

Support Vector Machine-Based Segmentation for Accurate Crowd Density Detection in Urban Spaces

Gourav Kalra^{1, †}, Rajeev Yadav^{2, †}, Satish Kumar Alaria^{3, *, †}

¹M. Tech. Scholar, Department of CSE, Arya College of Engineering, Jaipur, Rajasthan

²Professor, Department of CSE, Arya College of Engineering, Jaipur, Rajasthan

³Computer Instructor, Education Department, Government of Rajasthan, Rajasthan

Abstract

Estimating crowd density has become increasingly important in fields like public safety, event management, and urban planning. Accurate detection of crowd density helps in making informed decisions and ensuring safety in crowded areas. This study proposes a novel method for crowd density detection using segmentation and classification based on a Support Vector Machine (SVM). The method involves two key steps: crowd segmentation and density categorization. During segmentation, advanced image processing techniques like background removal and region-based segmentation extract crowd sections from input images or video frames. These segmented areas are then classified using an SVM model, known for handling complex data. The model is trained on a diverse dataset containing images with varying crowd densities. The approach captures crucial spatial and contextual information, and extensive testing on various datasets has demonstrated its accuracy and resilience in dynamic crowd scenarios. The proposed SVM-based method can be implemented in real-time, making it valuable for applications requiring quick decisions. This technique offers a reliable and efficient solution for crowd density detection, with significant implications for event management, public safety, and urban planning in congested environments.

Keywords

Crowd density detection, Support Vector Machine, crowd segmentation, image processing, real-time detection, region-based segmentation, urban planning, machine learning.

1. Introduction


The world has undergone rapid urbanization over the past two decades, leading to a significant increase in city populations. As cities become more crowded, the need for effective surveillance systems has grown, particularly to monitor people's movements and behaviors in public spaces, ensuring the safety and security of individuals and their possessions. Surveillance has become an integral part of maintaining public safety, with both public and private entities worldwide regularly employing video cameras for this purpose. However, traditional surveillance systems heavily rely on human operators, whose effectiveness can vary depending on their alertness and the available manpower. Given these limitations, modern surveillance is transitioning towards smart systems equipped with advanced technologies like intelligent video analysis, which enable automated decision-making without continuous human intervention.


Smart surveillance systems can be broadly categorized into two types: visual-based and multimodal. Visual-based systems utilize computer vision algorithms to process video data from cameras and drones in real time, offering solutions like facial recognition and license plate identification. On the other hand, multimodal systems integrate various data sources, including motion and audio sensors, alongside video data to provide comprehensive real-time insights. Companies like IBM and Intel have pioneered technologies that can detect traffic incidents, optimize routes, and even identify crime-related events using these advanced surveillance systems.

SCCTT-2024: International Symposium on Smart Cities, Challenges, Technologies and Trends, 29th Nov 2024, Delhi, India

* Corresponding author.

† These authors contributed equally.

 gkalra144@gmail.com (Gourav Kalra); rajeevtpo@gmail.com (Dr. Rajeev Yadav); satish.alaria@gmail.com (Satish Kumar Alaria)

 0009-0008-4926-1929 (Gourav Kalra); 0000-0002-1976-4065 (Dr. Rajeev Yadav); 0000-0001-8298-1364 (Satish Kumar Alaria)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

In today’s world, smart surveillance plays a critical role, particularly in monitoring crowds. This becomes especially relevant during large public gatherings, where the potential for disasters, accidents, or criminal activity increases. Effective crowd control is vital in these scenarios, as seen in airports, concert venues, and religious gatherings. As crime, terrorism, and natural disasters rise, smart surveillance systems must rely on robust algorithms to manage and predict crowd behavior. The analysis of crowd behavior is a key focus of this chapter. It begins by defining different types of crowds, highlighting their unique characteristics and behaviors in various contexts. A deeper exploration of collective crowd behavior from a psychological standpoint follows, offering insights into how crowds react in specific situations. From there, the discussion shifts to the challenges of analyzing crowd behavior through video footage, including the complexities involved in cognitive modeling for crowd behavior analysis. Ultimately, this chapter sets the stage for understanding the motivations behind this research and the primary contributions of the proposed approach.



Figure 1. Common applications of smart surveillance

A crowd is a large group of people gathered in one location, exhibiting a range of behaviors and attitudes. Based on movement patterns, crowds can generally be divided into two categories: dynamic and stationary. Dynamic crowds are constantly in motion and can be either organized or unstructured. In organized crowds, such as marathons or rallies, individuals move in the same direction, maintaining consistent behavior over time. In unstructured crowds, such as those seen in airports or stadiums, individuals move in various directions with varying spatiotemporal characteristics. Stationary crowds, on the other hand, include audiences at rallies, concerts, or plays, where people remain in one place for a period of time.

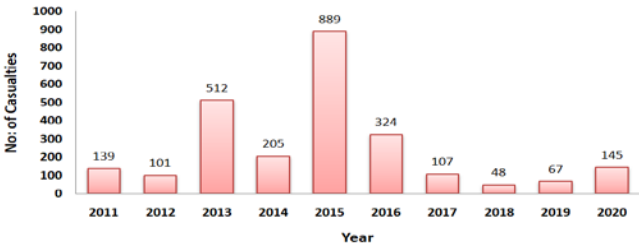


Figure 2: Statistics of crowd disasters

The characteristics of a crowd—such as its size, density, location, and time—are critical in understanding its behavior. Crowds can also be categorized into active and passive groups based on the behavior of their participants. While passive crowds primarily observe without engaging in

activities, active crowds may exhibit behaviors ranging from aggression to panic or expressive actions, such as cheering at a concert or participating in religious events.

