applied sciences

MDPI

Article

Target Identification via Multi-View Multi-Task Joint Sparse Representation

Jiawei Chen [1], Zhenshi Zhang [2] and Xupeng Wen [3,*]

[1] School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[2] Undergraduate School, National University of Defense Technology, Changsha 410073, China
[3] School of Traffic and Transportation Engineering, Central South University, Changsha 410073, China
* Correspondence: wenxupeng@csu.edu.cn; Tel.: +86-181-0125-1215

Abstract: Recently, the monitoring efficiency and accuracy of visible and infrared video have been relatively low. In this paper, we propose an automatic target identification method using surveillance video, which provides an effective solution for the surveillance video data. Specifically, a target identification method via multi-view and multi-task sparse learning is proposed, where multi-view includes various types of visual features such as textures, edges, and invariant features. Each view of a candidate is regarded as a template, and the potential relationship between different tasks and different views is considered. These multiple views are integrated into the multi-task spare learning framework. The proposed MVMT method can be applied to solve the ship's identification. Extensive experiments are conducted on public datasets, and custom sequence frames (i.e., six sequence frames from ship videos). The experimental results show that the proposed method is superior to other classical methods, qualitatively and quantitatively.

Keywords: target identification; multi-view multi-task; spare representation; feature extraction

check for updates

Citation: Chen, J.; Zhang, Z.; Wen, X. Target Identification via Multi-View Multi-Task Joint Sparse Representation. *Appl. Sci.* **2022**, *12*, 10955. https://doi.org/10.3390/app122110955

Academic Editors: Weitao Chen, Ailong Ma and Guohua Wu

Received: 6 September 2022
Accepted: 26 October 2022
Published: 28 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the emergence of various high-resolution visible light and infrared cameras, video monitoring technology has been rapidly developed and applied. Currently, many video cameras are deployed in streets, traffic, and ports, and these are mainly convenient for the staff at the monitoring center to observe and understand the site conditions of each monitoring station. However, the area covered by each monitoring point is relatively small. To expand the monitoring range, more and more video monitoring points are arranged, the workload of the personnel on duty becomes heavier, the work efficiency is relatively low, and they are no longer willing to use the monitoring video. Therefore, most video monitoring has become a kind of decoration at present. "How do I make the most of video surveillance?" It has become the focus and difficulty in the field of target identification. To solve the above problems, this paper studies the target identification method.

Recently, several computer vision and machine learning methods have shown competitive results in target identification and tracking [1,2], especially the deep learning-based method that has received extensive attention in recent years due to its excellent performance [3–7]. However, these methods rely on the parameter adjustment of the neural network model. The training time is long, and it is easy to consume a lot of computing resources [8]. Therefore, these models may be inefficient in the case of demanding tracking [9]. In addition, the visual tracking decomposition method [10] adopts the sparse principal component analysis method based on multiple features to build multiple basic trackers, but the tracking performance of this multi-feature method is not good. The sparse representation-based method [11–13] uses the particle filter framework [14] to target identification based on sparse representation. In addition, the multi-task learning [15] method is an efficient method, which learns the sparse representation of all particles in the particle filter framework [16]. Compared with the L1 method with sparse representation [17],

although the multi-task method (MTT) takes advantage of the interdependence between particles, the robustness of multi-task is not good due to the existence of outliers.

To identify targets more robustly, researchers have proposed a multi-task and multi-perspective learning method to optimize the target identification problem [18,19]. In this paper, we propose a target identification method based on multi-task multi-view sparse learning (MVMT).

The main contributions of this paper are summarised as follows:

(1) A multi-view multi-task method based on sparse representation, namely MVMT, is proposed for target identification. Compared with the previous related method, the proposed MVMT can not only use the learning based on sparse representation, but also introduces a variety of complementary perspective features to enhance the expression features of the targets;
(2) Each perspective of each particle is regarded as a separate task, and the potential connections between different perspectives and different particles are considered in a multi-task learning framework;
(3) To capture the outlier tasks that frequently occur in the particle sampling process, the coefficient matrix is decomposed into two cooperative parts to enhance the robustness of multi-task learning, and the posterior probability of outliers is set to zero to ensure that samples will not be taken in the resampling process.

2. Related Works

Recently, researchers have proposed various target identification methods: a typical method is finding the target position through the response map generated by the online learning correlation filter (CF) and determining the target scale using a fixed scale factor. The CF based trackers include Kernelized Correlation Filters (KCF) [20], multichannel features [21], adaptive scale estimation [22], the fusion of complementary learners [23], long short-term memory [24], support vector machines [25], sparse coding [26] etc. However, these CF-based trackers have some major shortcomings: First, the tracker relies excessively on the maximum response value when determining the target position. Second, the CF-based tracker uses a fixed scale factor to determine the size of the target, which is not suitable for the actual scale change of the moving target. Therefore, when the tracking scene is too complex, the generated model is not robust enough. Besides the optical flow method, Camshift and Meanshift are also popular target tracking methods used in recent years, showing their potential in identification in tracking applications [27,28]. However, these methods identify the target according to its single feature, and it is impossible to identify a common single feature that can be applied to different situations.

To improve the robustness of prediction, various ensemble methods have been studied. The Staple tracker [29] consists of a correlation filter and a model based on color histogram, which complement each other. In [30], three different feature trackers based on support vector machines are integrated, and the trackers are adaptively selected according to the consistency of front and rear tracks. In order to eliminate redundancy between weak models, efficient diverse ensemble co-tracking [31] trains different models by generating a set of effective manual data. Multi-cue correlation filters tracker [32] assigns the suitable weak classified based on their self-wise and pairwise relationships. Due to the limited training samples, sequent training convolutional tracking (STCT) solved the problem of overfitting existing convolutional neural networks (CNNs) in visual tracking [33]. STCT uses binary masks to force basic learners to learn different features. In the correlation filter method it is proposed to fuse response maps from different convolutional neural network levels. Therefore, integrating multiple individual CF [34] has become a common technology. Under the Siamese tracking pipeline, a twofold Siamese network trains an appearance model and a semantic model, respectively, for online combination [35]. The learning part-based model of joint tracking is also discussed in the Siamese network [36].

3. Problem Description and Modeling

The target can be represented by multiple types of visual features, including edge [37], intensity [38], color [39], and texture [40]. Using the complementary features of these multi-sources, information can significantly improve the target identification performance [39,41–43]. These different visual features are sampled from different feature spaces. Since the view data are extracted from the same object of interest, these data are interrelated, i.e., each view datum can be obtained from the sparse representation of a part of the samples under the view. In practical applications, gradient direction histogram (HOG) [44], local binary pattern (LBP) [45] and scale-invariant feature transformation (SIFT) [46,47] are highly expressive perspective features. Therefore, the robustness of the target identification problem is enhanced by combining the above multiple perspectives and training multiple samples at the same time. To express the scheme more intuitively and conveniently, the flow chart of the multi-view and multi-task learning method is presented in Figure 1.

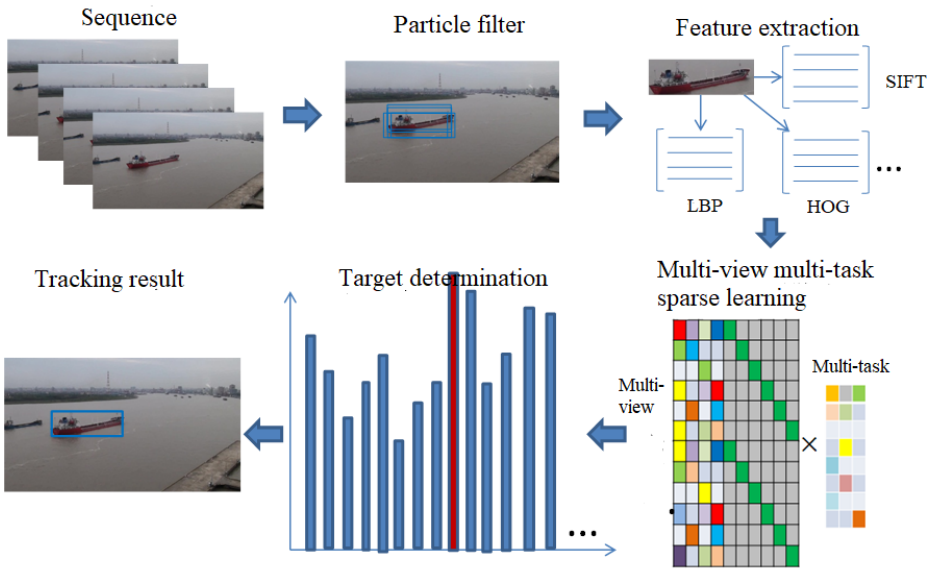


Figure 1. Target identification schematic of multi-view multi-task algorithm.

