating deep learning and pose estimation techniques to accurately capture visitor trajectories within museum spaces. We addressed perspective deformation introduced by camera angles, ensuring trajectory data are represented in the ground plane for an accurate analysis. Additionally, we evaluated the effectiveness of our trajectory extraction method, demonstrating its utility for the further analysis and clustering of visitor behaviors.

Furthermore, we integrated the described techniques into a comprehensive framework for real-time video analysis, enabling museums to proactively monitor security threats and visitor behaviors. We assessed the real-time performance of our framework, demonstrating its efficiency in processing and analyzing surveillance footage. Encouragingly, our results indicate promising prospects for anomaly detection through action recognition, as our methods were found to be effective in classifying actions and distinguishing between normal and anomalous behaviors. Similarly, our trajectory extraction method exhibited minimal approximation errors, indicating its suitability for extracting trajectories for subsequent analysis and behavioral categorization.

This research was conducted in partnership with the Museo dell'Opera della Metropolitana di Siena (OPA), which granted access to a dataset of footage recorded by surveillance

cameras installed in various rooms of the museum. This partnership offered full support to the research activities, enhancing the depth and breadth of the investigation.

The structure of this paper is as follows: In Section 1, we provide a comprehensive literature review of the field, discussing relevant works and delineating areas of research focus. Section 2 provides a detailed description of the techniques presented in this study, including the methodologies employed and the experiments conducted to evaluate their efficacy. In Sections 3 and 4, we describe the experimental setup and we present and discuss the results obtained from the experiments, detailing the performance of the proposed techniques. Finally, Section 5 summarizes the main highlights of this work, outlines open problems, and suggests future research directions.

Related Works

Video surveillance involves various tasks, from capturing and processing data to analyzing and applying it. This review explores the different facets of video surveillance and summarizes the main advances and methods in the field. In video surveillance, scene control is one of the main difficulties. Historically, this problem has been addressed using active sensors, such as sonar, laser, and radar sensors [19–22]. Thanks to technological progress, nowadays, video surveillance mostly relies on IP cameras, which enable direct monitoring of the relevant scenes. When unusual events occur, the operators who watch the live stream alert the authorities to deal with possible dangers or problems. To improve museum security, operators require the aid of automated or semi-automated surveillance systems capable of identifying anomalous events. Some solutions attempt to reach a trade-off between the system's ability to detect intruders, which must meet a minimum required level, and the size of the sensor, which should be minimized. Some examples can be found in [23], where a low-cost security device is proposed and visible light is employed for communication, and [24], where the authors described the use of a network of motion sensors placed in crucial positions. Ref. [25] illustrates how to track people and other objects in museum halls through the use of a stereo camera, which is also used to estimate a 3D model of such rooms, with the goal of notifying the human operator in charge of handling anomalous events. Refs. [26,27] offer extensive examinations of diverse methods for video anomaly detection. The authors compared different versions of the problem and methods for anomaly detection. They noted that some anomalies can be classified based on the context, such as human–artwork interaction in museums. They review the conventional techniques for anomaly detection, which include distance-based, probabilistic, and reconstruction-based methods, and highlight the recent trend of using deep learning approaches.

Various techniques have been proposed for anomaly detection using deep-learning-based methods to address the related challenges [28–31]. The recognition of actions is approached through various methods, each utilizing spatial and temporal information within videos in different ways. Models such as C3D [10,32] are based on the extraction of deep spatio-temporal features using architectures that employ 3D convolutions. These convolutions extend the standard convolutional filters used in CNNs to capture temporal information in addition to spatial features. To avoid the need to train the 3D CNNs from scratch, several techniques have been developed. The most relevant is I3D (Inflated 3D CNN) [33], where the authors proposed building the 3D convolutional filters starting from the weights of the convolutional layers of neural networks trained on image datasets. Some examples of architectures developed around the I3D idea are the Interaction Reduced Channel Separated Network (IRCSN) [34] and the Temporal Pyramid Network (TPN) [35]. In this work, we explored the capabilities and limitations of the IRCSN and TPN architectures in the context of museums.

Our attention was directed toward the task of anomaly detection in surveillance systems. As highlighted by Berroukham et al. [36], the task of detecting anomalies can be effectively approached using deep learning techniques, which are categorized into

four main strategies: reconstruction error, future frame prediction, classification, and scoring methods.

Reconstruction error techniques use a reconstruction error function to measure how likely a sample is to be an anomaly. The function's output score shows how much a sample differs from the training distribution. Usually, a lower score means a normal sample, as it is closer to the training data, while a higher score means an abnormal sample, as it is farther from the expected data distribution. Hasan et al. [37] proposed a minimally supervised approach to learning normal patterns using auto-encoders (AEs). The authors trained an AE on the spatio-temporal features of the samples and then used a fully convolutional AE to measure the reconstruction error. Sabokrou et al. [38] presented a method for detecting and localizing anomalies in videos using two types of cubic patches. One type uses an AE to reconstruct the input video patch, while the other type uses sparse representation to encode the input video patch.

Future frame prediction is a technique that tries to forecast the next image in a video sequence given the previous ones. It is based on the assumption that normal events follow a certain pattern or logic over time, which can be learned from the training data, while this behavior does not occur for an abnormal video. Refs. [39,40] show that a network can anticipate normal activities matching the ones in the training videos.

Classification proposed in work like that from Mohammad et al. [41] treats anomaly detection as a classification task. They used fully convolutional neural networks (FCNNs) and temporal information to analyze the deep layer outputs and find and locate anomalies in videos.

The main idea of the scoring-based method is to generate an anomaly score that may be used to determine whether a video segment or frame is abnormal. Sultani et al. [42] proposed an approach to learning anomalies by utilizing both normal and abnormal videos, where they suggested using weakly labeled training videos to learn anomalies. Another method was proposed by Xu et al. [43], which involves an unsupervised learning approach to automatically learn feature representations. A brand new double fusion architecture was created to take advantage of the complementing information contained in both the appearance and movement patterns, combining typical early fusion and late fusion advantages.

One of the most widespread action classification datasets is Kinetics-400 [44]. This is a collection of 400 classes (actions), each with 400 or more clips that last about 10 s on average. The actions include solo and group performances, some involving objects. The clips are extracted from YouTube videos and labeled with the corresponding action, duration, and the number of performers.

Regarding human trajectory estimation, there are different types of approaches. Ref. [45] faced the problem of trajectory prediction as a sequence generation task, where the goal was to predict the future trajectory of people based on their past positions. The authors proposed a long short-term memory (LSTM) model that can learn general human movement and predict their future trajectories. Zhao et al. [46] studied the task of trajectory estimation for assessing risks on a construction site, where they split the problem into the detection and after tracking. They used a YOLO (you only look once) model, namely, YOLOv3 [47], to detect a person in danger, and a tracking algorithm (in this case, a Kalman [48] filter and Hungarian algorithm [49]) to achieve pedestrian tracking. In [50], the authors present an unsupervised method for the representation of the activity taking place in a scene. This method is based on the detection of salient points in space and time that correspond to regions with a significant amount of activity. The method then tracks these points in time using a state estimation approach to reach a representation based on short trajectories.

Clustering is a technique that involves grouping data points based on their similarity or dissimilarity. It is one of the main unsupervised learning tasks, where the data have no labels or predefined categories. Clustering can be used for exploratory data analysis, feature engineering, pattern discovery, anomaly detection, and other similar applications.

Many clustering algorithms have been proposed for different types of data and objectives. Some of the well-known clustering algorithms are as follows:

K-means clustering, which is a centroid-based algorithm that partitions the data into K clusters, where each cluster is represented by its mean or center. The algorithm iterates until the cluster assignments do not change or a maximum number of iterations is reached. K-means clustering is simple and fast, but it requires specifying the number of clusters and it may not work well with non-spherical or noisy data [51].

DBSCAN, which is a density-based algorithm that identifies clusters as regions of high density separated by regions of low density. The algorithm does not need the number of clusters as an input, but it requires two parameters: the minimum number of points in a cluster and the maximum distance between two points to be considered neighbors. DBSCAN can handle arbitrarily shaped clusters and outliers, but it may not work well with a varying density or high-dimensional data [52].

Spectral clustering is a graph-based algorithm that uses the eigenvalues and eigenvectors of a similarity matrix to partition the data into clusters. The algorithm can capture complex structures and non-linear relationships, but it also requires the number of clusters as an input and it may be computationally expensive for large datasets [53].

Hierarchical clustering [54] consists of a family of algorithms that build a hierarchy of clusters by either merging smaller clusters into larger ones (agglomerative) or splitting larger clusters into smaller ones (divisive). The algorithm produces a dendrogram that shows the nested structure of the clusters and allows for choosing the optimal level of granularity. Hierarchical clustering does not need the number of clusters as an input, but it may be sensitive to the choice of distance measure and linkage criterion.

2. Materials and Methods

This section outlines the methodologies implemented in our study, organized into three primary segments. Initially, Section 2.1 discusses the technique utilized for the automatic identification of anomalies. Following this, Section 2.2 elaborates on the method for extracting trajectories and provides a practical application scenario. Lastly, Section 2.3 explains the integration of these two approaches and their application to video stream analysis.

2.1. Video Anomaly Detection

In this subsection, we elaborate on the video anomaly detection approach implemented in our study.

Method Description

Our approach to anomaly detection in surveillance videos leveraged a supervised, scene-based action recognition methodology. By integrating supervision into the traditionally unsupervised realm of anomaly detection, we gained vital insights into distinguishing between anomalous and normal behaviors. This integration allowed for a more nuanced understanding of the dynamics within surveillance footage, significantly enhancing our ability to identify deviations from expected patterns.

