applied sciences

MDPI

Article

Efficient Crowd Anomaly Detection Using Sparse Feature Tracking and Neural Network

Sarah Altowairqi [1,*], Suhuai Luo [1], Peter Greer [1] and Shan Chen [2]

[1] School of Information and Physical Sciences, The University of Newcastle, University Drive, Newcastle, NSW 2308, Australia; suhuai.luo@newcastle.edu.au (S.L.); peter.greer@newcastle.edu.au (P.G.)
[2] School of Computing, Macquarie University, 4 Research Park Drive, Sydney, NSW 2109, Australia; shan.chen@mq.edu.au
* Correspondence: sarah.altowairqi@uon.edu.au

Abstract: Crowd anomaly detection is crucial in enhancing surveillance and crowd management. This paper proposes an efficient approach that combines spatial and temporal visual descriptors, sparse feature tracking, and neural networks for efficient crowd anomaly detection. The proposed approach utilises diverse local feature extraction methods, including SIFT, FAST, and AKAZE, with a sparse feature tracking technique to ensure accurate and consistent tracking. Delaunay triangulation is employed to represent the spatial distribution of features in an efficient way. Visual descriptors are categorised into individual behaviour descriptors and interactive descriptors to capture the temporal and spatial characteristics of crowd dynamics and behaviour, respectively. Neural networks are then utilised to classify these descriptors and pinpoint anomalies, making use of their strong learning capabilities. A significant component of our study is the assessment of how dimensionality reduction methods, particularly autoencoders and PCA, affect the feature set's performance. This assessment aims to balance computational efficiency and detection accuracy. Tests conducted on benchmark crowd datasets highlight the effectiveness of our method in identifying anomalies. Our approach offers a nuanced understanding of crowd movement and patterns by emphasising both individual and collective characteristics. The visual and local descriptors facilitate high-level analysis by closely relating to semantic information and crowd behaviour. The analysis observed shows that this approach offers an efficient framework for crowd anomaly detection, contributing to improved crowd management and public safety. The proposed model achieves accuracy of 99.5 %, 96.1%, 99.0% and 88.5% in the UMN scenes 1, 2, and 3 and violence in crowds datasets, respectively.

Keywords: crowd anomaly detection; visual descriptor; sparse feature tracking; neural networks

Citation: Altowairqi, S.; Luo, S.; Greer, P.; Chen, S. Efficient Crowd Anomaly Detection Using Sparse Feature Tracking and Neural Network. *Appl. Sci.* **2024**, *14*, 3928. https://doi.org/10.3390/app14093928

Academic Editor: Antonio Fernández-Caballero

Received: 31 March 2024
Revised: 28 April 2024
Accepted: 30 April 2024
Published: 4 May 2024

1. Introduction

The phenomenon of crowds has garnered considerable academic interest in recent years owing to the proliferation of events that attract large gatherings [1]. The safety concerns associated with such events, particularly religious and sporting events, have underlined the importance of detecting crowd anomalies, which entails examining the actions and interactions of individuals in large groups. The study of crowd anomalies presents a formidable challenge due to the crowd dynamics' intricate and unpredictable nature. Fortunately, recent breakthroughs in computer vision, machine learning, and deep learning have opened up new possibilities for crowd behaviour analysis. Crowd anomaly detection (CAD) [2] involves identifying unusual or abnormal behaviour within a crowd. This capability is crucial in ensuring public safety, preventing accidents, and managing the flow of people in public spaces, especially at crowded events like religious or sporting events, where emotions can run high.

By detecting anomalies in real time, it is possible to respond quickly and effectively to potentially dangerous situations, preventing them from escalating. Local feature extraction

methods are pivotal in accurately representing crowd behaviour. These methods facilitate the identification of unique patterns of movement and behaviour within a crowd, simplifying the detection of abnormal or anomalous events. However, the effectiveness of these methods depends on several factors, including the features used and the specific anomaly being detected. This study aims to provide insights into the strengths and limitations of various methods and contribute to developing more robust and practical approaches to crowd anomaly detection. This study's findings are expected to contribute to the crowd behaviour analysis field and help to improve public safety in crowded spaces.

The rest of this paper is organised as follows. Section 2 briefly reviews recent works in crowd anomaly detection. Section 3 outlines the methodology employed in this study, including the extraction of local features, the utilisation of sparse feature tracking, and the spatial representation achieved through Delaunay triangulation. Additionally, the descriptors used for individual and interactive behaviours and their calculation using graph notations are also explained. Furthermore, we elaborate on this research's neural network-based classification method. Section 4 presents the details regarding the datasets utilised for the analysis and a thorough analysis of the results, including a comparison of the three local features with and without dimensionality reduction using the autoencoder and PCA techniques. Finally, Section 5 concludes the paper and outlines potential avenues for future research.

2. Related Work

Crowd anomaly detection aims to detect changes in and automatically identify crowd events in video sequences [3,4]. Currently, two main types of techniques are used in crowd anomaly detection: object-based and holistic techniques [5]. Object-based techniques require the breaking down of the crowd into smaller groups and the analysis of their extracted trajectories to infer crowd behaviour [6,7]. However, these methods struggle to recognise activities within a crowded scene due to occlusion and the loss of target object visibility. On the other hand, holistic approaches view the crowd as a single interconnected system and focus on exploiting low- and medium-level features to analyse crowd behaviours [8]. Optical flow fields are often used in these methods [9–12], which can detect various crowd events. Krausz and Bauckhage [9] introduced an automated method that utilises optical flow histograms to represent the overall motion of a crowd in a scene. These histograms were used to identify potentially hazardous situations within crowds, such as the Love Parade stampede in Germany in 2010. Benabbas et al. [10] utilised low-level motion features to create crowd models for direction and magnitude. Then, they generated unique motion sequences using a segmentation algorithm based on regions, which enabled the detection of various crowd events. Rao et al. [12] proposed a Riemannian probabilistic detection framework based on optical flow manifolds to detect various crowd events, such as running, walking, and local dispersion. Newer methods go beyond the frame-to-frame motion information used by earlier works by either tracking particularly interesting points [13–15] or employing particle advection [16–18]. These methods use trajectory information to capture long-term temporal dependencies and extract motion patterns for crowd anomaly detection. Mousavi et al. [13] proposed the histogram of oriented tracklets (HOT), a 2D histogram-based motion descriptor that encodes both the magnitude and direction of motion. It helps to spot unusual events in densely populated scenes.

CAD has been an active area of research in recent years, with various techniques and approaches being proposed to overcome the various challenges reported in the literature. A review of the recent advancements in crowd anomaly detection was presented in [2,19], which discussed the integration of multiple modalities and addressed challenges such as occlusion and scale variation. This paper concludes with a discussion of future research directions in crowd anomaly detection. A detailed review of the anomaly detection methods in crowd scenes from the computer vision perspective was presented in [1], which focused on studying the human crowd, specifically abnormal human behaviour. The paper summarised the state-of-the-art anomaly detection methods in crowd scenes and categorised

them based on their approach, anomaly scope, and processing target. It highlights the importance of intelligent monitoring systems for effective crowd management and discusses the role of computer vision, video analysis, and automated crowd anomaly detection in this context. Crowd anomaly detection using spatial constraints and meaningful perturbation was proposed in [20], which addressed the challenges of rare and diverse abnormal events in crowd scenes. The paper focused on detecting anomalies in crowded scenes to enhance automatic video surveillance systems. A novel approach to crowd anomaly detection was proposed in [21], which focused on detecting anomalies in crowded scenes to enhance automatic video surveillance systems. The proposed method used a combination of multiple optimised convolutional neural networks (ConvNets) to detect anomalies in video data showing crowded scenes. The approach was designed to be efficient and have a low computational cost, making it suitable for real-time applications. In [22], abnormal behaviours were detected in a two-step process. Initially, the Yolov5 model was employed for detection, while the DeepSORT model was utilised for tracking. Subsequently, the abnormal behaviour was classified by extracting features from each detected bounding box using the optical flow and other spatial features. These extracted features were then used to classify the behaviour by implementing a support vector machine (SVM). The SVM model demonstrated an average area under the curve (AUC) of 88.96%. The proposed framework significantly impacted the detection of anomalies in densely populated crowds, such as those observed during the Hajj pilgrimage. Furthermore, the framework exhibited promising outcomes when compared to the study conducted in [23], which categorised abnormal behaviours during Hajj into seven distinct categories. The proposed solution achieved superior performance, surpassing the previous AUC result by 12.88%. The combination of various modalities and the application of cutting-edge machine learning techniques continue to push the boundaries of what is possible in crowd anomaly detection.