Assuming that s_t and x_t represent the state variables and observation variables at time t, respectively, the target identification problem can be described as the estimation of a posteriori state probability $p_n(s_t|s_{t-1})$ through a limited set of n particles with importance weights ω_1, which $x_{1:t} = \{x_1, \cdots, x_t\}$ are the previously observed t frames of images. Particle samples s_t^i are sampled from the importance distribution $q(s_t|s_{1:t}, x_{1:t})$, i.e., the simplification of state transition probability $p(s_t|s_{t-1})$. In addition, the importance weight of particle i is updated through the observation probability, i.e., $w_t^i = w_{t-1}^i p(x_t|s_t^i)$

The sparse representation of feature x can be reconstructed by the regularization $\ell 1$ problem to minimize the error [3]:

$$\min_w ||MW - X||_2^2 + \lambda ||W||_1 \tag{1}$$

where $M = [D, I, -I]$ is a complete dictionary composed of target template set D and positive and negative background template sets I and $-I$. Each column is a target template in D, and the candidate region pixels are modified in the column vector. Each column in the background template set is a unit vector with one non-zero element.

$W = [a^T, e^{+T}, e^{-T}]^T$ is composed of target coefficient and positive and negative background coefficients e^{+T}, e^{-T}. Each column in a^T represents a classification plane.

The probability of this observation is derived from the reconstruction error of x

$$p(x|y) = \frac{1}{Y} \exp\{-\alpha ||Da - x||^2\} \tag{2}$$

where a is obtained by solving (1), α is constant and is used to control the shape of the Gaussian kernel, and Y is the normalization parameter.

4. Multiple Feature Extraction Methods for Target Identification

In this paper, we introduce four typical feature extraction methods to make the advantages of complementary features, including the scale-invariant feature transformation (SIFT), the histogram of oriented gradient (HOG) [9], the local binary pattern (LBP) [47], and the Hu invariant matrix. Specifically, HOG is edge distributions of objects captured based on gradient features. LBP has a powerful function of representing object texture. Hu moment invariants are fast and can better describe larger objects in the image. In addition, to ensure the quality of extracted features, a simple and effective illumination normalization method is adopted before feature extraction [19]. The unit norm normalization is applied to extract feature vectors for each particle's viewing angle [46]. To clearly show the feature extraction process, we adopt four feature extraction methods to extract features from the same image, which are elaborated as follows. Since the ship has significant texture features and edge features, we choose the ship image as an example, and the original image of the ship is presented in Figure 2.



Figure 2. Original image of ship.

4.1. Scale-Invariant Feature Transformation (SIFT) Feature Extraction Method

Scale-invariant feature transformation (SIFT) is a machine learning method for local feature detection. The feature points of adjacent image frames are matched by the features obtained from the size and direction of the feature points in the image and their related descriptors. The extraction result of the SIFT feature on key point descriptors of the ship is shown in Figure 3.

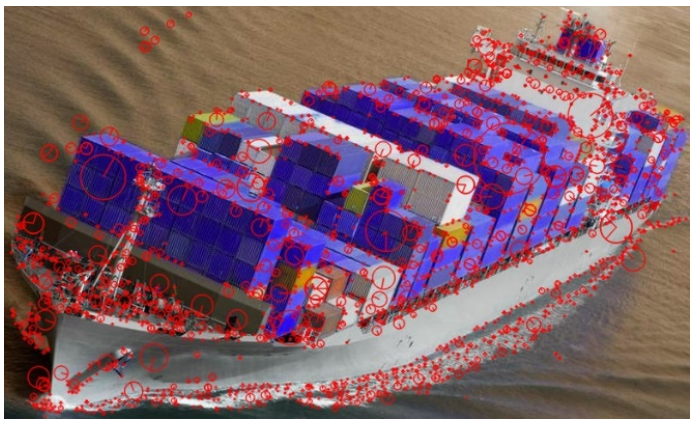


Figure 3. SIFT feature on key point descriptors of ship.

From Figure 3, it can be seen that the key point descriptors show the features of the scale size and direction of the descriptors. These features have a scale variance for changing image brightness, rotation angle, or shooting angle. These quantized features can be used to match the feature points of adjacent ship image frames, i.e., when the descriptors of the two adjacent ship frames are extracted, the key descriptors between the adjacent ship frames can be matched. When the direction and module length of the 128-dimensional vector are matched, the two frames are matched.

4.2. Gradient Direction Histogram (HOG) Feature Extraction Method

HOG is a gradient direction histogram feature that is used to detect objects in image processing in the field of computer vision. HOG is a histogram feature formed by counting the gradient direction of each pixel in the local neighborhood of the image. The implementation method of HOG is as follows:

Firstly, Gamma space and color space are normalized. In the texture feature of a ship image, the weight of the local surface exposure factor is large. Therefore, the Gamma compression algorithm can effectively reduce the local illumination variation of the image. The ship image is transformed into a gray scale, then the gamma compression formula $I(x,y) = I(x, y)^{gamma}$ is adopted, where gamma is the compression parameter, and gamma = 0.5 is desirable. The result is shown in Figure 4, which shows the normalized Gamma space and color space of the ship images, which can reduce the influence of light factors.



Figure 4. Gamma compression image of the ship.

Then, the image is divided into small unit intervals that are called cell units. The direction histogram of the gradient of each pixel is obtained from the cell unit. Finally, the collected histograms are combined to form a feature descriptor. The image of ship HOG feature extraction is shown in Figure 5.

From Figure 5, it can be seen that the information of the HOG gradient mainly exists in the edges of the image, which can describe the appearance and shape of the ship target with the edge or gradient directional density distribution. Therefore, the HOG features have good invariance to the geometric and optical changes of the image. In addition, under the conditions of strong local light normalization and rough spatial sampling and fine direction sampling, fine deformation can be allowed without affecting the detection result.

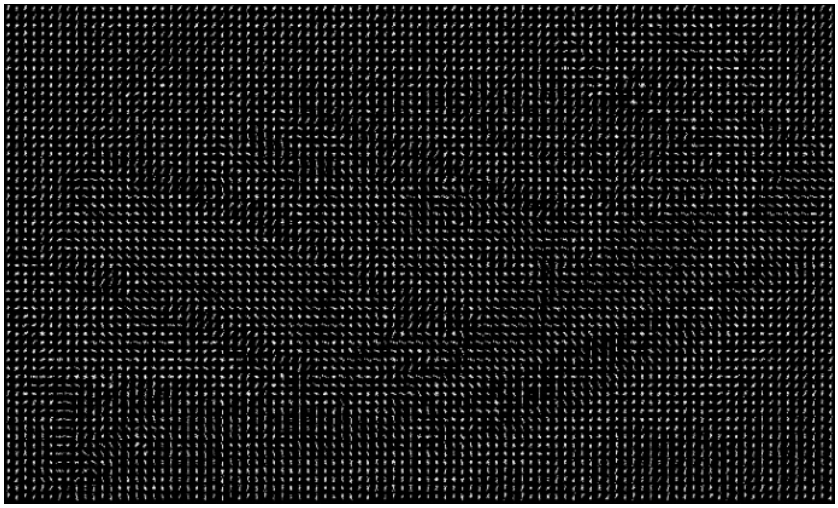


Figure 5. HOG feature extraction image of the ship.

4.3. Local Binary Pattern (LBP) Feature Extraction Method

LBP is a feature extraction method for local texture features in an image. Its essence is to describe the relationship between pixels in an image and pixels in its neighborhood. LBP has the advantages of gray constancy and rotation constancy. The algorithm for extracting LBP features from the ship database is as follows:

Step 1: Divide the manually drawn object of interest box into 16 × 16 cells;

Step 2: Compare each pixel of the ship sample with the gray value of the surrounding eight pixels, and convert the binary results into decimal numbers;

Step 3: Calculate the LBP value of each pixel in each cell, count the frequency of LBP value of each pixel, and obtain the histogram;

Step 4: Splice the statistical histograms of each cell into separate feature vectors.