Our methodology consisted of segmenting the videos into small clips and assigning them with the class representing the action contained in the clip. This labeling process posed challenges, particularly when multiple individuals simultaneously performed different actions in the camera frame. To address this complexity, we introduced a priority policy, which favored actions of higher significance—especially anomalies—to ensure precise classification. Our action repertoire covered typical behaviors in museum settings, such as walking, taking pictures, and sitting. We also identified critical behaviors that are indicative of potential security breaches, including touching artwork and crossing restricted areas. To capture atypical interactions with artworks not classified under standard categories, we established an encompassing category: abnormal interaction with artwork. This methodical classification aided in the accurate identification and analysis of both regular and anomalous behaviors.

To perform action recognition, our approach harnessed sophisticated deep learning frameworks based on the 3D CNN model [10,32], which is a specialized extension of CNNs designed for effectively analyzing 3D data structures. In this methodology, videos were viewed as sequences of 2D frames and aggregated to create a 3D tensor. This tensor became the input for our model, to which we applied 3D convolutional filters. These filters adeptly captured both spatial and temporal dimensions of the video data, enabling a comprehensive analysis of actions within the sequences through the intricate patterns and dynamics they unveiled. The essence of 3D CNNs lies in their capability to capture both spatial and temporal information inherent in video data. By extending the traditional 2D convolutional operations into the temporal dimension, these networks excel at extracting spatio-temporal features crucial for action-recognition tasks.

We employed the I3D method [33] to derive our 3D CNN architectures. I3D represents a technique for obtaining a 3D CNN without the need for extensive training, capitalizing on the effectiveness of established 2D CNN architectures. This approach involves inflating the 2D convolutional filters into their 3D counterparts. The process of inflation entails replicating the filters across the temporal dimension and subsequently normalizing them by dividing the resulting 3D filter by the size of the temporal dimension. Through this process, I3D effectively extends the capabilities of conventional CNNs to directly process video. This approach was shown to be highly effective in initializing the weights of a 3D CNN, leveraging the strong spatial features extracted by established CNNs for images, such as ResNet [55] and VGG [56].

In our investigation, we focused on two prominent architectures inspired by the I3D concept: the Interaction Reduced Channel Separated Network (IRCSN) [34] and the Temporal Pyramid Network (TPN) [35].

The IRCSN model exploits the concept of channel separation, which is typical of some convolutional architectures [57–59], and applies it to 3D CNNs. Channel separation in standard CNNs involves approximating the typical convolutional filters, which span across the spatial dimensions and consider all the channels, with a sequence of $1 \times 1 \times C$ point-wise filters and $N \times N \times 1$ filters. The first focuses exclusively on the channel dimension, neglecting spatial interactions, whereas the second concentrates only on the spatial dimensions, omitting channel interactions. This strategy of separation diminishes the interaction between various elements of the feature maps fed into the convolutional layer. Nonetheless, this method significantly enhances the computational efficiency, facilitating the development of architectures that are both deeper and more efficient. This balance between computational economy and architectural depth allows for advanced model designs that can process complex data more effectively without compromising on performance. This enhancement results in an overall performance gain, as the model can focus more effectively on relevant spatial and temporal cues that are essential for accurate action recognition.

Conversely, the TPN model capitalizes on the intrinsic tempo and rhythm inherent in action sequences. The TPN achieves this by extracting features at multiple levels of depth within a ResNet backbone. This multi-level feature extraction is advantageous because it allows the model to simultaneously capture both low-level details and high-level spatio-temporal patterns. The extracted features are then subject to a sequence of handcrafted transformations, including spatial and temporal modulations. By incorporating a temporal pyramid structure, the TPN can effectively capture temporal dependencies across different time scales, enhancing its ability to discern subtle variations in action sequences.

Both architectures leverage the I3D technique for initializing the weights of their convolutional filters. Notably, both architectures come in various versions with different bottleneck ResNet backbones. Furthermore, to enhance their performance, both architectures underwent fine-tuning on the Kinetics-400 dataset. Specifically, in our work, we employed the IRCSN with a ResNet-152 backbone and the TPN with a ResNet-50 backbone, both incorporating bottleneck structures.

2.2. Trajectory Extraction and Analysis

We present a methodology for extracting trajectories of museum visitors from video footage, along with the approach employed to analyze these trajectories. We represent the trajectory τ of a detected object as a set of 4-tuples: $\tau = \{(t_i, x_i, y_i, z_i), i = 1, 2, \dots, n\}$, where t_i represents the frame number and $x_i, y_i,$ and z_i represent the real-world coordinates assigned to the object at that frame. If the trajectory belongs to one of the axial planes, one of the coordinates is always equal to 0. In our work, we always project the trajectory into the xy plane (the ground plane). With this assumption, the z component of the position can be discarded and we consider $\tau = \{(t_i, x_i, y_i), i = 1, 2, \dots, n\}$.

Method Description

We decided to incorporate homographic transformations into our approach, enabling the extraction of trajectories within the world coordinate system rather than the image coordinate system. As stated, the trajectories were also projected onto the ground plane. We chose to use homographies for several reasons. First, they normalized the extracted trajectories, making them independent of the camera viewpoint. Such a normalization was useful when analyzing trajectories captured from different camera angles. Furthermore, this transformation ensured that distances within the trajectories are normalized, allowing for a more fair evaluation of the accuracy of the extraction method. Additionally, by projecting the trajectories onto the ground plane, we obtained metric coordinates that overcome the limitations of pixel coordinates and provide meaningful spatial information. A homographic transformation is a mathematical transformation that maps points from one plane to another while preserving straight lines [60]. In the context of trajectory extraction, it allows for switching between image, camera, and world reference systems. This transformation is achieved through the application of two matrices: the extrinsic matrix and the intrinsic matrix. The extrinsic matrix takes into account the camera pose, which transforms world coordinates into camera coordinates using the pinhole camera model and corresponding coordinates. The intrinsic matrix represents the internal parameters of the camera, such as the focal length and distortion coefficients. Applying these matrices performs the transformation from the camera coordinates to image coordinates [60]. Homographic transformations operate in projective spaces, and their application is performed through the use of homogeneous coordinates [60]. More in detail, given two planes in a projective space, there always exists a unique correspondence between their homologous points. This relationship is succinctly described by a homographic transformation, which establishes a bijective relationship between correspondent points on the two planes, enabling the mapping of points from one plane to the other.

To accurately determine the necessary transformation matrices, we performed a calibration of the camera. Our methodology encompassed two primary phases in the calibration process. Initially, we utilized the camera's intrinsic parameters to assemble the intrinsic matrix. Subsequently, we engaged in a detailed assessment of the real-world coordinates and their corresponding image coordinates, focusing on a specifically chosen set of points. This two-step approach ensured a comprehensive calibration, enabling precise transformations. For simplicity, our research narrowed its scope to a single camera and viewpoint. Specifically, we employed the Carrier TVGP-M01-0201-PTZ-G model (Carrier, Notting Hill, VIC, Australia), which is characterized by a 2.8 mm focal length and sensor dimensions of 7.2×5.4 mm. This camera is capable of capturing images in full HD (FHD) resolution, achieving a crisp 1920×1080 pixel clarity. This focused approach allowed for a detailed examination of the camera's performance under controlled conditions. Figure 1 shows the homographic transformation applied to the hall of statues in the museum.



Figure 1. (a) shows the original frame; (b) shows the frame distorted with homographic projection; (c) shows the planimetry of the same room. The colored circles provide insight into the transformation of positions and distances between points. (b) was obtained using only the extrinsic projection matrix, which took into consideration the pose (position and orientation) of the camera, and could be used to convert the positions of each of the pixels in (a) to an estimated position in the ground plane. Due to this choice for the reference plane, the pixels that did not represent objects on the ground were mapped to the wrong position, causing a higher distortion, which can be observed on the regions of (b) that represent the statues. Notably, the transformed pixel position was not arranged on a regular grid. For this reason, a grid had to be defined over the new space, and the intermediate pixel values were obtained by interpolation.

By referring to the previous approach of employing a single camera and viewpoint, this study introduced a comprehensive video trajectory extraction pipeline. The proposed extraction pipeline is based on a series of sequential steps that are executed for each frame of the video, ensuring a thorough analysis and processing of the captured footage:

1. Background learning and subtraction to exclude the static portions of the video frame from the analysis;
2. Detection of people in the scene;
3. Tracking and re-identification of detected people to maintain consistent identities across frames;
4. Pose estimations of the detected people to infer the positions of their heads and feet;
5. Estimations of the heights of the detected people, whenever possible, to project their position onto the ground plane, even if the feet are occluded in some frames;
6. Projection of the people's positions onto the ground plane using homographic transformations and exploiting the information about the feet and/or head positions, combined with the estimated height of each individual.

The described pipeline is visually represented in Figure 2.

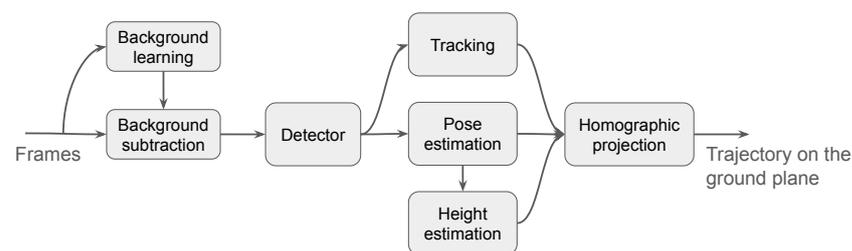


Figure 2. Trajectory extraction pipeline.

Background subtraction was performed using a Gaussian mixture model (GMM) [61,62], which models the background of the scene. Initially, during a warm-up phase at the beginning of video processing, the GMM model was built using the first 15 s of video, allowing it to adapt to the static elements in the scene and capture their composition. Then, once the model was built, it was frozen for the remaining part of the video processing. During the operational phase, each frame of the video was compared with the learned background model, allowing foreground objects to be identified by detecting deviations from the background statistics. In contrast to the conventional GMM approach, which continuously repeats the background estimation process to compensate for day–night changes,

our operational scenario exhibited negligible environmental variability throughout the day, enabling us to fix the background image following a specified time interval. Nevertheless, a monitoring process remained in place to ensure that no significant changes occur. Should substantial deviations be detected, the background learning process was initiated once more, as outlined earlier. Verification involved comparing the frozen static background with regions of the current image that were not occupied by detected targets. This process ensured that the background estimation remained consistent and accurate, providing a reliable foundation for analyzing changes or anomalies within the observed scene.

