IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 39, No. 4, August, 2022, pp. 1371-1377
Journal homepage: http://iieta.org/journals/ts

Automatic Recognition for IoT Supervision Images Based on Modal Decomposition

Yang Wang[1*], Yifeng Wang[1], Shengyu Zhang[1], Shimei Lin[2], Chao Chen[3]

[1] School of Electronics and Communication Engineering, Shenzhen Polytechnic, Shenzhen 518055, China
[2] EVOC Intelligent Technology Company Limited, Shenzhen 518055, China
[3] Innovation Center of Industrial Edge Intelligence, Shenzhen 518055, China

Corresponding Author Email: wyang@szpt.edu.cn

https://doi.org/10.18280/ts.390431

Received: 2 May 2022
Accepted: 1 August 2022

Keywords:
modal decomposition, Internet of Things, supervision images, automatic recognition

ABSTRACT

The automatic recognition for Internet of Things (IoT) supervision images is a prerequisite for the detection of abnormalities in monitoring images. This technology is a developmental trend in video surveillance. Various video image detection and recognition methods have certain weaknesses, such as poor generalization ability and poor anti-interference ability. In response, this paper conducts a study on automatic recognition for IoT supervision images based on modal decomposition. The paper presents an overall framework of the IoT supervision system. For the problems of poor real-time performance and few samples that commonly exist in video stream target recognition, the paper proposes a dynamic modal decomposition-based feature extraction algorithm for IoT supervision video stream to build a suitable platform for IoT supervision image foreground segmentation. The paper selects a dictionary with rich elements and exchanges a high computation time for minimizing the reconstruction error generated by applying the dynamic modal decomposition method. Experimental results validate the effectiveness of the proposed algorithm.

1. INTRODUCTION

The current video surveillance system has ushered in the era of intelligence [1-6]. If a variety of application scenarios in various industry sectors integrate the video surveillance system with the IoT technology, it can effectively solve the shortcomings of the traditional security control supervision mode and achieve the dual function of monitoring and communication. In this way, we will have more intelligent remote monitoring and emergency command to meet the demand in traffic, water conservancy, oil fields, banks, telecommunications and other areas [6-14]. The video image-based IoT supervision system has a variety of functions such as information collection, transmission and data analysis and processing. By real-time and accurate grasp of the normal state and abnormal situation of detection targets, it can provide relevant supervisors with direct real-time information of detection targets, and enhance the judgment and decision-making rate in response to abnormal events [15-21]. The automatic recognition for IoT supervision images is a prerequisite for the detection of abnormalities in monitoring images to improve supervision efficiency and reduce the occurrence of false alarm on abnormal situation. This technology is a developmental trend in video surveillance.

Managing distributed intelligent surveillance systems is considered to be a major challenge. Rajavel et al. [22] detailed cloud-based object tracking and behaviour recognition systems, an emerging research area for the IoT. It can bring robustness and intelligence in distributed video surveillance systems by minimising network bandwidth and response time between wireless cameras and cloud servers. Fathy and Saleh [23] investigated the integration of modal decomposition techniques with software-defined networking (SDN) architecture to support delay-sensitive applications in IoT environments. Weapon detection in real-time video surveillance applications was deployed as a case study upon which multiple deep learning-based models are trained and evaluated for detection using precision, recall, and mean absolute precision. Results revealed improvement of up to 75.0% in terms of average throughput, up to 14.7% in terms of mean jitter, and up to 32.5% in terms of packet loss. Even though there are several approaches for identifying moving objects in the video, background reduction is the one that is most often used. JayaSuaha et al. [24] used an adaptive background model to create a mean shift tracking technique. In this situation, the background model is provided and updated frame-by-frame, and therefore, the problem of occlusion is fully eliminated from the equation. In MATLAB, the works are simulated, and their performance is evaluated using image-based and video-based metrics to establish how well they operate in the real world. Akilan et al. [25] proposed a surveillance robot, which is integrated into any type of household devices. It watches the premises and delivers a notification to the authorized person about the video processing. This system made every user to feel safer using such a kind of device while the Authorised person is away from home or when they have left their children and elderly relatives alone at home. This device plays vital role in surveillance. Vision sensors in IoT-connected smart cities play a vital role in the exponential growth of video data. Muhammad et al. [26] carried out a survey of functional video summarization (VS) methods to understand their pros and cons for resource-constrained devices, with the ambition to provide a compact tutorial to the community of researchers in the field. Further, it presented a novel saliency-aware VS framework, incorporating 5G-enabled IoT devices, which keeps only important data, thereby saving storage resources and providing representative data for immediate exploration.