The LBP's texture feature of the ship is presented in Figure 6, and the histogram of the ship is shown in Figure 7.

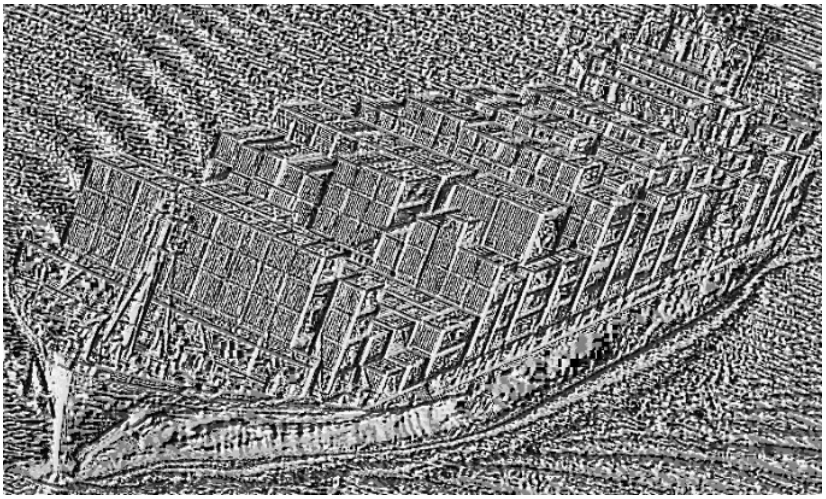


Figure 6. LBP texture image of the ship.

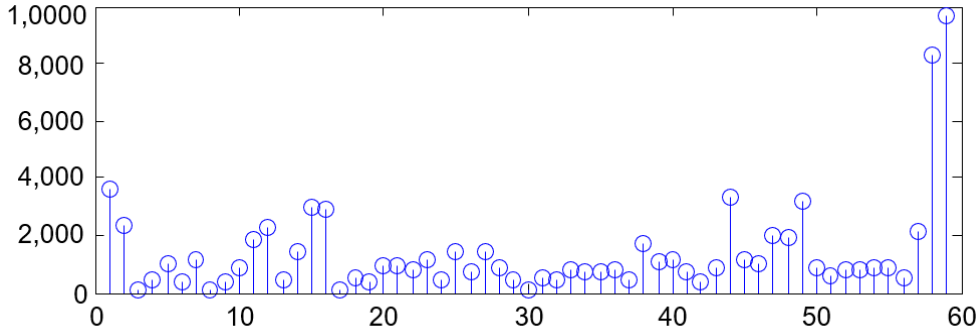


Figure 7. LBP histogram of ship.

In Figure 6, the principle of LBP is to compare the gray value of a pixel in the ship image with the gray value of the pixel around it, and then convert the comparison results into binary mode. From Figure 6 it can be seen that after LBP feature extraction the remarkable ship texture feature map is displayed, and it has the characteristics of gray scale constancy and rotation constancy. From Figure 7, most of the energy is distributed in 59 histograms with high probability. Taking into account the compression of ship eigenvectors, it is very reasonable for us to use these 59 histograms to describe the ship texture in this experiment.

4.4. Hu Invariant Matrix Feature Extraction Method

The Hu invariant matrix is a set of characteristic quantities composed of seven invariant matrices. In 1962, it was proven that these characteristic quantities are invariant to image scaling, translation and rotation [48]. Matrix invariants can well represent the structural features of images. The advantage of Hu matrix invariants for image feature extraction is that the calculation speed is fast, but the disadvantage is that the accuracy of image recognition is low, especially for images with rich texture. Especially, Hu invariant matrices are not sensitive to texture-rich images, and generally can only recognize large objects in the image. As the ship is a large object, so the texture is relatively simple. Therefore, Hu matrix invariants are used as auxiliary features to further enhance the robustness of ship identification in combination with other visual angle features. In particular, after being processed by the Hu invariant matrix feature method, it is a vector composed of seven scalars, rather than a processed ship image.

5. Target Identification Method Based on Multi-View Multi-Task Sparse Learning

Several previous research works have been studied for multi-view multi-task sparse learning. Chen et al. [49] applied the multi-view sparse learning method to the field of target identification. Mei et al. [50] proposed a generative tracker based on multi-view learning, with impressive tracking performance. The tracker proposed by Hong et al. [51] cannot be directly used to perform visual ship tracking tasks because the features to be tracked are color and intensity, which may be very similar between different ships.

To address the above issues, we propose a multi-view multi-task sparse learning for target identification. Suppose there are n candidates, each has v different perspectives (e.g., SIFT, HOG, LBP, hu feature). Defined $X^v \in R^{dv \times n}$ as the feature matrix, it is a normalized n-column d-dimensional particle image feature vector. The dv is the v-dimensional viewing angle and the n is the number of candidates. Especially, defined $D^v \in R^{dv \times n}$ as a target dictionary, where each column is a target of the v-th perspective, and n is the number of target templates. In this paper, it means four high-dimensional vectors of the SIFT, HOG, LBP, hu feature extraction methods. In this manner, it can effectively integrate multiple feature extraction methods. The target dictionary constructs a complete dictionary $A^v = [D^v, I_{dv}]$ by combining the background template I_{dv}. Each representation matrix C^v is composed of two cooperative parts P^v and Q^v. Component P captures shared features

among all tasks, and component Q captures outliers. The multi-view sparse representation can be obtained by the following cost function.

$$\min_{W,P,Q} \frac{1}{2}||A^v C^v - X^v||_F^2 + \omega_1||P||_{1,2} + \omega_2||Q^T||_{1,2} \tag{3}$$

$$||P||_{1,2} = \sum_i (\sum_j P_{i,j}^2)^{\frac{1}{2}} \tag{4}$$

where P_{ij} represents the elements in row i and column j of matrix P, $C^v = P^v + Q^v$, $P = [P^1, P^2, \ldots, P^v]$, $Q = [Q^1, Q^2, \ldots, Q^v]$, ω_1 and ω_2 are the parameters controlling the sparsity of P and Q, respectively. Figure 8 illustrates the structure of the learned matrices P and Q.

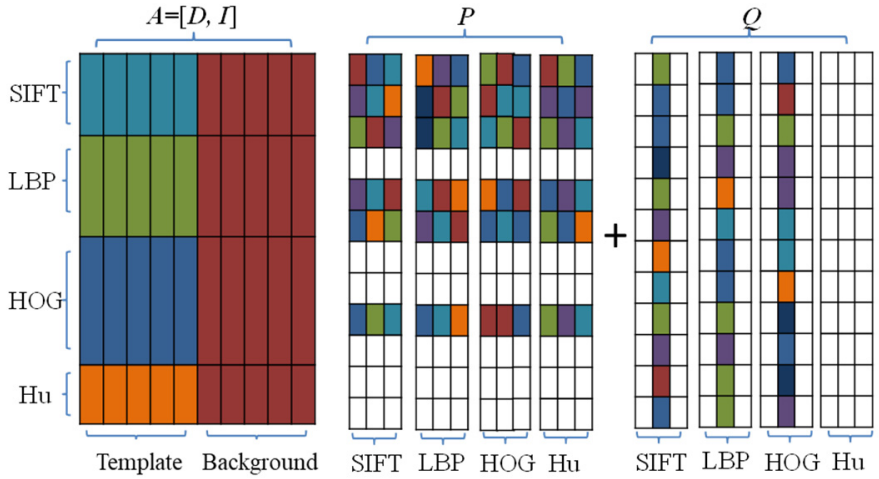


Figure 8. The structure of the matrices P, Q, and A.

