S & M 4007

Hidden Markov Models for Anomalous Behavior Detection in Surveillance Video with Depth Map

Jui-Feng Yeh,[*] Shu-Po Hsu, and Kai-Siang You

National Chiayi University, No. 300, Xuefu Rd., Chiayi City 600355, Taiwan (R.O.C.)

(Received August 5, 2024; accepted April 16, 2025)

Keywords: surveillance, behavior detection, depth map, hidden Markov model, video recognition

A real-time surveillance system is investigated on the basis of hidden Markov models (HMMs) using various features extracted from color images, human skeletons, and depth maps to sense anomalous behavior. Herein, the spatial and temporal features are included to enhance surveillance measurement accuracy by identifying and classifying anomalous activity. Hence, the proposed approach detects suspicious behaviors within a short time and achieves a better performance than traditional approaches. The HMM-based framework captures the underlying patterns or structures in sequential information when a human appears in the predefined monitoring area to detect anomalous behaviors. The highlights of the proposed system are its efficiency and practicality, balancing computational requirements and detection accuracy, making it suitable for real-time applications. For evaluating the proposed approach, a dataset collected by a Kinect camera is further divided into training and test data. Furthermore, the proposed approach significantly outperforms that based on naïve Bayes networks in precision rate according to the experimental results. As a result, evaluating observations demonstrates the potential of HMM-based systems to enhance security monitoring, providing reliable and effective solutions for instant anomalous behavior detection to ensure the security and protection of sensitive information and equipment in monitoring scopes.

1. Introduction

Surveillance video anomaly detection (SVAD) is a field in computer vision and security technology focused on identifying and sensing unusual or suspicious activities automatically within video footage. Anomalous behavior detection plays a crucial role in surveillance systems, not only serving general security measures such as theft prevention and visitor management but also reducing the need for physical guards in many scenarios, thereby lowering security costs over time. These systems enable the monitoring of human activities within alarm areas to ensure safety and appropriate behaviors. Additionally, remote monitoring systems utilize surveillance to observe activities without the need for continuous physical supervision by the Internet. Furthermore, cloud storage solutions facilitate the storage and retrieval of extensive video data from anywhere in the world, enhancing the flexibility and accessibility of surveillance

[*] Corresponding author: e-mail: ralph@mail.ncyu.edu.tw

https://doi.org/10.18494/SAM5278

recordings. The broad potential applications of SVAD have sparked increasing research interest.(1) SVAD is designed to detect and locate anomalous events within video footage. In general, such anomalous behaviors are rarer than normal activities. To avoid wasting human resources and time, there is an urgent need to develop advanced computer vision technologies capable of automatically identifying these anomalies.

Traditionally, surveillance relied on manual monitoring, a process prone to human error and fatigue. It is usually time-consuming, labor-intensive, and biased owing to subjective judgment. Anomaly detection in surveillance video has always been a challenging issue until now. Although existing methods(2) have demonstrated promising performance, they still suffer from three primary limitations. The first significant limitation is the absence of spatial information. Current surveillance systems primarily rely on RGB cameras, which capture images in only two dimensions. The lack of depth information often results in the loss of critical 3D data(3) especially in approaching and leaving the surveillance area. Furthermore, RGB-based detection systems are highly sensitive to varying lighting conditions. Changes in natural light during the day or artificial light adjustments can lead to false detections or missed detection.(4) There is an opportunity for intruders to take advantage of this. The second limitation concerns real-time processing. Immediate processing is crucial for timely anomaly detection in monitoring scenarios, enabling prompt responses to potential threats or emergencies. However, surveillance footage encompasses a vast array of information, requiring significant computational resources and time for processing and analysis. Achieving real-time performance without sacrificing accuracy poses a significant challenge. The third limitation involves sparse and imbalanced data. Anomalies are intrinsically rare events, which results in a scarcity of adequate training examples. Sparse data means that there may not be enough instances of anomalous behavior for the model to learn effectively. This can lead to issues such as overfitting or poor generalization, ultimately affecting the accuracy of anomaly detection.

To address these issues, we propose a real-time system using hidden Markov models (HMMs) to detect anomalous behaviors in RGB cameras with depth maps. Aside from the features obtained by RGB cameras, depth cameras provide depth information by capturing images and measuring the distance between objects and the camera, thereby enhancing the spatial analysis capability of surveillance systems. Herein, depth maps are usually generated using sensors that do not rely on visible light and are thus not affected by angles and lighting in determining human figures. HMMs are utilized as the primary architectural model; HMMs are a statistical-based model that describes the relationship between observable events and a series of unobservable internal states. Additionally, HMMs excel at handling temporal data, requiring fewer parameters and less memory space. This makes them highly suitable for scenarios with limited computational power, meeting the needs of surveillance video anomaly detection for real-time system monitoring. Additionally, HMMs' adaptability allows them to automatically adjust model parameters on the basis of new observational data, adapting to changes in behavioral patterns, and more accurately identifying unknown or evolving anomalous behaviors. HMMs are particularly well-suited for dealing with uncertainty or incomplete data often encountered in security monitoring contexts.