Analyzing crowd behavior is essential for smart surveillance systems, as it helps authorities understand crowd dynamics, develop control measures, and prevent crowd-related disasters. The behavior of a crowd is often influenced by the context in which it forms. For instance, in a shopping district, people might move peacefully alongside one another, while in a stadium, fans may express intense emotions in response to the game. These varying behaviors highlight the need for smart surveillance systems capable of monitoring and analyzing different crowd scenarios in real time. Crowd behavior is inherently complex, as it depends on the context and setting. Monitoring and understanding collective crowd behavior in both regular and emergency situations is challenging, particularly when individual identification is difficult in dense crowds. Over time, psychologists and sociologists have proposed numerous theories to explain crowd behavior. One of the earliest and most popular is Le Bon's Group Mind Theory, which suggests that crowd members lose their individual identity and are easily influenced by a leader. Freud's theories support the notion that individuals in a crowd open their unconscious minds, yet maintain control over their actions. McPhail's Pre-Disposition hypothesis posits that aggressive behavior in crowds is influenced by individual dispositions toward antisocial behavior. In contrast, the Emergent-Norm hypothesis suggests that crowds consist of people with common interests, leading to distinctive behavior patterns. These collective behaviors can often become impulsive, unpredictable, and volatile. Understanding these behaviors is crucial for developing smart surveillance systems that can anticipate and prevent crowd-related issues. Such systems must account for the social and psychological components of group behavior, including how crowd members concentrate their attention on a common cause, exchange ideas rapidly, and form homogenous groups based on shared beliefs and behaviors. Machine learning, particularly Support Vector Machines (SVM), is a key technology used in crowd behavior analysis. SVM models create distinct classes from input data features, enabling the classification of various crowd behaviors. Deep learning, especially Convolutional Neural Networks (CNN), is another powerful tool for crowd behavior research. CNN mimics the structure of neurons in the human visual cortex, allowing for the hierarchical processing of input data. Long Short-Term Memory (LSTM) networks, which resemble the brain's short-term memory, are also used to analyze and predict crowd behavior based on past events. These advanced AI models enable the system to learn from past examples, making it more effective in predicting crowd behaviors and detecting anomalies. This research is motivated by the need to develop smart surveillance systems capable of detecting crowd anomalies, evaluating behaviors in real-time, and providing timely alerts. The current pandemic crisis has also highlighted the importance of monitoring crowd behavior to ensure public safety, especially in terms of enforcing social distancing and detecting free-standing conversation groups. By combining video, audio, and other sensor data, this study aims to develop a comprehensive crowd behavior analysis system that can operate effectively in a variety of challenging scenarios.

In conclusion, the introduction of cognitive modeling and AI technologies into surveillance systems offers the potential to greatly improve crowd management, enhancing public safety and preventing disasters in crowded settings.

Related Works

Crowd behavior evaluation through computer vision techniques has been explored through various research studies, with each contributing to a broader understanding of how anomalies and movement patterns in large groups can be detected and analyzed. A review of these works highlights both advancements in this domain and the identification of gaps that future research must address. For instance, a framework [1] for video event identification that proved essential for high-level video indexing and retrieval. This framework addressed challenges such as skewed data distribution and loose video structure, automating the determination of crucial thresholds that were typically manually

set in conventional Association Rule Mining (ARM) techniques. The reduction in manual intervention in video analysis was a critical advancement towards fully autonomous video content analysis.

The Trajectory Segmentation and Multi-Instance Learning (TRASMIL) framework, which allowed for precise and adaptable local anomaly detection. This three-step method was found to outperform existing techniques in terms of identifying trajectories with local abnormalities [2]. TRASMIL emphasized the importance of trajectory-based anomaly detection for accurately understanding crowd movement and behaviors. Similarly, a semantic video [3] segmentation method that relied on One-Class Classification (OCC) techniques for identifying events through frame-by-frame processing. Their work highlighted the effectiveness of OCC in detecting unsupervised events, particularly through the use of Temporal Self-Similarity Maps (TSSMs), which were evaluated using a publicly available thermal video dataset. The use of OCC for unsupervised event detection opened new avenues for handling video data with minimal prior knowledge of the scene.

A dynamic time interval segmentation technique to improve item anomaly detection. Their segmentation approach dynamically validated the time interval length, grouping successive attack ratings [4]. While effective, [5] the robustness of anomaly detection methods had received limited attention in terms of accuracy and consistency, pointing to a gap that future research must address. Meanwhile, [6] contributed by proposing an unsupervised method for scene analysis and anomaly detection in traffic video data recorded by stationary security cameras. By using local Hierarchical Dirichlet Process (HDP) models, Kaltsa et al. were able to achieve improved accuracy with lower computational costs, emphasizing the need for efficient solutions in processing large amounts of traffic video data.

Other researchers have approached the problem from a probabilistic standpoint. A probabilistic framework for identifying [7] local spatiotemporal anomalies. This framework allowed for a more refined decision-making process by identifying ideal decision-making procedures based on score functions obtained from nearby neighbors' distances. The work emphasized the importance of spatiotemporal scales in accurately identifying anomalies. Spatiotemporal anomaly detection using scalable aggregation [8] and geolocated text visualization. They proposed a cluster analysis technique to automatically discover anomalies and presented these findings through a global map depiction. Their work demonstrated how scalable visualization could assist analysts in categorizing and evaluating event candidates on a global scale.