Table 1 below summarises various methods used for crowd anomaly detection, the datasets on which they were tested, and the accuracy achieved. These results demonstrate the variety of techniques applied to different datasets and the corresponding performance metrics, providing insights into the effectiveness of various approaches in crowd anomaly detection.

Table 1. Recent works on crowd anomaly detection.

Methods	Datasets Used	Performance Metrics
GAN [24]	ShanghaiTech	73.8% AUC
RNN, 2D CNN [25]	Violent-Flow	93.53% Accuracy
CNN, RNN KNN, Optical Flow [26]	ShanghaiTech	73.62% Accuracy
Optical Flow GAN [27]	Hajj datasets UMN	79.63% Accuracy 98.1% AUC
CNN Residual LSTM [28]	UCF-Crime	70.4% AUC
CNN [29]	ShanghaiTech	240.0 MAE, 260.5 MSE
CNN, Random Forest [23]	HajjV2	76.08% AUC
Optical Flow [30]	ShanghaiTech	89.29% AUC
CNN, Histogram of Optical Flow, SVM [22]	HajjV2	88.96% AUC
gKLT + Collectiveness Energy Index (CEI) [31]	UMN	Scene 1: 92.32%, Scene 3: 94.2%

3. Materials and Methods

Figure 1 provides a schematic representation of the proposed methodology, which is organised into two distinct sections: (i) crowd representation and behaviour descriptors and (ii) dimensionality reduction and classification. In the first section, the crowd is represented through the extraction of feature points from every L (default 20) frame, forming trajectories and Delaunay triangles. The visual descriptors capturing spatial and temporal information, such as the velocity, density, etc., are employed to elucidate the crowd's state. Moving

on to the next section, the descriptors for a frame are computed and they are aggregated into histograms. These histograms undergo a dimensionality reduction process using PCA/autoencoders and are subsequently subjected to classification using neural networks (NN) to categorise normal or abnormal behaviour accurately. The methods used in our approach are further elucidated in the following sub-sections.

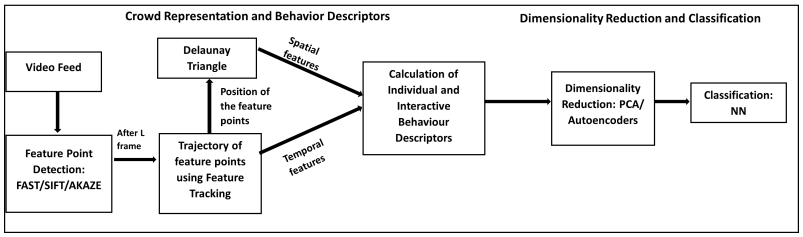

Figure 1. A schematic representation of the proposed approach.

3.1. Crowd Representation and Behaviour Descriptors

This sub-section explains the methods used to represent the crowd in detail. Firstly, we will explore the methods employed for local feature extraction and their significance in representing individuals within a scene, specifically within the context of crowd anomaly detection. Prominent local feature extraction techniques such as FAST, SIFT, and AKAZE are revisited in detail to emphasise that these algorithms can be utilised to extract a comprehensive set of local features, enabling us to analyse each frame's content effectively. These detected interest points serve as valuable observations to represent objects or people in the scene. Subsequently, we delve into the implementation of a sparse feature tracking framework. This framework allows us to track the identified local features over consecutive frames. As the tracking accuracy improves, the emphasis shifts from detecting new features to accurately tracing existing trajectories. Additionally, we explore the spatial representation aspect of crowds by employing the Delaunay triangulation method. This approach aids in creating a structured representation of the scene, facilitating the analysis of individual and interactive behaviours within the crowd. Furthermore, we investigate the utilisation of individual and interactive behaviour descriptors, which play a pivotal role in characterising and understanding the observed crowd dynamics. These descriptors provide valuable insights into anomalies or abnormal behaviours within the crowd. By combining the tracked features and descriptors, we aim to classify the anomalies in the video effectively.

3.1.1. Local Feature Extraction

This section outlines the local features incorporated into our methodology. Specifically, we utilise Features from the Accelerated Segment Test (FAST), Scale-Invariant Feature Transform (SIFT), and Accelerated-KAZE (AKAZE), each of which will be detailed in the following sub-sections.

1. Features from Accelerated Segment Test (FAST)
   The FAST algorithm was initially proposed by Rosten and Drummond [32] and has been used to identify interest points in an image. Interest points are pixels with well-defined positions that can be reliably detected. These points contain significant local information and should be consistently detected across different images. It scans the image using a circular neighbourhood around each pixel and identifies potential key points based on intensity differences. The algorithm employs acceleration techniques, such as a corner criterion using predetermined sample points, to reduce the computational overhead. Non-maximum suppression is applied to select the most salient key points, discarding redundant ones. Interest point detection finds application in image matching, object recognition, and tracking and can also be utilised for crowd anomaly detection. While there are established algorithms for corner detection, such as Harris and SUSAN, the FAST algorithm was developed to address the need for a

computationally efficient interest point detector suitable for real-time applications with limited computational resources, like SLAM on a mobile robot [33].

2. Scale Invariant Feature Transform (SIFT)
The SIFT algorithm, which was first introduced in [34], is one of the most widely known feature detection–description algorithms. It approximates the Laplacian-of-Gaussian (LoG) by utilising the Difference-of-Gaussians (DoG) operator. The DoG operator is employed to search for local maxima in images at various zoom levels, enabling the identification of feature points. To extract a robust descriptor, SIFT computes 128 bin values by considering a 16 × 16 neighbourhood around each detected feature and segmenting it into sub-blocks. Although SIFT exhibits robust invariance to image rotations, scales, and limited affine variations, its major drawback is its high computational cost. In Equation (1), the Difference-of-Gaussians (DoG) response at a given scale in the Scale Invariant Feature Transform (SIFT) algorithm is obtained by convolving the image $I(x, y)$ with the difference of two Gaussian filters calculated at different scales. The result of this convolution represents the DoG response, which is used to detect feature points in the image.

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \quad (1)$$

3. Accelerated-KAZE (AKAZE)
The Accelerated-KAZE (AKAZE) algorithm, introduced in [11] , is an extension of the KAZE algorithm that utilises a computationally efficient framework called Fast Explicit Diffusion (FED) to construct its non-linear scale spaces. AKAZE is based on non-linear diffusion filtering and employs the determinant of the Hessian matrix for feature detection. To enhance the rotation invariance, AKAZE utilises Scharr filters. The maximum responses obtained from the detectors indicate specific feature point locations. These feature points serve as the foundation for AKAZE's robust and distinctive feature detection. The AKAZE descriptor relies on the Modified Local Difference Binary (MLDB) algorithm, which is both powerful and efficient. AKAZE's scale spaces have a non-linear nature, resulting in invariance to scale, rotation, and limited affine transformations. Moreover, AKAZE's features become increasingly distinctive as they are scaled up or down.

In our approach, SIFT, AKAZE, and FAST were individually evaluated, emphasising their distinct feature extraction capabilities for anomaly detection within sparse feature tracking. SIFT is renowned for its robustness to changes in scale, rotation, and illumination; it excels in extracting distinctive key points from crowd images. Its capability to capture complex patterns enables the detection of anomalous behaviours that deviate from the expected crowd dynamics, making it particularly effective in detecting anomalies of varying sizes. AKAZE builds upon SIFT's principles while offering enhanced speed and robustness. AKAZE efficiently extracts key points across different scales and is less sensitive to noise and blur. Its adaptability to diverse environmental conditions makes it suitable for the detection of anomalies in challenging real-world scenarios, ensuring reliable performance in varied settings. FAST is designed for rapid corner detection. It can efficiently identify key features of corners within crowd images. Although not as invariant to scale and rotation as SIFT and AKAZE, FAST's speed makes it ideal for real-time anomaly detection applications, where a timely response is crucial. Its efficient detection of key points contributes to the overall effectiveness of sparse feature tracking in identifying anomalous crowd behaviours.