The main contributions of this study are summarized as follows:

(1) Development of a method using depth maps and color images to extract motion features: We introduced an innovative approach that utilizes depth maps to extract motion features, enhancing the capability of monitoring systems to analyze human behaviors in 3D space.

(2) Real-time system for anomalous behavior detection using HMMs: We developed a real-time system employing HMMs to detect anomalous activities. By processing temporal data efficiently, the system ensures high accuracy and prompt response in issuing alerts.

(3) Creation of a dataset to simulate real-world scenarios: We created a comprehensive dataset that simulates various scenarios likely to occur in real-life hospital settings. This dataset includes diverse behaviors, from normal to anomalous, enabling the system to be extensively trained and validated.

2. Related Works

Human activities can be categorized into four types: postures, actions, behaviors, and interactions. Posture refers to the control, manipulation, or communication performed using parts of the body;[5] the body's posture is described as a schematic of a collaborative manipulation task to achieve good performance. An action is any complex movement of the body, which can be decomposed into multiple fundamental movements.[6] Zhou and Wu[7] employed specialized intelligent machines for automated monitoring technology applied in supermarkets. In their study, a new method that focuses solely on moving hands was developed. An accurate localization of the palm is achieved by utilizing a linear prediction model to implement object tracking. In contrast to the approach adopted by Mor *et al.*,[8] our methodology extends beyond the analysis of hand movements. We utilize joint detection to facilitate the prediction of comprehensive human body movements, thereby capturing a broader spectrum of information. This enhancement ensures that our model achieves a higher degree of accuracy in the detection of anomalous behaviors. Mor *et al.*[8] discussed the broad applications of HMMs across various fields, emphasizing their role in solving detection problems through sequence analysis and state estimation. Gámiz *et al.*[9] utilized HMMs to dynamically analyze and predict the reliability of systems with Markovian signal processes, enhancing the understanding of system performance changes over time. This method is crucial for improving long-term system reliability and maintenance strategies. Feng and Liu[10] proposed a method called attentional temporal You Only Look Once (ATYOLO) that utilizes attention mechanisms and convolutional long short-term memory to detect and track humans and animals in videos. However, their method does not detect finer movements or make judgments about actions. Gao *et al.*[11] proposed a noncontact diagnostic system driven by deep-learning visual algorithms. Utilizing four RGB cameras, it captures human dynamics and employs a pose estimator to generate comprehensive 3D human posture behavior predictions. However, most surveillance systems do not have four RGB cameras to monitor the same area to build a 3D model.

Compared with traditional machine learning methods such as naïve Bayes, *K*-nearest neighbors, and support vector machines, HMMs demonstrate superior capabilities in handling temporal series data.[12] Although long short-term memory networks also excel in addressing time-series-related challenges, convolutional neural networks (CNNs) show remarkable performance in pattern recognition. However, these approaches require large models and high complexity, and they are difficult to implement in an embedded system, especially in real-time applications. Ovhal *et al.*(13) reviewed how HMMs are used to recognize and predict driving behaviors, highlighting their effectiveness and applications in vehicle safety systems. The output can be easily affected by slight variations in the application scenarios.

Agarwal *et al.*(14) used a variant of the slow-fast algorithm to detect and classify the unusual activity happening in the surveillance areas. Xue and Liu(15) discussed how HMMs are applied to human activity modeling in human activity recognition and fall detection. They concluded with a review of notable research works in the fields of smart home technologies and elderly care based on HMMs. San-Segundo *et al.*(16) proposed a human sensing system based on HMMs for classifying physical activities such as walking, climbing stairs, descending stairs, sitting, standing, and lying down. Liu and Datta(17) proposed a method that uses HMMs integrated with contextual information to dynamically predict agent interactions, addressing the challenge of establishing trust models in complex dynamic environments.

