applied sciences

MDPI

Article

An Efficient Person Search Method Using Spatio-Temporal Features for Surveillance Videos †

Deying Feng [1,2,*], Jie Yang [2,3], Yanxia Wei [1], Hairong Xiao [1] and Laigang Zhang [1]

[1] School of Mechanical and Automotive Engineering, Liaocheng University, Liaocheng 252000, China; sditwyx_1999@163.com (Y.W.); 13791075889@163.com (H.X.); zhanglaigang@lcu.edu.cn (L.Z.)
[2] Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai 200240, China; jieyang@sjtu.edu.cn
[3] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China
* Correspondence: fdy629@163.com

† This paper is an extended version of the paper "An Unsupervised Person Search Method for Video Surveillance" published in 2022 8th International Conference on Computing and Artificial Intelligence (ICCAI 2022).

Abstract: Existing person search methods mainly focus on searching for the target person using database images. However, this is different from real-world surveillance videos which involve a temporal relationship between video frames. To solve this problem, we propose an efficient person search method that employs spatio-temporal features in surveillance videos. This method not only considers the spatial features of persons in each frame, but also utilizes the temporal relationship of the same person between adjacent frames. For this purpose, the spatial features are extracted by combining Yolo network with Resnet-50 model, and the temporal relationship is processed by gated recurrent unit. The spatio-temporal features are generated by the following average pooling layer and used to represent persons in the videos. To ensure search efficiency, locality sensitive hashing is used to organize massive spatio-temporal features and calculate the similarity. A surveillance video database is also constructed to evaluate the proposed method, and the experimental results demonstrate that our method improves search accuracy while ensuring search efficiency.

Keywords: person search; spatio-temporal features; Yolo network; gated recurrent unit; locality sensitive hashing

check for updates

Citation: Feng, D.; Yang, J.; Wei, Y.; Xiao, H.; Zhang, L. An Efficient Person Search Method Using Spatio-Temporal Features for Surveillance Videos. *Appl. Sci.* **2022**, *12*, 7670. https://doi.org/10.3390/app12157670

Academic Editor: Antonio Fernàndez-Caballero

Received: 29 June 2022
Accepted: 27 July 2022
Published: 29 July 2022

1. Introduction

With the rapid development of smart city construction, more and more surveillance cameras are being widely distributed throughout cities. As a carrier for obtaining and storing information, surveillance video plays an important role in the disappearance of the elderly and children and the pursuit of criminal suspects. However, storing surveillance videos is not the same as finding the relevant information. Due to the large number of surveillance cameras and the long recording time, the number of videos is increasing exponentially. It often takes a lot of time and resources to find a specific person in massive surveillance videos. In addition, due to the different locations of surveillance cameras, the surveillance scenarios are also different, which brings greater challenges to person search. Therefore, it is of great significance to effectively and efficiently search for relevant persons in large-scale surveillance videos.

In recent years, many person re-identification methods [1–5] have been proposed to solve this problem. Given a query person, these methods find the same person across cameras, thereby avoiding the visual limitations of fixed cameras. These methods manually crop person images by bounding boxes [1,6,7] and separate them from the real scene. They then match the target person with a gallery of cropped pedestrian images, ignoring the influence of complex backgrounds. However, in real-world surveillance, these manually cropped bounding boxes are not available, and the target person is searched for in unknown surveillance videos, which limits the practical application of person re-identification methods.

To bridge the gap between person re-identification and practical application, some person search methods have been studied. Unlike person re-identification, person search methods need to detect persons in unknown images or videos and identify them in large-scale datasets. Recently, the end-to-end person search framework [8] has been widely used in person search methods. In this framework, person detection and person re-identification are jointly handled in a single convolutional neural network. An Online Instance Matching (OIM) loss function is also introduced to ensure training efficiency. After that, OIM is improved by some person search methods [9–11] to increase the similarity of the same person. Although these methods have achieved good performance to some extent, they are not suitable for surveillance videos. This is because these methods are trained on CHUK-SYSU [8] and PRW datasets [12], and only explore single images in the two datasets. They do not consider the temporal relationship of adjacent frames in the video, which reduces the search accuracy of surveillance videos.

In this work, we propose an efficient person search method that employs spatio-temporal features. Since surveillance videos contain a series of frames that are temporally correlated, our method not only employs the spatial features of persons in each frame, but also utilizes the temporal relationship of the same person between adjacent frames. For this purpose, the bounding boxes of persons are detected by Yolo network, and then the corresponding spatial features are obtained using ResNet-50 model. After that, Gated Recurrent Unit (GRU) is used to calculate the temporal correlation of spatial features between adjacent frames, and an average pooling layer is added to generate a spatio-temporal feature for each person. Finally, to ensure search efficiency, Locality Sensitive Hashing (LSH) is explored to organize a large number of spatio-temporal features. To verify our method, we also construct a real-world surveillance video database, which contains video clips of different backgrounds, viewpoints, and person identities.

The rest of this paper is organized as follows. Section 2 reviews the related work in the area of person re-identification and person search. Section 3 details our proposed methods using spatio-temporal features. Section 4 presents the experimental results to evaluate our method. Section 5 summarizes this paper.

2. Related Work

2.1. Person Re-Identification Method

Person re-identification methods originated from multi-camera tracking, and they are now considered a sub-direction of image retrieval. In 2014, deep learning methods were first introduced into person re-identification [13,14]. Later, spatio-temporal representation was used in some person re-identification methods, such as the deep spatio-temporal appearance descriptor [15], the hierarchical spatial-temporal model [16], the Batch Drop-Block network [17], the temporal attention network [18], the non-local video attention network [19] and spatio-temporal representation factorization [20].

Although the above person re-identification methods improve accuracy by incorporating spatial–temporal representation, they are trained on cropped image datasets, such as Market-1501 [1], CUHK03 [6], and MARS [7]. These datasets are manually annotated or detected by the Deformable Part Model (DBP) [21] to generate bounding boxes, without any real-world surveillance scenarios. Because most of the person re-identification methods do not involve pedestrian detection, they are not suitable for real-world surveillance videos, and they cannot be used in practical applications.

2.2. Person Search Method

Person search is developed based on person re-identification and fills the gap between person re-identification and practical applications. According to different person detection methods used in the search process, the existing person search methods can be divided into two-stage search methods [12,22–24] and end-to-end search methods [8–11,25–31].

Two-stage person search methods divide person detection and person re-identification into two separate stages. Zheng [12] explores region convolutional neural network as person detector and introduces the confidence weighted similarity metric into detection scores. Chen proposes a mask-guided two-stream CNN model [22] to obtain enriched representations. Similarly, Lan proposes a cross-level semantic alignment deep learning method [23] to learn more discriminative identity feature representation in the detection stage. Wang proposes a task-consistent two-stage framework [24], which consists of an identity-guided query detector and a detection results adapted re-ID model.

End-to-end person search methods integrate person detection and re-identification in the same framework. He proposes a detection and re-identification integration net [25] and utilizes Siamese architecture in the re-identification stage. Dong also uses a Siamese network with an additional instance-aware branch in the bi-directional interaction network [26]. Zhong proposes an align-and-park network [27] to enhance the robustness of person search. Yan presents a feature-aligned person search framework [28], which is the first anchor-free network to tackle misalignment issues. Han proposes a decoupled and memory-reinforced network [29] to reconcile the conflicts of multiple objectives between detection and identification sub-tasks. Zhang proposes diverse knowledge distillation and spatial-invariant augmentation [30] to assist end-to-end framework. Li proposes a sequential end-to-end network [31] to extract superior features. Yu introduces a cascade occluded attention transformer [32] into the end-to-end framework, which provides a discriminative coarse-to-fine representation for person detection and re-identification.