The visualization of social media data with a visual analytics technique [9], which allowed users to extract significant subjects from a chosen collection of communications. By applying Latent Dirichlet Allocation (LDA) and visualizing topic time series, analysts could better understand abnormal events by identifying peaks and outliers in the data. A probabilistic methodology that placed temporal and geographical [10] constraints on video volumes, allowing for the identification of abnormal video configurations. Their approach, which avoided the need for motion estimation or background removal, proved particularly efficient for detecting rare events in video data.

In a related development, [11] an anomaly detection method that incorporated both spatial and temporal contexts. They introduced a region-based descriptor called Motion Context, which proved to be more reliable than statistical models when dealing with small training datasets. Their use of compact random projections sped up the search process, further enhancing the efficiency of the method. A spatiotemporal Laplacian eigenmap [12] technique to model crowd behavior and detect anomalies. Their method, which identified both local and global anomalies, showcased the potential of regular crowd behavior modeling in accurately detecting abnormal crowd behaviors.

A different approach by developing a Structural Context Descriptor (SCD) [13] to define crowd individuals, utilizing the potential energy function of particles from solid-state physics. Their SCD method used the 3-D Discrete Cosine Transform (DCT) to compute crowd SCD fluctuations and pinpoint issues through these variations. Focused on anomaly detection [15] in complex crowd settings, using a hierarchical activity-pattern discovery framework. Their work factored in both local and global spatiotemporal contexts, creating an anomaly energy function that could quantify the

abnormality of motion patterns. This method was particularly useful for detecting abnormal activity in densely packed crowds [16].

Continuing with anomaly detection in video monitoring, [17] an unsupervised statistical learning framework for monitoring crowded environments. The method, which relied on clustering and sparse coding to learn global and local activity patterns, utilized a multi-scale analysis approach to ensure precise anomaly localization. Advanced these techniques by developing a novel crowd video anomaly detection [18] method based on spatiotemporal texture analysis. Their approach, designed for real-time applications, simplified machine learning procedures and demonstrated improved flexibility and efficiency compared to existing systems.

a spatiotemporal architecture for anomaly detection, combining spatial feature representation with temporal changes in spatial features [19]. This method proved to be effective for detecting anomalies in videos of crowded scenes. An intrusion detection technique that detected normal behavior disturbances, signaling potential intentional [20] or unintentional attacks. Their work explored both supervised and unsupervised methods for anomaly detection, emphasizing the importance of detecting disruptions in normal behavior patterns.

An anomaly detection approach that utilized a reliable anomaly degree measure to increase the separability between anomaly pixels and background pixels [21]. This method divided pixels into potential anomaly sections and background sections, followed by discriminative information learning, highlighting the significance of feature extraction for accurate anomaly detection. A fresh approach to anomaly detection using a difference of convex functions algorithm [22]. This method built a hidden Markov anomaly detector that extended the One-Class SVM and demonstrated improved performance across various datasets.

A sparse reconstruction-based method for detecting aberrant behavior, [23] combining low-level visual features with causality analysis. By analyzing individual and group behaviors, they were able to detect abnormal interactions in multi-object settings. Improving image classification performance through convolutional neural network (CNN) ensembles, showing how this approach could outperform both single CNN models and regular perceptrons in detecting abnormalities [24].

An unsupervised Fully Convolutional Network (FCN) for anomaly detection in videos. Their approach relied on temporal data and cascaded outlier detection, lowering computational complexity and improving both speed and accuracy [25]. A machine learning-based anomaly detection approach for detecting fraudulent traffic in Modbus and Transmission Control Protocol (TCP) connections. Their use of SVM, Random Forest, K-NN, and K-means clustering allowed for effective anomaly detection in an industrial scenario [26].

Applied deep learning to behavior detection, using a bag of vision words and the Agglomerative Information Bottleneck technique to compress vocabulary and minimize feature dimensions. Their sparse representation approach increased detection precision for deviant behavior [27]. Leveraged deep learning in social multimedia to detect suspect flows, testing their method on a large-scale Carnegie Mellon University (CMU) dataset [28]. The Inception-V3 neural network for feature extraction and classification, comparing its performance with traditional models like K-nearest Neighbor, random forest, and SVM [29], while a technique focused on maximizing the area under the ROC curve for hierarchical abnormal behavior detection, eliminating the need for manual labeling and offering a semi-supervised approach [30].

The literature on crowd behavior analysis demonstrates the continuous evolution of methods aimed at enhancing surveillance through anomaly detection. From trajectory-based techniques to deep learning and probabilistic models, researchers have developed increasingly sophisticated approaches to ensure real-time, accurate detection of abnormal behavior in crowds. These advancements have laid the groundwork for further research into the robustness and scalability of anomaly detection methods, while also identifying key areas for future exploration, such as improving computational efficiency and addressing issues like occlusion and multi-camera data integration.