Depth maps provide distance information for each pixel relative to the observer, enhancing object identification and scene understanding beyond the capabilities of traditional 2D imagery. In environments subject to variable conditions, such as changes in lighting(18) or visual obstructions,(19) the information provided by depth maps can assist systems in better adapting and responding to these challenges. Lee and Kim(20) proposed a novel algorithm for monocular depth estimation using relative depth maps, utilizing CNNs and estimating the relative depths between pairs of regions at various scales, as well as the absolute depths. Liu *et al.*(21) proposed an RGB posture-recognition network based on a two-stage CNN architecture. To enhance recognition performance from color images, they incorporated a hybrid loss function in the generation module for estimating depth posture images. They also introduced a depth estimation method applied to spatiotemporal recognition, which is used for dynamic action recognition. Meng *et al.*(22) used a novel neural network architecture to improve monocular depth estimation. This method more accurately predicts the depth of a single image by utilizing contextual information. This approach is particularly useful in applications where depth information is critical, such as in mobile devices or certain types of autonomous vehicle. Naeem *et al.*(23) presented a method of detecting abnormal or anomalous behavior in surveillance videos. The method uses human joint motion information from skeletal sequences to model human behavior, identifying patterns that deviate from normal activities, which is crucial for security in environments such as supermarkets, airports, and public spaces. Urtasun *et al.*(24) proposed a model-based approach that directly utilizes data reconstructed from depth maps. In this study, the method we use for reconstructing human posture is built upon this model-based approach.

3. Proposed Method

For evaluating the proposed approach, the proposed HMM-based system is developed and further divided into training and testing phases as shown in Fig. 1.

The system captures both the RGB and depth maps through the 3D camera while simultaneously detecting the human skeleton. From the detected skeleton, joint positions and

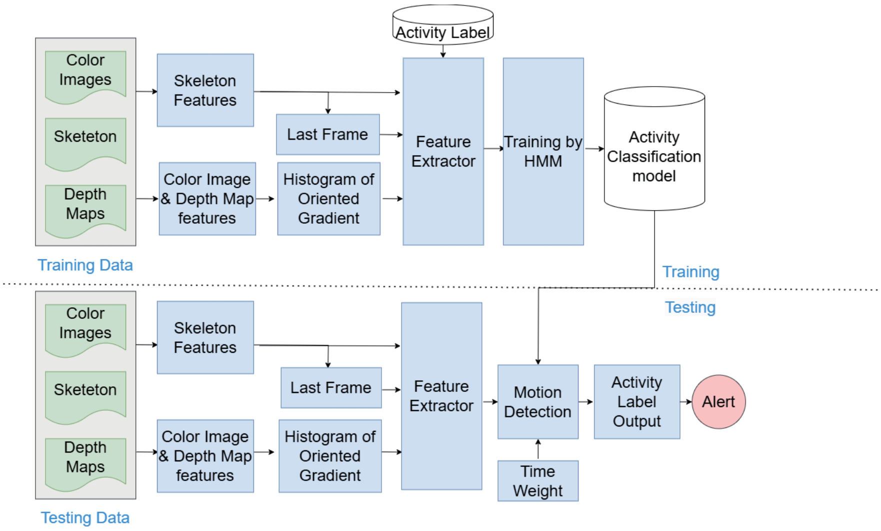


Fig. 1. (Color online) System framework of proposed HMM-based anomalous behavior detection.

joint orientations are extracted. To integrate these data sources captured from different sensors, the color and depth images are aligned to maintain spatial consistency. Subsequently, features are extracted from the aligned data using a histogram of oriented gradient (HOG) technology. After data processing, the proposed method can be divided into spatiotemporal feature extraction and anomalous behavior detection models. The spatiotemporal feature extraction is designed to extract both spatial and temporal features and assign the corresponding activity labels. Activity labels are employed to differentiate and identify various human motion patterns and behaviors, such as knocking, handling a door, and squatting. These labels enable the HMMs to learn the characteristics associated with each specific action. In the training of anomalous behavior detection models, extracted features and activity labels are used to train the HMMs. The state of the previous frame is utilized as a feature for training the model. This approach significantly enhances our understanding of sequence behaviors and aids in the prediction of future state transitions.

3.1 Spatiotemporal feature extraction

A person's activities can be recognized by capturing their motion posture and movement features over time by utilizing RGB-D images, RGB images, and joint information obtained by Kinect. Figure 2 is a summary of the features adopted in the proposed approach. To distinguish between the features obtained by data processing and a feature extractor, we denote the features obtained by the feature extractor as F. These features are then categorized into two main groups on the basis of spatial and temporal concepts. The spatial aspect is further subdivided into three subcategories: shape, posture, and behavior. The temporal aspect is divided into two subcategories: short time and long time. The shape feature, denoted as F_S, is used to identify individuals from differences in height, arm length, and leg length. The posture feature represents a static action at a key moment in time. In this study, the key action of opening a door is used for