3.1.2. Sparse Feature Tracking

Sparse feature tracking is an efficient technique used in computer vision to track a subset of features spanning in video frames. It addresses challenges such as occlusion and changes in crowd dynamics by focusing on tracking distinctive features within the crowd, making it less susceptible to occlusion and environmental changes compared to dense tracking methods. By selecting key points of interest, sparse feature tracking can

maintain robustness in tracking even when parts of the crowd are occluded or when there are dynamic changes in crowd movement. Additionally, sparse feature tracking algorithms often incorporate mechanisms for feature re-detection and association, allowing them to adapt to changes in crowd dynamics and maintain accurate tracking over time. This method utilises the most prominent features that are easy to track, such as the corners or any unique structures in the frames. Our approach is based on the local features explained above. The Lucas–Kanade approach [35] is employed to address the issue of an unrestricted optical flow. The Lucas–Kanade optical flow approach excels in crowd anomaly detection due to its accuracy, robustness, and adaptability to complex scenarios. By focusing on sparse feature tracking, it reduces the computational complexity while maintaining high tracking accuracy, making it suitable for real-time applications. Its ability to capture subtle motion variations within crowded scenes enables the detection of anomalous behaviours amidst complex crowd interactions. Leveraging temporal coherence ensures stability in feature tracking and reduces false positives, enhancing the reliability of anomaly detection. Overall, the Lucas–Kanade optical flow approach offers a simple, efficient, and effective solution for crowd anomaly detection in diverse surveillance and monitoring applications. This method anticipates that all neighbouring pixels within a highly restricted region will exhibit identical optical flow values. The optical flow is calculated using this group of pixels, in contrast to other optical flow algorithms, which utilise all pixels in the frame. Sparse optical flow offers several benefits, including faster computation and the rapid generation of training data. The optical flow constraint for a group of pixels moving at the same velocity can be mathematically represented as in Equation (2):

$$\begin{aligned} I_x(x_1, y_1) \cdot v_x + I_y(x_1, y_1) \cdot v_y &= -I_t(x_1, y_1) \\ I_x(x_2, y_2) \cdot v_x + I_y(x_2, y_2) \cdot v_y &= -I_t(x_2, y_2) \\ &\cdots \\ I_x(x_n, y_n) \cdot v_x + I_y(x_n, y_n) \cdot v_y &= -I_t(x_n, y_n) \end{aligned} \tag{2}$$

The above Equation (2) can be reformulated using matrix vector notation as represented in Equation (3):

$$\begin{pmatrix} I_x(x_1, y_1) & I_y(x_1, y_1) \\ I_x(x_2, y_2) & I_y(x_2, y_2) \\ I_x(x_n, y_n) & I_y(x_n, y_n) \end{pmatrix} \cdot \begin{pmatrix} v_x \\ v_y \end{pmatrix} = - \begin{pmatrix} I_t(x_1, y_1) \\ I_t(x_2, y_2) \\ I_t(x_n, y_n) \end{pmatrix} \tag{3}$$

Considering that this system usually presents more equations than variables, it is frequently over-determined. The Lucas–Kanade method applies the least squares technique to address this, thus finding a balanced solution. As a result, it offers a solution to the system, as illustrated below in Equation (4):

$$\begin{aligned} A^T A v &= A^T b \\ v &= \left(A^T A \right)^{-1} A^T b \end{aligned} \tag{4}$$

The Lucas–Kanade method [36], which was primarily developed for local optimisation, may perform poorly with severe object motion due to its reliance on neighbouring pixels for gradient determination. To address this, a pyramidal structure is utilised, using a coarse-to-fine technique in which the input images are down-sampled, first with a low-pass filter and then by a factor of 2 [35]. The optical flow computation begins with the lowest-quality images and proceeds to higher resolutions, improving the flow field's accuracy.

3.1.3. Spatial Representation Using Delaunay Triangulation

To encapsulate the spatial interactions between reliable tracklets, we employ Delaunay triangulation as a spatial representation technique [37]. Tracklets, which are short sequences of object locations in consecutive frames, are crucial in our approach. The terminal location

of each tracklet is considered as a vertex in a graph. These vertices are then interconnected using the Delaunay triangulation graph. The Delaunay graph facilitates the exploration of neighbouring nodes in any direction, thereby accurately representing both local and spatial interactions. It is denoted as $g^k\left(\vartheta^k, \varepsilon^k, \mathrm{F}^k\right)$, defined by the list of node connections ε^k and the list of triple indices $\left(\mathrm{F}^k\right)$, where each triplet represents a triangle. This graph effectively captures topological changes over time, preserving the overall shape of the graph despite noise or partial occlusion. From the constructed graph g^k, local entities called cliques can be extracted, which are groups of points connected by edges $\left(\varepsilon^k\right)$ and are defined around a seed point V_i^k $\mathcal{C}^k = \left\{C\left(V_1^k\right), \ldots C\left(V_{m_k}^k\right)\right\}$. Each seed point possesses a unique local configuration, forming a first-order clique, as represented in Equation (5):

$$C\left(V_i^k\right) = \left\{V_i^k\right\} \cup \left\{V_j^k, \forall\left(V_i^k, V_j^k\right) \in \mathcal{E}^k\right\}. \tag{5}$$

Higher-order connections can be extrapolated by considering indirect neighbours, thereby forming larger cliques, as in Equation (6) [38]

$$C_n\left(V_i^k\right) = C_{n-1}\left(V_i^k\right) \cup \left\{C_1\left(V_j^k\right), \forall V_j^k \in C_{n-1} \backslash C_{n-2}\right\} \tag{6}$$

with $C_0\left(V_i^k\right) = \left\{V_i^k\right\}$ and $C_1\left(V_i^k\right) = C\left(V_i^k\right)$. The local features are temporally interconnected through trajectories, capturing short-term and long-term patterns. Spatially, they are connected through cliques of varying order. By leveraging low-level features, a dynamic graph is constructed that illustrates the temporal and spatial distributions of individuals within the crowd. This comprehensive, localised scene model facilitates the measurement of diverse crowd attributes.

3.1.4. Visual Descriptors

A wide array of visual descriptors is extracted to depict various crowd characteristics, which are subsequently used for the analysis. These descriptors encapsulate diverse crowd-related semantic data, capturing the spatial and temporal aspects of the scene. By concurrently considering both individual and interactive properties, a comprehensive analysis of the crowd is enabled. The suggested visual descriptors are divided into individual behaviour descriptors and interactive descriptors. The individual behaviour descriptors exploit temporal features like tracklets and motion vectors, while the interactive descriptors integrate spatial features.

1. Individual Behaviours
   In order to scrutinise individual actions within the crowd, we utilise two descriptors that are specifically tailored towards capturing dynamics at the individual level. Descriptors like flow directions and velocity offer insights into various aspects of individual behaviour, thereby enabling the greater comprehension of crowd interactions and movement.

   - Flow Direction
     This descriptor characterises individual behaviours in terms of the motion direction, distinguishing between smooth and chaotic motions by capturing variations in tracklets' directions [37]. This is achieved by utilising the complete history of each trajectory, which is divided into F segments $\left\{S_i^k, S_i^{k-\tau_2}, \ldots, S_i^{k-(F-1)\tau_2}\right\}$, enabling a detailed analysis of directional changes over time. Following this, the variation in flow direction $D^{\mathrm{var}}\left(V_i^k\right)$ is determined by calculating the average of the angular differences across all trajectory segments, as given in

Equation (7). This calculation provides insights into the changes in the flow direction throughout the trajectory.

$$D^{\text{var}}\left(V_i^k\right) = \frac{1}{F} \cdot \sum_{f=0}^{F-2} d_\theta\left(S_i^{K-f\tau_2}, s_i^{k-(f+1)\tau_2}\right) \tag{7}$$

where $S_i^j = \overrightarrow{v_{I_1^{j-\tau_2} Vlj}}$, and τ_2.
$F = \left\lfloor \left(\Delta t_i^k / \tau_2\right) \right\rfloor$, and $d\theta$ is defined for vectors and b as in Equation (8),

$$d_\theta(a,b) = \Delta_\theta\left(\left|\theta_{(a)} - \theta_{(b)}\right|\right) \tag{8}$$

with $\theta_{(a)}$ representing the angle between the vector a and the x-axis, $\Delta_\theta(a) = \begin{cases} a, \text{ if } a < \pi \\ 2\pi - a, \text{ otherwise.} \end{cases}$

- Velocity
  The velocity of individuals is computed using the motion vectors as defined in Equation (4). To ensure accuracy, the motion vectors from the most recent frame history are considered, specifically those that exceed a predefined threshold for acceleration and velocity. By identifying these informative motion vectors, the velocity can be determined by dividing the vector norm by the total number of frames, as given in Equation (9).

$$D^{\text{veloc}}\left(V_i^k\right) = \frac{1}{\tau_1} \cdot \left\| \overrightarrow{V_i^{k-\tau_1} V_i^k} \right\|_2 \tag{9}$$

  To optimise the computational efficiency, the Euclidean distance is calculated between each motion vector's origin and current position, rather than summing the distances across fragments [37]. Empirical evidence indicates that both methods yield equivalent results, validating our decision to use the more computationally efficient method. It is worth noting that the descriptor parameter is adapted based on the video's frame rate. This adaptation effectively compensates for perspective distortions within and between videos by incorporating perspective map weights into the motion vector's norm. The velocity descriptor plays a vital role in capturing individuals' speed, proving particularly valuable in scenarios where individuals exhibit variations in their speed due to factors such as danger or urgency.