In addition, some person search methods are proposed for video datasets. Huang explores visual and temporal links of person identities by progressive propagation [33], but this method is only trained on a movie dataset. Alcazar presents a dataset for audiovisual person search [34], but this dataset consists of face tracks and voice segments which are not related to surveillance videos. Kumar provides a P-DESTRE dataset [35] collected by aerial devices for person detection, tracking, and re-identification. Rehman employs frame differencing and MACH filters [36] to detect persons in secure areas, and Malviya presents a multi-behavioral social particle filter [37] to track moving persons from a moving robot. Ma proposes an unsupervised video hashing method [38], but this method takes each video frame as a whole. Since this method does not detect persons in the videos, the spatial features of persons are affected by complex backgrounds.

Person search methods achieve good performance to some extent, but these methods still have some limitations. First, due to public privacy, there are few surveillance videos publicly available for person search, which limits the development of person search methods. Second, to the best of our knowledge, most of the person search methods are trained on CHUK-SYSU and PRW datasets. Because these two datasets do not possess the temporal property of videos, the person search methods trained on these datasets cannot be directly utilized in real-world surveillance. Third, compared with image-based person search methods, there are relatively few video-based person search methods. For surveillance videos, the performance of person search methods needs to be improved by incorporating the temporal relationship of persons between adjacent frames.

3. The Proposed Method

3.1. Framework

Figure 1 shows the framework of the proposed person search method for surveillance videos. This framework consists of three components: spatial feature extraction, spatio-temporal feature generation, and feature indexing. First, a surveillance video consists of continuous frames, which contain information about many pedestrians. Each frame is fed into Yolo network to detect the bounding boxes of persons, and the corresponding spatial features are extracted by Resnet-50 model for each person. After that, the process of person search is not affected by surveillance background. Next, to explore the temporal relationship of persons between adjacent frames, the spatial features of each person are successively imported into Gated Recurrent Unit (GRU). The hidden state generated by

GRU is used to calculate the temporal correlation of a person between adjacent frames. Since each spatial feature corresponds to a hidden state, it is not feasible to find the target person by directly computing the similarity between a large number of hidden state vectors. To solve this problem, an average pooling layer is added to the output of the GRU. A single spatio-temporal feature is generated by averaging the hidden state sequence for each person, and the computational cost is reduced in the search process. After generating spatio-temporal features, the spatial features within each frame and the temporal relationship between adjacent frames are fully exploited. Finally, due to a large number of video clips in the database, all the spatio-temporal features are organized by Locality Sensitive Hashing (LSH), thereby ensuring the efficiency of person search. The similarity between a target person and persons in the database is computed through LSH, and the persons with the greatest similarity are returned as the search results.

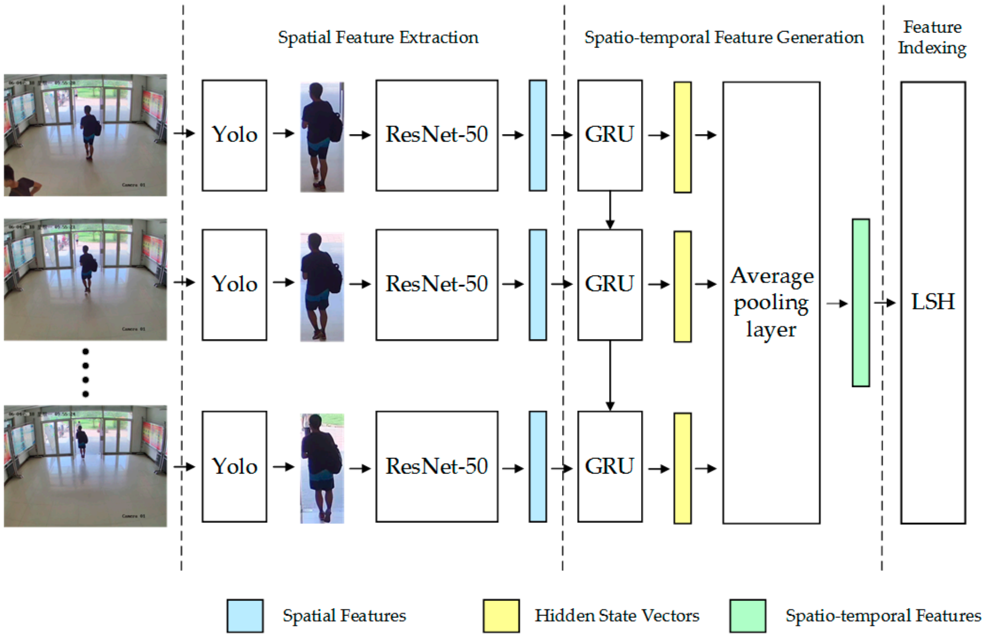


Figure 1. The framework of the proposed person search method.

3.2. Spatial Feature Extraction

Surveillance video contains a series of frames. To detect persons in each frame and avoid the influence of surveillance background, Yolo network [39] is employed to determine the bounding boxes of persons. Compared with Fast RCNN [40], the Yolo network can provide a faster detection speed, which is more suitable for surveillance videos. Let $V = \{v_1, v_2, \cdots, v_N\}$ be a collection of N frames in the video, and the i-th frame is represented as v_i. After the frame v_i is processed by the Yolo network, the bounding box is detected and represented as $r_{i,j} = \{x_{i,j}, y_{i,j}, l_{i,j}, w_{i,j}\}$, where $r_{i,j}$ is a bounding box of the j-th person in the frame, v_i, $x_{i,j}$, and $y_{i,j}$ are the central locations of the bounding box $r_{i,j}$, and $l_{i,j}$ and $w_{i,j}$ are the height and width of the bounding box $r_{i,j}$.

Based on the bounding boxes of persons in each frame, Resnet-50 model [41] is used to extract the corresponding spatial features of persons. Each bounding box is passed through the Resnet-50 model by using conv4_4 to conv5_3, and then sent into a global average pooling layer to generate a 2048-dimensional feature vector. After that, this high-dimensional feature vector is projected into an L2-normalized 256-dimensional subspace. Finally, each bounding box in one frame of the video can be represented by a 256-dimensional spatial feature.

For the bounding box $r_{i,j}$, the Resnet-50 model can be regarded as a non-linear function $f_{RES}(\cdot)$. After the bounding box $r_{i,j}$ is processed by the Resnet-50 model, the spatial feature $s_{i,j}$ can be represented as:

$$s_{i,j} = f_{RES}(v_i) \tag{1}$$

where $s_{i,j}$ is a 256-dimensional spatial feature of the j-th person in the frame v_i. All the spatial features of persons can be represented as $S = \{s_{i,j}\}(1 \le i \le n_j, 1 \le j \le M)$, in which n_j is the number of frames containing the j-th person, and M is the total number of persons in the video.