It should be noted that the superposition of P^v and Q^v requires the same number of columns as A^v, so the zero matrix is used to fill the A^v matrix. Definition $\hat{A}^v = [A^v, 0^v]$, where $0^v \in \Re^{dv \times (n_1 - n_v)}$, and each element of 0^v is zero, $n_1, \ldots, n_v$ is the dimension of each perspective. According to the target identification results, the observation probability of potential sample i is defined as:

$$p_i = \frac{1}{\Gamma} \exp\{-\alpha \sum_{v=1}^{V} ||A^v C_i^v - X_i^v||^2\} \tag{5}$$

where Γ is the normalization parameter, C_i^v is the coefficient corresponding to the v-th angle of view of the Q^i potential sample. The target identification result is the particle with the maximum observation probability. Update the target dictionary D step by step and weight the template.

In the process of target identification, some outlier tasks often exist. These sample particles are far from the target and have little overlap with other particles. The outlier task is captured by introducing the coefficient matrix. In particular, if the sum of coefficients corresponding to the particle of $\ell 1$ standard is greater than the adaptive threshold σ:

$$\sum_{v=1}^{V} |Q_i^v| > \sigma, \tag{6}$$

where Q_i^v is the i-th row of Q^v, it is defined as outliers, and the observation probability is set to zero, and the resampling process of outliers is ignored. Define the number of detected outlier tasks as N_0, and update the threshold σ as follows:

$$\begin{cases} \sigma new = \sigma old\kappa, n0 > N0 \\ \sigma new = \sigma old/\kappa, n0 = 0 \\ \sigma new = \sigma old, 0 < n0 \leq N0, \end{cases} \tag{7}$$

where κ is a scale factor, $N0$ is a threshold for the number of predefined outliers.

5.1. Multi View Multi-Task Sparse Learning Target Identification Method

In this paper, a new target identification method called multi-view multi-task learning method (MVMT), is proposed. The method mainly includes five steps: first, read the image sequence in the captured video and sample the particles according to the particle filter sampling and resampling methods; second, select candidate templates; third, extract multiple features of particles to increase the diversity of target representation from different feature spaces; fourth, use the multi-view sparse representation method for multi-task learning to train the optimal candidate template set; fifth, for each candidate template, transform the reconstruction error into a probability problem to determine the target; finally, calibrate the target and track it in a rectangular frame. To represent the process of a multi-view multi-task learning method more intuitively, the flow chart of a multi-view multi-task sparse learning target identification method is shown in Figure 9.

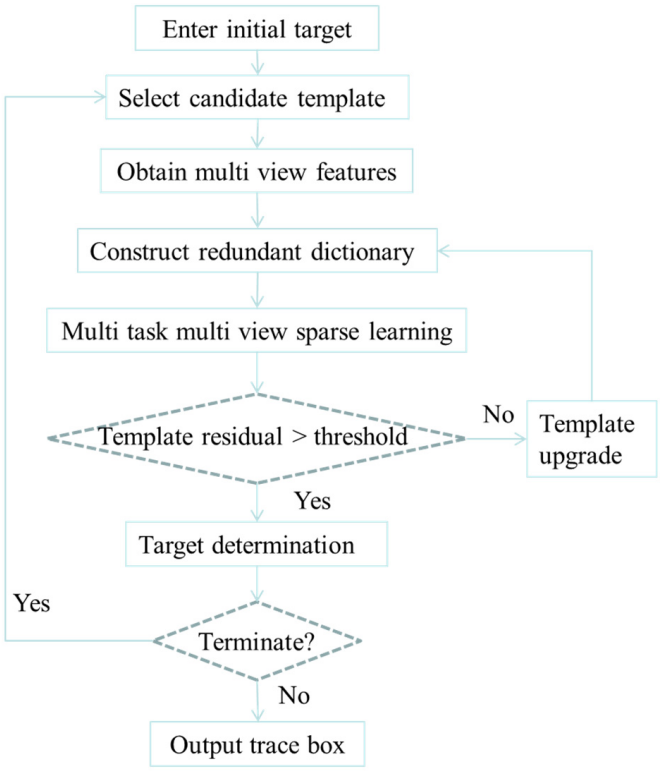


Figure 9. Flow chart of multi-view multi-task target identification method.

According to the flow chart in Figure 9, the target identification algorithm based on multi-task and multi-view can be divided into the following main steps:

(1) Initialize the target: Select the rectangular box of the target of interest on the first frame image;
(2) Select the candidate template: Select the candidate template according to the Gaussian distribution model and process its gray image level.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{8}$$

(3) Extract multi-view features: Extract the features of LBP, HOG, and SIFT view from the obtained candidate templates, and reduce the dimension of the extracted feature matrix;

(4) Construct the redundant dictionary: Assemble the obtained single column vectors of each view of a given template into a new column vector, i.e., template vector, and cycle through other candidate templates. All the template vectors obtained are arranged in columns, and the identity matrix is added to form a redundant dictionary;

(5) Multi-view multi-task sparse learning: Multi-view multi-task learning is performed on the templates in the redundant dictionary to solve the classification coefficient matrix C. For several particles selected by sampling around the target, select the candidate template with a small residual to replace the bad performance in the dictionary, and finally make the candidate template in the dictionary the optimal template set. For Equation (3), a gradient descent algorithm is proposed, so the target residual can be calculated as:

$$\cos t(W) = \frac{1}{2m}[\sum_{v=1}^{V}((A^v C^v - X^v)^2 + \omega_1 ||P||_{1,2} + \omega_2 ||Q^T||_{1,2} \tag{9}$$

For solution C, the following method is used for convergence to obtain the optimal solution:

$$c_j := c_j - \alpha \frac{1}{V} \sum_{v=1}^{V} ((A^v C^v - X^v) \cdot A_j^v) \tag{10}$$

(6) Update template: if the residual error between the training sample and the label is less than the given threshold, replace the corresponding template in the dictionary, as shown in Figure 10.

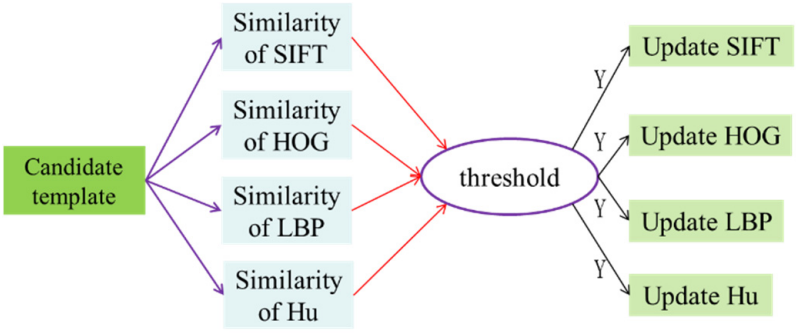


Figure 10. Template update schematic.

(7) Determine target: Carry out probability conversion for each classification plane (each column) in matrix C and take the classification plane with the largest probability as the final target to be tracked. For the t-th frame image, the determination criteria are as follows:

$$p(y_i|x_t) = \frac{1}{\Gamma} \exp\{-\alpha \sum_{v=1}^{V} ||Ta(1:n,i) - Y||\} \tag{11}$$

where Γ is the normalized factor, N represents the number of candidate templates in the dictionary, and Y represents the label. The flow chart of target judgment is shown in Figure 11.

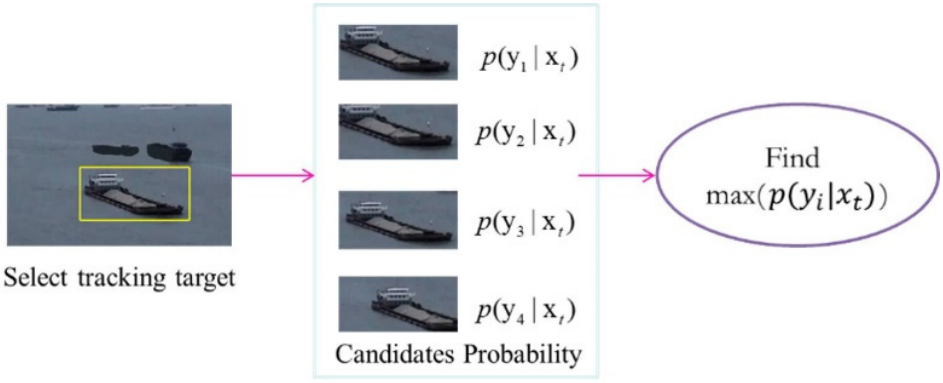


Figure 11. Flow chart target selection.