The detection and pose estimation steps in our pipeline were carried out using the YOLOv8 model developed by Ultralytics [63]. Specifically, we employed the XL version of the model, which is the largest and most powerful variant. Using the default pre-trained checkpoint, namely, yolov8x-pose, the model can simultaneously perform both detection and pose estimation tasks. In the detection phase, the model identifies individuals within the frame by delineating bounding boxes around them. Then, in the pose estimation phase, the model identifies and locates 17 body keypoints according to the COCO Keypoints val2017 [64] labeling scheme, which includes anatomical landmarks, such as the nose, ears, hands, and feet. Notably, the YOLOv8 model reliably handles challenging scenarios, such as partial occlusion or detection of only a subset of keypoints, ensuring robust performance in diverse real-world conditions [65–68].

The tracking and reidentification steps are crucial for maintaining continuity and identity across frames, especially in dynamic environments, such as museums, where multiple people are typically present at the same time. In museums, however, re-identification is even more challenging due to the typically poor lighting conditions used to preserve the artwork exhibited in the galleries. For this purpose, we used BoT-SORT [69]. This advanced multi-object tracking approach employs Kalman filtering [48] to analyze object motion and a sophisticated deep-learning-based appearance feature extractor for matching. This dual strategy allows for the precise association of objects across frames, leveraging both their motion and visual attributes. These attributes enable BoT-SORT to facilitate short-range re-identification, indicating that the re-identification process is particularly effective when the object to be tracked has moved minimally and/or only a short time has elapsed. The process facilitates BoT-SORT in generating tracklets, which depict the uninterrupted movement of an object over time. Additionally, BoT-SORT enhances the track accuracy through a post-processing stage, where the interpolation of points within each tracklet is performed to achieve seamless, continuous tracks.

In the final step of our pipeline, we projected the extracted tracks onto the ground plane using the two homographic transformation matrices obtained through camera calibration. Specifically, for each person, a representative point in the image coordinates was selected. We opted to utilize the midpoint of the segment connecting the two feet, provided they were visible within the frame. However, to accommodate scenarios where the feet may be occluded, our proposed method includes the estimation and tracking of each person's height. This height is estimated using the head keypoint and the same homographic transformations employed for projection. Consequently, even if the feet are obscured, as long as the head keypoint is visible, the person's position can be accurately projected onto the ground plane.

After the validation of the trajectory extraction approach, we moved on to use the extracted trajectories to perform a preliminary analysis of visitor behavior. Using the trajectories obtained from a one-hour recording within the same museum hall used for the previous experiments, we performed hierarchical clustering to identify patterns within the movements of the visitors. Using a bottom-up approach, the extracted trajectories were systematically grouped based on their similarities, which allowed for the identification of distinct behavioral patterns exhibited by visitors throughout the observation period.

2.3. Integration and Video Stream Processing

We integrated the approaches described in Sections 2.1 and 2.2 into a unified dataflow pipeline capable of processing video streams in real time. This pipeline manages all aspects of computation, including video acquisition, preprocessing, action recognition, trajectory extraction, annotation of video with processing results, and the visualization and/or storage of processed frames. In addition, the pipeline is designed to raise events in various forms, allowing for integration with various security management systems. This capability is essential for real-world applications, especially in scenarios where the pipeline is used for the real-time processing of video streams from security cameras.

The integrated system was constructed as a graph dataflow, where individual blocks perform elementary operations and are connected to form a unique processing pipeline. These blocks communicate with each other through a message-passing paradigm, where messages are referred to as packets. Upon their arrival, packets are enqueued in the input queue of the designated block responsible for their processing. These packets are subsequently processed sequentially by their corresponding blocks. To enhance both the efficiency and scalability, the system employs a multi-threaded architecture, enabling blocks to process packets in parallel. This is governed by a priority policy that effectively manages congestion within the system. Blocks experiencing higher levels of congestion are granted increased processing time, facilitating a quicker clearance of their packet queues. This strategic allocation of resources ensures smooth and efficient packet processing across the system.

Additionally, the system is also capable of controlling multiple cameras simultaneously, optimizing the use of existing computational blocks.

The integration of the trajectory extraction pipeline and the anomaly detection system allows for a synergy between the two: the person detector used in trajectory extraction can be exploited to crop the video around the detected people, allowing the action detection system to focus on each individual and detect only their actions. To preserve the important information about the visual context around each person, the bounding boxes are enlarged by 50% in width and 20% in height, before cropping.

Figure 3 graphically represents the entire system, showing the interconnected blocks and their respective dataflow. It also illustrates the message-passing mechanism used in the system, showing how packets are exchanged between interconnected blocks.

Figure 4 shows some examples of execution of the complete pipeline.

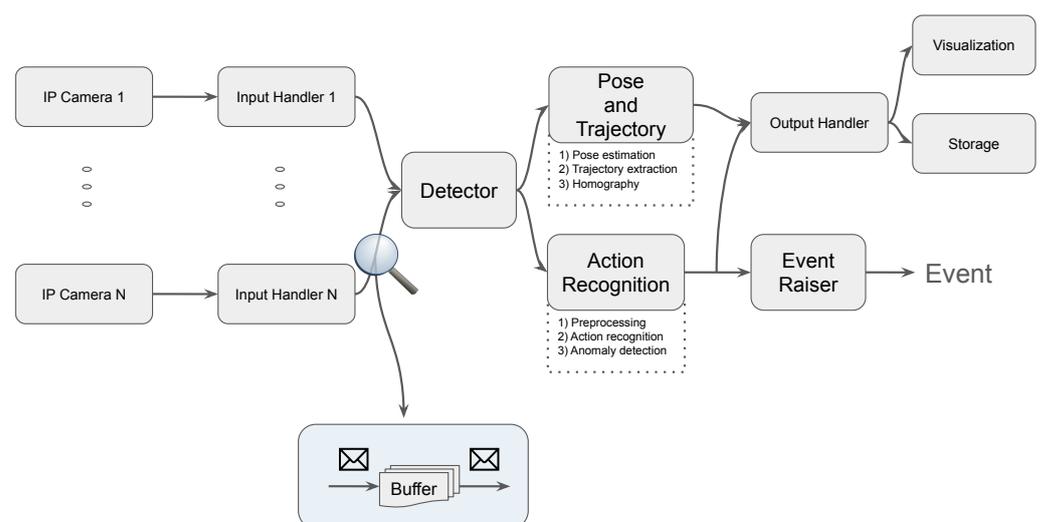
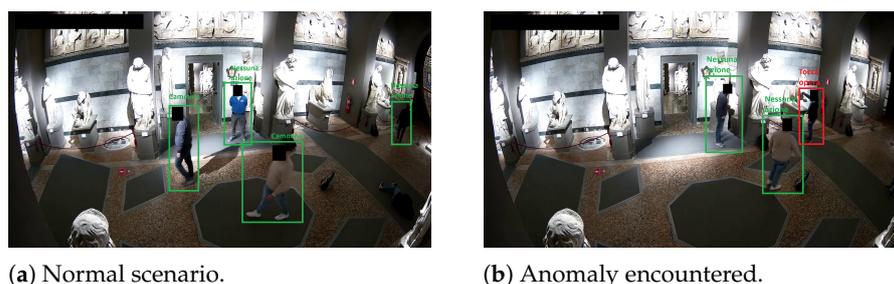


Figure 3. Integrated system pipeline.



(a) Normal scenario.

(b) Anomaly encountered.

Figure 4. This figure shows the system working in the control room of the OPA museum in Siena. Since the control room operators are Italian, the actions of each person are labeled in Italian. In the figure, we have *camminare*, which means walk and *nessuna azione*, which is no action, in green (a), and in red, the detected anomalies, in this case *tocca opera*, which means touch an artwork (b).

3. Experiments

For each of the three main video analysis contributions presented in this paper, we describe the experimental setup.

3.1. Video Anomaly Detection

We begin by detailing the dataset selected for the training and evaluation, highlighting its relevance and the rationale behind its choice. Following this, we discuss our experimental setup, which includes a series of steps to prepare the data for analysis. This preparation involved (a) preprocessing techniques to ensure the data were in a suitable format for processing, (b) augmentation methods to increase the diversity and robustness of our dataset, and (c) optimization strategies to enhance the performance of our anomaly detection model.

3.1.1. Dataset Composition

Anomaly detection in museum environments faces a significant challenge due to the scarcity of supervised datasets. Constructing a large-scale dataset for this purpose is hindered by privacy constraints and limitations on gathering surveillance footage from museums. To address this challenge, we collaborated with the OPA museum to collect a small dataset that comprised both real and staged video clips obtained from various museum rooms and camera viewpoints. We refer to this as the original dataset. We gathered video footage from two distinct rooms: the Sala del Duccio and the Sala delle Statue. In each room, footage was captured from four different perspectives. Cameras were strategically positioned at opposing ends to ensure comprehensive coverage, and each was mounted at varying elevations to optimize the field of view (for security purposes, as the cameras in question are integral to the museum's internal security infrastructure, we are unable to disclose details about them or their precise locations, as well as images captured from them during the visit time of the museum). This dataset encompassed recordings for a total of 277 min, with single-clip durations ranging from 2 to 20 min. The recordings were collected on eight different days in three different months to account for seasonal variations and they were scheduled at diverse times of the day to ensure a comprehensive capture of the variations in pedestrian volume, from the busiest to the quietest moments. Each recording day was assigned with an equal overall recording duration, with a maximum difference between different days of 2 min. Long recordings were subdivided into shorter clips lasting 5 to 10 s each to ensure the presence of at least one action in every clip. In the labeling process for the short video clips, the following rule was employed: if a clip showed only one person, the label reflected their action; for clips with multiple individuals, the label was determined by prioritizing the most significant action observed. Actions such as touching an artwork were prioritized over less significant actions, like walking, given the relatively small set of detectable actions within museum environments. The complete set of actions with their corresponding priorities is outlined in Table 1, while additional details on the dataset composition are presented in Figures 5 and 6.