3.3. Spatio-Temporal Feature Generation

Surveillance video is not only spatially invariant but also temporally continuous. To organize the temporal relationship of persons between adjacent frames, Gated Recurrent Unit (GRU) [42] is employed to process the spatial features of persons extracted from each frame. GRU is a type of Recurrent Neural Network (RNN) and solves the vanishing gradient problem in a standard RNN. Derived from Long-Short Term Memory (LSTM) [43], GRU combines a forget gate and an input gate into one single update gate and utilizes the hidden state to transfer information. Compared with LSTM, the parameters of GRU are fewer and the structure of GRU is relatively simple. Thus, GRU can be trained more easily than LSTM. Because GRU can process data sequences efficiently, it is used to model the temporal correlation of the same person between adjacent frames.

In Figure 2, the spatial feature $s_{i,j}$ is taken as an input vector and fed into GRU. Firstly, reset gate $c_{i,j}$ and update gate $d_{i,j}$ can be computed as:

$$c_{i,j} = \sigma(W_c s_{i,j} + U_c h_{i-1,j} + b_c) \tag{2}$$

$$d_{i,j} = \sigma(W_d s_{i,j} + U_d h_{i-1,j} + b_d) \tag{3}$$

where $h_{i-1,j}$ is the hidden state of the j-th person in the frame v_{i-1}, $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid non-linearity function that maps real-value inputs into the interval [0, 1], W_c, U_c, and b_c are the parameters of reset gates $c_{i,j}$, and W_d, U_d and b_d are the parameters of update gates $d_{i,j}$.

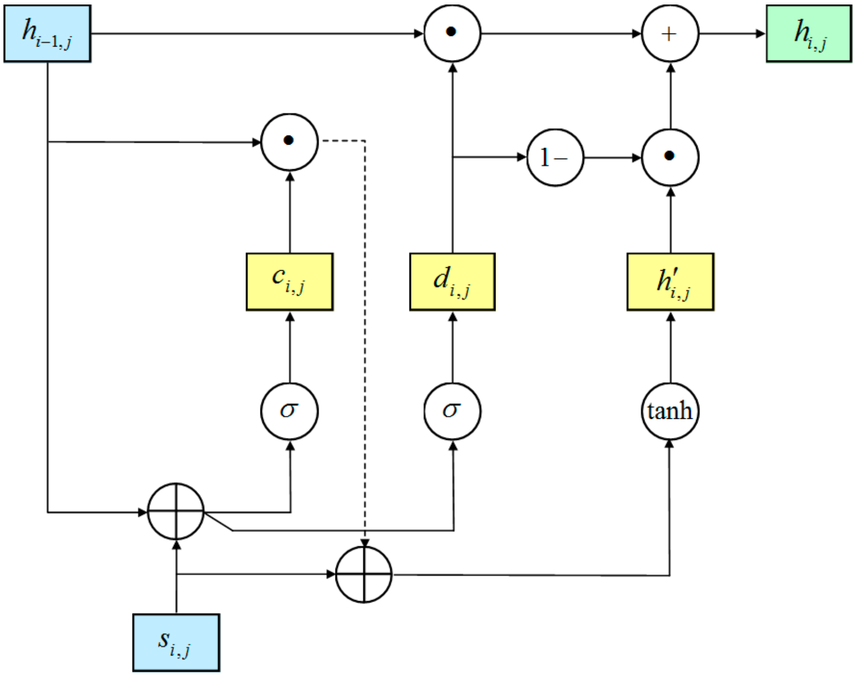


Figure 2. The structure of GRU.

Next, in a new gate, the candidate hidden state $h'_{i,j}$ is updated by spatial feature $s_{i,j}$, reset gate $c_{i,j}$, and hidden state $h_{i-1,j}$, and is computed as:

$$h'_{i,j} = \tanh(W_h s_{i,j} + U_h(c_{i,j} \odot h_{i-1,j}) + b_h) \tag{4}$$

where tanh is activation function, $\odot$ represents the element-wise product, and W_h, U_h and b_h are gate parameters.

Finally, the hidden state $h_{i,j}$ is generated by update gate $d_{i,j}$, hidden state $h_{i-1,j}$, and candidate hidden state $h'_{i,j}$, and is computed as:

$$h_{i,j} = d_{i,j} \odot h_{i-1,j} + (1 - d_{i,j}) \odot h'_{i,j} \tag{5}$$

from Equation (5), the hidden state $h_{i,j}$ not only considers the candidate hidden state $h'_{i,j}$ in the frame v_i, but also involves the hidden state $h_{i-1,j}$ in the frame v_{i-1}, thereby establishing the temporal relationship of the j-th person between frame v_{i-1} and v_i.

After all the spatial features of persons have been processed by GRU, for the j-th person, we can obtain a corresponding hidden state sequence $h_j = \{h_{i,j}\}(1 \leq i \leq n_j)$, which represents temporal correlation of persons in the n_j frames. Then, all the persons in the video can be represented by a collection of hidden state sequences $H = \{h_j\}(1 \leq j \leq M)$.

The hidden state $h_{i,j}$ is a 256-dimensional vector, thus the dimension of the hidden state sequence h_j is $256 \times n_j$, and the dimension of hidden state sequences H is $256 \times n_j \times M$. Due to the high dimensionality of h_j and H, the computational cost is increased by directly utilizing the hidden state sequences in the search process. To reduce the dimension of the hidden states and generate spatio-temporal features for each person, an average pooling layer is added to the output of GRU. After averaging the hidden state sequence h_j, for the j-th person, the spatio-temporal feature can be represented as:

$$p_j = \frac{1}{n_j} \sum_{i=1}^{n_j} h_{i,j} \tag{6}$$

from Equation (6), the average pooling layer finally transforms a $256 \times n_j$ dimensional hidden state sequence h_j into a 256-dimensional spatio-temporal feature p_j, thereby reducing the computational cost. For all the persons in the video, the spatio-temporal features can be represented as $P = \{p_j\}(1 \leq j \leq M)$.

3.4. Feature Indexing

After all the spatio-temporal features of persons are generated by GRU, each person corresponds to a spatio-temporal feature. Due to the large number of pedestrians in surveillance videos, it is difficult to ensure the efficiency of person search by directly calculating the similarity between a large number of spatio-temporal features. Because Locality Sensitive Hashing (LSH) [44] can perform a fast approximate search of the massive high-dimensional data, it is employed to encode spatio-temporal features into short binary codes, and then compute the similarity between different features in the Hamming space.

All the spatio-temporal features are input into LSH and projected into a low-dimensional Hamming space. For the spatio-temporal feature p_j, it is mapped to a b-bit hash code by applying b binary-valued hash functions H, and represented as:

$$H : p_j^{256\times 1} \rightarrow \{0,1\}^b \tag{7}$$

In the search process, the spatio-temporal feature of the query person is generated and projected into the Hamming space, and the similarity between the query person and persons in the surveillance videos can be computed as:

$$sim(q, p_j) = \Pr[H(q) = H(p_j)] \tag{8}$$

where q is the spatio-temporal feature of the query person.

After the query person is searched for in the video database through LSH, the persons in the videos are sorted by the similarity of spatio-temporal features, and the top-K persons with the greatest similarity are taken as the final search results.