As can be seen from the existing literature, IoT supervision images are usually camera-collected images, which are usually compressed and processed to improve data transmission efficiency. However, the low clarity and high noise level of the processed images pose a great challenge to the determination of the normal status and abnormalities of the detection targets. Various video image detection and recognition methods each have their own disadvantages, such as poor generalisation ability and poor anti-interference ability, and these methods are usually designed for specific small data sets, which are less accurate when applied to the detection and recognition of IoT supervision images. In response, this paper conducts a study on automatic recognition for IoT supervision images based on modal decomposition. The paper presents an overall framework of the IoT supervision system in Chapter 2. For the problems of poor real-time performance and few samples that commonly exist in video stream target recognition, the paper proposes a dynamic modal decomposition-based feature extraction algorithm for IoT supervision video stream to build a suitable platform for IoT supervision image foreground segmentation. In Chapter 3, the paper selects a dictionary with rich elements and exchanges a high computation time for minimizing the reconstruction error generated by applying the dynamic modal decomposition method. Experimental results validate the effectiveness of the proposed algorithm.

2. FOREGROUND SEGMENTATION OF IOT SUPERVISION IMAGES

Considering the industrial IoT system architecture, this paper divides the IoT supervision system into three parts according to its functional division: IoT supervision terminal, data transmission system and IoT remote monitoring cloud platform. Figure 1 shows the overall framework of the IoT supervision system.

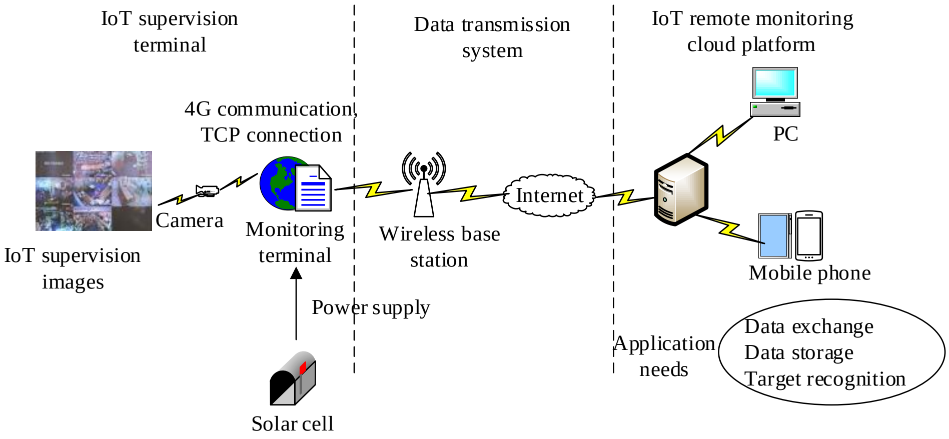


Figure 1. The overall framework of the IoT supervision system

Dynamic feature extraction of monitoring moving targets is a basic task of automatic recognition of IoT supervision images, which can be achieved by effectively separating foreground and background of IoT monitoring video streams. To reduce the impact of dynamic changes in illumination, background and foreground in real application scenes on the recognition effect, a variety of methods have been proposed by domestic and foreign scholars, but most of them are unable to process the local dynamic information in the spatiotemporal data of IoT supervision. Meanwhile, in the practical application world, automatic recognition of targets will become difficult if there is no complete background frame in the IoT supervision images with complex environment. Therefore, this paper conducts a study on automatic recognition for IoT supervision images based on modal decomposition. For the problems of poor real-time performance and few samples that commonly exist in video stream target recognition, the paper proposes a dynamic modal decomposition-based feature extraction algorithm for IoT supervision video stream to build a suitable platform for IoT supervision image foreground segmentation.

The dimensionality of video surveillance image data is very high in many applications. This paper uses a dynamic modal decomposition method to perform a spatiotemporal decomposition of IoT supervision image sequences, so as to achieve dimensionality reduction of the image data, which can be specifically achieved based on a data-driven decomposition of the Koopman operator spectrum. If the A pixel matrix characterising IoT supervision images is $L \times M$ and $L >> 1$, it is difficult to compute the eigenvectors of $A^{(m)}$ $(A^{(m\text{-}1)})^+$. The eigenvectors of $\dot{X}_s$ can be computed by first decomposing them into singular values and retaining only the first $s(<< L)$ orders, which is more computationally efficient to handle. Assuming that the approximate matrix is represented by X and $\dot{X}_s$, we had:

$$\left(A^{(m-1)}\right) = V_s \sum\nolimits_s U_s^+ \tag{1}$$

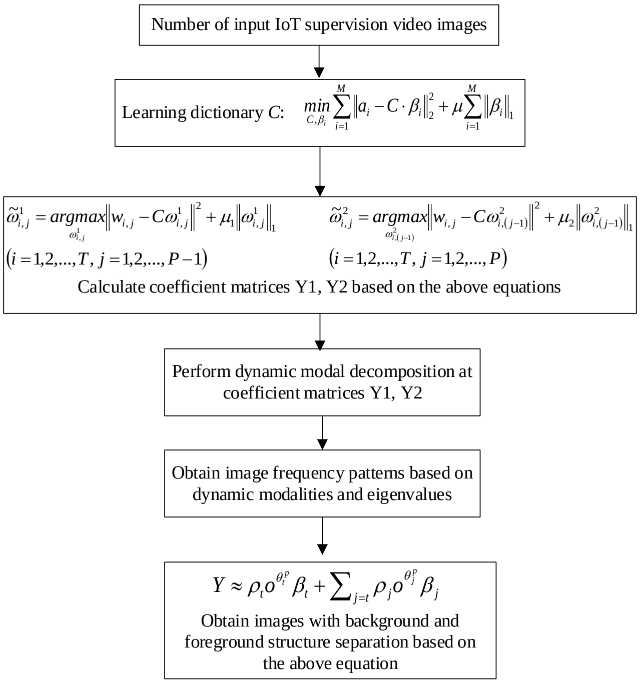
$$\dot{X}_s = V_s^+ \sum\nolimits_s U_s \tag{2}$$