(8) Output box to calculate the target identification error.

5.2. Simplified MVMT Methods

To further verify the superiority of multi-view sparse learning performance compared with single-view single-feature tracker, the single-task single-view target identification method (STSV), single-task multi-view target identification method (STMV), and single-view multi-task target identification method (MTSV) are designed, which are three special cases of multi-view multi-task (MVMT) method.

For single-view multi-task learning algorithm (MTSV), it can be obtained from the following formula:

$$\min_{W,P,Q} ||AC - X||_2^2 + \omega_1 ||P||_{1,2} + \omega_2 ||Q^T||_{1,2} \tag{12}$$

The difference between single-view multi-task learning algorithm and multi-view multi-task learning algorithm is that when constructing a dictionary, only a single-view feature is used to describe a template and train multiple particles at the same time.

$$D = [T^1, T^2, \cdots, T^n] \tag{13}$$

where T_k (k = 1, 2, ..., n) is the extracted single-view template. By comparing Equations (3) and (12), it can be obtained that SVMT is a special case of MVMT.

6. Experiment and Result Analysis

To verify the effectiveness of the proposed method (MVMT), several classical trackers with excellent performance are selected as comparison methods. The experimental performance of these algorithms is compared on the image sequence frames dataset.

To make full use of the advantages of complementary features, three popular features are used in the experiment: gradient direction histogram (HOG), local binary pattern (LBP), and scale-invariant feature transformation (SIFT). HOG is edge distributions of objects captured based on gradient features. LBP has a powerful function of representing object texture. SIFT has scale invariance for changing image brightness, rotation angle or shooting angle. In addition, to ensure the quality of the extracted features, the illumination normalization method is used to extract the feature vectors from multiple perspectives of each particle before feature extraction.

MVMT (multi-view multi-task) is compared with the other three special cases: SVST, MVST, and SVMT algorithms. The source code of all algorithms runs in MTLAB. The experimental hardware environment is as follows: the CPU is i5-3612qe; the memory is 8GB; the video memory is 2GB; the operating system is windows 10; and the software development environment is matlab2018.

The number of initial candidate templates for MVMT algorithm, L1T, and MTT proposed in the algorithm parameter setting is n = 10, and the number of training templates is

100. Gaussian distribution is used to sample the particles. After selecting the candidate template, the sparse learning method is used to construct the redundant dictionary for later sample training and updating. Scale-invariant feature transformation (SIFT), gradient direction histogram (HOG), local binary pattern (LBP), and Hu invariant matrix are used. The gradient descent iterative optimal algorithm is adopted to continuously update the candidate templates in the dictionary. Finally, the template with the smallest residual error is selected as the method of the next frame of image.

The experimental data are divided into two groups of data for experimental verification. One group is a public dataset, including video picture frame sequence. The other group is custom dataset, including the video sequence of ships in key waters and port.

6.1. Experimental Results and Analysis of Public Datasets

To evaluate the effectiveness of the proposed MVMT method, three complementary features are adopted. Four widely used sequence diagrams (car4, faceocc2, david_interior, sylv) are selected. The experimental results of the common dataset are shown in Figure 12.

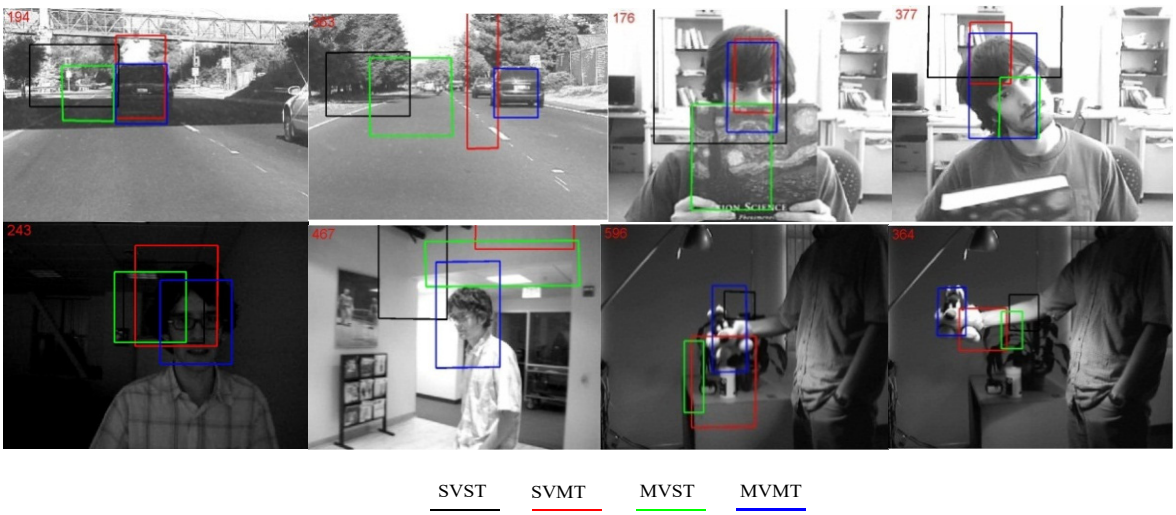


Figure 12. The identification results of the proposed MVMT and SVST, MVST, SVMT algorithms.

In the car4 sequence, the main influencing factors are illumination change and size change. The MVMT and the fast-moving vehicle are tracked from beginning to end, while the SVST and MVST are affected by illumination and size change of the vehicle. They contain more and more background areas, and the target is gradually lost. The target has been completely lost in 194 frames. The SVMT has good performance on the influence of illumination, but at about the 310th frame the vehicle passes the bridge and suddenly loses the target. According to the experimental results, MVMT ratio can better deal with the factors of light change and target size change.

In the faceocc2 sequence, the task is to identify the swing of David's head in the room. The main factors of this sequence are swing and occlusion. MVMT and MVST can identify the target successfully in the whole sequence, but MVMT can only identify a small part of the target after the target is occluded many times. In addition, after the head target swings, the appearance and angle change. SVST and SVMT algorithms lose some information of the target. Then after being occluded many times, SVST completely loses the target. The experimental results show that MVMT can better deal with the occlusion, appearance change and angle change in face tasks.

In the David_indoor sequence, the influence of different illumination is mainly studied. The experimental results show that in the case of dark light, that is, in the early stage of target identification, SVST has been separated from the target because the texture features or edge features are not obvious. With the significant enhancement of illumination, the

SVMT and MVST algorithms also slowly lose their target. In the whole target identification process, only MVMT algorithm has captured the target. The experimental results show that MVMT can better deal with the problems caused by light changes.

In the owl sequence, the task is to identify the owl doll. The owl sequence is relatively more fixed, and fuzzy scenes and fast motion are considered. At the early stage of the sequence frame the shape of the target changes and receives different illumination. The SVST and SVMT algorithms lose the target. With the rapid movement, the target becomes fuzzy. The MVST algorithm gradually loses the target information and cannot capture the target. In the whole owl sequence, only MVMT algorithm has captured the target. The experimental results show that MVMT can better deal with the problem of fuzzy targets.

Figure 13 shows the experimental error comparison curve. It can be seen that the identification methods of SVST and MVST are very poor, and the SVST method is the worst. The reason is that a single task lacks associated information with other tasks and is artificially divided. At the same time, the identification performance of both SVMT method and MVMT method is generally better, because the multi-task learning method cannot only capture the common features among multiple tasks but also eliminate outliers, which makes the performance of identification better. Among them, the MVMT method has the best identification performance, because the multi-view and multi-task simultaneously contain a variety of feature vectors in different state spaces, provide multiple views for the same target object, and learn together, which has very good stability and strong robustness.