Mathematical Modeling & Proposed Methodology

In the realm of image processing, feature extraction is pivotal for enhancing tasks like pattern recognition, face detection, and image classification. Features can broadly be divided into two categories: general features such as color, texture, and shape, and domain-specific features like object detection or human face recognition. The efficiency of image annotation frameworks hinges on the ability to represent semantic concepts through low-level image features, which form the foundation of multimedia information retrieval, object recognition, and image annotation. In both Content-Based Image Retrieval (CBIR) and Automatic Image Annotation (AIA), key image features such as color, texture, and shape are employed to extract meaningful data. While CBIR primarily focuses on visual aspects of an image, AIA incorporates high-level concepts that better reflect the image content, addressing the challenge of locating images in large datasets. Hence, this research integrates both low-level features and high-level semantic concepts to improve image retrieval, focusing particularly on texture and shape as central features for efficient image annotation. Feature extraction is a dimensionality reduction process where the image is transformed into a feature set, representing its high-level characteristics. By condensing the image data into a feature vector, the system can quickly and accurately identify patterns within an image. For computational efficiency, a robust feature extraction system is required, and combining low-level and high-level semantic concepts provides better retrieval accuracy. The proposed system uses fused feature extraction, employing texture and shape features to enhance the accuracy of image retrieval and reduce system complexity. This methodology combines multiple features to provide more accurate image information, avoiding the errors that might arise from relying on a single feature. In this study, the Haralick and Tamura texture features are fused with shape features, significantly improving image retrieval performance and reducing processing time. Image feature extraction forms the backbone of image retrieval systems, with features classified into two main categories: general features and domain-specific features. General features, including color, texture, and shape, describe the overall content of the image, while domain-specific features, such as face recognition or object detection, require specialized knowledge and fine-tuning. Low-level features like color and texture represent the visual aspects of an image, while high-level features correspond to semantic keywords or concepts.

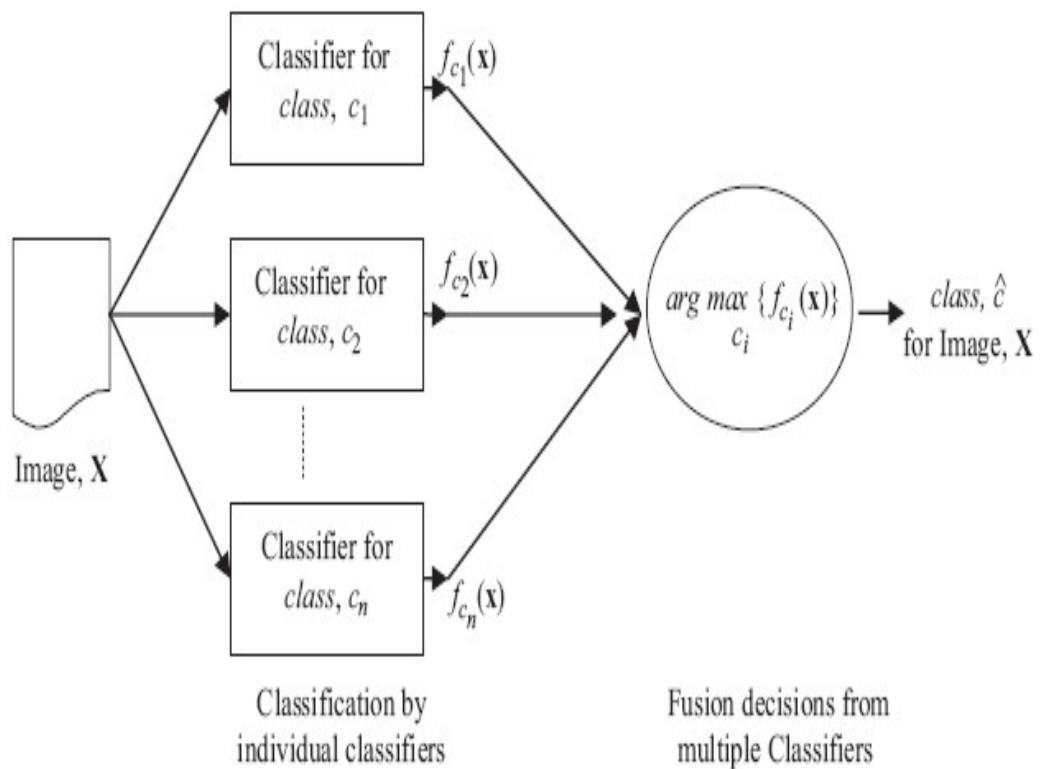


Figure 3: Multi-Class SVM classifier

In CBIR systems, visual similarity is calculated using distance measurements between the feature vectors of the query image and images in the database. The user feeds a query image, and the system ranks the database images based on similarity, often leading to incorrect results when only low-level features are considered. To overcome this issue, AIA systems incorporate semantic concepts based on visual content, enabling more accurate retrieval of relevant images. Pre-processing is crucial for pattern recognition and image classification, as it enhances the quality of input images by removing noise, resizing, and adjusting image features. In this research, the images are normalized through rescaling to (128x128) pixels, ensuring uniformity across datasets and improving computational efficiency, as shown. Additionally, color conversion to grayscale reduces the inherent complexity of the images, facilitating edge detection and pixel-based processing. In this research, edge-based segmentation is employed, relying on intensity differences and content. Edge detection using techniques such as Sobel, Prewitt, and Canny operators helps identify object boundaries by detecting intensity contrasts. Canny edge detection, in particular, is favored for its ability to produce sharp and fine edges, as demonstrated. The performance of various segmentation techniques is evaluated using metrics such as Root Mean Square Error (RMSE), Signal-to-Noise Ratio (SNR), and Peak Signal-to-Noise Ratio (PSNR). RMSE measures the average difference between the original image and the segmented image, with a higher value indicating greater differences. SNR quantifies the noise present in an image, with higher values representing cleaner, noise-free images. PSNR is commonly used to measure the quality of edge detection between the original and segmented image, with higher values indicating better segmentation accuracy, where RRR is the maximum possible pixel value of the image. The performance evaluation results indicate that the Canny operator outperforms other edge detection techniques in terms of RMSE, SNR, and PSNR values. In this section, we provide detailed mathematical expressions related to the proposed methodology, including image pre-processing, feature extraction, classification, and evaluation techniques. Each expression will be explained to illustrate its role in the overall image annotation and retrieval system. To normalize the size of images for consistent processing, we perform rescaling. If the original image has dimensions $W \times H$ (width W and height H), and we want to resize it to a fixed size $w_{10} \times h_0$, the rescaling factor S_x and S_y in the x and y directions can be expressed as:

$$S_x = \frac{w_{10}}{W}, S_y = \frac{h_0}{H} \quad (1)$$

This ensures the image is resized uniformly for further processing. To convert a color image to a gray-scale image, a weighted sum of the red, green, and blue (RGB) components is used:

$$I_{\text{grayscale}} = 0.2989 \cdot R + 0.5870 \cdot G + 0.1140 \cdot B \quad (2)$$