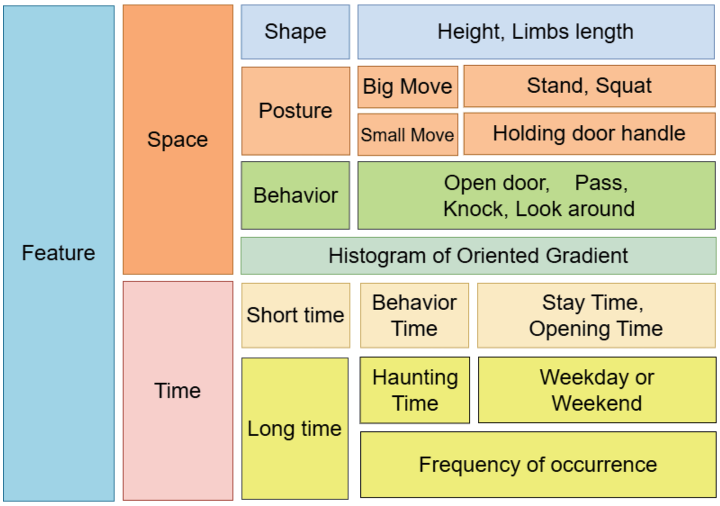


Fig. 2. (Color online) Feature set including spatial and temporal ones used in the proposed approach.

posture judgment. Therefore, posture is subdivided into large and small actions. We denote the posture feature as F_p. The behavior feature, denoted as F_b, refers to a series of action changes. We define five behaviors, which are opening, unlocking, and knocking on the door, passing by, and head turns. The HOG feature here is the same as the feature extracted during the data processing stage, and we denote it as F_h. The short-time feature, denoted as F_{st}, indicates a person's actions within the surveillance area, which is further divided into stay time and door-opening time. Compared with short-time features, which are measured on a time scale of minutes, long-time features are evaluated on a scale of days or weeks. Long-time features assess the times of presence and combine this information with frequency metrics. We denote the long-time feature as F_{lt}.

The Kinect camera can simultaneously capture color images $\boldsymbol{I_c^t}$, human skeletons $\boldsymbol{I_s^t}$, and depth maps $\boldsymbol{I_d^t}$. The data $\boldsymbol{f_{last}}$ captured by the last frame are denoted as $\boldsymbol{I_c^{t-1}}$, $\boldsymbol{I_s^{t-1}}$, and $\boldsymbol{I_d^{t-1}}$. Moreover, the joint position from the human skeletal data is obtained as shown in Ref. 15. Given the joint positions $\boldsymbol{P_n}$ for each $n = \{1, 2,\ldots, 20\}$, each point has three spatial values, namely, X, Y, and Z, which indicate horizontal, vertical, and depth positions, respectively. The human skeleton position $\boldsymbol{f_{sp}} = \{\boldsymbol{P_1}, \boldsymbol{P_2}, ..., \boldsymbol{P_{20}}\}$ is further obtained according to the spatial values. Aside from the skeleton position, the joint vectors are formed according to eleven joint positions.(15) Each joint vector is a 3 × 3 orthogonal matrix. The rows represent the three vectors for the X-, Y-, and Z-axes. Given the 11 orthogonal matrices $\boldsymbol{M_j} \in \boldsymbol{R}^{\boldsymbol{3} \times \boldsymbol{3}}$ for each joint position $j = \{1, 2, \ldots , 11\}$, each matrix $\boldsymbol{M_j}$ must satisfy the orthogonality condition:

$$M_j^T M_j = I, \tag{1}$$

where I is the identity matrix ensuring orthogonality and normalization. This representation allows for precise calculations of joint angles and the orientation of body parts in various activities or movements. Finally, the human skeleton vector $\boldsymbol{f_{sv}} = \{M_1, M_2,\ldots, M_{11}\}$ is obtained. To simultaneously capture features from color images and depth maps, the enhancement of detailed reconstructions and robust object detection, the integration of color images and depth maps, followed by the application of HOG are proposed. The formula is

$$\boldsymbol{I_{cd}}\left(x,\ y\right) = concat(\boldsymbol{I_c}\left(x,\ y\right),\ \boldsymbol{I_d}\left(x,\ y\right)), \tag{2}$$

$$\boldsymbol{f_h} = hog(\ \boldsymbol{I_{cd}}\ ). \tag{3}$$

To achieve strong robustness against changes in illumination and slight shape variations, the HOG is used here. The HOG is a feature extraction algorithm, using the distribution of gradient or edge directions, which is robust against multiple postures, complex backgrounds, and varying light sources. The gradient at the pixel point $\boldsymbol{I_{cd}}(x, y)$ is calculated as

$$G_x(\boldsymbol{I_{cd}}(x, y)) = H(\boldsymbol{I_{cd}}(x+1, y)) - H(\boldsymbol{I_{cd}}(x-1, y)), \tag{4}$$

$$G_y(\boldsymbol{I_{cd}}(x, y)) = H(\boldsymbol{I_{cd}}(x, y+1)) - H(\boldsymbol{I_{cd}}(x, y-1)). \tag{5}$$