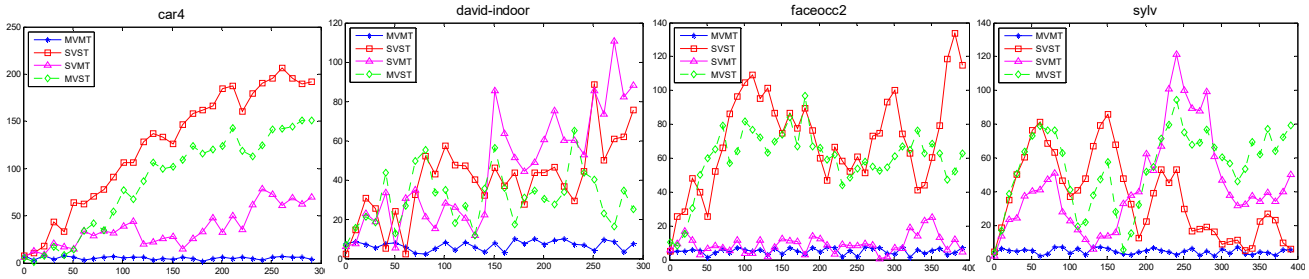


Figure 13. The errors of the proposed MVMT algorithm and SVST, MVST, SVMT algorithms (pixel).

6.2. Experiment and Analysis of Custom Datasets

To further verify the application performance of the proposed algorithm in the actual scene, we adopt the targets from the user-defined sequence frames and compare the performance of different algorithms. The custom dataset comes from the field shooting in a port. Six sequence frames are selected from the ship video Ship-1, ship video Ship-2, ship video Ship-3, and ship video Ship-4, ship video Ship-5, and ship video Ship-6. Among them, Ship-1, Ship-2, Ship-3, and Ship-4 are dangerous watercourse scenes taken in different places during the day, Ship-5 is the scene taken at night, and Ship-6 is the video sequence obtained through infrared rays. As with the public dataset, all pictures are resized to 320 $\times$ 240 (pixels). In this experiment, the source code provided by L1T and MTT and the proposed MTMV source code are tested in MTLAB. Meanwhile, the above STSV (single-view single-task), STMV (multi-view single-task), and MTSV (single-view multi-task) algorithms are compared in the experiment. The identification results of these seven algorithms are shown in Figure 14.

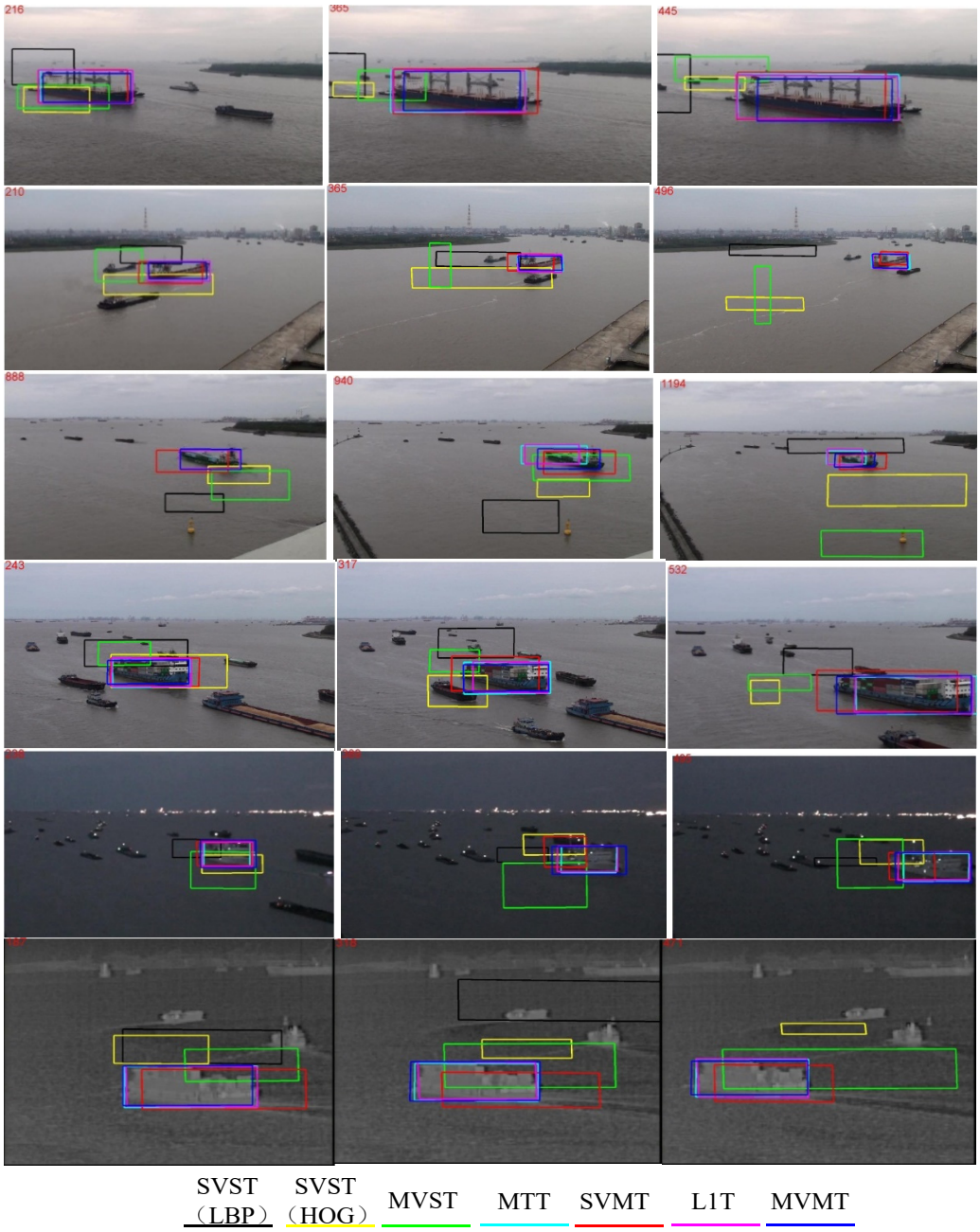


Figure 14. Identification results of 7 algorithms in 6 groups of ship sequence pictures.

6.3. Qualitative Comparison and Analysis

In the Ship-1 sequence, the influence of the size change of the target on the identification process is mainly considered. At the 216th frame, the SVST and MVST gradually separated from the target when the target size just began to increase. At the 365th frame, the target size has doubled from the initial target size due to the continuous increase invariati the ship size, and the SVST and MVST have lost the target. In the whole process, although L1T, MTT and SVMT kept tracking the target, the effect was not as good as MVMT. Experiments show that Mvmt has strong robustness to target size changes.

In the Ship-2 sequence, the main interference factors are the appearance change and angle change of the target. At the 241st frame, MVST soon lost its target. At the 389th and 436th frames, MTT, SVMT, L1T, and MVMT tracked the target all the way, but SVMT only captured part of the target, which was not as effective as MVMT. The experimental results show that MVMT can deal with the problems of size change, appearance change

and angle change. L1T and MVST fail to track the whole sequence, which shows that the mechanism of multi-task joint learning makes MTT, SVMT, and MVMT more robust. However, MTT and SVMT are less robust than MVMT because MVMT uses the advantages of complementary features and can detect outliers.

In the Ship-3 sequence, the main factors are fast motion and blur. At the 821st frame, the MVST quickly loses the target. At the 931st frame, after the camera lens moves rapidly, the image becomes blurred. L1T and SVMT only capture some targets, while MVMT keeps locking the targets. Compared with L1T and MVST, MVMT can fully track the target under different changes because of the robustness of additional features. In the whole database-3 sequence, only MVMT can successfully track the target in the whole sequence, while other trackers either completely lose the target or contain most of the background. It is concluded from the experiment that the single-task tracker is easily affected by the appearance change.

The Ship-4 sequence is mainly tested for occlusion factors. At the 243rd frame, the target is unobstructed, and all methods lock the target well. However, At the 317th frame, SVST and MVST have been separated from the target due to partial occlusion of the target. SVMT has included part of the background, and L1T has only captured part of the target. With the occlusion leaving, SVMT, L1T and MTT can track the target at the 532nd frame, but their performance is not as good as MVMT. This shows that MVMT can better deal with the influence of occlusion factors than other methods.