In the second phase of our experimental work, we collected an additional dataset of staged videos captured in many non-museum settings. The primary goal of this second step was to expand the set of video clips that included anomalous actions and, more broadly, actions of lower frequency. This strategic decision was made to strengthen the diversity and richness of our dataset, thereby increasing the robustness and effectiveness of our anomaly detection algorithms. We refer to the combination of this newly acquired dataset with the original dataset as the extended dataset to denote its expanded scope and enriched content. The videos constituting the extended dataset were collected in a total of seven locations, more precisely, four house rooms and three university rooms. Being controllable recording setups, we implemented different strategies to improve the variability of the recordings to improve the resulting generalization. First, we used a set of five cameras and a total of 21 viewpoints, 3 for each location. Then, we employed a sample of 12 participants of different sexes, ages, and appearances to act in the footage. In addition to ensuring a broad spectrum of clothing, we implemented a specific protocol for staged activities. Each participant was instructed to modify their clothing by incorporating distinct elements, such as hats or hand-carried items for each action. This protocol was established to introduce variability and promote the robustness of our dataset against variations in apparel, thereby facilitating greater generalization. The total duration of the footage was 132 min and the dataset construction step was the same as for the original dataset.

The original dataset was split into training, validation, and test sets. The training set was composed of the videos recorded in 4 days, while the validation and test sets were built over 2 days of recording each. This constituted the dataset setup of the first experimental phase. The extended dataset was completely added to the training set in the second experimental phase.

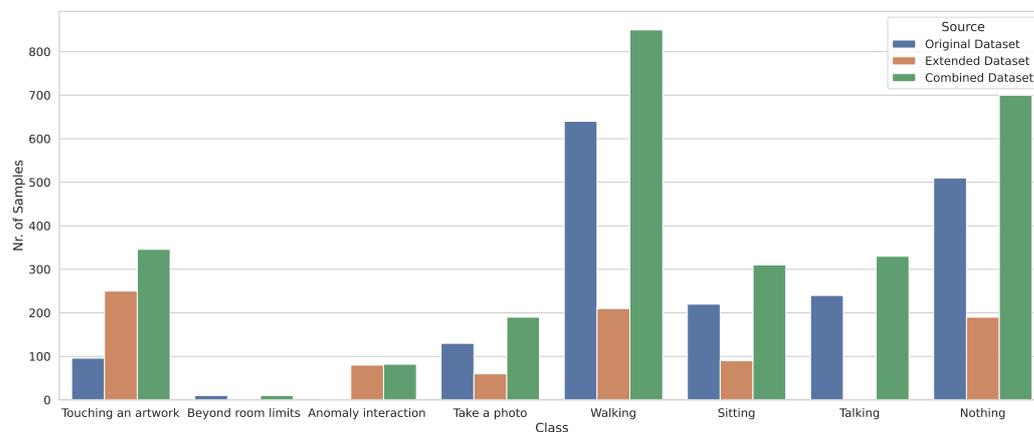


Figure 5. Class samples distribution with respect to original and extended datasets, grouped by assigned action.

Table 1. Set of predictable actions, along with their priority (higher priority is given to actions with a lower value).

Action	Anomaly	Priority
Touching an artwork	✓	1
Beyond room limits	✓	2
Anomaly interaction	✓	3
Take a photo		4
Walking		5
Talking		6
Sitting		7
Nothing		8

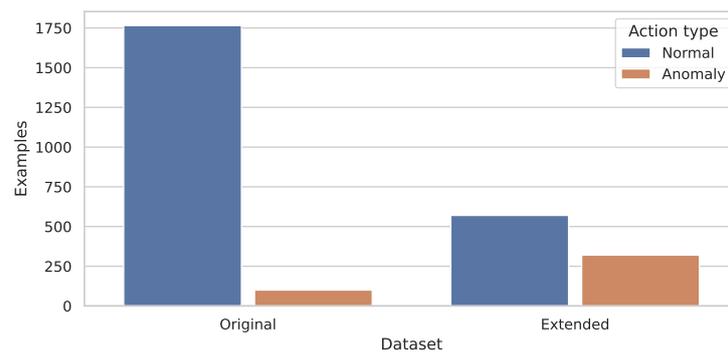


Figure 6. Class samples distribution of the action type (normal or anomalous action), grouped by the experimental stage of employment (original or extended dataset).

3.1.2. Experimental Setup

For the implementation of the model, we relied on the MMAAction2 framework developed by the openMMLab team [70].

Before providing the video clips to the model, they were subject to a normalization procedure to adapt them to the expected input of the model. This preprocessing was applied to RGB clips spanning durations of 5 to 10 s, recorded at a frame rate of 30 frames per second (FPS). The procedure is graphically represented in Figure 7 and was composed of the following steps:

1. Decoding: extracting the corresponding sequence of frames;
2. Time window sampling: randomly selecting a frame window of 64 frames;
3. Uniform frame sampling: dividing the obtained frame window into eight equally long parts and selecting the first frame of each part while ignoring the others;
4. Resizing: adjusting the spatial dimensions to ensure their shortest side is 256 pixels;
5. Cropping: center-cropping the frames to obtain eight RGB square frames with a side length of 224 pixels;
6. Normalization: channel-wise normalization of the resulting images by subtracting the mean and dividing by the standard deviation of the color distribution.

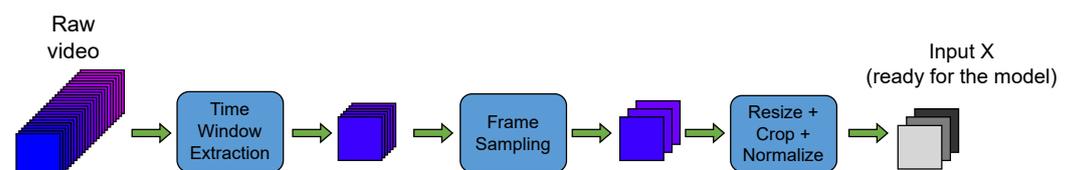


Figure 7. Video preprocessing pipeline.

After each video clip was normalized, it was fed into the feature extraction architecture, which consisted of either the IRCSN or TPN. These architectures produce a 2048-dimensional feature vector that represents salient aspects of the spatio-temporal video content. This feature vector is then passed through a final classification layer that encodes the relationship between the extracted features and corresponding actions. The linear classifier includes a softmax normalization step, which makes the output values interpretable as action probabilities. To determine whether a given clip should be classified as a normal action or an anomaly, we selected the action with the maximum conditional probability and we used the mapping described in Table 1, which prioritized actions based on predefined criteria. In constructing our system, we leveraged the transfer of features extracted by existing checkpoints of IRCSN and TPN that were pre-trained on the Kinetics-400 action recognition dataset. Specifically, we froze these networks during training and only fine-tuned the parameters of the final classification head. The action recognition pipeline is illustrated in Figure 8.

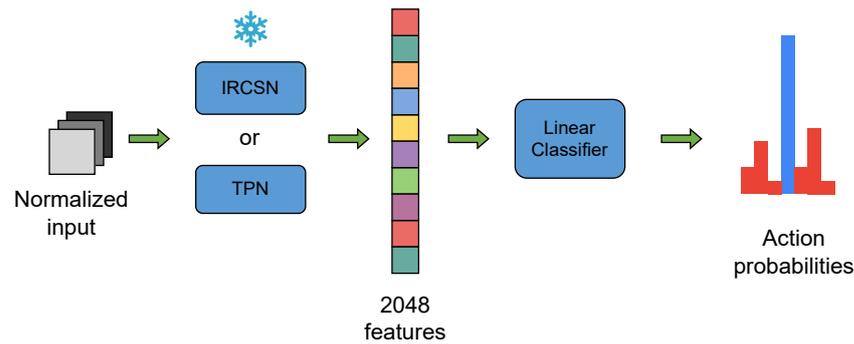


Figure 8. Action recognition pipeline.

In our study, we examined the effect of using augmentation versus not using augmentation. Data augmentation is the process of creating additional training examples that are obtained by applying minimal transformations to the original data, such as rotations and translations. Its objective is to increase the variability of the dataset and improve the model performance. We applied several data-augmentation techniques, including random rotations (videos randomly rotated between -30° and $+30^\circ$ to introduce orientation variations), random horizontal flipping, and Gaussian noise (added to 50% of the input videos with a randomly selected standard deviation between 0 and 3 to diversify the pixel values and increase the model's robustness to noise). Additionally, during training, each loaded video underwent a different color jitter to further improve the model's robustness to color variations. These transformations were applied to both the training and validation examples, effectively increasing their size and variability. Moreover, to correct for the class imbalance, we examined the effect of removing a portion of the most common classes to ensure that no action was excessively prevalent compared with the others during training. In our study, we compared the outcomes obtained with and without using class-balancing techniques.

Throughout the training phase, we investigated the effectiveness of two learning algorithms, namely, stochastic gradient descent (SGD) and Adam [71]. We maintained consistency with the previous studies [34,35] by using the same set of hyperparameters. In particular, we set the learning rate to 0.01 for both optimization algorithms. The learning rate determines the step size during optimization. Furthermore, to mitigate overfitting, we applied a weight decay regularization term in the loss function, with a regularization weight of 10^{-4} . We used the standard Adam parameters [71]. These parameters modulate the dynamics of the adaptive learning rates and the momentum computation. For the SGD, a momentum rate of 0.9 was incorporated to allow for accelerated optimization by accumulating gradient updates over time. Each experiment spanned 50 epochs, with the best model selected based on maximizing the validation acc@1 metric, which was evaluated every five epochs. In addition, a multistep learning rate scheduler [72] was employed to improve the convergence of the training. The base learning rate was multiplied by a gamma factor of 0.1 at epoch 20 and epoch 40.