X and $\dot{X}_s$ are $L \times L$ and $s \times s$ respectively, with the first s eigenvectors being the same. Assume that the feature vectors of $\dot{X}_s$ are $\dot{X}_s = Q\Gamma_s Q^+$. The feature vectors corresponding to the normal and abnormal conditions of different detection targets can be classified based on the feature values in Γ_r.

To extract features of IoT supervision images with complex environments and without complete background frames, this paper optimizes the traditional dynamic modal decomposition method, which is based on dictionary learning to extract the most essential dynamic features of IoT supervision video streams and reduce the interference of information less relevant to the monitoring target on the automatic recognition effect. Figure 2 gives the IoT supervision image foreground extraction process. A good dictionary has a sufficiently sparse model. Figure 3 shows the principle of background and foreground separation of IoT supervision images. Assuming that the input IoT supervision video image of dimension c_p is represented by a, the dictionary model of dimension c_p * L *is* represented by C, and the sparse matrix of dimension L *is* represented by β, we had:

$$a = C \cdot \beta \tag{3}$$

Assuming that the open square after the square of each component vector of β *is* represented by ||β||, the purpose of the IoT supervision image feature extraction is to find min||β||.



Number of input IoT supervision video images
Learning dictionary C: $\min_{C,\beta_i} \sum_{i=1}^{M} \|a_i - C \cdot \beta_i\|_2^2 + \mu \sum_{i=1}^{M} \|\beta_i\|_1$
$\tilde{\omega}_{i,j}^1 = \underset{\omega_{i,j}^1}{argmax} \left\| w_{i,j} - C\omega_{i,j}^1 \right\|^2 + \mu_1 \left\| \omega_{i,j}^1 \right\|_1$
$(i = 1,2,...,T,\ j = 1,2,...,P-1)$
$\tilde{\omega}_{i,j}^2 = \underset{\omega_{i,(j-1)}^2}{argmax} \left\| w_{i,j} - C\omega_{i,(j-1)}^2 \right\|^2 + \mu_2 \left\| \omega_{i,(j-1)}^2 \right\|_1$
$(i = 1,2,...,T,\ j = 1,2,...,P)$
Calculate coefficient matrices Y1, Y2 based on the above equations
Perform dynamic modal decomposition at coefficient matrices Y1, Y2
Obtain image frequency patterns based on dynamic modalities and eigenvalues
$Y \approx \rho_t o^{\theta_t^p} \beta_t + \sum_{j=t} \rho_j o^{\theta_j^p} \beta_j$
Obtain images with background and foreground structure separation based on the above equation

Figure 2. IoT supervision image foreground extraction process

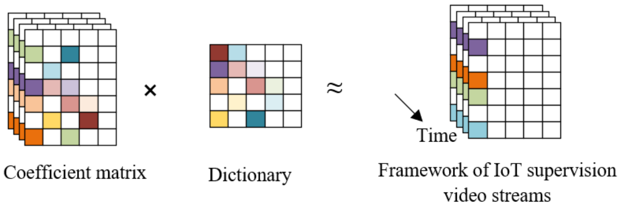


Figure 3. Principle of background and foreground separation of IoT supervision images

The problem of constructing the dictionary model C *is* approximated as a bi-objective optimization problem with the objectives of ensuring that C and β_i can reconstruct a_i as distortion-free as possible and ensuring that x_i is as sparse as possible.

$$\min_{C,\beta_i} \sum_{i=1}^{M} \|a_i - C \cdot \beta_i\|_2^2 + \mu \sum_{i=1}^{M} \|\beta_i\|_1 \tag{4}$$

The m image samples are randomly selected from the IoT supervision image sample set A as the initial samples for the dictionary model C, and β *is* set to 0.

The solution to a_i is discussed in this paper for a sample of IoT supervision images. The sample is assumed to be an a-vector and the sparse code to be a β-vector. If a and C are known, solving for β and needing it to be as sparse as possible means that the fewer non-zero elements of the matrix are required, the better.

Assume that C *is* represented by [$c1$, $c2$, $c3$, $c4$, $c5$] containing five matrix elements. This paper first finds the closest element to a. Assuming that it is c_4, then it derives β = [0, 0, 0, $d4$, 0], where the size of d_4 characterizes the weight of the matrix elements. Assuming that $a = d_4 * c_4$, the value of d_4 can be obtained by calculation. Based on the result of the calculation, find the residual vector $a' = a - d_3 * d3$ and stop the algorithm when a' is less than the pre-set threshold, and go to the next step if a' is greater than the pre-set threshold.