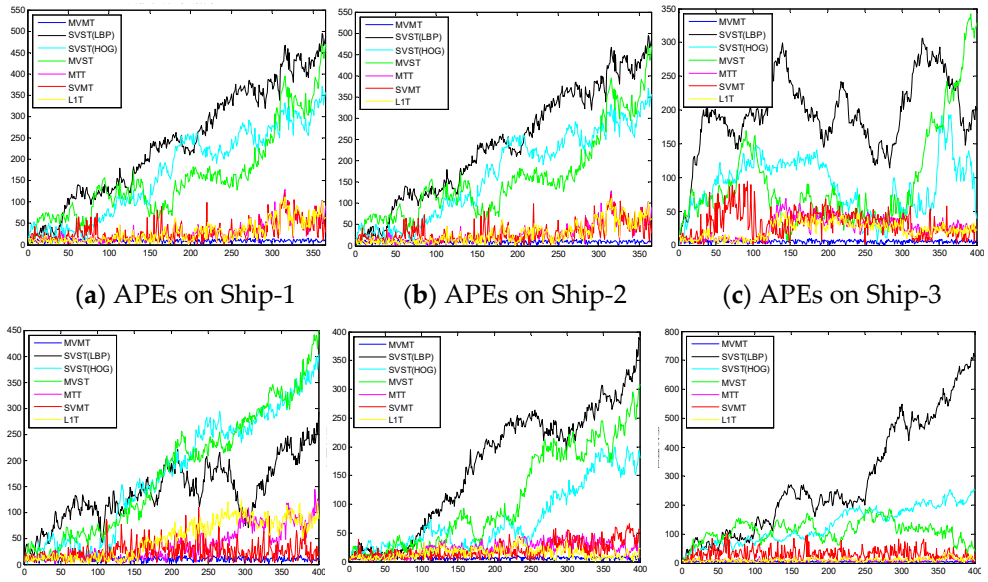
In the Ship-5 sequence, the influence of illumination change factor is mainly considered. At the 233rd frame, MVMT, MTT, and L1T can fully track the sailing ship and only MVST is separated from the target. With the gradual dimming of the light, MVST contains more and more background areas at the 346th frame. According to the experimental results, MVMT has better robustness than other methods.

In the Ship-6 sequence, it mainly detects the stability and robustness of target identification on infrared video. At the 222nd frame, MVST has lost the target. MTT, SVMT, L1T, and MVMT can keep locking the target, but SVMT is easy to contain more background. MVMT can faithfully track the sailing ship, which shows that MVMT is better than other methods and can track infrared video stably.

6.4. Quantitative Comparison and Analysis

To quantitatively evaluate the performance of these comparison methods, this experiment calculates the distance between the center of the result and the real position of each frame; it then draws the variation diagram of the center position error and the number of frames. Due to space limitations, only the error comparison of the six sequence diagrams provided here by the seven methods is listed. For a perfect result, the positioning error should be zero. For a more intuitive comparison, the average position errors (APEs) of the six sequences are shown in Figure 15.

To further illustrate the comparison between different video sequence diagrams, Table 1 shows the average error between the results and the real position on six sequence diagrams for the seven methods. Table 2 shows the weighted standard deviation between the results and the real position for the seven methods, and the specific weights of each particle are equal. The following two statistical tables show that the average error and standard deviation of the MVMT method are the smallest in both visible and infrared video sequences. Therefore, in general, MVMT has the best performance.



(**a**) APEs on Ship-1 (**b**) APEs on Ship-2 (**c**) APEs on Ship-3

(**d**) APEs on Ship-4 (**e**) APEs on Ship-5 (**f**) APEs on Ship-6

Figure 15. 7 Average position error (pixel) of 5 comparison methods.

Table 1. Average error between results of five comparison methods and real position (pixels).

Sequence	Ship-1	Ship-2	Ship-3	Ship-4	Ship-5	Ship-6
STMV	160.4	280.4	96.1	179.2	109.1	110.8
MTT	28.57	7.12	25.52	38.06	20.52	10.42
MTSV	30.9	20.33	32.82	24.9	21.46	32.57
L1T	28.67	6.69	21.99	50.76	12.76	11.71
MTMV	**9.21**	**4.56**	**7.00**	**11.04**	**3.01**	**5.99**

The bold indicates that the experimental value of this algorithm is superior to other algorithms.

Table 2. Weighted standard deviation between the results of the five comparison methods and the real position (pixels).

Sequence	Ship-1	Ship-2	Ship-3	Ship-4	Ship-5	Ship-6
STMV	9.95	15.64	6.11	10.7	6.98	5.82
MTT	1.98	0.41	1.46	2.46	1.18	0.60
MTSV	2.01	1.22	1.89	1.52	1.31	1.88
L1T	1.99	0.39	1.23	3.08	0.75	0.66
MTMV	**0.51**	**0.24**	**0.37**	**0.58**	**0.16**	**0.32**

The bold indicates that the experimental value of this algorithm is superior to other algorithms.

7. Conclusions

In this paper, a multi-view multi-task target identification method based on sparse learning is proposed. Extensive experiments were conducted on public datasets and custom video datasets on ship sequence frames. Under different challenging interference factors such as size, illumination, blur, deformation, and occlusion, the experimental results show that the proposed MVMT can better identify the target in both visible and infrared video. Compared with the comparisons method, the performance of the proposed MVMT is the best. In summary, the conclusions of this paper are presented as follows:

(1) A target identification method based on machine learning is proposed. The traditional target identification method makes use of less information. In the case of many targets and complex backgrounds, tracking loss and error are common. To break through the limitations of traditional methods, we introduce a machine learning approach and provide a new idea for target identification by using sparse representation, dictionary learning method and particle filter framework;
(2) By analyzing the features of targets in the surveillance video, the texture, edge, and geometry invariant features of targets are automatically extracted, and the multi-view feature learning method is introduced to realize the information fusion of multiple features. To obtain better results, a multi-task learning method is proposed based on the multi-view feature for joint learning, which makes full use of the information transfer function between different tasks, makes up for the deficiency of the traditional single learning task, improves the classification and recognition accuracy, and provides a theoretical basis for the accurate and stable identification.

Author Contributions: J.C.: writing—original draft preparation, methodology, and conceptualization, and X.W.: writing—original draft preparation, visualization, validation, and software, and Z.Z.: funding acquisition, supervision and project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This work has no funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Tang, J.; Liu, F.; Zou, Y.; Zhang, W.; Wang, Y. An Improved Fuzzy Neural Network for Traffic Speed Prediction Considering Periodic Characteristic. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 2340–2350. [CrossRef]
2. Yan, Y.; Zhang, S.; Tang, J.; Wang, X. Understanding characteristics in multivariate traffic flow time series from complex network structure. *Phys. A Stat. Mech. Its Appl.* **2017**, *477*, 149–160. [CrossRef]
3. Chen, W.; Zhou, G.; Liu, Z.; Li, X.; Zheng, X.; Wang, L. NIGAN: A Framework for Mountain Road Extraction Integrating Remote Sensing Road-Scene Neighborhood Probability Enhancements and Improved Conditional Generative Adversarial Network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
4. Chen, W.; Li, X.; Wang, L. Target Detection for Mine Remote Sensing Using Deep Learning. In *Remote Sensing Intelligent Interpretation for Mine Geological Environment*; Springer: Singapore, 2022; pp. 127–164.
5. Tong, W.; Chen, W.; Han, W.; Li, X.; Wang, L. Channel-attention-based DenseNet network for remote sensing image scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4121–4132. [CrossRef]
6. Liu, M.; Hu, Q.; Wang, C.; Tian, T.; Chen, W. Daff-Net: Dual Attention Feature Fusion Network for Aircraft Detection in Remote Sensing Images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 4196–4199.
7. Ouyang, S.; Xu, J.; Chen, W.; Dong, Y.; Li, X.; Li, J. A Fine-Grained Genetic Landform Classification Network Based on Multimodal Feature-Extraction and Regional Geological Context. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
8. Craswell, N.; Mitra, B.; Yilmaz, E.; Campos, D.; Voorhees, E.M. Overview of the TREC 2019 deep learning track. *arXiv* **2020**, arXiv:2003.07820.
9. Zhang, S.; Tang, J.; Wang, H.; Wang, Y.; An, S. Revealing intra-urban travel patterns and service ranges from taxi trajectories. *J. Transp. Geogr.* **2017**, *61*, 72–86. [CrossRef]
10. Kwon, J.; Lee, K.M. Visual tracking decomposition. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 1269–1276.
11. Wang, Y.; Luo, X.; Ding, L.; Hu, S. Visual tracking via robust multi-task multi-feature joint sparse representation. *Multimed. Tools Appl.* **2018**, *77*, 31447–31467. [CrossRef]
12. Lan, X.; Ye, M.; Zhang, S.; Zhou, H.; Yuen, P.C. Modality-correlation-aware sparse representation for RGB-infrared object tracking. *Pattern Recognit. Lett.* **2020**, *130*, 12–20. [CrossRef]
13. Wan, M.; Gu, G.; Qian, W.; Ren, K.; Maldague, X.; Chen, Q. School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. Unmanned aerial vehicle video-based target tracking algorithm using sparse representation. *IEEE Internet Things J.* **2019**, *6*, 9689–9706. [CrossRef]