3.5. Computational Complexity Analysis

For one video clip V in the dataset, as described in Sections 3.2–3.4, $n_j \times M$ spatial features are extracted for all the persons, which has a computational complexity of $O(n_j M)$. After the spatial features are imported into GRU and the average pooling layer, M spatio-temporal features are generated. Compared with the number of spatial features, the number of spatio-temporal features decreases n_j times, and the computational complexity reduces to $O(M)$. Suppose the dataset contains N_V video clips, the total number of spatio-temporal features is $N_V \times M$, and the computational complexity is $O(N_V M)$. Due to the large number of spatio-temporal features in the dataset, searching for a target person by directly computing the similarity between spatio-temporal features increases the computational cost. Thus, all the spatio-temporal features are organized by LSH, and the query time for top-K results reduces to $O((N_V M)^{1/K})$ for the Hamming distance, thereby reducing the computational cost of the search process and improving search efficiency. In Section 4.4, the experimental results verify the search efficiency of the proposed method.

4. Experiments and Results

4.1. Datasets

For these experiments, we built a surveillance system to capture videos. This system consisted of nine cameras, which were deployed indoors and outdoors. Different cameras exhibited different backgrounds, illuminations, viewpoints, time periods, and person identities, which led to differences in the appearance of the same person. In addition, some pedestrians were occluded by complex backgrounds or crowded environments.

From the surveillance system, we collected 15k video clips with 897 person identities and created a surveillance video database. To evaluate the performance of the person search method, we selected 11,546 video clips as training data and the remaining 3454 video clips as testing data. The statistics are summarized in Table 1, and a few examples of the video database are shown in Figure 3.

Table 1. The statistics of surveillance video database.

Database	Video Clips	Person Identities	Training Data	Testing Data
Surveillance Video	15,000	897	11,546	3454

4.2. Evaluation Criteria

In the surveillance video database, two different criteria were used to evaluate the performance of the proposed method. First of all, Average Precision (AP) was utilized as the performance metric for a query person, and Mean Average Precision (MAP) was employed to evaluate the overall performance and computed as:

$$MAP = \frac{1}{N_q} \sum_{i=1}^{N_q} AP_i \tag{9}$$

where N_q is the total number of query persons.

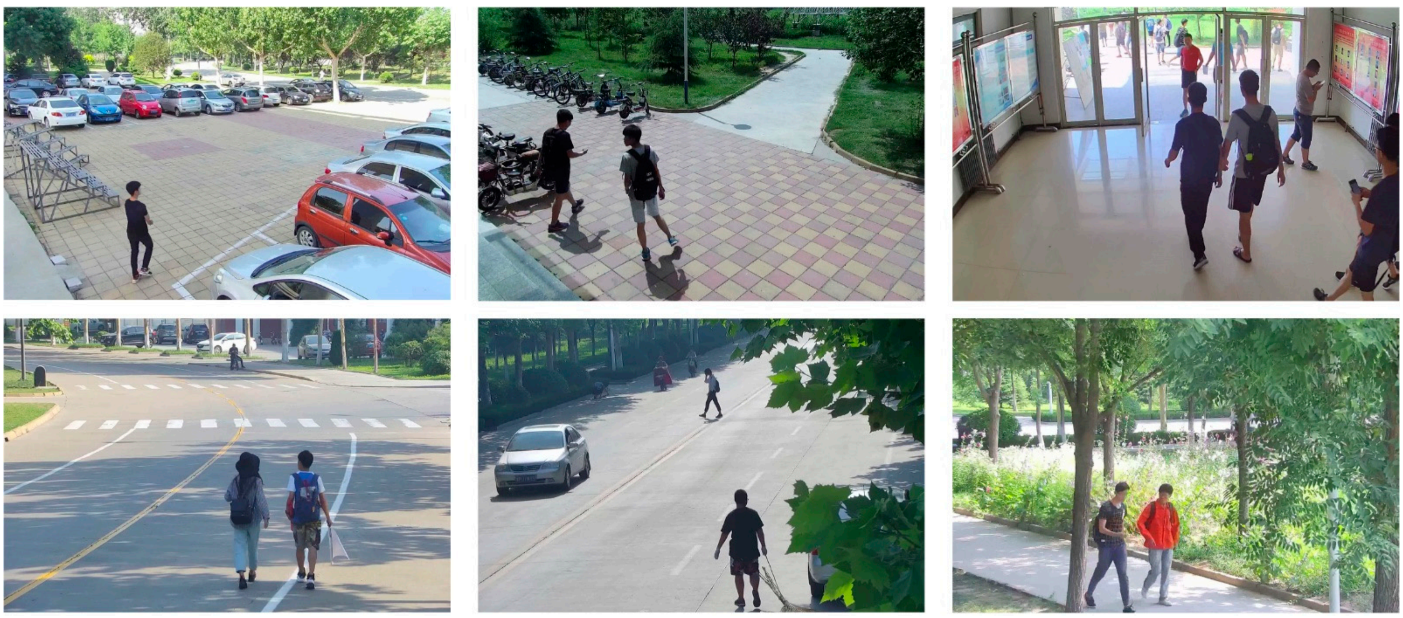


Figure 3. Examples of the surveillance video database.

In real-world surveillance, we pay more attention to the top-K search results. Thus, the top-K accuracy is defined as:

$$Accuracy = \frac{D_q}{N_q} \tag{10}$$

where D_q is the number of query persons that obtain top-K positive search results, and N_q is the total number of query persons. In the following experimental results, the value of K is set to 1, and top-1 accuracy is used to evaluate the performance.

To evaluate search efficiency, the average time for N_q query persons is used to compare our proposed method with the existing person search methods.

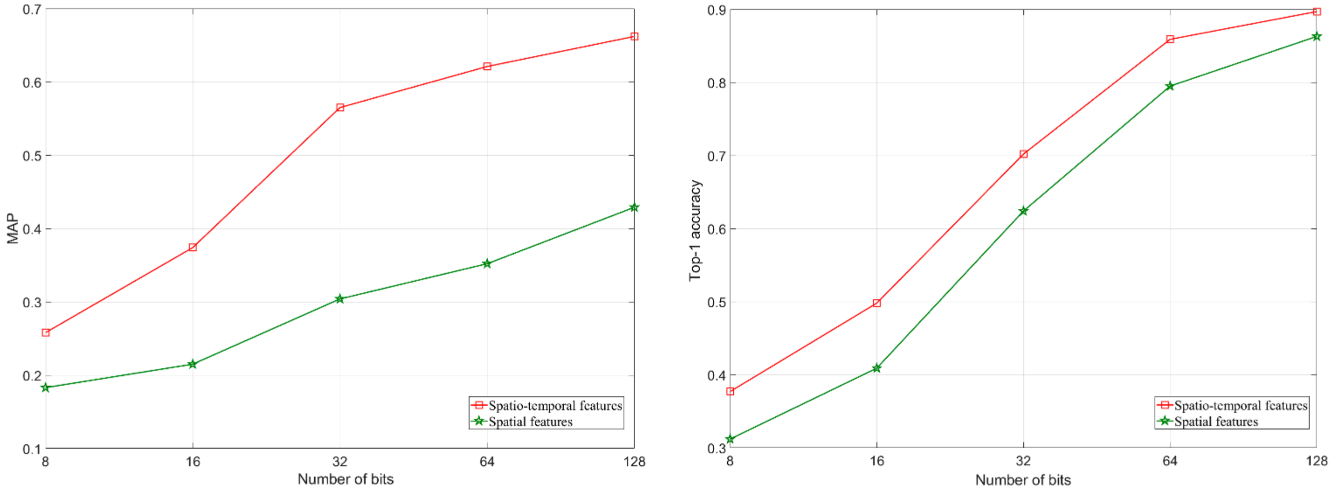
4.3. Experimental Design

In the surveillance video database, the RGB frames of video clips were employed as the inputs for the proposed person search method. A pretrained Yolov5 network [45] was used to detect the bounding boxes of persons in each frame, and an ImageNet-pretrained Resnet-50 model [42] was used to extract the spatial features of persons. After that, GRU and average pooling layer were added to the network and finetuned together with the Resnet-50 model. Finally, to ensure search efficiency, LSH was used to organize all the 256-dimensional spatio-temporal features in the video database.