2. Interactive Behaviours
   Besides individual behaviour descriptors, we highlight the significance of integrating interactive descriptors for a comprehensive analysis. In this study, we employ a set of five interactive descriptors, three of which capture spatiotemporal information, while the remaining two specifically target spatial properties. These descriptors draw inspiration from [14] but have unique formulations as they incorporate the local crowd representation as a fundamental element in their computation. The descriptors used in this analysis include stability, collectiveness, conflict, local density, and uniformity. They are computed locally, enabling a detailed examination of the crowd characteristics and offering valuable insights for crowd analysis.

   - Stability
     The concept of stability, as defined in [14], captures the degree of consistency in the topological structure of a crowd over time. It measures the tendency of individuals in a crowd to maintain their proximity to the same set of neighbours as time progresses. By assessing the stability property, valuable insights can be gained into the persistent patterns and relationships within the crowd, providing

a deeper understanding of its dynamics and behaviour. This characteristic is established by drawing a parallel between a Delaunay graph's topological structure and a crowd's evolving structure. To be more precise, the stability of a graph g^k at time k is calculated [37] via its graphical distance to the corresponding graph $g^{k-\tau_2}$ at time $k-\tau_2$, where τ_2 represents the interval used to fragment the trajectory for the computation of the D^{var} descriptor, as defined in Equation (7). In order to establish temporal matching between cliques, the proposed graphical distance utilises the temporal aspect of the model of the tracklets and is locally computed for each vertex. The stability of a given vertex V_i^k is calculated as the strain between the two adjacent cliques $C_n\left(V_i^k\right)$ and $C_n\left(V_i^{k-\tau_2}\right)$, computed as in the following Equation (10):

$$D^{stab}\left(V_i^k\right) = \text{dist}_g\left(C_n\left(V_i^k\right), C_n\left(V_i^{k-\tau_2}\right)\right) \tag{10}$$

Each clique is represented by a set of clockwise-oriented triangles, and this is used to define the graphical distance between two cliques $C_n\left(V_i^k\right)$ and $C_n\left(V_i^{k-\tau_2}\right)$, represented by Equation (11):

$$\begin{aligned}\text{dist}_g\left(C_n\left(V_i^{t_1}\right), C_n\left(V_i^{t_2}\right)\right) = {} & \frac{1}{\left|C_n\left(V_i^{t_1}\right)\right|} \\ & \cdot \sum_{r_{i\alpha_1} \in R_n\left(V_i^{t_1}\right)} \sum_{r_{i\beta_1} \in R_n\left(V_i^{t_2}\right)} g\left(r_{i\alpha_1}, r_{i\beta_1}\right) \\ & + \sum_{r_{i\alpha_2} \in R_n\left(V_i^{t_1}\right)} \min_{r_{i\beta_2} \in R_n\left(V_i^{t_2}\right)} g\left(r_{i\alpha_2}, r_{i\beta_2}\right)\end{aligned} \tag{11}$$

The term $\left(C_n\left(V_i^{t_1}\right)\right.$ pertains to the quantity of neighbours within the clique. The computation is carried out in two steps, as currently formulated. Initially, we determine the dissimilarity between the triangles that are matched and indexed as $i\alpha_1$ and $i\beta_1$. For the remaining triangles on both sides, where no matching is achieved through tracklets, we estimate the distance by selecting the most similar triangle as a potential corresponding candidate. The function $g(\cdot)$ denotes the measure of the distance between triangles, which is mathematically defined as the discrepancy in the cross ratio, taking into account the relative size of each triangle.

$$g\left(r_{i\alpha}, r_{i\beta}\right) = \left\|a_{i\alpha} - a_{i\beta}\right\| \cdot \left\|c_{i\alpha} - c_{i\beta}\right\| \tag{12}$$

where $a_{i\alpha}$ and $c_{i\alpha}$ in Equation (12) are the area and the cross ratio of a triangle indexed by $i\alpha$. To calculate the cross ratio, we use Equation (13):

$$c_{i\alpha} = f_{cr}\left(f_o\left(V_\alpha, v'_{i\alpha}, v'_{i\alpha'}, V_{\alpha'}\right)\right) \tag{13}$$

The cross ratio is used to measure the shape difference since it is invariant to a projective transform. For a triangle $r_{i\alpha}$, it is computed using the two ends of the boundary edges (V_α and $V_{\alpha'}$), and the projections ($v'_{i\alpha}$ and $v'_{i\alpha'}$) of the midpoints of two sides ($[V_iV_\alpha]$ and $[V_iV_{\alpha'}]$) on the boundary line. Since the cross ratio is not affected by a projective transformation, it can be used to quantify the degree of shape dissimilarity between two figures. This ratio is calculated for a triangle $r_{i\alpha}$, by locating the endpoints of the boundary edges (V_α and $V_{\alpha'}$) and the projections of the midpoints of the two sides ($[V_iV_\alpha]$ and $[V_iV_{\alpha'}]$) onto the boundary line ($v'_{i\alpha}$ and $v'_{i\alpha'}$).

- Collectiveness

The collectiveness property in crowd analysis refers to how pedestrians move together as a group. In [14], this property is quantified by computing each

individual's directional deviation from the group's global motion. Traditionally, coherent motion has been determined using predefined collective transitions. However, an alternative approach is used in this work by utilising cliques for the local computation of this descriptor [37]. Specifically, the collectiveness of a set of seed points is defined based on the degree of motion deviation from the global motion exhibited by their neighbouring points, all moving cohesively towards a common goal, as defined by the clique. By considering the local interactions within the clique, we can capture the collectiveness of the pedestrians, providing valuable insights into their coordinated movement patterns and group dynamics. The collectiveness can be computed using the following Equation (14):

$$D^{\text{collec}}\left(V_i^k\right) = \frac{1}{\left|C_n\left(V_i^k\right)\right|} \cdot \sum_{V_j^k \epsilon C_n\left(V_i^k\right)} h\left(\overrightarrow{V_i^{k-\tau_1} V_i^k}, \overrightarrow{V_j^{k-\tau_1} V_j^k}\right) \tag{14}$$

where $h(a,b) = \begin{cases} d_\theta(a,b), & \text{if } d_\theta(a,b) < T_1 \\ 0, & \text{otherwise.} \end{cases}$

- Conflict

  Conflict is an important property that captures human interactions in crowded environments, particularly when individuals are near each other. Like the approach used to compute the collectiveness descriptor, the conflict property is also computed locally [37]. A neighbour point from the corresponding clique is considered as a potential conflict point candidate for each seed point only if both of their motion vectors converge, indicating movement towards their origins. Consequently, the set of conflict points, denoted as $C_n'\left(V_i^k\right)$, forms a subset of the neighbours. Once the set of conflict points is determined, the conflict level of the central point is calculated by considering the angular difference and distance from the other points. The calculation of the conflict level is represented by Equation (15), in which the angular difference and distance from the other points are used to determine the conflict level of the central point. This helps to gain insights into the level of interpersonal interaction and potential congestion in crowded scenes, enabling a more comprehensive understanding of the dynamics and social behaviours within the crowd.

$$D^{\text{conf}}\left(V_i^k\right) = \frac{1}{\left|c\left(V_i^k\right)\right|} \sum_{V_j^k \in C'\left(V_i^k\right)} \frac{d_\theta\left(\overrightarrow{V_i^{k-\tau_1} V_i^k}, \overrightarrow{V_j^{k-\tau_1} V_j^k}\right)}{\left\|V_i^k V_j^k\right\|_2} \tag{15}$$

- Local Density

  The local density descriptor focuses solely on the spatial aspect of the model, distinguishing it from the previous interactive descriptors. It captures a critical characteristic of crowd behaviour, specifically how individuals are distributed within the scene. An approximate measure of the local density can be obtained by assessing the proximity of nearby features, as defined in [15]. This is based on the observation that when nearby features move closer together, it indicates a higher likelihood of a larger crowd gathering in that area. After removing static tracklets, the remaining raw tracklets are utilised for this purpose. Each vertex's local density, denoted as V_i^k, is estimated by applying a kernel density function to the relative positions of the vertices within their respective neighbourhood sets. Instead of using a clique as in [15], a clique from the Delaunay graph is

employed to define the neighbourhood set [37]. The calculation of the local density descriptor is shown in Equation (16) as follows:

$$D^{\text{dens}}\left(V_i^k\right) = \frac{1}{\sqrt{2\pi\sigma}} \sum_{V_j^k \in C_n\left(V_i^k\right)} \exp - \frac{\left\| V_i^k V_j^k \right\|_2}{2\sigma^2} \tag{16}$$

The contribution of each neighbouring point to the density calculation is determined by the bandwidth of the 2D Gaussian kernel, represented as σ. It is crucial to select an appropriate value for σ to ensure that feature points in close proximity to V_i^k are adequately considered in the density estimation. A larger σ is required for objects that are closer together due to the influence of perspective distortions on the detected feature points. A normalisation process is applied to the perspective map to address this issue, and the Euclidean distances between vertices are adjusted accordingly. This normalisation guarantees that the computation of the local density remains consistent regardless of the scale or resolution used, providing reliable and comparable results.