This study assessed the effectiveness of our machine learning models using a range of metrics, primarily including the following:

1. Accuracy (acc@1): This metric measures the percentage of video clips in which the action is correctly recognized by the model. It serves as a reliable indicator of the model's ability to identify different actions in the videos.
2. Area under the ROC curve (ROCAUC): The goal of anomaly detection is to distinguish between normal and abnormal actions. To evaluate the performance in scenarios with unbalanced datasets, where anomalies occur much less frequently than normal instances, the ROCAUC was adopted for the binary classification task. The ROCAUC is a suitable method for this purpose.
3. Mean class accuracy (MCA): The MCA represents the mean accuracy of the system and it is calculated as the average of the acc@1 metrics computed on the examples of

each class individually. This metric provides valuable insight into the performance of the model for each specific action category.

4. Confusion matrix: In addition, an in-depth analysis of the confusion matrix was performed to better understand the model's ability to minimize false positives and false negatives, which is a critical aspect, especially in real-world museum settings.
5. F1-score: The F1-score is suitable to assess the overall performance of the binary classification system that is supposed to distinguish between normal and anomalous clips. This metric gives a trade-off between precision and recall, providing a measure of the minimization of false positives and false negatives.

We conducted two sets of experiments: First, we focused on refining the optimization setup by evaluating the different optimization algorithms to select the most appropriate optimizer. Subsequently, we evaluated the impact of dataset augmentation and balancing techniques on the accuracy and robustness of the model. The main results of the training and evaluation experiments are reported in Section 4.

3.2. Trajectory Extraction and Analysis

3.2.1. Dataset Composition

To validate the accuracy of our trajectory extraction pipeline, we conducted experiments on a dataset consisting of 100 tracks. These tracks were extracted from video clips recorded by the same camera described in Section 2.2. The dataset consisted of samples extracted from 20 video clips with a total duration of 30 min. For evaluation purposes, the ground truth was established through meticulous manual annotation of each video clip, with labels consistently applied to every fourth frame. Initially, these labels were assigned in the image space and then projected onto the ground plane of the world coordinate system using the homographic transformations described in Section 2.2. Figure 9 illustrates a selection of the extracted trajectories in image space, which were superimposed on the planimetry of the museum hall. The similarity between Figure 9a and Figure 9b was due to the quality of the automatic trajectory extraction method, which returned an accurate approximation of the ground truth trajectories extracted by a human.

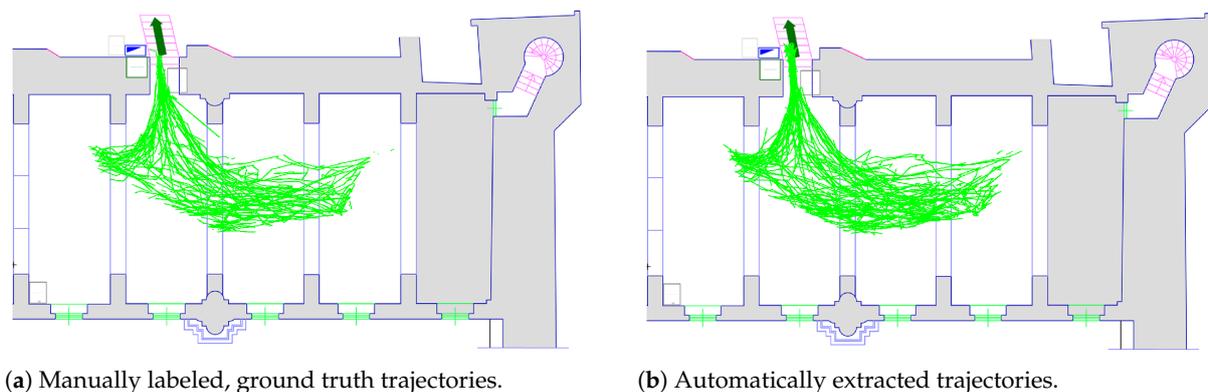


Figure 9. The green lines represent a random subset of the manually labeled (a) and corresponding automatically extracted (b) trajectories.

3.2.2. Experimental Setup

In the first experimental phase, the automatically extracted trajectories were compared with the manually labeled ground truth trajectories. To ensure compatibility between the two sets of trajectories, we took care to synchronize the manually and automatically labeled trajectory points by subsampling frames at the same rate. This ensured that corresponding points in both trajectories were aligned in time. Next, the distance between the corresponding nodes of each pair of trajectories was measured using a simple Euclidean distance metric. These node distances were then aggregated to compute the mean absolute error (MAE), root-mean-squared error (RMSE), and maximum absolute error (MaxAE),

along with the standard deviation of the error distribution (stdAE). The use of the Euclidean distance had the advantage of returning distances expressed in meters in the same reference system of the trajectories, which allowed for a qualitative evaluation of the quality of the trajectory approximation.

For the clustering, we used a distance measure based on the mean absolute Euclidean distance between two trajectories, ignoring time information to focus only on the paths that the visitors took during their stay in the hall. To quantify the distance between trajectories of varying lengths, characterized by differing numbers of nodes, we employed a symmetric spline approximation technique. Initially, for two distinct trajectories, namely, τ and τ' , which comprised n and n' nodes, respectively, we approximated these as polylines. This was achieved by interpolating between nodes using a first-degree spline, thus generating continuous curves $\hat{\tau}$ and $\hat{\tau}'$. Subsequently, we calculated the distance of each node in the discrete trajectory τ to the approximated polyline $\hat{\tau}'$ and computed the mean distance across all nodes. This process was repeated for the trajectory τ' , assessing its nodes' distances to the polyline $\hat{\tau}$. The final distance measure between the two trajectories was determined by averaging these mean distances.

Once the clustering process was complete, we manually determined the number of clusters based on their intra-cluster distance. This selection allowed us to identify meaningful groupings of trajectories that captured different movement patterns within the museum hall. In Section 4.2, we present detailed results of this analysis.

3.3. Integration and Video Stream Processing

To assess the overall performance of the system, we calculated the average interval between the completion of the processing for one sub-window of frames and the subsequent sub-window. This metric shed light on the system's efficiency and its competency in managing real-time video processing demands.

4. Results and Discussion

4.1. Video Anomaly Detection

In the first experimental phase, we used only the original dataset. The results of this set of experiments are presented in Table 2.

First, we consistently observed high top-five accuracy. However, given the limited number of classes in our problem (eight), it was relatively easy to include the target class in the top five predicted classes. As a result, we decided not to consider the top-five accuracy in subsequent experiments. Our investigation revealed that models trained with Adam consistently performed worse than those trained with other optimizers in terms of the ROCAUC. This finding guided our subsequent experiments. Notably, no significant difference between the IRCSN and TPN models was observed, except for the reduced performance of TPN when using the Adam optimizer. Regarding dataset balancing, we expected that a more balanced dataset would result in a better internal description of the actions and improve the reliability of the test results. However, in practice, class balancing greatly limited the set of available examples for training, thereby compromising the generalization capability of the model and the accuracy of the test results. This consideration justified extending the dataset with new real (staged) and artificial data samples.

The results obtained through data augmentation are promising and surpassed those of previous experiments. Moreover, this conclusion is supported by a larger and more diverse test set, which increased the confidence in our measurements. The results of the second experimental phase are reported in Table 3.

Table 2. Results of the first experimental phase, based on the original dataset. The bold results are the best according to the considered metric, while the underlined results are the second best.

Model	Optimizer	Balancing	acc@1 ↑	MCA ↑	ROCAUC ↑
IRCSN	Adam		<u>68.6%</u>	55.2%	95.1%
IRCSN	SGD		71.2%	61.7%	98.8%
TPN	Adam		45.6%	30.8%	68.4%
TPN	SGD		66.0%	<u>59.7%</u>	<u>98.3%</u>
IRCSN	Adam	✓	59.0%	58.1%	84.4%
IRCSN	SGD	✓	57.4%	51.1%	88.73%
TPN	Adam	✓	27.9%	25.4%	88.2%
TPN	SGD	✓	57.4%	56.8%	88.7%

Table 3. Results of the second experimental phase, based on the extended dataset. The bold results are the best according to the considered metric, while the underlined results are the second best.

Model	Balancing	Augmentation	acc@1 ↑	MCA ↑	ROCAUC ↑
IRCSN		Staged	<u>72.7%</u>	62.2%	95.3%
TPN		Staged	70.7%	63.1%	94.1%
IRCSN	✓	Staged	67.1%	51.6%	97.1%
TPN	✓	Staged	60.6%	52.6%	94.0%
IRCSN		Staged + augmented	73.0%	53.4%	<u>96.9%</u>
TPN		Staged + augmented	72.1%	66.9%	95.1%
IRCSN	✓	Staged + augmented	66.1%	59.2%	96.7%
TPN	✓	Staged + augmented	69.2%	<u>66.8%</u>	93.6%

The fact that the ROCAUC metric consistently reached above 90% indicates the validity of the approach. Its variation between different configurations was negligible. Although the effect of dataset balancing remained slightly negative, it was less pronounced than in previous experiments, suggesting potential benefits from further data enrichment efforts. In particular, the IRCSN demonstrated an ability to discriminate between normal actions and anomalies, while the highest accuracy in recognizing the precise actions was achieved by the TPN.