In the given equations, $G_x(x, y)$, $G_y(x, y)$, and $H(x, y)$ represent the horizontal gradient, vertical gradient, and pixel intensity at position (x, y), respectively.

$$G(\boldsymbol{I_{cd}}(x, y)) = \sqrt{G_x\left(I_{cd}\left(x,y\right)\right)^2 + G_y\left(I_{cd}\left(x,y\right)\right)^2} \tag{6}$$

$$\alpha(\boldsymbol{I_{cd}}(x, y)) = \tan^{-1}\left(\frac{G_y\left(I_{cd}\left(x,y\right)\right)}{G_x\left(I_{cd}\left(x,y\right)\right)}\right) \tag{7}$$

Here, $G(x, y)$ represents the magnitude of the gradient at pixel (x, y), and $\alpha(x, y)$ represents the direction of the gradient at pixel (x, y). These values are essential for constructing the HOG features used in object detection and recognition tasks. Actually, many visualization techniques for vector signal gradient fields by multivariate data analysis have been proposed to obtain the optimal solution.(25)

3.2 HMMs with three stages

An extended HMM is adopted to analyze time-series information within surveillance scenarios as depicted in Fig. 3. The model architecture incorporates multiple hidden states $x = \{x_1, x_2, \ldots, x_t\}$ and corresponding observation states $y = \{y^1, y^2, \ldots, y^t\}$. The observation sequence is derived from the features obtained during the feature extractor stage. At each time t, the feature vector is represented as $y^t = \{F_s^t, F_p^t, F_b^t, F_{st}^t, F_{lt}^t, f_h^t\}$. Each hidden state affects not only its subsequent state but also multiple related observation values, reflecting the complex dependences between states and observations.

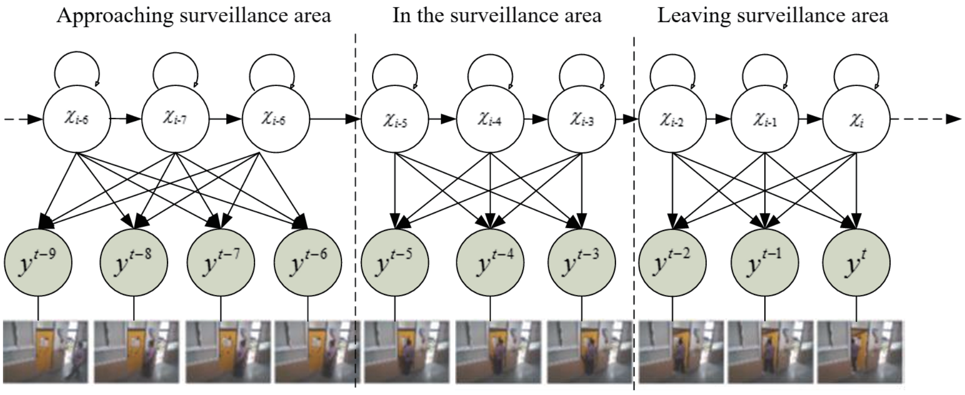


Fig. 3. (Color online) Proposed HMM-based approach with three stages.

The transition probabilities for a state sequence x of length t are represented as

$$P(x_t \,|x_1,x_2,\dots,x_{t-1}) = P(x_t \,|x_{t-1}). \tag{8}$$

The conditional independence of the observation parameter y is

$$P(y^t \,|y^1,y^2,\dots,y^l,x_1,x_2,\dots,x_{t-1}) = P(y^t|x_t)\,. \tag{9}$$

ω_T represents the weight of anomalous behavior occurring at time T. By combining the transition probabilities and the conditional independence of the observation parameters, we can obtain the most likely hidden state sequence probability.

$$P(y,x) = \prod_{t=1}^{T} P(x_t \,|x_{t-1})P(y^t \,|x_t)\omega_T \tag{10}$$

The proposed HMMs are partitioned into three distinct sections: approaching the surveillance area, entering the surveillance area, and leaving the surveillance area. The surveillance area is defined as the region within 1.5 m of doors or windows. Segmenting the surveillance process into three distinct phases aids in more accurately simulating and understanding the behavior patterns of objects in/out of the surveillance area. This segmentation improves the model's sensitivity to specific activities at different stages and allows for the focused allocation of computational resources, thereby enhancing the efficiency of the model. To enhance the model's capability to capture the diversity of each state more effectively, the model processes the features and information obtained from each section to form the observation sequence. Note that the features obtained at each stage vary, with certain features being collected only within the surveillance area. These include posture and behavior characteristics, as well as short-term features such as the duration of an individual's stay within the surveillance area. This design facilitates the capture of dynamics across time steps and the long-term effects of latent states.