To verify the spatio-temporal features and evaluate the search performance, our proposed method was compared with Region Convolutional Neural Network (RCNN) [8], which is widely used in person search methods, and the proposed method was also compared with the Progressive Propagation Person Search (PPPS) method [34]. The comparisons included: (i) the accuracy comparison between spatio-temporal features and spatial features, (ii) the accuracy comparison between our method and RCNN using spatio-temporal features, (iii) the accuracy comparison between GRU and LSTM with different pooling methods, (iv) the accuracy comparison between our method and the RCNN method with different numbers of video clips, (v) the accuracy comparison between our method and the PPPS method with different numbers of video clips, (vi) the efficiency comparison between our method, the RCNN method and the PPPS method with different numbers of video clips. All the methods were performed on machines with an Intel Xeon E5-1620 3.5 GHz processor, 256 GB of RAM, and NVIDIA Geforce GTX 1080Ti GPU. All the experiments are implemented based on PyTorch [46].

4.4. Experimental Results

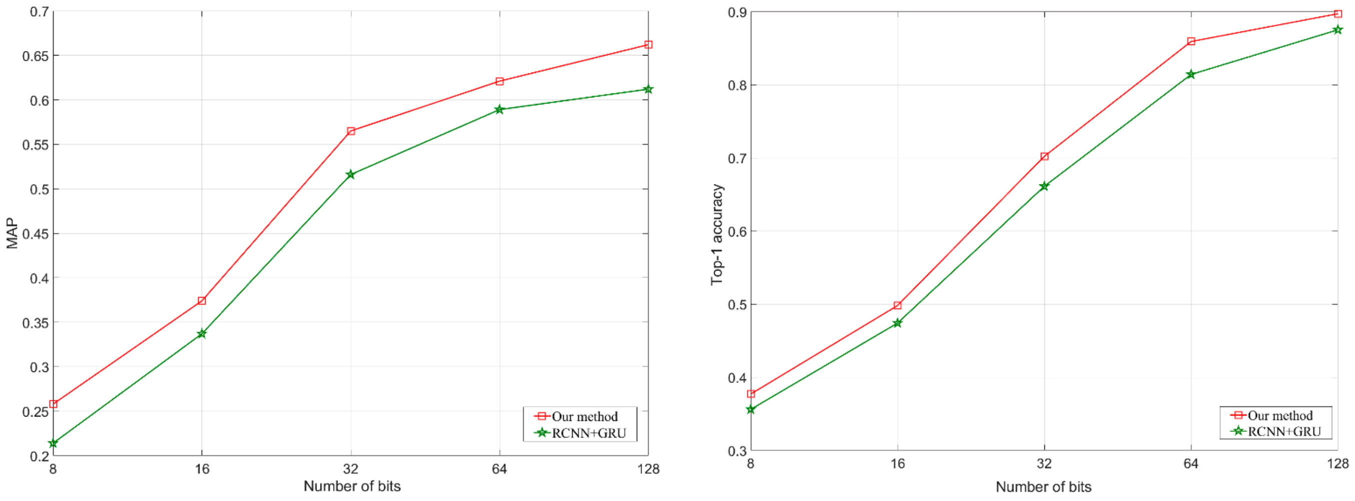
To evaluate search accuracy, we compared spatio-temporal features with spatial features in the framework of the proposed method. In order to compare them with spatio-temporal features, the bounding boxes of spatial features were also detected by the Yolov5 network, and the spatial features were extracted from key frames. As the number of bits in the hash code increased from 8 to 128, the MAP value of our method increased from 0.258 to 0.662 in Figure 4a, and the top-1 accuracy of our method increased from 0.377 to 0.897 in Figure 4b. This was because the large number of hash bits can increase the distinction between spatio-temporal features. Although the search accuracy of using only spatial features increases to some extent, it is still lower than the search accuracy of our method. Compared with the spatial features extracted from key frames, the spatio-temporal features not only contain the spatial features within each frame, but also represent the temporal relationship of spatial features between adjacent frames. The representative ability of spatio-temporal features is better than that of spatial features, thereby increasing the distinction between different persons in the surveillance videos. Thus, the search accuracy is improved by utilizing the spatio-temporal features.



(**a**) (**b**)

Figure 4. The accuracy comparison between spatio-temporal features and spatial features: (**a**) MAP comparison; (**b**) Top-1 accuracy comparison.

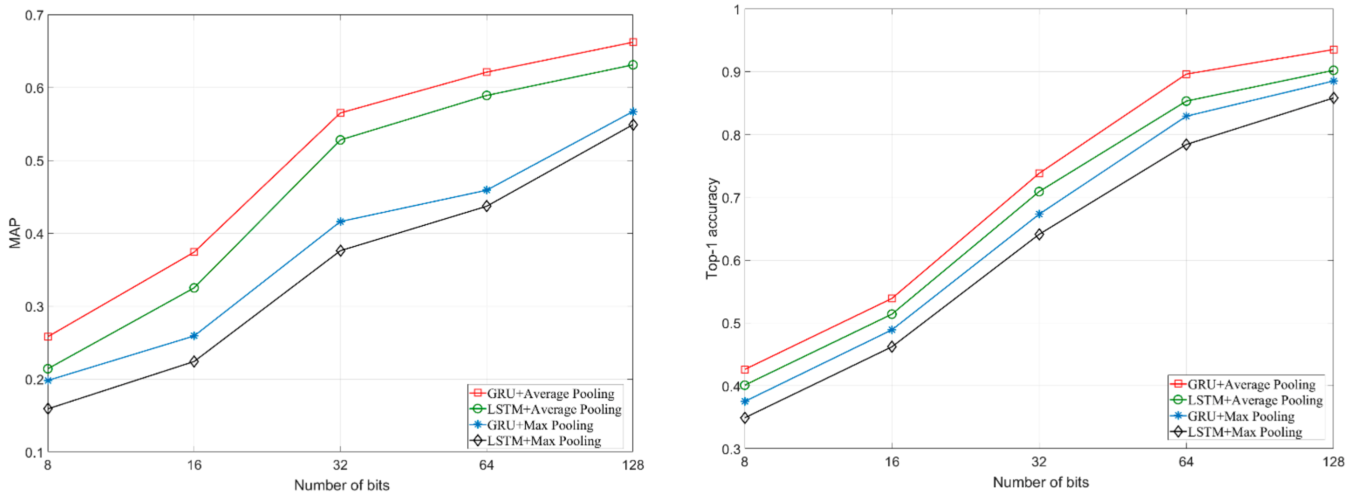
To further verify the effectiveness of the proposed method, our method was compared with RCNN, which is also employed to generate spatio-temporal features. In the structure of RCNN, the pedestrian proposal net is used to select the top 128 bounding boxes, and the identification net is used to extract spatial features. For the same comparison with our method, these spatial features were also fed into GRU to generate 256-dimensional spatio-temporal features, and then organized by LSH. As the number of bits in the hash code increases, the MAP value of our method is larger than that of RCNN in Figure 5a, and the top-1 accuracy of our method is also better in Figure 5b. This is because the pedestrian proposal net in RCNN would inevitably generate some false alarms and misalignments in the bounding boxes, which reduce the representative ability of spatio-temporal features and decrease the search accuracy. Compared with RCNN, the Yolo network used in our method can avoid the false alarms in the candidate bounding boxes and provide fast detection speed for a large number of frames. Therefore, the search accuracy of our method is improved by combining the Yolo network with GRU.