- Uniformity

The uniformity descriptor is employed to assess the coherence of the spatial distribution of regional features. It indicates whether a group exhibits a tendency to cluster together (uniform) or to fragment into smaller subgroups (nonuniform), as described in [14]. This descriptor operates at a semi-local level, focusing on the characteristics of groups rather than individual points. To achieve this, a clustering algorithm is applied to visually distinguish different types of people. Distance-based clustering, which identifies clusters based on the proximity of points, is a suitable approach as it does not require prior knowledge of the number of clusters [37]. Subsequently, for a set of p clusters denoted as clusters $\mathcal{N} = \{\mathcal{N}_1, \ldots, \mathcal{N}_P\}$, the modularity function is computed for each cluster to quantify its consistency. This evaluation considers both internal and external relationships within the clusters, providing insights into the level of coherence exhibited by the group. It is calculated as in Equation (17) given below:

$$D^{\text{unif}}\left(N_i\right) = \frac{A(N_i, N_i)}{A(N, N)} - \left(\frac{A(N_i, N)}{A(N, N)}\right) \tag{17}$$

The computation of the uniformity descriptor involves graph-based calculations. After applying the clustering procedure, each vertex V_i^k is assigned to a specific cluster. The distances between connected points are considered to determine the inter-cluster and intra-cluster relationships. If a connected point belongs to the same cluster as the seed point, the distance is used to enhance the intra-cluster weight. Conversely, if the connected point belongs to a different cluster, the distance contributes to the inter-cluster weight. This analysis is performed using a first-order clique $C_1\left(V_i^k\right)$, which facilitates the assessment of the spatial relationships between vertices within and across clusters [37].

Hence, the terms stated in Equation (17) can be reformulated as in Equation (18):

$$\begin{cases} A(N_i, N_i) = \sum_{p \in N_i} \sum_{q \in C_{1(P) q \in N_i}} \frac{1}{\|pq\|2} \\ A(N_i, N) = \sum_{p \in N_i} \sum_{q \in C_{1(P)} \notin \notin N_i} \frac{1}{\|pq\|2} \\ A(N, N) = \sum_{i \in N} \sum_{p \in N_i} \sum_{q \epsilon C_{1(p)}) } \frac{1}{\|pq\|2} \end{cases} \tag{18}$$

Shorter within-class distances and longer between-class distances are indicative of a high level of spatial uniformity within each grouping.

The proposed visual and local descriptors capture both interactive and individual properties, offering valuable insights into crowds' spatial distributions and

movements. These descriptors are particularly relevant for high-level analysis as they closely align with the crowd behaviour and semantic information. Each descriptor is encoded using a 1D histogram with 16 bins, enabling statistical computations at both local patch and global frame levels. The histogram effectively represents the distribution of the descriptor values within specific regions, making it a suitable choice for our analysis. After scaling, the histograms can be concatenated to form a feature vector. The current methodology serves as a robust foundation for the detection of crowd anomalies across diverse scenarios.

3.2. Dimensionality Reduction and Classification

In the process of our crowd anomaly detection analysis, we adopted dimensionality reduction techniques to handle the gathered descriptors. These techniques aim to reduce the complexity of the feature set while retaining essential information. We explored two distinct approaches for dimensionality reduction: principal component analysis (PCA) [39] and autoencoders [40].

The first method employed was principal component analysis (PCA). PCA seeks to transform the original features into a new set of uncorrelated variables, known as principal components, while retaining as much variance as possible. By projecting the data onto these components, we effectively compress the information into a lower-dimensional space.

The second approach involves utilising autoencoders, a type of neural network designed to learn efficient representations of the input data. Autoencoders consist of an encoder network that compresses the input into a latent space representation and a decoder network that reconstructs the original input from this representation. By training the autoencoder to minimise the reconstruction error, it learns to capture the most salient features of the data in the latent space.

After applying dimensionality reduction using PCA and autoencoders, we move on to the classification stage. In this step, we leverage the reduced-dimensional feature set to train neural network classifiers. Specifically, we utilise neural networks like the multi-layer perceptron (MLP) [41]. Neural networks are well suited to handle non-linear problems and extract intricate patterns from the input data.

The versatility of neural networks is particularly beneficial in detecting anomalies in crowd behaviour. They excel in uncovering nuanced relationships within data and identifying unusual crowd dynamics that might remain unnoticed by traditional methods. Due to their hierarchical architectures, neural networks can capture both intricate details and higher-level representations, providing a comprehensive understanding of crowd behaviour.

Throughout our study, we conducted evaluations using neural networks with varying numbers of hidden layers to determine the optimal configuration for our analysis. Through systematic testing and performance comparisons, we identified the most effective approach for crowd anomaly detection. This thorough approach ensured that our methodology was robust and capable of addressing the complexities of detecting anomalies within crowd dynamics.

4. Results

4.1. Datasets

The effectiveness of the proposed visual descriptors for crowd anomaly detection in challenging crowded environments is evaluated using two state-of-the-art crowd datasets. Two widely recognised public datasets, the University of Minnesota (UMN) dataset [16] and the violence in crowds dataset [42], are employed for this evaluation. The UMN dataset is highly regarded in the field. This dataset offers a comprehensive collection of videos that capture various crowd behaviours, including both normal and abnormal instances. It consists of eleven videos recorded in diverse indoor and outdoor settings. The videos are categorised into three scenes: Scene 1 (videos 1 to 2), Scene 2 (videos 3 to 8), and Scene 3 (videos 9 to 11). These videos depict scenarios such as crowds running in a single direction

or dispersing from a centralised location, providing a wide range of typical and out-of-the-ordinary segments. Researchers can extract valuable insights and evaluate the performance of crowd anomaly detection algorithms using this dataset.

The violence in crowds dataset is sourced from YouTube. This dataset offers a diverse collection of challenging real-world viewing conditions. It includes footage compiled from various settings and surveillance operations, specifically focusing on crime and violence. The dataset has been meticulously curated, following established standards for violence classification. It consists of videos divided into distinct groups, with an equal number of violent and peaceful videos, ensuring a balanced representation of different scenarios. Researchers can thoroughly evaluate the proposed visual descriptors for crowd anomaly detection by utilising these state-of-the-art datasets.

The UMN dataset comprehensively assesses various crowd behaviours, while the violence in crowds dataset offers realistic and challenging real-world scenarios related explicitly to crime and violence. Leveraging these datasets enables the analysis and comparison of different approaches, facilitating advancements in the field of crowd anomaly detection.

4.2. Crowd Representation

In this section, we present the effectiveness of the proposed visual descriptors, as depicted in Figure 1, for crowd anomaly detection. The utilisation of the Delaunay triangulation method to capture the spatial proximity and enhance the neighbourhood representation is illustrated in Figure 2.

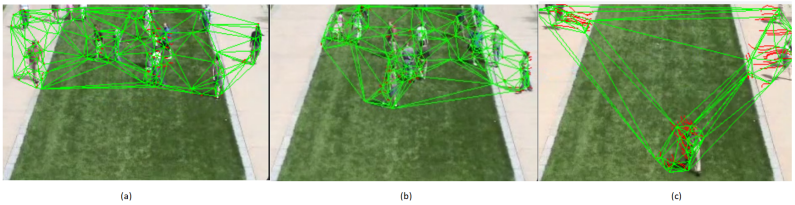

Figure 2. Spatial proximity using Delaunay triangulation to better represent the neighbourhood. (**a**,**b**) represent Delaunay triangles on crowd scenes with no abnormal activities. (**c**) represents Delaunay triangles on crowd scene with abnormal activity. Green lines: Delaunay triangle edges,structuring crowd area. Red dots: indicate anomaly hotspots in crowd dynamics.