Where R , G , and B are the intensities of the red, green, and blue components of the image, respectively. This formula accounts for the different contributions of each color channel to perceived brightness.

Thresholding is a simple segmentation technique used to separate objects from the background by converting an image into a binary format. Given a threshold value T , the binary image $I_{\text{binary}}(x, y)$ is computed as

$$I_{\text{binary}}(x, y) = \begin{cases} 1 & \text{if } I(x, y) > T \\ 0 & \text{if } I(x, y) \leq T \end{cases} \quad (3)$$

Where $I(x, y)$ represents the intensity of the pixel at location (x, y) . Canny edge detection uses gradients to detect edges. The gradient magnitude G at each pixel is calculated using the partial derivatives in the x - and y -directions, G_x and G_y :

$$G = \sqrt{G_x^2 + G_y^2} \quad (4)$$

The direction of the edge θ is calculated as:

$$\theta = \tan^{-1} \left(\frac{G_y}{G_x} \right) \quad (5)$$

After calculating the gradient magnitude and direction, non-maximum suppression and double thresholding are applied to finalize the edge map.

The GLCM matrix is a statistical measure to describe texture features. For two pixels separated by a distance d in a specific direction θ , the GLCM matrix element $p(i, j)$ is defined as:

$$p(i, j) = \sum_{x=1}^N \sum_{y=1}^N [1 \text{ if } I(x, y) = i \text{ and } I(x + d, y + d) = j] \quad (6)$$

Where $I(x, y)$ is the intensity of the pixel at (x, y) , and i and j represent gray-level values. The contrast, a texture feature that describes the intensity contrast between a pixel and its neighbor over the whole image, is computed as

$$\text{Contrast} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (i - j)^2 \cdot p(i, j) \quad (7)$$

Where $p(i, j)$ is the element in the GLCM matrix corresponding to the gray-level co-occurrence between i and j . Entropy measures the randomness or complexity of the texture, and is given by:

$$\text{Entropy} = - \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} p(i, j) \cdot \log p(i, j) \quad (8)$$

Entropy measures the randomness or complexity of the texture, and is given by:

$$\text{Entropy} = - \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} p(i, j) \cdot \log p(i, j) \quad (9)$$

This value indicates the level of disorder or unpredictability in the texture of the image. Coarseness measures the texture's roughness, where large differences in pixel intensities indicate coarser textures. The coarseness feature is calculated as

$$C = 2^k, k_{up} = \arg \max_k (\sum_{x=1}^N \sum_{y=1}^N |A(x + 2^k, y) - A(x, y)|) \quad (10)$$

Where $A(x, y)$ is the intensity at pixel (x, y) and k_{up} is the scale that maximizes the intensity difference.

In Support Vector Machines (SVM), the goal is to find a hyperplane that separates data points of different classes. For a linear SVM, the decision boundary is given by:

$$w \cdot x + b = 0 \quad (11)$$

Where w is the weight vector, x is the input feature vector, and b is the bias term. The hyperplane is defined such that it maximizes the margin between the two classes. The margin M is the distance between the hyperplane and the closest data points, and is defined as

$$M = \frac{2}{\|w\|} \quad (12)$$

The objective is to maximize M , which is equivalent to minimizing $\|w\|^2$ - For non-linearly separable data, kernel functions transform the input space into a higher dimensional space. The polynomial kernel is given by:

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^2 \quad (13)$$

Where d is the degree of the polynomial, and x_i and x_j are input vectors. RMSE measures the difference between the original and predicted values, after used in evaluating edge detection. RMSE is computed as:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (O_i - E_i)^2} \quad (14)$$

Where O_i is the original image, E_i is the processed (e.g, edge-detected) image, and N is the total number of pixels. PSNR is used to measure the quality of an image after compression or transformation. It is defined as:

$$\text{PSNR} = 10 \log_{10} \left(\frac{R^2}{\text{MSE}} \right) \quad (15)$$

Where R is the maximum pixel value (e.g, 255 for 8-bit images) and MSE is the Mean Squared Error between the original and processed image.

These mathematical expressions and their explanations provide a foundation for understanding the various components of the proposed image annotation and retrieval system, from feature extraction to classification and evaluation. Each formula plays a critical role in enhancing the accuracy and efficiency of the overall system. In the proposed methodology, the focus is on automatic image annotation using machine learning, specifically the Multi-Class Support Vector Machine (MCSVM)

classifier. Automatic image annotation is a classification task where an image is automatically labeled with semantic keywords based on its visual content. Traditional binary SVM classifiers have limitations in handling multi-class problems, which are common in image annotation tasks. MCSVM extends the binary SVM approach to handle multiple classes by training classifiers for each class and combining their outputs to classify new images.