(**a**) (**b**)

Figure 5. The accuracy comparison between our method and RCNN method using spatio-temporal features: (**a**) MAP comparison; (**b**) Top-1 accuracy comparison.

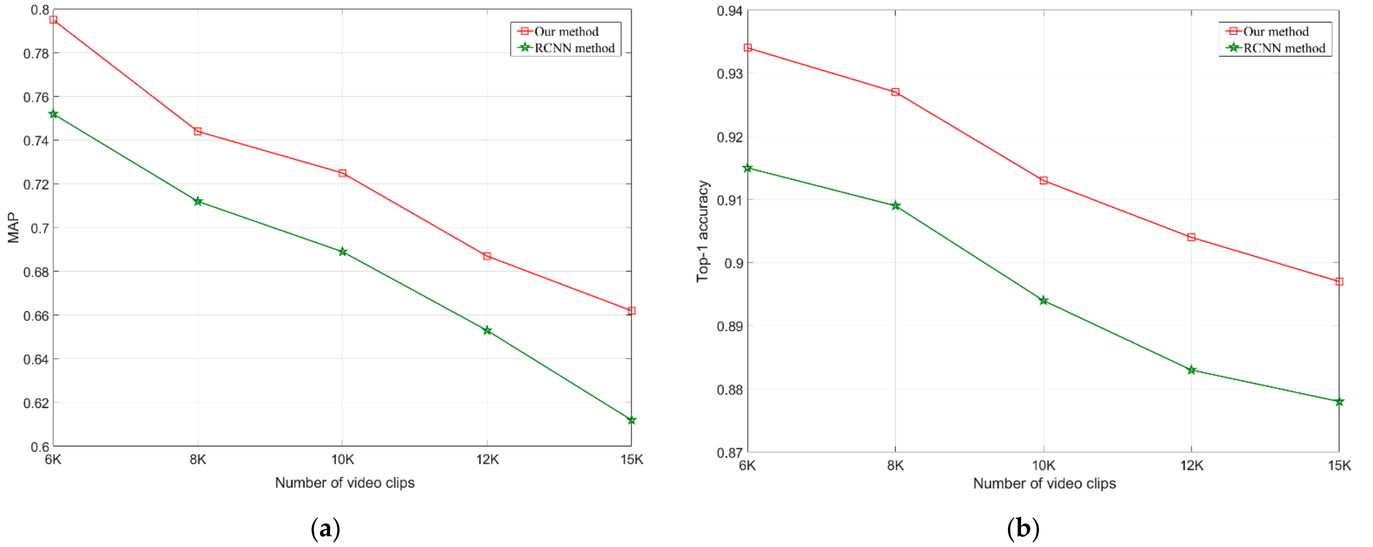
To further evaluate search accuracy, accuracy comparison was performed between GRU and LSTM with different pooling methods. In Figure 6, as the number of bits in the hash code increases, the search accuracy of four different methods increases. However, with the same pooling method, the search accuracy of GRU is better than that of LSTM. This is because GRU can be trained more easily and effectively than LSTM, and the parameters of GRU are fewer. Thus, GRU can better calculate the temporal correlation of the same person between adjacent frames and achieves better performance than LSTM in the search process. On the other hand, when using GRU or LSTM in the same situation, the search accuracy of the average pooling method is also better than that of the max pooling method, as shown in Figure 6. Compared with the max pooling method, the average pooling method can generate more representative spatio-temporal features. Of the four different methods, GRU and the average pooling method achieve the largest search accuracy. Therefore, in our proposed method, we employ GRU and the average pooling method to increase search accuracy.



(**a**) (**b**)

Figure 6. The accuracy comparison between GRU and LSTM with different pooling methods: (**a**) MAP comparison; (**b**) Top-1 accuracy comparison.

We compared the search accuracy between our method and the RCNN method with different numbers of video clips. For the same comparison, the RCNN method also uses GRU to generate spatio-temporal features. As the number of video clips increases from 6K to 15K, the MAP value of our method decreases from 0.795 to 0.662 in Figure 7a, and the top-1 accuracy of our method decreases from 0.934 to 0.897 in Figure 7b. This is because more video clips generate more spatio-temporal features which are affected by complex backgrounds, illuminations, viewpoints, occlusions, and various appearances of persons. Due to dimension reduction in our method and the increasing number of spatio-temporal features, the distinction between different spatio-temporal features may be decreased to some extent, which reduces the search accuracy. Although the search accuracy of our method is decreased, it is still better than that of the RCNN method, which proves that the spatio-temporal features generated by our method are more representative and ensure search accuracy.

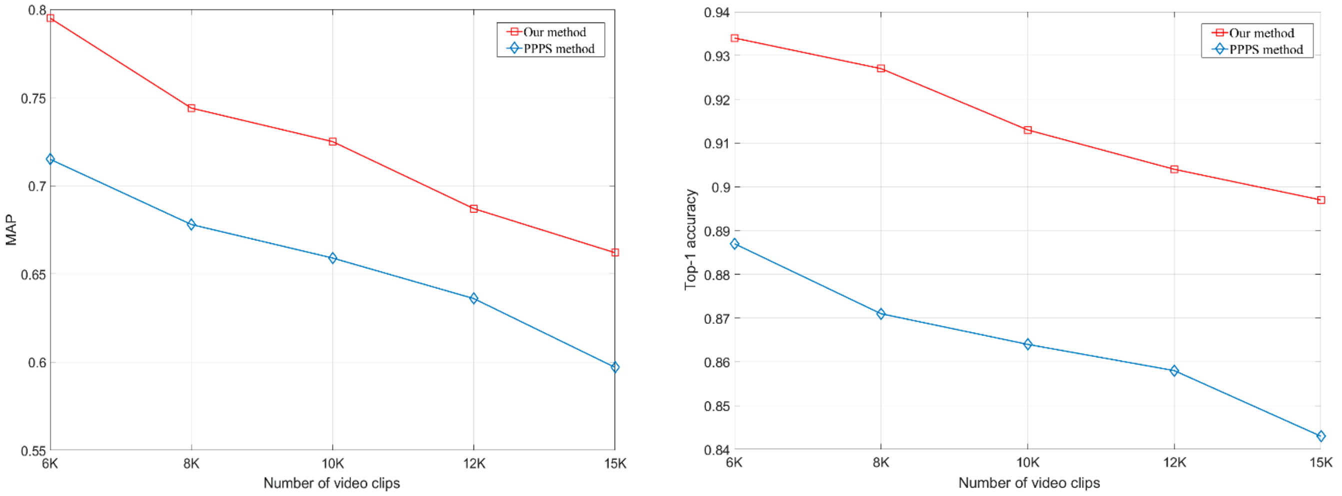


(**a**) (**b**)

Figure 7. The accuracy comparison between our method and the RCNN method with different numbers of video clips: (**a**) MAP comparison; (**b**) Top-1 accuracy comparison.