Additionally, the structure of the model accounts for potential nonsequential relationships between observation values and hidden states, which is particularly advantageous for handling data with high temporal dependence and sparsity. Such a framework significantly enhances the model's predictive capability for anomalous behaviors where anomalies may manifest over multiple time steps.

4. Implementation and Experimental Results

4.1 Data preparation

To satisfy the requirement of practical applications, a dataset is gathered and labeled by utilizing a Kinect v1 depth camera to capture depth and color images with a resolution of 640 × 480. Actually, videos of the behaviors of two individuals are gathered as training data. They are recorded as 1736 frames across 77 video sequences and 2530 frames across 87 video sequences, making a total of 164 video sequences used as training data. Because anomaly intruders are usually strangers, videos of six other individuals simulating anomaly behavior are gathered as test data at the same time. There is basic information about the six individuals that we observed for testing data in Table 1. Finally, 4266 and 6894 videos are used as the training and test data, respectively, as shown in Table 2. For a more detailed evaluation, four categories of the dataset, namely, non-anomalous, thief, sneaky, and anomalous with distributions of 39, 18, 29, and 14%, respectively, are obtained by splitting the original anomalous data into the three categories. The dataset is designed to thoroughly train and evaluate the model, ensuring that it can accurately differentiate between anomalous and non-anomalous behaviors. By including a variety of behavior types and a significant amount of data for both training and testing, we developed a model that is better equipped to handle real-world scenarios and enhance security measures effectively.

Table 1
Basic information of six individuals for test dataset.

ID	High	Educational qualification	Dominant hand	Number of videos
A	177	Undergrad	Right hand	44
B	181	Undergrad	Right hand	35
C	178	Undergrad	Left hand	20
D	171	Undergrad	Right hand	30
E	162	Undergrad	Right hand	20
F	172	Graduate	Right hand	15

Table 2
Numbers of videos of training and test datasets.

	Anomalous	Non-anomalous
Training data	2502	1764
Test data	4018	2876

4.2 Experiments of anomalous behavior detection and anomalous classification

Table 3 shows the results of the proposed HMM-based and naïve Bayes-based approaches for anomalous behavior detection.

According to the observations, the proposed HMM-based approach with the time-series analysis capability demonstrates higher accuracy and reliability in detecting anomalous behavior than the naïve Bayes-based approaches. It effectively identifies a significant number of actual threats, with 186 true positives and only 6 false negatives, ensuring that most anomalous behaviors are detected and addressed promptly. The proposed system also distinguishes non-anomalous behaviors very well, with 61 true negatives, thereby reducing false alarms and unnecessary disruptions. The HMM-based approach shows a balanced performance with a precision of 76.2%, a recall of 96.9%, and an F1-score of 85.3%, indicating strong overall effectiveness. The naïve Bayes-based approaches also perform well, with a precision of 67.6%, a recall of 95.8%, and an F1-score of 79.3%. These results suggest that the system is highly effective in enhancing surveillance by accurately detecting and differentiating between anomalous and non-anomalous behaviors, thereby improving overall safety and security. Compared with the deep-learning-based approaches such as the use of the transformer or CNN, the proposed HMM-based approach does not lose its correctness significantly and achieves a real-time performance, especially for an embedded system with limited computational power. This result indicates that the proposed HMM-based approach is practical in real life.

For a more detailed analysis, four categories of anomalous behavior classification are used by the HMM-based approach proposed here. The confusion matrix reveals several positive aspects of the system's performance shown in Table 4. Notably, the system excels at identifying non-anomalous behaviors, with 61 correct classifications, highlighting its reliability in recognizing normal, nonthreatening actions.

It also effectively detects sneaky behaviors, correctly identifying 33 instances, which enhances security by catching subtle, potentially suspicious activities. Additionally, the system shows a moderate capability to detect thief behaviors, correctly classifying 16 instances,

Table 3
Performance characteristics of proposed HMM-based and naïve Bayes-based approaches.

	Precision (%)	Recall (%)	F1-score (%)
HMMs	76.2	96.9	85.3
naïve Bayes	67.6	95.8	79.3

Table 4
Confusion matrix of identification results.