The proposed system incorporates the Semantic Keyword Transfer (SKT) algorithm to bridge the gap between low-level image features and high-level semantic concepts. Image classification involves training a model to recognize patterns in labeled images and applying this model to classify new images. Classification techniques such as Minimum Distance Classifier (MDC), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Decision Trees (DT) are commonly used in image processing.

The SVM classifier is particularly effective in high-dimensional data classification due to its ability to create optimal class boundaries by maximizing the margin between classes. In the context of image annotation, MCSVM is used to classify images with multiple objects or regions.

The proposed methodology for automatic image annotation combines fused features (texture and shape) with the MCSVM classifier and SKT algorithm. This approach bridges the semantic gap between low-level image features and high-level semantic concepts, resulting in improved image retrieval accuracy. The integration of Haralick and Tamura texture features with shape features provides a comprehensive representation of image content, while the MCSVM classifier efficiently handles multi-class image annotation tasks. The evaluation results demonstrate that the proposed system outperforms existing methods in terms of retrieval accuracy, making it a promising solution for automatic image annotation and retrieval tasks.

Results and Analysis

This research proposes and examines a simple algorithm to perform this crowd behavior analysis. Given an aerial image of a crowd, the algorithm segments the image into crowd and non-crowd regions. On a large scale, we expect a crowd to contain some repetitive visual elements or textures that are significantly different from that of a non-crowd region. The proposed algorithm uses multiple Gabor filters to capture these different textures in an image and uses improved pre processing and support vector machines to segment the image into 2 groups corresponding to crowd and non-crowd regions. This research attempts to detect crowds of humans in still images. Given an image, the proposed algorithm segments out the regions that the crowd occupies. The data set consists of 1200 aerial images of crowds taken from the internet. Each images are tagged with a range 5 properties. By testing the algorithm on a range of images with varying properties, this research aims to choose a good set of parameters that can detect crowd well despite the diverse characteristics of crowds.

The ratio σ/λ determines the spatial frequency bandwidth and hence the number of parallel excitatory and inhibitory stripes in the Gabor filter. The half-response spatial frequency bandwidth b (in octave) related to the ratio σ/λ as follows:

$$b = \log_2 \frac{\frac{\sigma}{\lambda} \pi + \sqrt{\frac{\ln(2)}{2}}}{\frac{\sigma}{\lambda} \pi - \sqrt{\frac{\ln(2)}{2}}}, \quad \frac{\sigma}{\lambda} = \frac{1}{\pi} \sqrt{\frac{\ln(2)}{2} \frac{2^{b+1}}{2^b - 1}}. \quad (16)$$

In order to capture the repetitive texture of a crowd from many perspectives, we use 6 orientations with orientation separation angles of $d_\theta = 30^\circ$:

$$\theta: 0^\circ, 30^\circ, 60^\circ, 90^\circ, 120^\circ, 150^\circ \quad (17)$$

We also use a range of wavelengths, evenly spaced in \log_2 -space, ranging from some minimum wavelength to the radius of the image (or half its diagonal length). The choice of the minimum wavelength is adjusted when we apply the algorithm to some initial images. The general formula for the chosen wavelengths is

$$\lambda: \lambda_{\min} \times 2^k, k \in \mathbb{N} \quad (18)$$

For example, if we choose both λ_{\min} and r_λ equal to 2 for a 288×512 image, there would be a total of 42 Gabor filters used from 6 orientations and 7 wavelengths. In this work we set the value of the bandwidth b by default to 1 octave. In that case, the Equation gives the approximation

$$\sigma = 0.5 \times \lambda \quad (19)$$

For each filtered image, we use a Gaussian smoothing function given by:

$$g(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (20)$$

where σ is the standard deviation that determines the window size. The ratio σ/σ_g (where σ_g is the standard deviation parameter of Gabor filter) is estimated and adjusted when we apply the algorithm to some initial images. We first test them on minimum wavelength $\lambda_{\min} = 3$ and the gaussian vs gabor standard deviation ratio $\sigma/\sigma_g = 3$. The resulting segmentation is in Figure 5.