To measure the quantity of false negative and false positive examples, while also considering the proportion between the two, we calculated the F1-score on the validation set for the model with the best accuracy (acc@1). We tested various decision thresholds and selected 0.6 as the threshold resulting in the best F1-score. We then computed the F1-score of the selected model on the test set using the chosen threshold and obtained a value of 86.7%. Using the same model and threshold selection, we computed the confusion matrix, which is shown in Figure 10.

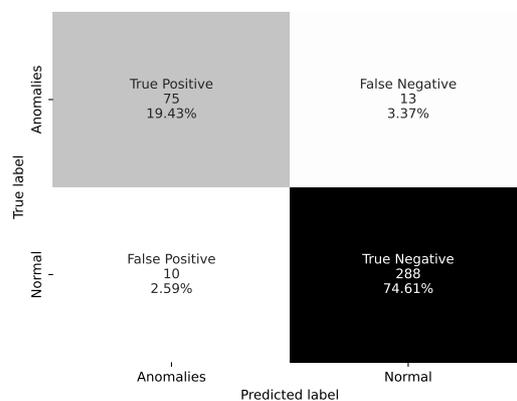


Figure 10. Action recognition confusion matrix.

In this situation, when the MCA was lower than the acc@1 metric, it reflected superior performance in classifying actions that were more frequent in the dataset. This result was advantageous considering that the most frequent actions were normal actions. A smaller MCA compared with the corresponding acc@1 reduced the false positives, increasing the system usability and reliability.

4.2. Trajectory Extraction and Analysis

The aggregated results of the comparison of the extracted trajectories with the ground truth are reported in Table 4, while the distribution of the distance between the predicted and real nodes of the trajectories is shown in Figure 11.

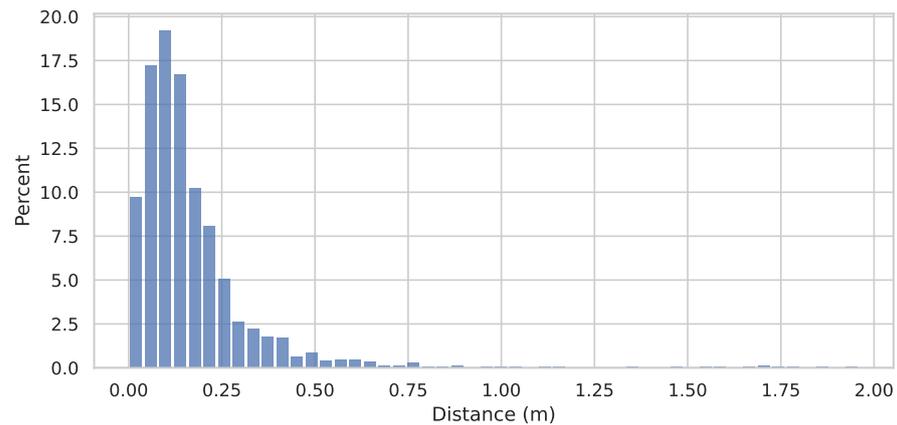


Figure 11. Distribution of the distance between the predicted and real nodes of the trajectories.

Table 4. Aggregated trajectory estimation errors. All the metrics represent distances, and thus, they are expressed in meters.

MAE [m] ↓	RMSE [m] ↓	MaxAE [m] ↓	STD [m] ↓
0.18	0.30	4.36	0.23

The test results on the accuracy of the trajectory extraction method show promising results for its applicability within the museum environment. With a mean error below 20 cm, the achieved accuracy level was considered highly acceptable for the specific environment under investigation. Such a low error rate allows for the effective use of trajectory information for both analytical insights and security purposes. On average, the results show satisfactory performance, as characterized by a tightly concentrated distribution of errors around the mean. This concentration was further emphasized by the significantly low standard deviation, which quantitatively expressed the consistency and reliability of the trajectory extraction method across different scenarios within the museum environment.

In our clustering experiment, the hierarchical clustering algorithm revealed two primary clusters, as shown in Figure 12. The intuitive interpretation of these clusters suggests a division between those visitors who lingered in the museum hall and explored a portion of it and those individuals who used the space as a thoroughfare. While this preliminary analysis provides initial insights into visitor behavior, further exploration with a larger dataset could yield a more comprehensive and informative analysis. With additional data, we expect to be able to reveal more nuanced patterns and behaviors among museum visitors, enhancing our understanding of their interactions within the space.

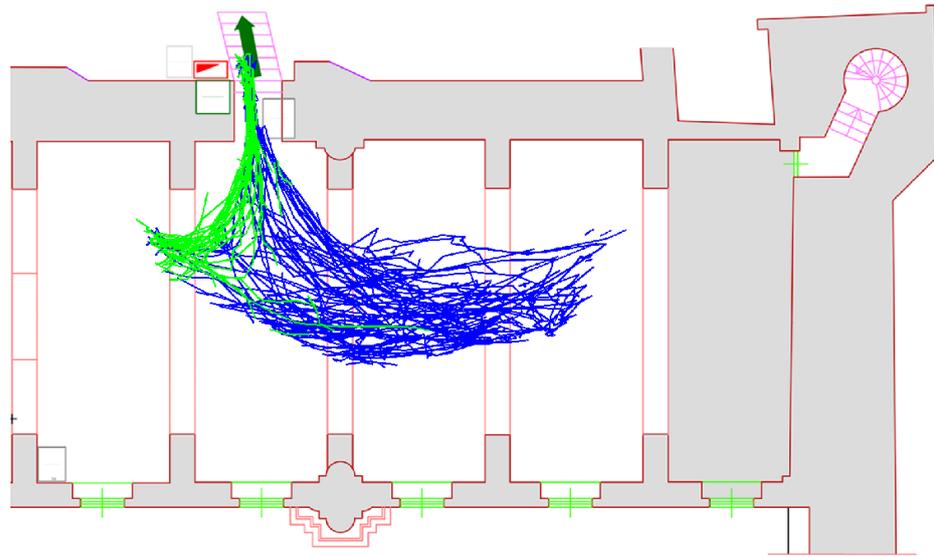


Figure 12. The different trajectory colors identify the different clusters. By choosing to terminate the clustering algorithm after identifying two clusters, the trajectory segmentation highlighted different behavioral patterns exhibited by visitors within the museum hall.

4.3. Integration and Video Stream Processing

To evaluate the time efficiency of the integrated pipeline, we conducted experiments using video footage with a frame rate of 30 FPS and a full-HD resolution. The analysis showed that the average time between consecutive sub-windows of frames was 1471 ms, with a standard deviation of 338 ms. The performance of the pipeline was significantly affected by two main factors. First, the real-time arrival of frames while processing a live video stream imposed constraints on the overall pipeline speed. Second, the computational requirements of the action recognition network, which is the most resource-intensive component of the pipeline, also affected the performance (368 ms on average). Person detection and pose estimation took, respectively, 41 ms and 24 ms for their inferences on average (Table 5). Despite these challenges, it is noteworthy that the observed average time was close to the expected ideal time. Considering that each frame sub-window consists of 64 frames, the expected ideal time was approximately $64/30 \times 1000 \simeq 2133$ ms. The result obtained was consistent with this expectation, taking into account the potential overhead in the processing pipeline. Additionally, we examined how the time required to analyze a packet changed over time. We first considered the three main computational blocks (Figure 13) where we demonstrated that the pipeline achieved a steady state in terms of computational time within five packets. Figure 13a–c shows the execution time required to perform inference over a packet over time; while the action recognition maintained the same inference time over time, Figure 13b shows that the method reached a steady state after a few packets. Then we examined the total execution time (Figure 14).

Table 5. Mean and standard deviation referring to Figure 13a–c, respectively.

Operation	Execution Time (ms)	
	Mean	Std
Action recognition	386.72	145.75
Person detection	41.31	7.14
Pose estimation	24.36	1.61

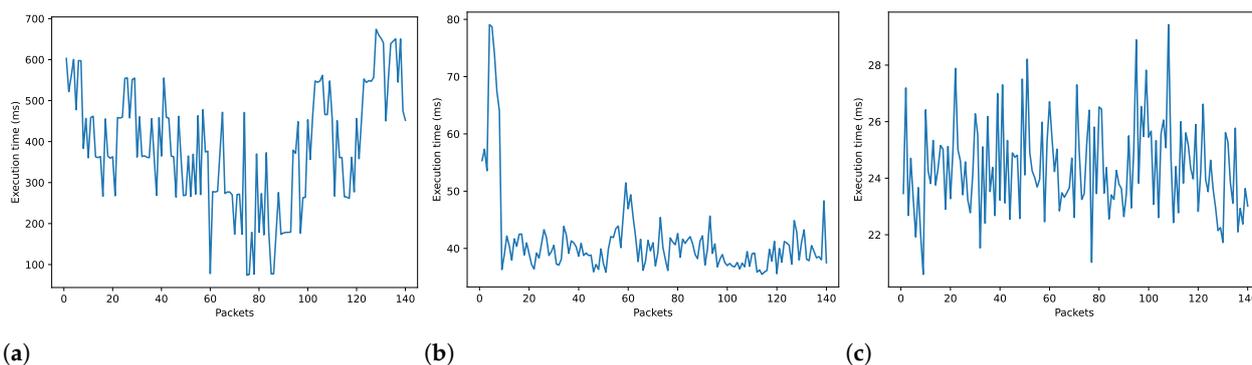


Figure 13. Neural models inference time as time goes on: (a)—action recognition model; (b)—person detection model; (c)—pose estimation model.

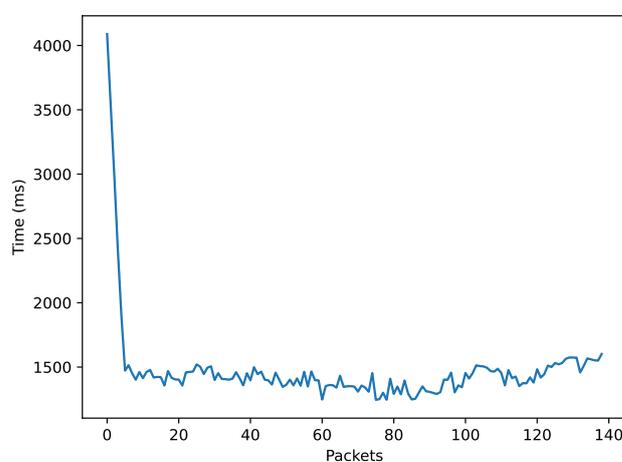


Figure 14. Evolution of time needed to analyze a packet.