In Figure 2, we illustrate the concept of spatial proximity within a crowd using Delaunay triangulation, a technique that enhances the understanding of the crowd's neighbourhood dynamics. The figure serves to visually explain how the crowd's behaviour is interpreted through feature points and visual descriptors, which are calculated and then represented using Delaunay triangles.

Figure 2a,b display Delaunay triangles overlaid onto a crowd scene where no abnormal activities are taking place. The crowd's behaviour is analyzed through sparse feature tracking, allowing us to define significant points that represent the scene. Delaunay triangulation helps to delineate the spatial relationships between these points, forming triangles that reflect the crowd's structure and arrangement.

In Figure 2c, we observe the impact of abnormal behaviour within the crowd. As an anomalous event occurs, the crowd's typical cohesion breaks down, causing the individuals to scatter and move in various directions. This change in behaviour is effectively captured and illustrated through Delaunay triangulation. The method of sparse feature tracking allows us to follow these scattered movements and depict them as changes in the arrangement of the Delaunay triangles.

A noteworthy feature that aids in discerning abnormal behaviour is the length of the edges within the Delaunay triangles. In a normal crowd scenario, the edges maintain a consistent length, reflecting the crowd's relatively uniform distribution. However, when an abnormal event disrupts the crowd, such as scattering due to a sudden disturbance, the edges' lengths can alter significantly and rapidly. This edge length variation

is clearly depicted in the Delaunay triangles illustrated in Figure 2c. The uniformity of the edge lengths is disrupted, with certain edges becoming notably longer than others. This disparity in the edge lengths signifies the occurrence of abnormal behaviour within the crowd. The magnitude and rapidity of these changes are visually evident in the triangulation representation, providing a valuable tool to understand the dynamics of the crowd behaviour.

4.3. Classification Results

The following tables display the accuracy and area under the curve (AUC) scores obtained for each combination of feature point extraction methods (SIFT, FAST, AKAZE) applied to the two datasets utilised in this analysis. The results are presented for three different configurations of the neural network: NN, 128/NN, and PCA/NN. The NN configuration uses the neural network without any dimensionality reduction. In the 128/NN configuration, dimensionality reduction is performed using an autoencoder to reduce the feature space to 128 dimensions. Lastly, in the PCA/NN configuration, dimensionality reduction is achieved by applying principal component analysis (PCA) with a variance threshold of 95. The results obtained for the UMN dataset and the violence in crowds dataset are presented in Tables 2 and 3, respectively. Five-fold cross-validation is used in the classification. There are a total of five tests taken, with four training sets used for each. The average prediction accuracy (ACC) and the area under the curve (AUC) are reported for the results. The results of Scene 1 of the UMN dataset show that all three feature point extraction methods (SIFT, FAST, and AKAZE) achieve high accuracy and AUC scores across different neural network configurations. The NN configuration generally performs slightly better than the 128/NN and PCA/NN configurations. Among the feature point extraction methods, SIFT consistently exhibits the highest accuracy and AUC scores, followed closely by FAST and AKAZE. This indicates that SIFT is particularly effective in capturing spatial information and detecting anomalies in crowd behaviour for the Scene 1 dataset. The results on Scene 2 of the UMN dataset show comparable performance among the feature point extraction methods and neural network configurations. SIFT consistently achieves the highest accuracy and AUC scores, followed by FAST and AKAZE. However, the overall accuracy and AUC scores are slightly lower compared to Scene 1. This suggests that Scene 2 may pose more significant challenges in crowd anomaly detection, requiring more robust feature extraction methods and network configurations.

The results for Scene 3 indicate consistent performance across the feature point extraction methods and neural network configurations. AKAZE demonstrates the highest accuracy and AUC scores, followed by SIFT and FAST. The overall performance in Scene 3 is similar to that in Scene 1, with slightly lower scores. This suggests that the crowd behaviour in Scene 3 may exhibit similar characteristics to that in Scene 1 but with some variations. The results reveal that SIFT consistently outperforms FAST and AKAZE in terms of accuracy and AUC scores across all scenes. Notably, AKAZE shows promise, particularly in Scene 3. These findings highlight the robustness of SIFT and AKAZE as feature point extraction methods for crowd anomaly detection in various scenes. In terms of the neural network configurations, the NN configuration consistently achieves the highest accuracy and AUC scores, followed by the 128/NN and PCA/NN configurations. This suggests that employing a deeper neural network without dimensionality reduction captures more intricate patterns in crowd behaviour. However, the differences between the configurations are relatively minor, indicating that dimensionality reduction techniques such as 128/NN and PCA/NN can yield competitive results with lower computational complexity. Overall, combining SIFT with the NN configuration emerges as a particularly effective approach for crowd anomaly detection, delivering superior performance and enabling the identification of complex crowd behaviours.

Table 2. Classification results for the UMN dataset (Scene 1, Scene 2, and Scene 3) with descriptors and methods. Best accuracy is given in bold.

Scene 1	**NN**	**128/NN**	**PCA/NN**
SIFT	ACC: **0.995** AUC: **0.990**	ACC: 0.985 AUC: 0.979	ACC: 0.986 AUC: 0.983
FAST	ACC: 0.988 AUC: 0.977	ACC: 0.986 AUC: 0.973	ACC: 0.988 AUC: 0.978
AKAZE	ACC: 0.953 AUC: 0.904	ACC: 0.952 AUC: 0.943	ACC: 0.966 AUC: 0.955
Scene 2	**NN**	**128/NN**	**PCA/NN**
SIFT	ACC: 0.936 AUC: 0.901	ACC: 0.926 AUC: 0.893	ACC: 0.924 AUC: 0.885
FAST	ACC: 0.954 AUC: 0.937	ACC: **0.961** AUC: **0.943**	ACC: 0.951 AUC: 0.931
AKAZE	ACC: 0.937 AUC: 0.892	ACC: 0.918 AUC: 0.892	ACC: 0.917 AUC: 0.872
Scene 3	**NN**	**128/NN**	**PCA/NN**
SIFT	ACC: **0.990** AUC: **0.976**	ACC: 0.983 AUC: 0.980	ACC: 0.985 AUC: 0.968
FAST	ACC: 0.981 AUC: 0.956	ACC: 0.983 AUC: 0.973	ACC: 0.983 AUC: 0.959
AKAZE	ACC: 0.975 AUC: 0.914	ACC: **0.990** AUC: **0.988**	ACC: 0.985 AUC: 0.968

Table 3. Classification results for the violence in crowds dataset (with descriptors and methods). Best accuracy is given in bold.

Methods	**NN**	**128/NN**	**PCA/NN**
SIFT	ACC: 0.803 AUC: 0.805	ACC: 0.799 AUC: 0.804	ACC: 0.848 AUC: 0.846
FAST	ACC: 0.848 AUC: 0.851	ACC: 0.844 AUC: 0.847	ACC: 0.844 AUC: 0.851
AKAZE	ACC: 0.856 AUC: 0.862	ACC: **0.885** AUC: **0.895**	ACC: 0.873 AUC: 0.878

The classification results for the violence in crowds dataset, as presented in Table 2, exhibit some variations compared to the UMN dataset. In the case of the violence in crowds dataset, the performance of all three feature point extraction methods, namely SIFT, FAST, and AKAZE, shows slightly lower accuracy and AUC scores compared to the UMN dataset. AKAZE consistently outperforms FAST and SIFT, demonstrating superior accuracy and AUC scores. This observation suggests that AKAZE possesses inherent capabilities that enable it to effectively capture the distinctive characteristics of the crowd anomalies within the violence in crowds dataset. Shifting our focus to the neural network configurations, we observe that the PCA/NN configuration consistently outperforms both the NN and 128/NN configurations in terms of accuracy and AUC scores. This outcome highlights the significance of employing dimensionality reduction techniques, specifically through principal component analysis (PCA), to enhance the performance of crowd anomaly detection on the violence in crowds dataset. By reducing the dimensionality of the data, PCA facilitates the extraction of salient features, thereby leading to improved accuracy and AUC scores. The obtained results for the violence in crowds dataset emphasise the importance of carefully selecting and evaluating the appropriate combination of feature point extraction methods and neural network configurations for each dataset. It is evident that the performance of these methods can be subject to variations based on the dataset's characteristics and the specific nature of crowd anomalies. Therefore, it is important to

exercise caution and conduct thorough evaluations to identify the optimal combination of methods that will yield the highest performance in crowd anomaly detection tasks.