Calculate the closest distance to a' in c_1, c_2, c_3, c_4, c_5. If assuming it is c_2, then update x = [0, d_2, 0, $d4$, 0]. Assume $a = d_2 *c_2 + d_4 *c_4$, then update the residual vector $a' = a - d_2 *c_2 - d_4 *c_4$ based on the result of the d_2 calculation. Determine if it is less than the threshold, and if not, continue to find the coefficient of the next closest element. The number of coefficients of the solved elements can also be specified as a constant, for example 3, that would represent 3 iterations. Once all the x_i have been found, C can be updated.

Coefficient matrix Y_1=$\{\tilde{\omega}^1{}_{i,1}, \tilde{\omega}^1{}_{i,2}, ..., \tilde{\omega}\}^1{}_{i,P}{}^t{}_{i=1}$ and Y_2=$\{\tilde{\omega}^2{}_{i,2}, \tilde{\omega}^2{}_{i,2}, ..., \tilde{\omega}\}^2{}_{i,P-}{}^t{}_{i=1}$, ($Y_1$,$Y_2 \in R^{M\times(P\text{-}1)}$) are all approximated IoT supervision image sequence block coefficient matrices W_1=$\{w_{i,1}, w_{i,2}, ..., w\}_{i,P\text{-}1}{}^t{}_{i=1}$ and W_2=$\{w_{i,2}, w_{i,3}, ..., w\}_{i,P}{}^t{}_{i=1}$, ($W_1$, $W_2 \in R^{M\times(P\text{-}1)}$), which are obtained through dictionary learning training.

Assuming that the column vector containing the overlapping patches T is denoted by $\{.\}^t{}_{i=1}$, the number of rows of the frame sequence matrix and the coefficient matrix is denoted by M and L respectively, the colour block along all frame sequences is denoted by W=$\{w_{i,j}\}^t{}_{i=1}(W \in R^{M\times P})$, the regularization parameters used to control the sparsity in the coefficient matrices Y_1 and Y_2 are denoted by μ_1 and μ_2, and the coefficient matrix approximation can be obtained by solving the minimization problem shown in the following equation:

$$\begin{gathered} \tilde{\omega}_{i,j}^1 = \underset{\omega_{i,j}^1}{argmax} \left\| w_{i,j} - C\omega_{i,j}^1 \right\|^2 + \mu_1 \left\| \omega_{i,j}^1 \right\|_1 \\ (i = 1,2,...,T,\ j = 1,2,...,P-1) \end{gathered} \tag{5}$$

$$\begin{gathered} \tilde{\omega}_{i,j}^2 = \underset{\omega_{i,(j-1)}^2}{argmax} \left\| w_{i,j} - C\omega_{i,(j-1)}^2 \right\|^2 + \mu_2 \left\| \omega_{i,(j-1)}^2 \right\|_1 \\ (i = 1,2,...,T,\ j = 1,2,...,P) \end{gathered} \tag{6}$$

By considering the coefficient matrices Y_1 and Y_2 as the solution of the basis functions, this is similar to the extended dynamic modal decomposition method.

Suppose that a set of dynamic modes of IoT supervision images is represented by $\rho = \{\rho_1, ..., \rho_2\}$ and the corresponding feature values are represented by Γ=$\{\Gamma_1, ..., \Gamma_s\}$. Based on both ρ and Γ, this paper reconstructs the IoT supervision image sequences, and the number of feature vectors used is denoted by s. $\rho = \{\rho_1, ..., \rho_2\}$ denotes that in the IoT supervision video streams, the monitoring target with changes at the time point $p \in$ {0, 1, 2, ..., P-1} has an associated continuous time frequency, i.e.:

$$\theta_j = \frac{log\left(\Gamma_j\right)}{\Gamma p} \tag{7}$$

Suppose that the column vector of the *j*-th dynamic mode containing the spatial structure information is represented by ρ_j and the initial amplitude of the corresponding dynamic mode decomposition method mode is represented by β_j. Then the approximate video frames of different frequency modes at any time point can be reconstructed in the following way:

$$Y\left(p\right) \approx \sum_{j=1}^{s} \rho_j o^{\theta_j^p} \beta_j = \Omega o^{\chi p} \beta \tag{8}$$