5. Conclusions

This work tested various deep learning models to detect anomalous behavior inside museums. First, we introduced a prioritization approach, which effectively addressed the challenges associated with multiple actors appearing within each frame. We then applied various data-augmentation techniques, including examples from various sources and implementing dataset-balancing strategies. These efforts led to significant outcomes and improved the models' abilities to generalize across different scenarios.

Moreover, we proposed an approach for trajectory extraction and validated it by measuring its accuracy against ground truth data. Additionally, we used the extracted trajectories for a preliminary analysis based on clustering to identify different trajectory patterns and interpret them, trying to infer the behavior of the visitors. Finally, we integrated the two proposed techniques into a unique video-processing system.

This study's most significant findings can be summarized as follows:

1. Identification of effective models: The investigation successfully identified and validated models that are valuable in a museum setting. These models demonstrated their utility in addressing the specific anomaly detection challenges encountered in this environment.
2. Priority policy for actions: Our study introduced a novel approach to prioritizing actions within the anomaly detection framework. This policy addresses the challenge of ambiguous action prediction by establishing a hierarchical system for determining the importance of detected actions. This approach effectively allows for the use of scene-based action recognition to detect anomalies, also in the presence of multiple individuals.

3. Dataset improvement techniques: The experiments that involved data-augmentation techniques, dataset balancing, and the introduction of new and out-of-domain examples resulted in positive outcomes of model generalization. These techniques enhanced the model's ability to adapt to various scenarios, improving its performance and robustness.
4. Ground plane trajectory extraction: a novel approach for ground plane trajectory extraction was proposed and validated, demonstrating a degree of accuracy sufficient for most tasks required in a museum environment.
5. Real-time video processing system: The anomaly detection and trajectory extraction systems were seamlessly integrated into a unified video processing architecture capable of processing a video stream in real-time. The performance of the system was tested to assess its ability to operate in real-time, demonstrating its potential for practical deployment in real-world scenarios.

5.1. Limitations

Although our approach shows promise in detecting anomalies within museum environments, it is important to acknowledge its limitations that may affect its effectiveness in real-world scenarios. First, our scene-based approach struggles to handle multiple simultaneous actions within a scene, risking missed detections or false alarms. Second, while our prioritization policy was designed to deal with prediction ambiguity, it lacks full control over the predicted actions. In complex scenarios, prioritization may consider not only the action but also contextual factors, leading to potentially suboptimal decisions, especially in crowded environments. This vulnerability could be exploited by well-prepared adversaries, compromising the system's effectiveness in detecting anomalies and potentially resulting in damage to artworks. Accurate action selection and prioritization are essential to mitigate this risk, ensuring effective threat response and asset protection.

5.2. Future Work

The results of this study are encouraging for the advancement of automated anomaly detection through surveillance camera footage. However, to further enhance the effectiveness and accuracy of these detection systems, several avenues for future research were identified.

First, there is a significant opportunity to tailor anomaly detection models specifically for the unique environment of museums. This customization could involve refining existing models or experimenting with innovative approaches, such as vision transformers and skeleton-based action recognition techniques. Such explorations are anticipated to not only boost the model performance but also uncover novel insights into the dynamics of anomaly detection within specialized settings. Second, the efficiency of models focusing on individual actions warrants thorough evaluation. This is because a more refined understanding and improvement of person-focused action recognition could potentially offer superior results over current scene-based methodologies. Delving deeper into the nuances of individual behaviors could unlock new dimensions in surveillance analytics, offering a more granular and precise detection capability. Moreover, more advanced video processing approaches should be used in place of standard background subtraction. Intelligent video segmentation could represent the next step toward complete scene understanding [73,74]. Lastly, the exploration of advanced video stream processing techniques presents a promising frontier. By leveraging strategies that capitalize on the temporal continuity of video footage, such as utilizing overlapping frame windows and extracting additional contextual information, there is the potential to significantly enhance the detection framework. These approaches could lead to more robust and responsive anomaly detection systems that are better equipped to identify and respond to irregular activities in real time. From the point of view of the extracted trajectory, a straightforward application lies in refining our anomaly detection algorithm. By leveraging information derived from specific trajectory patterns, the latter can provide valuable insights for recognizing threats as they occur or

even predicting them by behavioral analysis. In addition to refining our anomaly detection algorithm based on extracted trajectory data, further directions for future work include extending our capabilities to multi-camera tracking [75]. By integrating data from cameras positioned in different rooms within the museum, we can track individuals over larger areas, resulting in more comprehensive or extended trajectories. This broader scope allows for the detection of more complex behaviors. In addition, the exploration of alternative approaches to object detection and tracking, such as those based on vision transformers, represents another promising direction for advancing our analytic capabilities and improving the accuracy of anomaly detection algorithms [76,77]. These efforts will contribute to a more robust and sophisticated system for monitoring and securing museum environments.

In conclusion, while the initial findings are promising, the path ahead is rich with potential for innovation and refinement in the field of anomaly detection. By pursuing these suggested research directions, there is a strong possibility of achieving more sophisticated, efficient, and accurate surveillance systems that are capable of safeguarding our public spaces with unprecedented effectiveness.

Author Contributions: Conceptualization, C.D.M., G.N. and A.M.; methodology, C.D.M., G.N. and A.M.; software, C.D.M. and G.N.; validation, C.D.M. and G.N.; formal analysis, C.D.M. and G.N.; investigation, C.D.M. and G.N.; resources, A.M.; data curation, C.D.M. and G.N.; writing—original draft preparation, C.D.M. and G.N.; writing—review and editing, A.M.; visualization, C.D.M. and G.N.; supervision, A.M.; project administration, A.M.; funding acquisition, A.M. All authors read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets presented in this article are not readily available due to privacy limitations and norms that prevent their publication. Requests to access the datasets should be directed to a corresponding author.

Acknowledgments: We wish to extend our deepest appreciation to the Museo dell’Opera della Metropolitana di Siena and the VISLab laboratory at the University of Siena for their significant contributions to this research project. Their assistance, resources, and expertise were instrumental in advancing our work, and we deeply appreciate their partnership throughout this endeavor. We also extend our heartfelt gratitude to Luca Storai for his invaluable support in facilitating crucial technical tasks. His expertise and assistance were instrumental in extracting essential data, greatly enhancing the efficiency of our research endeavors.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Sulman, N.; Sanocki, T.; Goldgof, D.; Kasturi, R. How effective is human video surveillance performance? In Proceedings of the 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–3.
2. Shindell, L.M. Provenance and title risks in the art industry: Mitigating these risks in museum management and curatorship. *Mus. Manag. Curatorship* **2016**, *31*, 406–417. [[CrossRef](#)]
3. Sharma, V.; Gupta, M.; Kumar, A.; Mishra, D. Video processing using deep learning techniques: A systematic literature review. *IEEE Access* **2021**, *9*, 139489–139507. [[CrossRef](#)]
4. Sreenu, G.; Durai, S. Intelligent video surveillance: A review through deep learning techniques for crowd analysis. *J. Big Data* **2019**, *6*, 48. [[CrossRef](#)]
5. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [[CrossRef](#)] [[PubMed](#)]
6. Chai, J.; Zeng, H.; Li, A.; Ngai, E.W. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Mach. Learn. Appl.* **2021**, *6*, 100134. [[CrossRef](#)]
7. O’Mahony, N.; Campbell, S.; Carvalho, A.; Harapanahalli, S.; Hernandez, G.V.; Krpalkova, L.; Riordan, D.; Walsh, J. Deep learning vs. traditional computer vision. In *Advances in Computer Vision, CVC 2019; Advances in Intelligent Systems and Computing*; Springer: Cham, Switzerland, 2020; Volume 943, pp. 128–144.
8. LeCun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; Jackel, L. Handwritten digit recognition with a back-propagation network. *Adv. Neural Inf. Process. Syst.* **1989**, *2*, 396–404.

9. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [\[CrossRef\]](#)
10. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Li, F.-F. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
11. Diba, A.; Fayyaz, M.; Sharma, V.; Karami, A.H.; Arzani, M.M.; Yousefzadeh, R.; Van Gool, L. Temporal 3d convnets: New architecture and transfer learning for video classification. *arXiv* **2017**, arXiv:1711.08200.
12. Bansod, S.; Nandedkar, A. Transfer learning for video anomaly detection. *J. Intell. Fuzzy Syst.* **2019**, *36*, 1967–1975. [\[CrossRef\]](#)
13. Cauli, N.; Reforgiato Recupero, D. Survey on videos data augmentation for deep learning models. *Future Internet* **2022**, *14*, 93. [\[CrossRef\]](#)
14. Polson, N.; Sokolov, V. Deep learning: Computational aspects. *Wiley Interdiscip. Rev. Comput. Stat.* **2020**, *12*, e1500. [\[CrossRef\]](#)
15. Reuther, A.; Michaleas, P.; Jones, M.; Gadepally, V.; Samsi, S.; Kepner, J. Survey and benchmarking of machine learning accelerators. In Proceedings of the 2019 IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA, 24–26 September 2019; pp. 1–9.
16. Reuther, A.; Michaleas, P.; Jones, M.; Gadepally, V.; Samsi, S.; Kepner, J. Survey of machine learning accelerators. In Proceedings of the 2020 IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA, 22–24 September 2020; pp. 1–12.
17. Saiyeda, A.; Mir, M.A. Cloud computing for deep learning analytics: A survey of current trends and challenges. *Int. J. Adv. Res. Comput. Sci.* **2017**, *8*, 68.
18. Jauro, F.; Chiroma, H.; Gital, A.Y.; Almutairi, M.; Shafi'i, M.A.; Abawajy, J.H. Deep learning architectures in emerging cloud computing architectures: Recent development, challenges and next research trend. *Appl. Soft Comput.* **2020**, *96*, 106582. [\[CrossRef\]](#)
19. Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S. A review of video surveillance systems. *J. Vis. Commun. Image Represent.* **2021**, *77*, 103116. [\[CrossRef\]](#)
20. Xu, Z.; Liu, Y.; Mei, L.; Hu, C.; Chen, L. Semantic based representing and organizing surveillance big data using video structural description technology. *J. Syst. Softw.* **2015**, *102*, 217–225. [\[CrossRef\]](#)
21. Xu, Z.; Hu, C.; Mei, L. Video structured description technology based intelligence analysis of surveillance videos for public security applications. *Multimed. Tools Appl.* **2016**, *75*, 12155–12172. [\[CrossRef\]](#)
22. El Harrouss, O.; Moujahid, D.; Tairi, H. Motion detection based on the combining of the background subtraction and spatial color information. In Proceedings of the 2015 Intelligent Systems and Computer Vision (ISCV), Fez, Morocco, 25–26 March 2015; pp. 1–4.
23. Kim, M.; Suh, T. A Low-Cost Surveillance and Information System for Museum Using Visible Light Communication. *IEEE Sens. J.* **2019**, *19*, 1533–1541. [\[CrossRef\]](#)
24. Viani, F.; Salucci, M.; Rocca, P.; Oliveri, G.; Massa, A. A multi-sensor WSN backbone for museum monitoring and surveillance. In Proceedings of the 2012 6th European Conference on Antennas and Propagation (EUCAP), Prague, Czech Republic, 26–30 March 2012; pp. 51–52. [\[CrossRef\]](#)
25. Bahadori, S.; Iocchi, L. A stereo vision system for 3d reconstruction and semi-automatic surveillance of museum areas. In Proceedings of the AI*IA 2003: Advances in Artificial Intelligence, Pisa, Italy, 23–26 September 2003; Volume 147, p. 148.
26. Ramachandra, B.; Jones, M.J.; Vatsavai, R.R. A Survey of Single-Scene Video Anomaly Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 2293–2312. [\[CrossRef\]](#)
27. Nayak, R.; Pati, U.C.; Das, S.K. A comprehensive review on deep learning-based methods for video anomaly detection. *Image Vis. Comput.* **2021**, *106*, 104078. [\[CrossRef\]](#)
28. Li, S.; Liu, F.; Jiao, L. Self-Training Multi-Sequence Learning with Transformer for Weakly Supervised Video Anomaly Detection. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 1395–1403. [\[CrossRef\]](#)
29. Doshi, K.; Yilmaz, Y. Rethinking video anomaly detection—A continual learning approach. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 3961–3970.
30. Chang, Y.; Tu, Z.; Xie, W.; Luo, B.; Zhang, S.; Sui, H.; Yuan, J. Video anomaly detection with spatio-temporal dissociation. *Pattern Recognit.* **2022**, *122*, 108213. [\[CrossRef\]](#)
31. Hao, Y.; Li, J.; Wang, N.; Wang, X.; Gao, X. Spatiotemporal consistency-enhanced network for video anomaly detection. *Pattern Recognit.* **2022**, *121*, 108232. [\[CrossRef\]](#)
32. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
33. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
34. Tran, D.; Wang, H.; Torresani, L.; Feiszli, M. Video classification with channel-separated convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5552–5561.
35. Yang, C.; Xu, Y.; Shi, J.; Dai, B.; Zhou, B. Temporal pyramid network for action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 591–600.
36. Berroukham, A.; Housni, K.; Lahraichi, M.; Boulfrifi, I. Deep learning-based methods for anomaly detection in video surveillance: A review. *Bull. Electr. Eng. Inform.* **2023**, *12*, 314–327. [\[CrossRef\]](#)

37. Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning temporal regularity in video sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 733–742.
38. Sabokrou, M.; Fathy, M.; Hoseini, M. Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. *Electron. Lett.* **2016**, *52*, 1122–1124. [[CrossRef](#)]
39. Liu, W.; Luo, W.; Lian, D.; Gao, S. Future frame prediction for anomaly detection—a new baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6536–6545.
40. Medel, J.R.; Savakis, A. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv* **2016**, arXiv:1612.00390.
41. Sabokrou, M.; Fayyaz, M.; Fathy, M.; Moayed, Z.; Klette, R. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Comput. Vis. Image Underst.* **2018**, *172*, 88–97. [[CrossRef](#)]
42. Sultani, W.; Chen, C.; Shah, M. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6479–6488.
43. Xu, D.; Yan, Y.; Ricci, E.; Sebe, N. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Comput. Vis. Image Underst.* **2017**, *156*, 117–127. [[CrossRef](#)]
44. Carreira, J.; Noland, E.; Hillier, C.; Zisserman, A. A Short Note on the Kinetics-700 Human Action Dataset. *arXiv* **2019**, arXiv:1907.06987.
45. Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Li, F.-F.; Savarese, S. Social lstm: Human trajectory prediction in crowded spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 961–971.
46. Zhao, Y.; Chen, Q.; Cao, W.; Yang, J.; Xiong, J.; Gui, G. Deep learning for risk detection and trajectory tracking at construction sites. *IEEE Access* **2019**, *7*, 30905–30912. [[CrossRef](#)]
47. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
48. Kalman, R.E. A new approach to linear filtering and prediction problems. *J. Basic Eng. Mar.* **1960**, *82*, 35–45. [[CrossRef](#)]
49. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [[CrossRef](#)]
50. Oikonomopoulos, A.; Patras, I.; Pantic, M.; Paragios, N. Trajectory-based representation of human actions. In *Artificial Intelligence for Human Computing: ICMI 2006 and IJCAI 2007 International Workshops, Banff, Canada, November 3, 2006, Hyderabad, India, January 6, 2007, Revised Selected and Invited Papers*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 133–154.
51. Ahmed, M.; Seraj, R.; Islam, S.M.S. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics* **2020**, *9*, 1295. [[CrossRef](#)]
52. Khan, K.; Rehman, S.U.; Aziz, K.; Fong, S.; Sarasvady, S. DBSCAN: Past, present and future. In Proceedings of the Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014), Chennai, India, 17–19 February 2014; pp. 232–238.
53. Bach, F.; Jordan, M. Learning spectral clustering. *Adv. Neural Inf. Process. Syst.* **2003**, *16*, 305–312.
54. Nielsen, F.; Nielsen, F. Hierarchical clustering. In *Introduction to HPC with MPI for Data Science*; Springer: Cham, Switzerland, 2016; pp. 195–211.
55. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
56. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
57. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
58. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
59. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
60. Szeliski, R. *Computer Vision: Algorithms and Applications*; Springer Nature: Berlin, Germany, 2022.
61. Zivkovic, Z. Improved adaptive Gaussian mixture model for background subtraction. In Proceedings of the 17th International Conference on Pattern Recognition, 2004, ICPR 2004, Cambridge, UK, 26–26 August 2004; Volume 2, pp. 28–31.
62. Lee, D.S. Effective Gaussian mixture learning for video background subtraction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 827–832. [[PubMed](#)]
63. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 5 February 2024).
64. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
65. Terven, J.; Cordova-Esparza, D. A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond. *arXiv* **2023**, arXiv:2304.00501.

66. Aboah, A.; Wang, B.; Bagci, U.; Adu-Gyamfi, Y. Real-time multi-class helmet violation detection using few-shot data sampling technique and yolov8. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5349–5357.
67. Saeed, S.M.; Akbar, H.; Nawaz, T.; Elahi, H.; Khan, U.S. Body-Pose-Guided Action Recognition with Convolutional Long Short-Term Memory (LSTM) in Aerial Videos. *Appl. Sci.* **2023**, *13*, 9384. [[CrossRef](#)]
68. Elbarrany, A.M.; Mohialdin, A.; Atia, A. The Use of Pose Estimation for Abnormal Behavior Analysis in Poultry Farms. In Proceedings of the 2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, 21–23 October 2023; pp. 33–36.
69. Aharon, N.; Orfaig, R.; Bobrovsky, B.Z. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv* **2022**, arXiv:2206.14651.
70. MMAAction2 Contributors. OpenMMLab’s Next Generation Video Understanding Toolbox and Benchmark. 2020. Available online: <https://github.com/open-mmlab/mmaaction2> (accessed on 5 February 2024).
71. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
72. Ge, R.; Kakade, S.M.; Kidambi, R.; Netrapalli, P. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 14977–14988.
73. Lu, X.; Wang, W.; Ma, C.; Shen, J.; Shao, L.; Porikli, F. See More, Know More: Unsupervised Video Object Segmentation With Co-Attention Siamese Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
74. Lu, X.; Wang, W.; Shen, J.; Crandall, D.; Luo, J. Zero-Shot Video Object Segmentation with Co-Attention Siamese Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2228–2242. [[CrossRef](#)] [[PubMed](#)]
75. Im, S.K.; Chan, K.H. Distributed Spatial Transformer for Object Tracking in Multi-Camera. In Proceedings of the 2023 25th International Conference on Advanced Communication Technology (ICACT), Pyeongchang, Republic of Korea, 19–22 February 2023; pp. 122–125.
76. Chan, K.H.; Im, S.K.; Ian, V.K.; Chan, K.M.; Ke, W. Enhancement spatial transformer networks for text classification. In Proceedings of the 4th International Conference on Graphics and Signal Processing, Nagoya, Japan, 26–29 June 2020; pp. 5–10.
77. Raisi, Z.; Younes, G.; Zelek, J. Arbitrary shape text detection using transformers. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 3238–3245.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.