We also compared the search accuracy of our method with the PPPS method with different numbers of video clips. In Figure 8, as the number of video clips increases from 6K to 15K, the search accuracy of our method and the PPPS method decreases, but the accuracy of our method is better than that of the PPPS method. Unlike our method, the PPPS method uses a different way to organize the temporal relationship of video frames. In the PPPS method, the IDE feature is used to represent the person in each video frame, and then it is used to explore visual and temporal links of person identities and compute the similarity. Compared with the IDE feature representing a person in a single frame, the spatio-temporal features in our method represent persons between adjacent frames, which can better represent persons in the surveillance videos and increase the discrimination between different persons. Therefore, our method achieves greater search accuracy than the PPPS method.

To evaluate search efficiency, the average search time was used to compare our method with the RCNN method and the PPPS method. In Figure 9, as the number of video clips increases from 6 K to 15 K, all the search times of the three methods increase because more video clips increase the computational cost of the search process. However, the search time of our method is lower than that of the RCNN method and the PPPS method. Compared with the RCNN method using the pedestrian proposal net, our method generates fewer spatial features by utilizing the Yolo network in the stage of person detection, thereby reducing the computational cost of the subsequent process and obtaining a shorter search time. Different from our method and the RCNN method, the PPPS method explores the visual and temporal links to search for a target person, thus it has to calculate the similarity of these links for each frame, which greatly increases the computational cost of the search process. In contrast, our method employs LSH to organize the spatio-temporal features and compute the similarity, thereby reducing the computational cost. Compared with the RCNN method and the PPPS method, our method achieves the shortest search time and improves search efficiency.



(**a**) (**b**)

Figure 8. The accuracy comparison between our method and the PPPS method with different numbers of video clips: (**a**) MAP comparison; (**b**) Top-1 accuracy comparison.

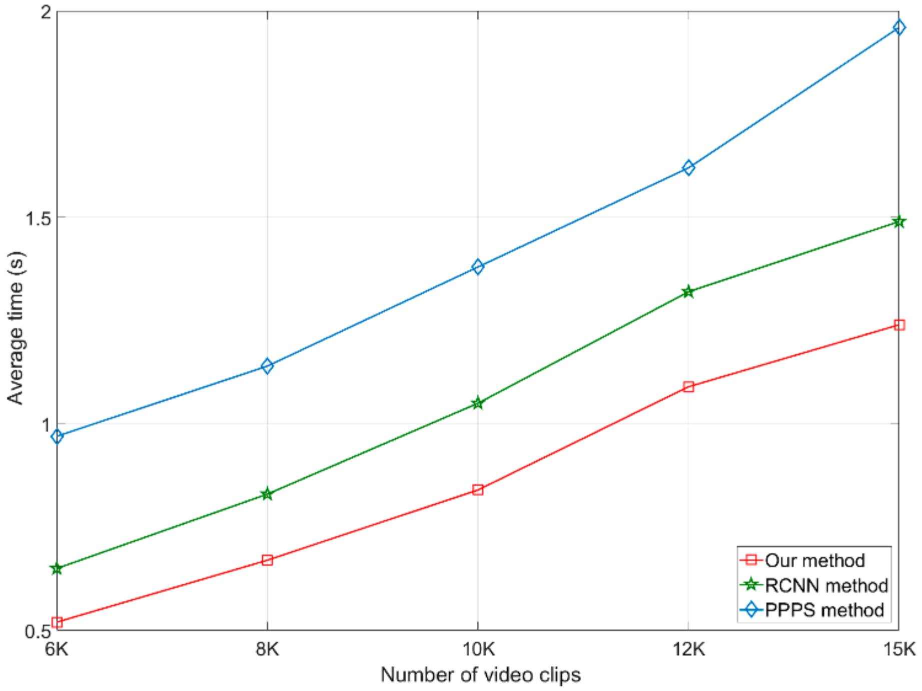


Figure 9. The efficiency comparison between our method, the RCNN method, and the PPPS method with different numbers of video clips.

4.5. Experimental Summary and Discussion

To fully explore the property of spatial invariance and temporal continuity of videos, the proposed method generates spatio-temporal features to represent persons and organizes these features to achieve person search of surveillance videos. Compared with conventional methods [37,38] utilizing frame differencing to detect persons in each frame, our method can detect persons more accurately with the Yolo network and better represent persons using spatial features. Simultaneously, unlike previous person search methods that only use spatial features, the spatio-temporal features not only contain the spatial features of persons in each frame, but also involve the temporal correlation of spatial features between adjacent frames. Thus, the spatio-temporal features are more representative than spatial features and are more suitable for surveillance videos. The spatio-temporal features increase the discrimination between different persons, thereby improving the search accuracy. As shown in the experimental results, the MAP value of our methods is 0.662 and the top-1 accuracy is 0.897, which are both better than the RCNN and PPPS methods.

Although the proposed method achieves better search performance compared with some state-of-the-art methods, it still suffers from some limitations. Firstly, our method can achieve good performance in single person surveillance scenarios, but the search accuracy drops in crowded environments. Because the persons are occluded by each other, the Yolo network cannot detect a complete bounding box for each person, thereby reducing the description of spatio-temporal features. Furthermore, our method only uses one GRU to calculate the temporal correlation of spatial features between adjacent frames, which limits the representative ability of spatial-temporal features. Finally, due to the high dimensionality of the hidden state sequence, our method adopts an average pooling layer to reduce dimensionality and generate spatio-temporal features, but it may lose some information on hidden state sequence and reduce the discrimination between different persons.

5. Conclusions

This paper proposes an efficient person search method that employs spatio-temporal features in surveillance videos. The spatial features of persons are extracted from each frame, avoiding the influence of complex surveillance background. GRU is then utilized to process the temporal relationship of spatial features between adjacent frames, followed by an average pooling layer to generate spatio-temporal features. Compared with spatial features, the spatio-temporal features are more representative for surveillance videos. To ensure search efficiency, LSH is used to organize massive spatio-temporal features and calculate similarity. We also constructed a surveillance video dataset to evaluate the proposed method, and the experimental results demonstrate that the proposed method improves the search accuracy while ensuring search efficiency.

In future work, we aim to apply this method to more video surveillance scenarios, such as the P-DESTRE dataset [36]. Therefore, we will explore more video frames to organize the temporal relationship of spatial features and increase the distinction of spatio-temporal features between different persons.

Author Contributions: Conceptualization and methodology, D.F. and J.Y.; formal analysis, D.F., Y.W. and L.Z.; writing—original draft preparation, D.F.; writing—review and editing, J.Y. and H.X.; supervision, J.Y. and H.X.; project administration, D.F. and J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China, China (No. 61603171), the Natural Science Foundation of Shandong Province, China (No. ZR2021MF091), and the Open Project of the Key Laboratory of System Control and Information Processing, Ministry of Education, China (No. Scip202011).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors are grateful to the Editor-in-Chief, Editor, and anonymous reviewers for their valuable reviews.