Vector β can be obtained by taking the supervision video image at the starting time point. Figure 4 shows the principle of reconstructing the coefficient matrix. Such a process effectively reduces the computational process of $\{\tilde{\omega}^1{}_{i,1}\}^t{}_{i=1}=\rho x$. Since the eigenvector matrix involved in the calculations in this paper is not a square matrix, β can be calculated by the pseudo-inverse procedure shown in the following equation:

$$\beta = \Omega^{+} \left\{\tilde{\gamma}_{i,1}^{1}\right\}_{i=1}^{t} \tag{9}$$

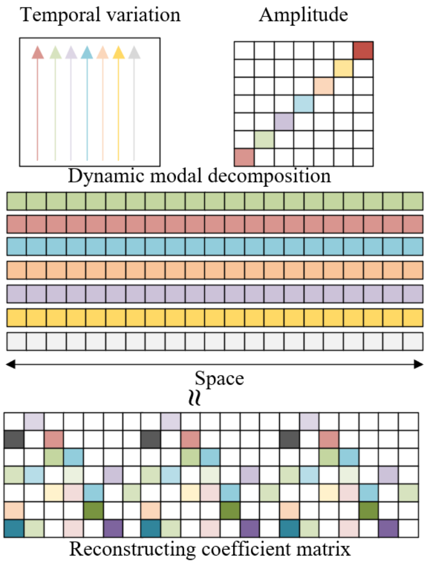

Temporal variation
Amplitude
Dynamic modal decomposition
Space
≈
Reconstructing coefficient matrix

Figure 4. Principle of reconstructing coefficient matrix

The key operation for separating IoT supervision video images into foreground and background is to threshold the images in low frequency mode based on the feature values of the image foreground and background. Basically, the image block representing the background in the IoT supervision video image is constant in the video stream; and when $t \in \{1,2,..., s\}$, it satisfies the equation $|\theta_t| \approx 0$. In particular, it is important to note that background structure features near the spatial origin in real surveillance scenes are characterised through a single modality. $|\theta_t$ | denotes the eigenvalues of foreground structures far from the spatial origin.

Assume that the background part of the image is represented by $\rho_t o^{\theta}{}_t{}^p x_t$, the foreground part is represented by $\Sigma_{j=T}\rho_j o^{\theta}{}_j{}^p x_j$, and the reconstructed coefficient matrix is represented by $Y^{\sim}=\{\tilde{\omega}^1{}_{i,1},\tilde{\omega}^1{}_{i,2},...,\tilde{\omega}^1{}_{i,P}\}^t{}_{i=1}$, and $p=\{0,1,2,... ,P\text{-}1\}$ is the time index to the (P-1) image frame. Thus, the IoT supervision video image that is divided into background and foreground structures can be represented as:

$$Y \approx \rho_t o^{\theta_t^p} \beta_t + \sum_{j=t} \rho_j o^{\theta_j^p} \beta_j \tag{10}$$

The initial amplitude value $x_j=\rho_j{}^{+}\{\tilde{\omega}^1{}_{i,1}\}^t{}_{i=1}$ of the stationary background is constant and the initial amplitude of the changing foreground structure is represented by $x_j=\rho_j{}^{+}\{\tilde{\omega}^1{}_{i,1}\}^t{}_{i=1}$, $\forall j \neq t$. Then the fully-spreading approximate IoT supervision video image sequence B^* can go through dictionary reconstruction based on the following equation:

$$\left\{b_{i,j}{}^{*}\right\}_{i=1,j=1}^{T,P} = C\left\{\tilde{\omega}_{i,j}\right\}_{i=1,j=1}^{T,P} \tag{11}$$

3. DICTIONARY LEARNING AND SIGNAL APPROXIMATION

The learning dictionary has two learning strategies based on large datasets for offline learning or on current estimation for adaptive online learning. This paper chooses the former learning strategy for the approximation of spatiotemporal information of IoT supervision image sequences. The training process of the model is only once and delivers the advantage of a high run rate.

Random blocks are selected from a randomly selected stream of IoT supervision image and video to construct the training data for the model. In practical monitoring application scenarios, supervision image sequences contain mainly background information when no or only momentary dynamic detection targets are present. For such images, the input signal can be approximated based on images that contain both foreground and background information. Dictionaries with a large number of elements trade off high computation time for minimising the reconstruction error generated by applying dynamic modal decomposition methods, while dictionaries with fewer elements contain less information, which leads to a higher reconstruction error generated by dynamic modal decomposition methods.

Assuming that the total number of pixels in the input image sequence is represented by O and the reconstruction error is represented by U, the formula for the total reconstruction error RE calculated using the original IoT supervision image sequence is as follows:

$$RE = \sqrt{\frac{1}{O}\sum_{i=1}^{O}\left|U\left(i\right) - B^{*}\left(i\right)\right|} \tag{12}$$

This paper introduces a correction matrix D to minimise the reconstruction error arising from the application of the dynamic modal decomposition method, which is obtained by solving for the minimum value shown in the following equation:

$$\min_{D} \left\|U - DB^{*}\right\|_{G}^{2} + \mu_D \left\|B^{*}\right\|_{1} \tag{13}$$

4. EXPERIMENTAL RESULTS AND ANALYSIS

Figure 5 shows the different eigenvalues corresponding to the foreground and background parts present in the IoT supervision video image. The background position in a static state corresponds to the vicinity of the origin with a feature value of 0. The feature values of other dynamic points and moving targets correspond to positions far from the origin.

