Automatic Recognition for IoT Supervision Images Based on Modal Decomposition

Yang Wang1*, Yifeng Wang1, Shengyu Zhang1, Shimei Lin2, Chao Chen3

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 surveillance 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](image)

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 $L \gg 1$, it is difficult to compute the eigenvectors of $A^{(m)} (A^{(m-1)})^{-1}$. The eigenvectors of $\hat{X}$ 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 $\hat{X}$, we had:

\[
\begin{align*}
(A^{(m-1)}) &= V \sum U^T, \quad (1) \\
\hat{X} &= V' \sum U, \quad (2)
\end{align*}
\]

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

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...
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 \( \beta \)-vector. If \( a \) and \( C \) are known, solving for \( \beta \) 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 \([c_1, c_2, c_3, c_4, c_5]\) containing five matrix elements. This paper finds the closest element to \( a \). Assuming that it is \( c_4 \), then it derives \( \beta = [0, 0, 0, d_4, 0] \), 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 \( a' = a - d_4 * c_4 \) 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 \( x = [0, d_2, d_2, d_4, 0] \). 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 \( d_4 \) 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 \( x \) have been found, \( C \) can be updated.

Coefficient matrix \( Y_1=\{\tilde{o}_{1,1}, \tilde{o}_{1,2}, ..., \tilde{o}_{1,l,1}\} \) and \( Y_2=\{\tilde{o}_{2,1}, \tilde{o}_{2,2}, ..., \tilde{o}_{2,l,1}\} \) are all approximated IoT supervision image sequence block coefficient matrices \( W_1=\{w_{1,1}, w_{1,2}, ..., w_{1,l,1}\} \) and \( W_2=\{w_{2,1}, w_{2,2}, ..., w_{2,l,1}\} \), which are obtained through dictionary learning training.

Assuming that the column vector containing the overlapping patches \( T \) is denoted by \( \{w_{1,j}\}_{i=1} \), 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 \( W=\{w_{1,j}\}_{i=1} \), the regularization parameters used to control the sparsity in the coefficient matrices \( Y_1 \) and \( Y_2 \) are denoted by \( \mu_1 \) and \( \mu_2 \), and the coefficient matrix approximation can be obtained by solving the minimization problem shown in the following equation:

\[
\tilde{o}_{i,j} = \text{argmax} \left\{ \sum_{i=1}^{M} \left( \| c_j - C \tilde{o}_{i,j} \|_2^2 + \mu_1 \| \beta \|_1 \right) \right\} \\
(i = 1, 2, ..., T, j = 1, 2, ..., P-1)
\]

\[
\tilde{o}_{i,j} = \text{argmax} \left\{ \sum_{i=1}^{M} \left( \| c_j - C \tilde{o}_{i,j} \|_2^2 + \mu_2 \| \beta \|_1 \right) \right\} \\
(i = 1, 2, ..., T, j = 1, 2, ..., P)
\]

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 \( \rho = \{\rho_1, ..., \rho_l\} \) and the corresponding feature values are represented by \( \Gamma = \{\Gamma_1, ..., \Gamma_l\} \). Based on both \( \rho \) and \( \Gamma \), this paper reconstructs the IoT supervision image sequences, and the number of feature vectors used is denoted by \( s \). \( p = \{p_1, ..., p_z\} \) 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.:
\[ \theta_j = \frac{\log(\Gamma_{i,j})}{\Gamma_p} \]  

(7)

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

\[ Y(p) = \sum_{j=1}^{n} \rho_j o^{\theta_j} \beta_j = \Omega o^{\theta_j} \beta \]  

(8)

Vector \( \beta \) 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 \( \{\hat{\omega}_j\} \). Since the eigenvector matrix involved in the calculations in this paper is not a square matrix, \( \beta \) can be calculated by the pseudo-inverse procedure shown in the following equation:

\[ \beta = \Omega^+ \{p_i\}_{i=1}^p \]  

(9)

**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 \( |\theta_t| = 0 \). 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 \( \rho o^{\theta} x_o \), the foreground part is represented by \( \Sigma_{j=1}^{n} p_j o^{\theta_j} x_o \), and the reconstructed coefficient matrix is represented by \( Y = \{\hat{\omega}_j\} \), and \( p = 0, 1, 2, ..., p-1 \) 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:

\[ Y = \rho o^{\theta} \beta + \sum_{j=1}^{n} \rho_j o^{\theta_j} \beta_j \]  