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

References

1. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1116–1124.
2. Xu, J.; Zhao, R.; Zhu, F.; Wang, H.; Ouyang, W. Attention-aware compositional network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2119–2128.
3. Borgia, A.; Hua, Y.; Kodirov, E.; Robertson, N.M. Cross-view discriminative feature learning for person re-identification. *Proc. IEEE Trans. Image Process.* **2018**, *27*, 5338–5349. [CrossRef]
4. Sun, X.; Zheng, L. Dissecting person re-identification from viewpoint of viewpoint. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 608–617.
5. Zheng, K.; Liu, W.; He, L.; Mei, T.; Luo, J.; Zha, Z.J. Group-aware label transfer for domain adaptive person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 5310–5319.
6. Matsukawa, T.; Okabe, T.; Suzuki, E.; Sato, Y. Hierarchical Gaussian descriptor for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1363–1372.
7. Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; Tian, Q. Mars: A video benchmark for large-scale person re-identification. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 868–884.
8. Xiao, T.; Li, S.; Wang, B.; Lin, L.; Wang, X. Joint detection and identification feature learning for person search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3376–3385.
9. Shi, W.; Liu, H.; Meng, F.; Huang, W. Instance enhancing loss: Deep identity-sensitive feature embedding for person search. In Proceedings of the IEEE International Conference on Image Processing, Athens, Greece, 7–10 October 2018; pp. 4108–4112.
10. Dai, J.; Zhang, P.; Lu, H.; Wang, H. Dynamic imposter based online instance matching for person search. *Pattern Recognit.* **2020**, *100*, 107120. [CrossRef]
11. Munjal, B.; Amin, S.; Tombari, F.; Galasso, F. Query-guided end-to-end person search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 811–820.
12. Zheng, L.; Zheng, H.; Sun, S.; Chandraker, M.; Yang, Y.; Tian, Q. Person re-identification in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honlulu, HI, USA, 21–26 July 2017; pp. 3346–3355.
13. Yi, D.; Liao, S.; Li, S.Z. Deep metric learning for person re-identification. In Proceedings of the International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 34–39.
14. Li, W.; Zhao, R.; Xiao, T.; Wang, X. DeepReID: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 152–159.
15. Ksibi, S.; Mejdoub, M.; Amar, C.B. Deep salient-Gaussian Fisher vector encoding of the spatio-temporal trajectory structures for person re-identification. *Multimed. Tools Appl.* **2018**, *78*, 1583–1611. [CrossRef]
16. Li, M.; Shen, F.; Wang, J.; Guan, C.; Tang, J. Person re-identification with activity prediction based on hierarchical spatial-temporal model. *Neurocomputing* **2018**, *275*, 1200–1207. [CrossRef]
17. Dai, Z.; Chen, M.; Gu, X.; Zhu, S.; Tan, P. Batch DropBlock network for person re-identification and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 3691–3701.
18. Rambhatla, S.S.; Jones, M. Body part alignment and temporal attention for video-based person re-Identification. In Proceedings of the British Machine Vision Conference, Cardiff, UK, 25 September 2019; pp. 1–12.
19. Liu, C.T.; Wu, C.W.; Wang, Y.C.F.; Chien, S.Y. Spatially and temporally efficient non-local attention network for video-based person re-identification. In Proceedings of the British Machine Vision Conference, Cardiff, UK, 5 August 2019; pp. 1–13.
20. Aich, A.; Zheng, M.; Karanam, S.; Chen, T.; Roy-Chowdhury, A.K.; Wu, Z. Spatio-Temporal Representation Factorization for Video-based Person Re-Identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 152–162.
21. Felzenszwalb, P.; McAllester, D.; Ramanan, D. A discriminatively trained, multiscale, deformable part model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
22. Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; Tai, Y. Person search via a mask-guided two-stream CNN model. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 764–781.
23. Lan, X.; Zhu, X.; Gong, S. Person search by multi-scale matching. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 553–569.
24. Wang, C.; Ma, B.; Chang, H.; Shan, S.; Chen, X. TCTS: A task-consistent two-stage framework for person search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11952–11961.
25. He, Z.; Zhang, L.; Jia, W. End-to-end detection and re-identification integrated net for person search. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; pp. 264–349.
26. Dong, W.; Zhang, Z.; Song, C.; Tan, T. Bi-directional interaction network for person search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2839–2848.
27. Zhong, Y.; Wang, X.; Zhang, S. Robust partial matching for person search in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6827–6835.
28. Yan, Y.; Li, J.; Qin, J.; Bai, S.; Liao, S.; Liu, L.; Zhu, F.; Shao, L. Anchor-free person search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 7690–7699.

29. Han, C.; Zheng, Z.; Gao, C.; Sang, N.; Yang, Y. Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search. In Proceedings of the 35th AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021.
30. Zhang, X.; Wang, X.; Bian, J.W.; Shen, C.; You, M. Diverse knowledge distillation for end-to-end person search. In Proceedings of the 35th AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021.
31. Li, Z.; Miao, D. Sequential end-to-end network for efficient person search. In Proceedings of the 35th AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021.
32. Yu, R.; Du, D.; LaLonde, R.; Davila, D.; Funk, C.; Hoogs, A.; Clipp, B. Cascade Transformers for End-to-End Person Search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 7267–7276.
33. Huang, Q.; Liu, W.; Lin, D. Person search in videos with one portrait through visual and temporal links. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 425–441.
34. Alcazaar, J.L.; Heilbron, F.C.; Mai, L.; Perazzi, F. APES: Audiovisual person search in untrimmed video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Virtual, 19–25 June 2021; pp. 1720–1729.
35. Kumar, S.; Yaghoubi, E.; Das, A.; Harish, B.; Proenca, H. The P-DESTRE: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. *IEEE Trans. Inf. Forensics Secur.* **2020**, *16*, 1696–1708. [CrossRef]
36. Rehman, S.; Riaz, F.; Hassan, A.; Liaquat, M.; Young, R. Human detection in sensitive security areas through recognition of omega shapes using Mach filters. In *Optical Pattern Recognition XXVI*; SPIE: Bellingham, WA, USA, 2015; p. 947708.
37. Malviya, V.; Kala, R. Trajectory prediction and tracking using a multi-behaviour social particle filter. *Appl. Intell.* **2022**, *52*, 7158–7200. [CrossRef]
38. Ma, C.; Gu, Y.; Gong, C.; Yang, J.; Feng, D. Unsupervised video hashing via deep neural network. *Neural Process. Lett.* **2018**, *47*, 877–890. [CrossRef]
39. Redom, J.; Divvla, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
40. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
41. He, K.; Zhang, X.; Ren, S.; Sun, S. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
42. Cho, K.; Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014.
43. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
44. Gionis, A.; Indyk, P.; Motwani, R. Similarity Search in High Dimensions via Hashing. In Proceedings of the 25th International Conference on Very Large Data Bases, Edinburgh, UK, 7–10 September 1999; pp. 518–529.
45. ultralytics. yolov5. Available online: https://github.com/ultralytics/yolov5 (accessed on 5 December 2021).
46. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 8024–8035.