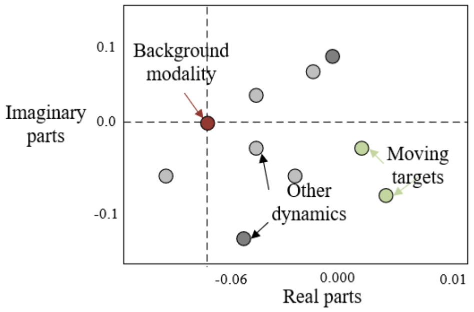


Figure 5. Scenario of different eigenvalues corresponding to foreground and background

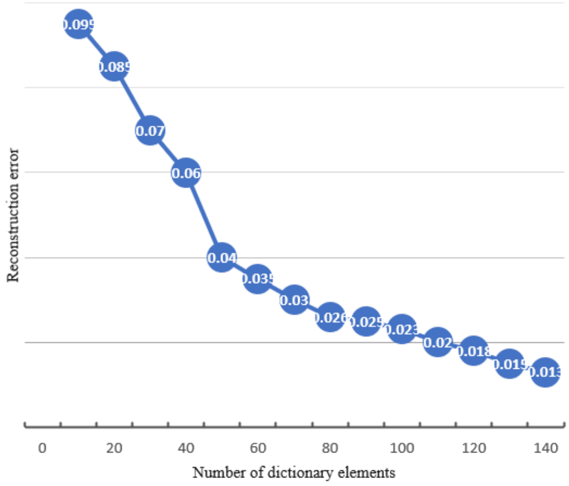


Figure 6. Variation curve of reconstruction error for different number of dictionary elements

The approximation B_1 and B_2 of the spatiotemporal information of the IoT supervision image sequences depends on the estimation of the coefficient matrices Y_1 and Y_2. To achieve these objectives, this paper uses the $L1$-parametric regularization method and a fast iterative shrinkage threshold algorithm to solve equations 5 and 6. The selection of μ_1 and μ_2 determines the number of non-zero coefficients in the sparse matrix. In the dynamic modal decomposition method used in this paper, the above parameters are all set manually to obtain the desired approximation of the signal. Figure 6 shows the variation curves of reconstruction error for different numbers of dictionary elements. To better approximate the input image sequence, more reasonable μ_1 and μ_2 are set to perform noise filtering on the IoT supervision image sequence based on a dictionary containing fewer elements.

Table 1. Automatic recognition results for different sample sets

Sample set number	Recall rate	Accuracy	*F1* value
1	0.517	0.569	0.537
2	0.795	0.735	0.769
3	0.831	0.913	0.451
4	0.852	0.658	0.701
5	0.416	0.537	0.537
6	0.629	0.618	0.684
7	0.537	0.635	0.795
8	0.428	0.594	0.537
9	0.730	0.861	0.729

Table 1 shows the results of the automatic recognition of different sample sets. As can be seen from the results, when the detection target is static for a long time, it is difficult to detect any changes that occur in it in the future period, which reduces the F1 value of the sample's automatic recognition. However, video images with minimal background changes have a high F1 value. The optimised dynamic modal decomposition method used in this paper has desirable F1 values for sample sets (2), (4), (7) and (9), as these sample sets are from different surveillance locations but are almost static throughout the surveillance period. The optimised dynamic modal decomposition method used in this paper can detect volumetrically different target objects while obtaining more desirable F1 values.

Table 2 gives the target recognition speed statistics for continuous images. It can be seen that the algorithm of this paper used for automatic recognition in continuous IoT supervision video images takes less than 40ms, and the performance is more satisfactory. In the training process of the proposed algorithm, the spatiotemporal decomposition of the image samples of the sample set has been carried out to complete the dimensionality reduction of the image data, which in turn effectively reduces the time of automatic image recognition and increases the efficiency of automatic recognition by more than one fifth.

To visually and clearly verify the recognition effectiveness of the algorithm proposed in this paper on the IoT supervision video images, the recognition results of four different products in the industrial IoT pipeline were compared in this paper, with the outputs corresponding to 1, 2, 3 and 4 in turn. The actual results were compared as shown by Figure 7, which reveals that 45 out of 50 product 1 samples were identified, delivering a recognition rate of 90%. Product 1 identified 44 with a recognition rate of 88%; product 3 identified 47 with a recognition rate of 94%; and product 4 identified 49 with a recognition rate of 98%, making the recognition accuracy rate quite desirable.