14. Elfring, J.; Torta, E.; van de Molengraft, R. Particle filters: A hands-on tutorial. *Sensors* **2021**, *21*, 438. [CrossRef]
15. Zhang, T.; Xu, C.; Yang, M.H. Learning multi-task correlation particle filters for visual tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 365–378. [CrossRef] [PubMed]
16. Nai, K.; Li, Z.; Gan, Y.; Wang, Q. Robust Visual Tracking via Multitask Sparse Correlation Filters Learning. *IEEE Trans. Neural Netw. Learn. Syst.* 2021, in press. [CrossRef] [PubMed]
17. Gurkan, F.; Gunsel, B. Integration of regularized l1 tracking and instance segmentation for video object tracking. *Neurocomputing* **2021**, *423*, 284–300. [CrossRef]
18. Javanmardi, M.; Qi, X. Robust structured multi-task multi-view sparse tracking. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
19. Lu, R.; Liu, J.; Lian, S.; Zuo, X. Multi-view representation learning in multi-task scene. *Neural Comput. Appl.* **2020**, *32*, 10403–10422. [CrossRef]
20. Yan, P.; Yao, S.; Zhu, Q.; Zhang, T.; Cui, W. Real-time detection and tracking of infrared small targets based on grid fast density peaks searching and improved KCF. *Infrared Phys. Technol.* **2022**, *123*, 104181. [CrossRef]
21. Lukezic, A.; Vojir, T.; Zajc, L.C.; Matas, J.; Kristan, M. Discriminative correlation filter with channel and spatial reliability. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4847–4856.
22. Yuan, D.; Kang, W.; He, Z. Robust visual tracking with correlation filters and metric learning. *Knowl. Based Syst.* **2020**, *195*, 105697. [CrossRef]
23. Cheng, G.; Li, R.; Lang, C.; Han, J. Task-wise attention guided part complementary learning for few-shot image classification. *Sci. China Inf. Sci.* **2021**, *64*, 120104. [CrossRef]
24. Yang, Y.; Xing, W.; Zhang, S.; Gao, L.; Yu, Q.; Che, X.; Lu, W. Visual tracking with long-short term based correlation filter. *IEEE Access* **2020**, *8*, 20257–20269. [CrossRef]
25. Kurani, A.; Doshi, P.; Vakharia, A.; Shah, M. A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. *Ann. Data Sci.* **2021**, 1–26. [CrossRef]
26. Abbass, M.Y.; Kwon, K.C.; Kim, N.; Abdelwahab, S.A.; El-Samie, F.E.A.; Khalaf, A.A.M. Visual tracking using convolutional features with sparse coding. *Artif. Intell. Rev.* **2021**, *54*, 3349–3360. [CrossRef]
27. Bardow, P.; Davison, A.J.; Leutenegger, S. Simultaneous optical flow and intensity estimation from an event camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 884–892.
28. Tripathi, R.P.; Ghosh, S.; Chandle, J.O. Tracking of object using optimal adaptive Kalman filter. In Proceedings of the IEEE International Conference on Engineering and Technology, Coimbatore, India, 17–18 March 2016; pp. 1128–1131.
29. Zhang, S.; Yang, Y.; Zhang, M.; Mi, P. An Efficient Tracker via Multi-feature Adaptive Correlation Filter. *J. Syst. Simul.* **2022**, *34*, 1864.
30. Weng, X.; Ivanovic, B.; Pavone, M. Mtp: Multi-hypothesis tracking and prediction for reduced error propagation. In Proceedings of the 33rd IEEE Intelligent Vehicles Symposium (IV), Aachen, Germany, 5–9 June 2022; pp. 1218–1225.
31. Meshgi, K.; Oba, S.; Ishii, S. Efficient diverse ensemble for discriminative co-tracking. In Proceedings of the Conference Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4814–4823.
32. Wang, N.; Zhou, W.; Tian, Q.; Hong, R.; Wang, M.; Li, H. Multi-cue correlation filters for robust visual tracking. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4844–4853.
33. Jiao, L.; Wang, D.; Bai, Y.; Chen, P.; Liu, F. Deep learning in visual tracking: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**. [CrossRef] [PubMed]
34. Dai, K.; Wang, D.; Lu, H.; Sun, C.; Li, J. Visual tracking via adaptive spatially-regularized correlation filters. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4670–4679.
35. He, A.; Luo, C.; Tian, X.; Zeng, W. A twofold siamese network for real-time object tracking. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4834–4843.
36. Gao, J.; Zhang, T.; Xu, C. Graph convolutional tracking. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4649–4659.
37. Dong, B.; Zhou, Y.; Hu, C.; Fu, K.; Chen, G. BCNet: Bidirectional collaboration network for edge-guided salient object detection. *Neurocomputing* **2021**, *437*, 58–71. [CrossRef]
38. Wu, L.; Fang, S.; Ma, Y.; Fan, F.; Huang, J. Infrared small target detection based on gray intensity descent and local gradient watershed. *Infrared Phys. Technol.* **2022**, *123*, 104171. [CrossRef]
39. Liu, J.; Zhong, X. An object tracking method based on Mean Shift algorithm with HSV color space and texture features. *Clust. Comput.* **2019**, *22*, 6079–6090. [CrossRef]
40. Humeau-Heurtier, A. Texture feature extraction methods: A. survey. *IEEE Access* **2019**, *7*, 8975–9000. [CrossRef]
41. Chen, X.; Chen, H.; Wu, H.; Huang, Y.; Yang, Y.; Zhang, W.; Xiong, P. Robust visual ship tracking with an ensemble framework via multi-view learning and wavelet filter. *Sensors* **2020**, *20*, 932. [CrossRef] [PubMed]
42. Mao, J.L. Adaptive multi-view learning and its application in image classification. *J. Comput. Appl.* **2013**, *33*, 1955–1959. [CrossRef]
43. Deng, J.; Czarnecki, K. MLOD: A multi-view 3D object detection based on robust feature fusion method. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; pp. 279–284.
44. Zhou, W.; Gao, S.; Zhang, L.; Lou, X. Histogram of oriented gradients feature extraction from raw bayer pattern images. *IEEE Trans. Circ. Syst. II Express Briefs* **2020**, *67*, 946–950. [CrossRef]

45. Hassaballah, M.; Alshazly, H.A.; Ali, A.A. Ear recognition using local binary patterns: A comparative experimental study. *Expert Syst. Appl.* **2019**, *118*, 182–200. [CrossRef]
46. Tareen, S.A.K.; Saleem, Z. A comparative analysis of sift, surf, kaze, akaze, orb, and brisk. In Proceedings of the IEEE International Conference on Computing, Mathematics and Engineering Technologies, Sukkur, Pakistan, 3–4 March 2018; pp. 1–10.
47. Li, G.; Feng, Y. Moving object detection based on SIFT feature matching and K-means clustering. *J. Comput. Appl.* **2012**, *32*, 2824–2826.
48. Hu, M.K. Visual pattern recognition by moment invariants. *IRE Trans. Inf. Theory* **1962**, *8*, 179–187.
49. Chen, X.; Wang, S.; Shi, C.; Wu, H.; Zhao, J.; Fu, J. Robust ship tracking via multi-view learning and sparse representation. *J. Navig.* **2019**, *72*, 176–192. [CrossRef]
50. Mei, X.; Hong, Z.; Prokhorov, D. Robust Multitask Multiview Tracking in Videos. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *26*, 2874–2890. [CrossRef] [PubMed]
51. Hong, Z.; Mei, X.; Prokhorov, D.; Tao, D. Tracking via robust multi-task multi-view joint sparse representation. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 649–656.