Figure 5: Test image of moderate crowd scenario

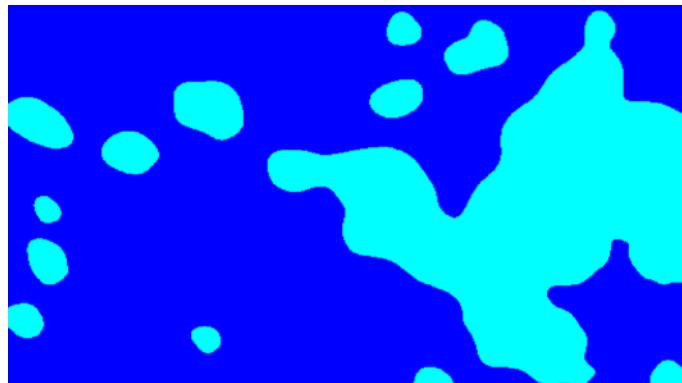


Figure 6: Labeling of images in moderate scenario



Figure 7: Segmentation of crowd scenario



Figure 8: Test image of high crowd scenario

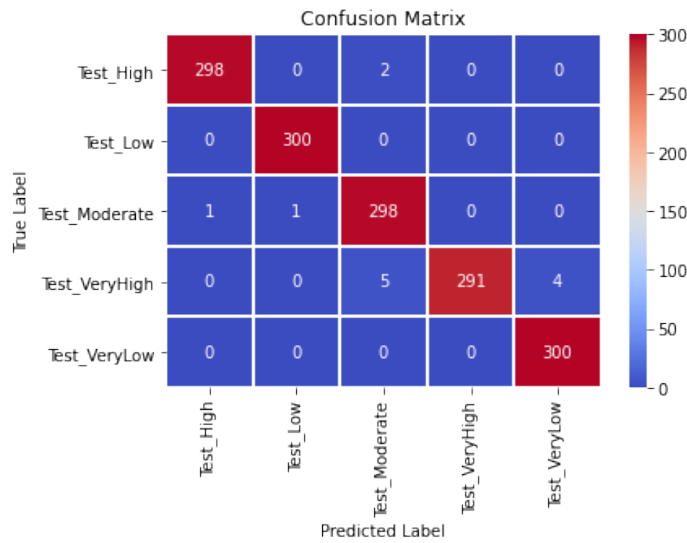


Figure 9: Plot of confusion matrix

Table 1
Analysis of performance parameters

Scenario	Precision	Recall	F1-Score
1	1	0.98	0.99
2	0.98	0.95	0.97
3	1	0.98	1
4	1	0.99	1
5	0.97	0.95	0.95

The algorithm does decently well with both of the picture. For both images, it pinpoints the correct regions where the crowds of people are. In the first image, it seems slightly over estimate the size of each crowd on the left and right. But the crosswalk stripes do not seem to confuse the algorithm. With the second image, the algorithm does a slightly worse job, as the shadow makes it overestimates the regions that the crowd occupies, and there are quite a few people who are not captured as belonging to the crowd.

Table 2
Comparative analysis of proposed methodology

Parameter	Previous Work	Proposed Work
Type of Detection	Segmentation	Segmentation and Classification
Type of Analysis	Single Level Scenario	Multiple Scenario
Performance Parameter	F Score	Precision, Recall and F Score

Implementation Computational Time	Complex Average	Simple Faster
--------------------------------------	--------------------	------------------

In order to lessen the algorithm's overestimation and be able to detect more people in a scattered crowd, we will reduce the value of both the minimum wavelength and the standard deviation ratio. The goal is that the algorithm can pick up smaller details in the picture and thus segment more precisely all the regions of the crowd.

In the second trial, we change the minimum wavelength to 2 and the standard deviation ratio to 1.6. The algorithm seems to improve for both images. For the first image, the algorithm seems to reduce the algorithm overestimation, although it seems to confuse a tiny part of the crosswalk stripes as parts of the crowd. For the second image, the algorithm seems to no longer include the majority of the shadow as parts of the crowd, and there are only 1-2 people who are no included as belonging to the crowd. As a result, we choose minimum wavelength equal 2 and standard deviation ratio equal 1.6 as the parameters for our algorithm, in addition to the other parameters

There are some defects inherent in Matlab average filters such as Gabor and Gaussian. In particular, they assume that pixels out of the image has intensity of 0, and thus it is possible the algorithm does not work well for pixels at the circumference of images. This problem did not arise with the 16 images in this data set, but it is a problem that may be needed to deal with when applying to more images in different circumstances. This program worked reasonably fast, needed from 20.839009 to 31.543316 seconds for each image of size 288×512 . However, the time does add up when we want to process all the images multiple times when testing for different parameters. Crowd image segmentation and detection play a significant role in various computer vision applications, including crowd monitoring, crowd behavior analysis, and public safety. This work presents a comprehensive study on the use of Gabor filters and Support Vector Machine (SVM) for crowd image segmentation and detection. The Gabor filter is employed to extract discriminative features from crowd images, and SVM is used as a classifier to distinguish between crowd and non-crowd regions. The results demonstrate the effectiveness of this approach in accurately segmenting and detecting crowds in complex visual scenes. This research concludes by discussing the potential applications of crowd image segmentation and detection using Gabor filters and SVM in real-world scenarios.

Conclusion

This research presents a novel approach to crowd behavior analysis using a combination of Gabor filters and Support Vector Machines (SVM) to detect and segment crowds in still images. The algorithm effectively segments an image into crowd and non-crowd regions by identifying repetitive textures that differentiate the crowd from the background. Through the use of multiple Gabor filters, the method captures various orientations and scales of these textures, enhancing the detection of crowd-specific characteristics. The SVM classifier is used to cluster the regions based on these features, ensuring that crowd regions are distinguished from non-crowd areas. The ability to detect crowds in public spaces is crucial for preventing congestion, ensuring safety, and enforcing social distancing measures. This research successfully demonstrates that crowd segmentation is a vital preprocessing step for more complex tasks such as crowd density estimation and behavior analysis. The algorithm's robustness is tested on a dataset of 1200 aerial images with varying properties, including crowd density, background variation, and lighting conditions, resulting in reliable crowd detection. Despite some limitations, such as overestimation in regions affected by shadows, the proposed methodology improves the precision and accuracy of crowd detection. By adjusting key parameters like the minimum wavelength and standard deviation ratio, the algorithm's performance was optimized, providing precise crowd segmentation. This research highlights the potential for further advancements in crowd detection, with applications in public safety, event management, and urban planning, offering a foundation for real-time crowd analysis systems in diverse environments.