Table 2. Statistics of recognition speed for continuous images

Continuous images No.	Self-recognition time	Optimised recognition time	Continuous images No.	Self- recognition time	Optimised recognition time
66	41.2595	33.6294	76	42.5187	37.5984
67	47.5182	31.5219	77	45.6493	39.5281
68	43.6218	33.6285	78	41.5207	35.6094
69	40.1252	37.4158	79	46.2513	37.5681
70	43.6059	30.2519	80	48.5971	30.4162
71	45.1284	34.5182	81	43.6295	39.0528
72	43.2741	39.6251	82	44.5083	34.5102
73	46.0418	33.5302	83	46.5618	36.2911
74	49.5281	34.8152	84	42.5196	39.1625
75	43.6258	30.2619	85	48.6273	35.0217

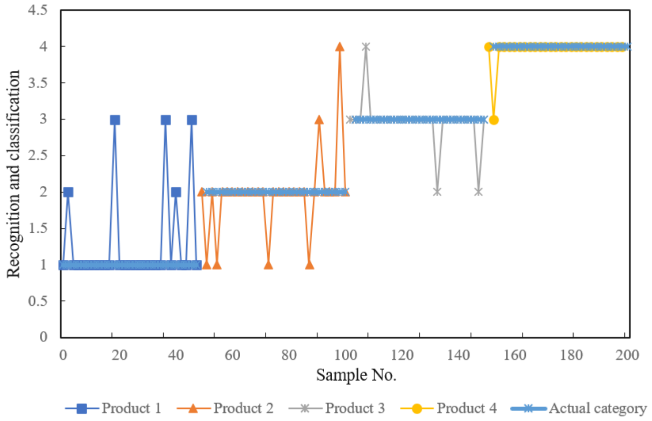


Figure 7. Results of the automatic recognition and classification of different products

5. CONCLUSION

This paper conducts a study on automatic recognition for IoT supervision images based on modal decomposition. The paper presents an overall framework of the IoT supervision system. For the problems of poor real-time performance and few samples that commonly exist in video stream target recognition, the paper proposes a dynamic modal decomposition-based feature extraction algorithm for IoT supervision video stream to build a suitable platform for IoT supervision image foreground segmentation. The paper selects a dictionary with rich elements and exchanges a high computation time for minimizing the reconstruction error generated by applying the dynamic modal decomposition method. Experimental results demonstrate the different feature values corresponding to the foreground and background parts present in the IoT supervision video images, plot the reconstruction error variation curves for different dictionary element numbers, and count the automatic recognition results for different sample sets. The results verify that the optimized dynamic modal decomposition method adopted in this paper can detect target objects with different volumes, while obtaining desirable F1 values. The results also give the target recognition speed statistics for continuous images, and verify that the algorithm of this paper performs better in automatic image recognition. Finally, the recognition results of four different products in the industrial IoT pipeline are compared to more visually and clearly verify the effectiveness of the algorithm proposed in this paper for the recognition of IoT supervision video images.

ACKNOWLEDGMENT

This study was supported by the Project of Science and Technology of Shenzhen (Grand No. GJHZ20200731095412038).

REFERENCES

[1] Deepak Raj, S., Ramesh Babu, H.S. (2022). Identification of intelligence requirements of military surveillance for a WSN framework and design of a situation aware selective resource use algorithm. Revue d'Intelligence Artificielle, 36(2): 251-261. https://doi.org/10.18280/ria.360209

[2] Sun, P., Liu, Q. (2022). Intelligent traffic accident detection system using surveillance video. In Proceedings of China SAE Congress 2020, pp. 995-1005. https://doi.org/10.1007/978-981-16-2090-4_61

[3] Li, P., Zhou, Z.J., Liu, Q.J., Sun, X.Y., Chen, F.M., Xue, W. (2021). Machine learning-based emotional recognition in surveillance video images in the context of smart city safety. Traitement du Signal, 38(2): 359-368. https://doi.org/10.18280/ts.380213

[4] Cheng, L., Wang, J., Li, Y. (2021). ViTrack: Efficient tracking on the edge for commodity video surveillance systems. IEEE Transactions on Parallel and Distributed Systems, 33(3): 723-735. https://doi.org/10.1109/TPDS.2021.3081254

[5] Liu, Y., Zhang, C. (2022). The auxiliary system of video surveillance in smart substation. In Journal of Physics: Conference Series, 2195(1): 012030. https://doi.org/10.1088/1742-6596/2195/1/012030

[6] Gayal, B.S., Patil, S.R. (2022). Detection and localization of abnormal events for smart surveillance. Ingénierie des Systèmes d'Information, 27(2): 233-241. https://doi.org/10.18280/isi.270207

[7] Surya Priya, M., Diana Josephine, D., Abinaya, P. (2021). IOT based smart and secure surveillance system using video summarization. In Advances in Computing and Network Communications, 423-435. https://doi.org/10.1007/978-981-33-6977-1_32

[8] Liu, Y., Kong, L., Chen, G., Xu, F., Wang, Z. (2021). Light-weight AI and IoT collaboration for surveillance video pre-processing. Journal of Systems Architecture, 114: 101934. https://doi.org/10.1016/j.sysarc.2020.101934

[9] Khudhair, A.B., Ghani, R.F. (2020). IoT based smart video surveillance system using convolutional neural network. In 2020 6th International Engineering Conference "Sustainable Technology and Development"(IEC), pp. 163-168. https://doi.org/10.1109/IEC49899.2020.9122901