(10)

The initial amplitude value \( x_j = p_j \{\hat{\omega}_j\} \) of the stationary background is constant and the initial amplitude of the changing foreground structure is represented by \( y_j = p_j \{\hat{\omega}_j\} \). Then the fully-spreading approximate IoT supervision video image sequence \( B' \) can go through dictionary reconstruction based on the following equation:

\[ \{b_{,j}\} = C \{\hat{\omega}_{,j}\} \]  

(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:

\[ RE = \sqrt{\frac{1}{O} \sum_{i=1}^{O} [U(i) - B'(i)]^2} \]  

(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:

\[ \min_{B'} \|U - DB'\|_{\omega} + \mu\|B'\| \]  

(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.
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 $L_1$-parametric regularization method and a fast iterative shrinkage threshold algorithm to solve equations 5 and 6. The selection of $\mu_1$ and $\mu_2$ 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 $\mu_1$ and $\mu_2$ 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

<table>
<thead>
<tr>
<th>Sample set number</th>
<th>Recall rate</th>
<th>Accuracy</th>
<th>$F1$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.517</td>
<td>0.569</td>
<td>0.537</td>
</tr>
<tr>
<td>2</td>
<td>0.795</td>
<td>0.735</td>
<td>0.769</td>
</tr>
<tr>
<td>3</td>
<td>0.831</td>
<td>0.913</td>
<td>0.451</td>
</tr>
<tr>
<td>4</td>
<td>0.852</td>
<td>0.658</td>
<td>0.701</td>
</tr>
<tr>
<td>5</td>
<td>0.416</td>
<td>0.537</td>
<td>0.537</td>
</tr>
<tr>
<td>6</td>
<td>0.629</td>
<td>0.618</td>
<td>0.684</td>
</tr>
<tr>
<td>7</td>
<td>0.537</td>
<td>0.635</td>
<td>0.795</td>
</tr>
<tr>
<td>8</td>
<td>0.428</td>
<td>0.594</td>
<td>0.537</td>
</tr>
<tr>
<td>9</td>
<td>0.730</td>
<td>0.861</td>
<td>0.729</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Continuous images</th>
<th>Self-recognition</th>
<th>Optimised recognition</th>
<th>Continuous images</th>
<th>Self-recognition</th>
<th>Optimised recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>time</td>
<td>time</td>
<td>No.</td>
<td>time</td>
<td>time</td>
</tr>
<tr>
<td>66</td>
<td>41.2595</td>
<td>33.6294</td>
<td>76</td>
<td>42.5187</td>
<td>37.5984</td>
</tr>
<tr>
<td>67</td>
<td>47.5182</td>
<td>31.5219</td>
<td>77</td>
<td>45.6493</td>
<td>39.5281</td>
</tr>
<tr>
<td>68</td>
<td>43.6218</td>
<td>33.6285</td>
<td>78</td>
<td>41.5207</td>
<td>35.6094</td>
</tr>
<tr>
<td>69</td>
<td>40.1252</td>
<td>37.4158</td>
<td>79</td>
<td>46.2513</td>
<td>37.5681</td>
</tr>
<tr>
<td>70</td>
<td>43.6059</td>
<td>30.2519</td>
<td>80</td>
<td>48.5971</td>
<td>30.4162</td>
</tr>
<tr>
<td>71</td>
<td>45.1284</td>
<td>34.5182</td>
<td>81</td>
<td>43.6295</td>
<td>39.0528</td>
</tr>
<tr>
<td>72</td>
<td>43.2741</td>
<td>39.6251</td>
<td>82</td>
<td>44.5083</td>
<td>34.5102</td>
</tr>
<tr>
<td>73</td>
<td>46.0418</td>
<td>33.5020</td>
<td>83</td>
<td>46.5618</td>
<td>36.2911</td>
</tr>
<tr>
<td>74</td>
<td>49.5281</td>
<td>34.8152</td>
<td>84</td>
<td>42.5196</td>
<td>39.1625</td>
</tr>
<tr>
<td>75</td>
<td>43.6258</td>
<td>30.2619</td>
<td>85</td>
<td>48.6273</td>
<td>35.0217</td>
</tr>
</tbody>
</table>
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