	Sneaky	Anomalous	Thief	Non-anomalous
Sneaky	**33**	17	14	3
Anomalous	19	**10**	34	1
Thief	21	22	**16**	2
Non-anomalous	20	18	21	**61**

indicating a foundational capability in this area. The balanced distribution of misclassifications suggests that the system does not exhibit significant bias towards any particular type of error.

5. Conclusions

We investigated an HMM-based approach that can successfully sense various forms of anomalous behavior utilizing RGB-D images with a depth map that can capture intricate motion in 3D video details. To obtain human action information, both spatial and temporal features are used. The spatial features include shape, posture, behavior, and HOGs. Temporal features including short- and long-time observations are also extracted. Short-time observations are for behavior time within one day. Long-time observations aim at the haunting time and frequency of occurrence. The inclusion of comprehensive training data significantly improved the system's accuracy, ensuring the reliable detection of both subtle and overt suspicious activities. Compared with the deep-learning-based approach, the proposed approach can achieve real-time detection with acceptable computation complexity. The proposed three-stage HMM-based approach provided a balance between computational efficiency and detection accuracy, making the system suitable for real-time applications. According to the experimental results, the proposed HMM-based approach outperforms the naïve Bayes-based approaches for anomalous behavior detection. This is particularly important in environments with limited computational resources, such as an embedded system. The system accurately detects anomalous behavior in a security-sensitive environment.

The main contributions of this research are as follows. First, the HMMs with 3D related information offer a practical, reliable, and efficient solution to defining three stages according to the behaviors, namely, approaching, entering, and leaving the surveillance area. Second, the proposed method adopted in this paper can indeed identify anomalous behaviors as a real-time application in embedded systems. Finally, the dataset was collected under the premise of single-person scenarios with simple backgrounds captured by 3D cameras.

In future research, the HMM-based surveillance system should be expanded to more applications. By incorporating a broader range of spatial features, such as 3D point cloud data and advanced texture analysis techniques, the system can be used to identify and analyze objects beyond human figures, including props and weapons, thereby increasing the system's application scope and flexibility. Currently, although the system has achieved the extraction of short- and long-term observational features, the analysis of temporal behavioral features remains somewhat coarse. Future studies will focus on refining and subdividing temporal features to enhance the overall accuracy of behavior identification. Additionally, given the importance of accuracy in monitoring security systems operating in environments with limited computational resources, we will continue to explore algorithmic optimizations that maintain low computational demands while improving accuracy.

Acknowledgments

This research was funded by the National Science and Technology Council, Taiwan, under grant numbers 103-2622-E-415 -003 -CC3 and NSTC 111-2221-E-415-012-MY3.

References

1 A. Mukherjee, V. Hassija, and V. Chamola: IEEE Trans. Consum. Electron. **70** (2024) 3. https://doi.org/10.1109/TCE.2024.3440520

2 R. Raja, P. C. Sharma, M. R. Mahmood, and D. K. Saini: Multimedia Tools Appl. **82** (2023) 8. https://doi.org/10.1007/s11042-022-13954-1

3 Y. Wang, D. Cao, S. L. Chen, Y. M. Li, Y. W. Zheng, and N. Ohkohchi: World J. Gastrointestinal Surgery **13** (2021) 9. https://doi.org/10.4240/wjgs.v13.i9.904

4 Y. Lu, J. Gao, Q. Yu, Y. Li, and Y. Lv, and H. Qiao: IEEE Trans. Intell. Transp. Syst. **24** (2023) 7. https://doi.org/10.1109/TITS.2023.3264573

5 P. Culbertson, J. Slotine, and M. Schwager: IEEE Trans. Rob. **37** (2021) 6. https://doi.org/10.1109/TRO.2021.3072021

6 P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea: IEEE Trans. Circuits Syst. Video Technol. **18** (2008) 11. https://doi.org/10.1109/TCSVT.2008.2005594

7 G. Zhou and Y. Wu: Proc. 2009 Inter. Conf. Information Engineering and Computer Science (IEEE, 2009) 1–4. https://doi.org/10.1109/ICIECS.2009.5364586

8 B. Mor, S. Garhwal, and A. Kumar: Arch. Comput. Methods Eng. **28** (2021) 1429. https://doi.org/10.1007/s11831-020-09422-4

9 M. L. Gámiz, F. Navas-Gómez, R. Raya-Miranda, and M. C. Segovia-García: Reliab. Eng. Syst. Saf. **239** (2023) 109498. https://doi.org/10.1016/j.ress.2023.109498