[10] Gagliardi, A., Saponara, S. (2019). Distributed video antifire surveillance system based on IoT embedded computing nodes. In International Conference on Applications in Electronics Pervading Industry, Environment and Society, pp. 405-411. https://doi.org/10.1007/978-3-030-37277-4_47

[11] Muhammad, K., Hussain, T., Tanveer, M., Sannino, G., de Albuquerque, V.H.C. (2019). Cost-effective video summarization using deep CNN with hierarchical weighted fusion for IoT surveillance networks. IEEE Internet of Things Journal, 7(5): 4455-4463. https://doi.org/10.1109/JIOT.2019.2950469

[12] Che, R., Wang, L., Wang, Y., Lin, Q. (2019). Research on intelligent video surveillance system in remote area based on NB-IoT. In Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence, pp. 255-259. https://doi.org/10.1145/3377713.3377750

[13] Sultana, T., Wahid, K.A. (2019). IoT-guard: Event-driven fog-based video surveillance system for real-time security management. IEEE Access, 7: 134881-134894. https://doi.org/10.1109/ACCESS.2019.2941978

[14] Lee, D.G., Lee, D., Kwon, K. (2019). A CMOS wideband RF energy harvester employing tunable

impedance matching network for video surveillance disposable IoT applications. The Transactions of The Korean Institute of Electrical Engineers, 68(2): 304-309.

[15] Rego, A., Canovas, A., Jiménez, J.M., Lloret, J. (2018). An intelligent system for video surveillance in IoT environments. IEEE Access, 6: 31580-31598. https://doi.org/10.1109/ACCESS.2018.2842034

[16] Gallo, P., Pongnumkul, S., Nguyen, U.Q. (2018). BlockSee: Blockchain for IoT video surveillance in smart cities. In 2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), pp. 1-6. https://doi.org/10.1109/EEEIC.2018.8493895

[17] Vela-Medina, J.C., Guerrero-Sánchez, A.E., Rivas-Araiza, J.E., Rivas-Araiza, E.A. (2018). Face detection for efficient video-surveillance IoT based embedded system. In 2018 IEEE International Conference on Automation/XXIII Congress of the Chilean Association of Automatic Control (ICA-ACCA), pp. 1-6. https://doi.org/10.1109/ICA-ACCA.2018.8609835

[18] Lakshya, L., Kota, V.S., Voleti, M.R., Singh, S. (2021). Compressed domain consistent motion based frame scoring for IoT edge surveillance videos. In International Symposium on Visual Computing, pp. 534-545. https://doi.org/10.1007/978-3-030-90439-5_42

[19] Gulve, S.P., Khoje, S.A., Pardeshi, P. (2017). Implementation of IoT-based smart video surveillance system. In Computational Intelligence in Data Mining, 771-780. https://doi.org/10.1007/978-981-10-3874-7_73

[20] Hasan, R., Mohammed, S.K., Khan, A.H., Wahid, K.A. (2017). A color frame reproduction technique for IoT-based video surveillance application. In 2017 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-4. https://doi.org/10.1109/ISCAS.2017.8050236

[21] Feng, X., Ye, M., Swaminathan, V., Wei, S. (2017). Towards the security of motion detection-based video surveillance on IoT devices. In Proceedings of the on Thematic Workshops of ACM Multimedia 2017, pp. 228-235. https://doi.org/10.1145/3126686.3126713

[22] Rajavel, R., Ravichandran, S.K., Harimoorthy, K., Nagappan, P., Gobichettipalayam, K.R. (2022). IoT-based smart healthcare video surveillance system using edge computing. Journal of Ambient Intelligence and Humanized Computing, 13(6): 3195-3207. https://doi.org/10.1007/s12652-021-03157-1

[23] Fathy, C., Saleh, S.N. (2022). Integrating deep learning-based IoT and fog computing with software-defined networking for detecting weapons in video surveillance systems. Sensors, 22(14): 5075. https://doi.org/10.3390/s22145075

[24] JayaSudha, A.R., Dadheech, P., Prasad, K.R., Hemalatha, S., Sharma, M., Jamal, S.S., Krah, D. (2022). Intelligent wearable devices enabled automatic vehicle detection and tracking system with video-enabled UAV networks using deep convolutional neural network and IoT surveillance. Journal of Healthcare Engineering, 2022: 2592365. https://doi.org/10.1155/2022/2592365

[25] Akilan, T., Srivastava, R., Chandraprabha, M., Chaudhary, A., Garg, A., Verma, A.K. (2021). High secure wireless video surveillance robot using IOT technology. In 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 734-739. https://doi.org/10.1109/ICAC3N53548.2021.9725609

[26] Muhammad, K., Hussain, T., Rodrigues, J.J., Bellavista, P., de Macêdo, A.R.L., de Albuquerque, V.H.C. (2020). Efficient and privacy preserving video transmission in 5G-enabled IoT surveillance networks: Current challenges and future directions. IEEE Network, 35(2): 26-33. https://doi.org/10.1109/MNET.011.1900514