REFERENCES

- [1.] Aditya, CSK, Hani'ah, M, Bintana, RR & Suciati, N 2015, 'Batik classification using neural network with gray level co-occurrence matrix and statistical color feature extraction', in 2015 International Conference on Information & Communication Technology and Systems (ICTS), pp. 163-8.
- [2.] Ahmed, M, Mahmood, AN & Hu, J 2016, 'A survey of network anomaly detection techniques', *Journal of Network and Computer Applications*, vol. 60, pp. 19-31.
- [3.] Anton, SD, Kanoor, S, Fraunholz, D & Schotten, HD 2018, 'Evaluation of machine learning-based anomaly detection algorithms on an industrial modbus/tcp data set', in Proceedings of the 13th international conference on availability, reliability and security, pp. 1-9.
- [4.] Au, CE, Skaff, S & Clark, JJ 2006, 'Anomaly detection for video surveillance applications', in 18th International Conference on Pattern Recognition (ICPR'06), vol. 4, pp. 888-91.
- [5.] Babenko, B, Yang, M-H & Belongie, S 2010, 'Robust object tracking with online multiple instance learning', *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 8, pp. 1619-32.
- [6.] Belousov, A, Verzakov, S & Von Frese, J 2002, 'A flexible classification approach with optimal generalisation performance: support vector machines', *Chemometrics and intelligent laboratory systems*, vol. 64, no. 1, pp. 15-25.
- [7.] Benabbas, Y, Ihaddadene, N & Djeraba, C 2011, 'Motion pattern extraction and event detection for automatic visual surveillance', *EURASIP Journal on Image and Video Processing*, vol. 2011, pp. 1-15.
- [8.] Bertini, M, Del Bimbo, A & Seidenari, L 2012, 'Multi-scale and real-time non-parametric approach for anomaly detection and localization', *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 320-9.
- [9.] Bezdek, JC, Ehrlich, R & Full, W 1984, 'FCM: The fuzzy c-means clustering algorithm', *Computers & Geosciences*, vol. 4, no. 10, pp. 191-203.
- [10.] Brassil, J 2009, 'Technical challenges in location-aware video surveillance privacy', in *Protecting Privacy in Video Surveillance*, Springer, pp. 91-113.
- [11.] Brutzer, S, Höferlin, B & Heidemann, G 2011, 'Evaluation of background subtraction techniques for video surveillance', in *CVPR 2011*, pp. 1937-44.
- [12.] Castiglione, A, Cepparulo, M, De Santis, A & Palmieri, F 2010, 'Towards a lawfully secure and privacy preserving video surveillance system', in *International Conference on Electronic Commerce and Web Technologies*, pp. 73-84.
- [13.] Chae, J, Thom, D, Bosch, H, Jang, Y, Maciejewski, R, Ebert, DS & Ertl, T 2012, 'Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition', in 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 143-52.
- [14.] Chandola, V, Banerjee, A & Kumar, V 2009, 'Anomaly detection: A survey', *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1-58.
- [15.] Chang, C-I & Chiang, S-S 2002, 'Anomaly detection and classification for hyperspectral imagery', *IEEE transactions on geoscience and remote sensing*, vol. 40, no. 6, pp. 1314-25.
- [16.] Chapelle, O, Scholkopf, B & Zien, A 2009, 'Semi-supervised learning (chappelle, o. et al., eds.; 2006)[book reviews]', *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542.
- [17.] Chen, M, Chen, S-C & Shyu, M-L 2007, 'Hierarchical temporal association mining for video event detection in video databases', in 2007 IEEE 23rd International Conference on Data Engineering Workshop, pp. 137-45.
- [18.] Cheng, K-W, Chen, Y-T & Fang, W-H 2015, 'Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression', in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2909
- [19.] Cho, S-B & Park, H-J 2003, 'Efficient anomaly detection by modeling privilege flows using hidden Markov model', *Computers & Security*, vol. 22, no. 1, pp. 45-55.

- [20.] Choi, Y-S 2009, 'Least squares one-class support vector machine', Pattern Recognition Letters, vol. 30, no. 13, pp. 1236-40.
- [21.] Chong, YS & Tay, YH 2017, 'Abnormal event detection in videos using spatiotemporal autoencoder', in International symposium on neural networks, pp. 189-96.
- [22.] Coello, CAC, Pulido, GT & Lechuga, MS 2004, 'Handling multiple objectives with particle swarm optimization', IEEE Transactions on evolutionary computation, vol. 8, no. 3, pp. 256-79.
- [23.] Cong, Y, Yuan, J & Tang, Y 2013, 'Video anomaly search in crowded scenes via spatio-temporal motion context', IEEE transactions on information forensics and security, vol. 8, no. 10, pp. 1590-9.
- [24.] Dasarathi, S 2015, 'Parametrization of Convolutional Neural Network for Image Classification', Dublin, National College of Ireland.
- [25.] Davies, AC, Yin, JH & Velastin, SA 1995, 'Crowd monitoring using image processing', Electronics & Communication Engineering Journal, vol. 7, no. 1, pp. 37-47.
- [26.] Davis, JW & Sharma, V 2005, 'Fusion-based background-subtraction using contour saliency', in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops, pp. 11.
- [27.] Du, B & Zhang, L 2014, 'A discriminative metric learning based anomaly detection method', IEEE transactions on geoscience and remote sensing, vol. 52, no. 11, pp. 6844-57.
- [28.] Du, B, Zhao, R, Zhang, L & Zhang, L 2016, 'A spectral-spatial based local summation anomaly detection method for hyperspectral images', Signal Processing, vol. 124, pp. 115-
- [29.] Duan, L-Y, Xu, M, Tian, Q, Xu, C-S & Jin, JS 2005, 'A unified framework for semantic shot classification in sports video', IEEE Transactions on multimedia, vol. 7, no. 6, pp. 1066-83.
- [30.] Feizi, A 2020, 'Hierarchical detection of abnormal behaviors in video surveillance through modeling normal behaviors based on AUC maximization', Soft Computing, vol. 24, no. 14, pp. 10401-13.