Figure 3 encapsulates the results derived from the application of three classification approaches used in this work on the UMN and violence in crowds datasets, using the SIFT, FAST, and AKAZE descriptors. The performance of the crowd anomaly detection methods is generally higher on the UMN dataset compared to the violence in crowds dataset. The UMN dataset provides well-defined scenarios, while the violence in crowds dataset has more diverse and challenging real-world conditions. The controlled settings in the UMN dataset and the more precise differentiation between normal and abnormal behaviours contribute to the higher performance. On the other hand, the lower performance on the violence in crowds dataset may be attributed to factors such as poor video quality, occlusion, and variability in the surveillance scenarios. It is essential to address these specific challenges in each dataset to improve the performance of crowd anomaly detection methods. When examining the three scenarios individually, it was observed that Scenes 1 and 3 (outdoor) led to perfect accuracy, whereas Scene 2 (indoor) did not. This result is intriguing; however, the limited complexity of the videos hindered our ability to conduct a more in-depth investigation.

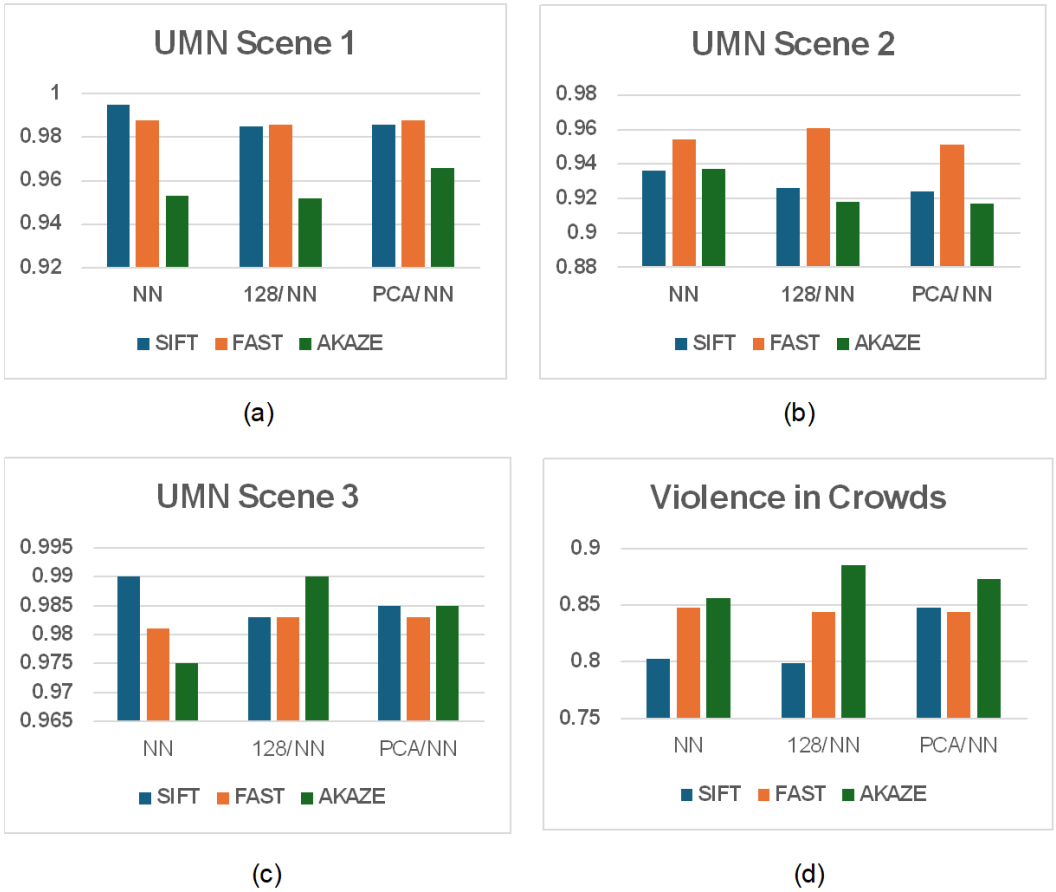

Figure 3. Summary of the performance of CAD on the UMN dataset (**a**–**c**) and on violence in crowds (**d**) using the SIFT, FAST, and AKAZE descriptors.

4.4. Result Comparison

In an exhaustive analysis of the UMN and violence in crowds datasets, a meticulous comparison between the proposed approach and the top-performing techniques from recent studies has been conducted, unveiling subtle distinctions and noteworthy achievements in their respective performance. Our approach has been rigorously compared with recent advancements in the literature, as delineated in Tables 4 and 5. These tables articulate

the AUC and accuracy metrics for the CAD methodology applied to both the UMN and violence in crowds datasets. Within the context of the UMN and violence in crowds datasets, our approach consistently manifests superior performance when juxtaposed with the best alternative method. Specifically, for the UMN dataset, our approach attains an AUC of 99.0%, eclipsing the best results procured with [37], which include an AUC of 98.72%, 95.21%, and 97.50% for Scene 1, Scene 2, and Scene 3, respectively. To further underscore the consistency in the method's superior performance, we have delineated the AUC and accuracy results for all three scenes of the UMN dataset. The comparative analysis of the accuracy also corroborates the superior performance relative to other methods. Our approach surpassed the other techniques, achieving 99.5%, 96.5%, and 99% for Scene 1, Scene 2, and Scene 3 of the UMN dataset, respectively. In contrast, the advantage of our approach is more pronounced in the violence in crowds dataset, where it transcends the top results with an AUC of 89.5% compared to 88% and accuracy of 88.5% compared to 84.44%. The consistent performance of our approach across both datasets indicates a higher level of effectiveness in classification tasks, effectively leveraging unique features and techniques that contribute to its advantages.

Table 4. A comparison of the results on the UMN dataset.

Methods	AUC %	Accuracy %
Optical Flow [16]	84.0	/
SFM [16]	96.0	/
Sparse Reconstruction [43]	97.0	/
Visual Descriptors [37]	Scene 1: 0.98.72, Scene 2: 95.21, Scene 3: 97.50	/
Optical Flow, GAN [27]	98.1	/
Our Approach	Scene 1: 99.0, Scene 2: 94.3, Scene 3: 98.8	Scene 1: 99.5, Scene 2: 96.1, Scene 3: 99.0

Table 5. A comparison of the results on the violence in crowds dataset.

Method	AUC %	Accuracy %
VIF [42]	85	81.30
Visual Descriptors [37]	88	84.44
Our Approach	89.5	88.5

5. Conclusions and Future Study

This study presents a pioneering approach to crowd anomaly detection, capitalising on the fusion of visual descriptors and neural networks. The method's efficacy, showcased through rigorous experiments on the established UMN and violence in crowds datasets, underscores its potential in accurately identifying aberrant crowd behaviours. The superiority of the SIFT and AKAZE descriptors, combined with the competitive performance of the neural network configurations, highlights the robustness of our approach. As such, this research contributes significantly to the field's evolving landscape, providing a strong foundation for the future development of crowd anomaly detection systems.

As the research progresses, several promising avenues for further exploration in the field have been identified:

1. Transitioning from conventional neural networks to advanced deep learning architectures promises enhanced performance;
2. Incorporating cutting-edge feature extraction methods can provide more comprehensive insights into crowd behaviour patterns;
3. To enable real-world applicability, rigorous testing on real-time datasets that encompass distortions and complexities is necessary;
4. The fusion of multimodal data and the extension of the methodology to detect various types of anomalies hold substantial potential.

We believe that these future directions will collectively pave the way for refined, adaptable, and multifaceted crowd anomaly detection systems.