10 Y. Feng and J. Liu: Sens. Mater. **36** (2024) 1. https://sensors.myu-group.co.jp/sm_pdf/SS4423.pdf

11 H. Gao, X. Wang, Z. Liu, and Y. Jiang: Sens. Mater. **36** (2024) 8. https://doi.org/10.18494/SAM4786

12 D. Lindsay and S. Cox: Pattern Recognition and Data Mining (Springer, Heidelberg, 2005) pp. 35–44. https://doi.org/10.1007/11551188_4

13 K. B. Ovhal, S. S. Patange, R. S. Shinde, V. K. Tarange, and V. A. Kotkar: Proc. 2017 Inter. Conf. Intelligent Sustainable Systems Conf. (IEEE, 2017) 596–601. https://doi.org/10.1109/ISS1.2017.8389240

14 M. Agarwal, P. Parashar, A. Mathur, K. Utkarsh, and A. Sinha: Proc. Adv. Data Computing, Communication and Security Conf. (I3CS, 2021) 647–658. https://doi.org/10.56726/IRJMETS50202

15 T. Xue and H. Liu: Proc. Communications, Signal Processing, and Systems (CSPS, 2021) 863–869. https://doi.org/10.1007/978-981-19-0390-8_108

16 R. San-Segundo, J. D. Echeverry-Corre, C. Salamea, and J. M. Pardo: IEEE Instrum. Meas. Mag. **19** (2016) 6. https://doi.org/10.1109/MIM.2016.7777649

17 X. Liu, and A. Datta: Proc. AAAI 26th Artificial Intelligence (AAAI, 2021) 1938–1944. https://doi.org/10.1609/aaai.v26i1.8395

18 A. Beghdadi and M. Mallem: Mach. Vision Appl. **33** (2022) 54. https://doi.org/10.1007/s00138-022-01306-w

19 V. Bansal, K. Balasubramanian, and P. Natarajan: SN Appl. Sci. **2** (2020) 1131. https://doi.org/10.1007/s42452-020-2815-z

20 J. Lee and C. Kim: Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition (IEEE, 2019) 9729–9738. https://doi.org/10.1109/CVPR.2019.00996

21 J. Liu, S. Tsujinaga, S. Chai, H. Sun, T. Tateyama, and Y. Iwamoto: IEEE Sens. J. **21** (2021) 23. https://doi.org/10.1109/JSEN.2021.3122128

22 X. Meng, C. Fan, Y. Ming, and H. Yu: IEEE Trans. Circuits Syst. Video Technol. **32** (2022) 7. https://doi.org/10.1109/TCSVT.2021.3128505

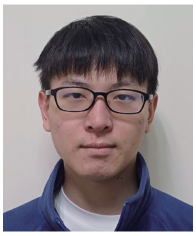
23 H. B. Naeem, M. H. Yousaf, F. H. Khan, and A. Yasin: Proc. 2021 Inter. Conf. Artificial Intelligence (IEEE, 2021) 193–197. https://doi.org/10.1109/ICAI52203.2021.9445205

24 R. Urtasun, D. J. Fleet, and P. Fua: Comput. Vision Image Understanding **104** (2006) 2. https://doi.org/10.1016/j.cviu.2006.08.006

25 X. Du, H. Liu, H.-W. Tseng, and T.-H. Meen: Symmetry **12** (2020) 724. https://doi.org/10.3390/sym12050724

About the Authors

Jui-Feng Yeh received his B.S. degree in computer science and information engineering from National Chiao-Tung University, Hsinchu, Taiwan, in 1993, and his M.S. and Ph.D. degrees in computer science and information engineering from National Cheng Kung University, Tainan, Taiwan, in 1995 and 2006, respectively. He is currently a professor and chairman of the Department of Computer Science and Information Engineering, National Chiayi University, Chiayi, Taiwan, Republic of China (ROC). His research interests include artificial intelligence, speech signal processing, natural language processing, and affective computing. (ralph@mail.ncyu.edu.tw)



Shu-Po Hsu received his B.S. degree in computer science and information engineering from Tunghai University, Taichung, Taiwan, in 2022. He is currently a graduate student at National Chiayi University. His research interests include real-time image processing, medical signal processing, and multimodal machine learning. (super213578642@gmail.com)

Kai-Hsiang Yu received his B.S. degree in computer science and information engineering from I-Shou University in 2014. He then received his M.S. degree in computer science and information engineering from National Chiayi University, Chiayi, Taiwan, in 2016. He is currently working as a software engineer for Formosa Technologies Corporation. His research focuses on the development of software and firmware for computer vision and factory automation. (s1030495@mail.ncyu.edu.tw)