Author Contributions: Conceptualization and methodology, S.A. and S.L.; software, S.A.; Formal analysis, S.A. and S.L.; Writing—original draft preparation, S.A.; Writing—review and editing, S.A., S.L., P.G. and S.C.; Supervision, S.L., P.G. and S.C.; Project administration, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://mha.cs.umn.edu/proj_events.shtml.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CAD	Crowd Anomaly Detection
PCA	Principle Component Analysis
FAST	Features from the Accelerated Segment Test
SIFT	Scale-Invariant Feature Transform
AKAZE	Accelerated-KAZE
NN	Neural Network
ACC	Accuracy
AUC	Area Under the Curve

References

1. Aldayri, A.; Albattah, W. Taxonomy of Anomaly Detection Techniques in Crowd Scenes. *Sensors* **2022**, *22*, 6080. [CrossRef] [PubMed]
2. Altowairqi, S.; Luo, S.; Greer, P. A Review of the Recent Progress on Crowd Anomaly Detection. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, 3448–3470. [CrossRef]
3. Kaltsa, V.; Briassouli, A.; Kompatsiaris, I.; Hadjileontiadis, L.J.; Strintzis, M.G. Swarm intelligence for detecting interesting events in crowded environments. *IEEE Trans. Image Process.* **2015**, *24*, 2153–2166. [CrossRef] [PubMed]
4. Ribeiro, P.C.; Audigier, R.; Pham, Q.C. RIMOC, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance. *Comput. Vis. Image Underst.* **2016**, *144*, 121–143. [CrossRef]
5. Li, T.; Chang, H.; Wang, M.; Ni, B.; Hong, R.; Yan, S. Crowded scene analysis: A survey. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *25*, 367–386. [CrossRef]
6. Loy, C.C.; Xiang, T.; Gong, S. Detecting and discriminating behavioural anomalies. *Pattern Recognit.* **2011**, *44*, 117–132. [CrossRef]
7. Choi, W.; Savarese, S. A unified framework for multi-target tracking and collective activity recognition. In Proceedings of the Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Proceedings, Part IV 12; Springer: Berlin/Heidelberg, Germany, 2012; pp. 215–230.
8. Real-Time Crowd Simulation: A Review. Available online: http://www.leggettnet.org.uk/docs/crowdsimulation.pdf (accessed on 31 December 2023).
9. Krausz, B.; Bauckhage, C. Loveparade 2010: Automatic video analysis of a crowd disaster. *Comput. Vis. Image Underst.* **2012**, *116*, 307–319. [CrossRef]
10. Benabbas, Y.; Ihaddadene, N.; Djeraba, C. Motion pattern extraction and event detection for automatic visual surveillance. *EURASIP J. Image Video Process.* **2010**, *2011*, 163682. [CrossRef]
11. Alcantarilla, P.F.; Solutions, T. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell* **2011**, *34*, 1281–1298.
12. Rao, A.S.; Gubbi, J.; Marusic, S.; Palaniswami, M. Crowd event detection on optical flow manifolds. *IEEE Trans. Cybern.* **2015**, *46*, 1524–1537. [CrossRef]
13. Mousavi, H.; Mohammadi, S.; Perina, A.; Chellali, R.; Murino, V. Analyzing tracklets for the detection of abnormal crowd behavior. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 148–155.
14. Shao, J.; Change Loy, C.; Wang, X. Scene-independent group profiling in crowd. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2219–2226.

15. Fradi, H.; Dugelay, J.L. Spatial and temporal variations of feature tracks for crowd behavior analysis. *J. Multimodal User Interfaces* **2016**, *10*, 307–317. [CrossRef]
16. Mehran, R.; Oyama, A.; Shah, M. Abnormal crowd behavior detection using social force model. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 935–942.
17. Wu, S.; Moore, B.E.; Shah, M. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 2054–2060.
18. Mehran, R.; Moore, B.E.; Shah, M. A Streakline Representation of Flow in Crowded Scenes. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010*; Proceedings, Part III 11; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6313, pp. 439–452.
19. Bendali-Braham, M.; Weber, J.; Forestier, G.; Idoumghar, L.; Muller, P.A. Recent trends in crowd analysis: A review. *Mach. Learn. Appl.* **2021**, *4*, 100023. [CrossRef]
20. Feng, J.; Wang, D.; Zhang, L. Crowd Anomaly Detection via Spatial Constraints and Meaningful Perturbation. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 205. [CrossRef]
21. Singh, K.; Rajora, S.; Vishwakarma, D.K.; Tripathi, G.; Kumar, S.; Walia, G.S. Crowd anomaly detection using Aggregation of Ensembles of fine-tuned ConvNets. *Neurocomputing* **2020**, *371*, 188–198. [CrossRef]
22. Alhothali, A.; Balabid, A.; Alharthi, R.; Alzahrani, B.; Alotaibi, R.; Barnawi, A. Anomalous event detection and localization in dense crowd scenes. *Multimed. Tools Appl.* **2023**, *82*, 15673–15694. [CrossRef]
23. Alafif, T.; Hadi, A.; Allahyani, M.; Alzahrani, B.; Alhothali, A.; Alotaibi, R.; Barnawi, A. Hybrid classifiers for spatio-temporal real-time abnormal behaviors detection, tracking, and recognition in massive hajj crowds. *arXiv* **2022**, arXiv:2207.11931.
24. Hao, Y.; Li, J.; Wang, N.; Wang, X.; Gao, X. Spatiotemporal consistency-enhanced network for video anomaly detection. *Pattern Recognit.* **2022**, *121*, 108232. [CrossRef]
25. Traoré, A.; Akhloufi, M.A. Violence detection in videos using deep recurrent and convolutional neural networks. In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 11–14 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 154–159.
26. Doshi, K.; Yilmaz, Y. A modular and unified framework for detecting and localizing video anomalies. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 3982–3991.
27. Alafif, T.; Alzahrani, B.; Cao, Y.; Alotaibi, R.; Barnawi, A.; Chen, M. Generative adversarial network based abnormal behavior detection in massive crowd videos: A Hajj case study. *J. Ambient Intell. Humaniz. Comput.* **2021**, *13*, 4077–4088. [CrossRef]
28. Ullah, W.; Ullah, A.; Hussain, T.; Khan, Z.A.; Baik, S.W. An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos. *Sensors* **2021**, *21*, 2811. [CrossRef]
29. Bhuiyan, M.R.; Abdullah, J.; Hashim, N.; Al Farid, F.; Samsudin, M.A.; Abdullah, N.; Uddin, J. Hajj pilgrimage video analytics using CNN. *Bull. Electr. Eng. Inform.* **2021**, *10*, 2598–2606. [CrossRef]
30. Sikdar, A.; Chowdhury, A.S. An adaptive training-less framework for anomaly detection in crowd scenes. *Neurocomputing* **2020**, *415*, 317–331. [CrossRef]
31. Xiao, X. Abnormal Event Detection and Localization Based on Crowd Analysis in Video Surveillance. *J. Artif. Intell. Pract.* **2023**, *6*, 58–65.
32. Rosten, E.; Porter, R.; Drummond, T. Faster and better: A machine learning approach to corner detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *32*, 105–119. [CrossRef] [PubMed]
33. Rosten, E.; Drummond, T. Fusing points and lines for high performance tracking. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), Beijing, China, 17–21 October 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 2, pp. 1508–1515.
34. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
35. Sharmin, N.; Brad, R. Optimal filter estimation for Lucas-Kanade optical flow. *Sensors* **2012**, *12*, 12694–12709. [CrossRef]
36. Lucas, B.D.; Kanade, T. An iterative image registration technique with an application to stereo vision. In Proceedings of the 7th International Joint Conference on Artificial Intelligence, IJCAI'81, San Francisco, CA, USA, 24–28 August 1981; Volume 2, pp. 674–679.
37. Fradi, H.; Luvison, B.; Pham, Q.C. Crowd behavior analysis using local mid-level visual descriptors. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 589–602. [CrossRef]
38. Shin, D.; Tjahjadi, T. Similarity invariant delaunay graph matching. In Proceedings of the Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, SSPR & SPR 2008, Orlando, FL, USA, 4–6 December 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 25–34.
39. Partridge, M.; Calvo, R.A. Fast dimensionality reduction and simple PCA. *Intell. Data Anal.* **1998**, *2*, 203–214. [CrossRef]
40. Wang, Y.; Yao, H.; Zhao, S. Auto-encoder based dimensionality reduction. *Neurocomputing* **2016**, *184*, 232–242. [CrossRef]
41. Aljuaid, H.; Akhter, I.; Alsufyani, N.; Shorfuzzaman, M.; Alarfaj, M.; Alnowaiser, K.; Jalal, A.; Park, J. Postures anomaly tracking and prediction learning model over crowd data analytics. *PeerJ Comput. Sci.* **2023**, *9*, e1355. [CrossRef]

42. Hassner, T.; Itcher, Y.; Kliper-Gross, O. Violent flows: Real-time detection of violent crowd behavior. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1–6.
43. Cong, Y.; Yuan, J.; Liu, J. Sparse reconstruction cost for abnormal event detection. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 3449–3456.