Hindawi
Advances in Multimedia
Volume 2022, Article ID 4726450, 6 pages
https://doi.org/10.1155/2022/4726450

Research Article

Research on Intelligent Evaluation System of Sports Training based on Video Image Acquisition and Scene Semantics

Yong Ma

Wuhan Huaxia University of Technology, Wuhan, Hubei 430023, China

Correspondence should be addressed to Yong Ma; 20152201165@m.scnu.edu.cn

Received 15 December 2021; Revised 6 February 2022; Accepted 18 February 2022; Published 26 March 2022

Academic Editor: Qiangyi Li

This paper proposes to check the travel target of the dynamic background in the video surveillance with a fixed camera. A travel target detection method based on video picture acquisition and scene semantics for surveillance video was proposed. First, on the basis of combing the concepts and methods of picture recognition, the semantic information of the scene was fused to eliminate the interference factors in the unnecessary detection area. Secondly, a remote sensing picture visual feature representation method containing a semantic recognition method of remote sensing picture scenes and CSIFT features based on PLSA was presented. 10 types of typical remote sensing picture scenes are used for tests, and the visual vocabulary extraction method remains the same. The fixed visual vocabulary was 600, and the potential semantic subjects changes between 8~50. The test results indicated that the highest average recognition rate was obtained when the latent semantic topics were 20. Inappropriate latent semantic topics will lead to a decline in recognition rates. The effectiveness of this method was fully verified.

1. Introduction

Development of video surveillance technology can be defined as two stages, namely, the analog system stage and the digital system stage. After the digital video surveillance system, the network and intelligence began, the networking and intelligence belong to the category of digitalization [1]. The simulation system stage was the initial stage, and its main feature was that all the signals flowing in the system are analog signals, and the processing of the signals adopts the simulation technology [2]. Due to the limitations of analog technology, as digital technology developed, the video surveillance system began to develop in a digital direction. A semidigital system was a transition stage from an analog system to a digital system. It was characterized by the use of analog technology in the stage of video signal acquisition and digital technology in the stage of processing. After a period of time, it eventually developed into a fully digital video surveillance system [3]. Full digitization has opened a door for video surveillance technology and entered a broader development space [4]. After digitization comes networking and intelligence, which can realize some functions that cannot be imagined in the analog technology stage.

Moving target detection was at the bottom of the intelligent video system, which provides the basis for moving target tracking, moving target classification, and target behavior recognition. All subsequent processing was only directed at the region where the moving target was located, while other backgrounds were ignored [5]. Therefore, the quality of moving target detection results directly affects the accuracy of subsequent processing. The methods of background subtraction in moving target detection mainly consist of two categories: pixel-level background modeling methods and region level background modeling methods. Pixel-level background modeling methods can be divided into two categories: one of them is to model a single pixel in isolation without considering the relationship between it and its neighboring pixels; the other group takes the correlation between pixels into full consideration and takes the correlation information as a part of the features of pixels for the detection of the front and background. For the first type of modeling method, it first establishes a model for each pixel, such as a single Gaussian model, mixed Gaussian model, etc.,

then matches the current sampling value of the pixel with the background model, and judges whether the pixel was in the foreground or the background according to a certain matching algorithm. The foreground means there was movement, and the background means there was no movement. For the second type of modeling method, it was not only necessary to know the sampling value of the current pixel but also the sampling value of the surrounding pixel. A variable describing the correlation between pixels was calculated by a certain algorithm, which was used to distinguish the front background. The advantage of pixel-level background modeling was that the moving target detected was more detailed. But it also has its disadvantages: (1) since each pixel in the picture was modeled, more memory resources and computing resources were consumed. Meanwhile, the processing time was longer, compared with the first method. The second method consumes more resources because it needs to consider the correlation of adjacent pixels; (2) for the first type, the isolated modeling of a single pixel was often susceptible to the influence of noise, resulting in incomplete moving targets detected with holes. In order to realize timely prediction and early tourism emergency warning, computer vision and intelligent video processing technology are applied in Geng's work to check the singular incident in tourism surveillance video. At present, most video-based singular incident check methods perform well in normal scenarios, but they still cannot prevent low check rate and high miss rate in complex motion scenes, which means that they cannot be applied to real-time check of singular incidents. As a solution, a tourism video anomaly event detection model based on highlighting spatiotemporal features and sparse combinatorial learning was proposed in the paper. Excellent robustness and real-time performance of this model can be obtained in complex motion scenes, which can adapt to real-time singular incident checking in practical application. The three-dimensional gradient feature on the foreground target of the video sequence is extracted with a spatiotemporal gradient model combined with foreground check as the prominent spatiotemporal feature to eliminate the background interference. The abnormal event detection model was established based on sparse combinatorial learning algorithms to realize the real-time check of singular incidents. A new ScenicSpot dataset containing 18 video clips (5964 frames) was established as well, which includes ordinary and special events. The ScenicSpot dataset and two standard benchmark data sets are run in this method. Results of the tests show that the method can automatically detect and identify tourists' abnormal behavior, which performs better than the classical model [6].

Therefore, this research presents a monitoring video motion object detection method based on scene semantics. First, the related knowledge and technology of picture scene semantic recognition are summarized. Then, a remote sensing picture visual feature representation method and a scene semantic recognition method based on PLSA are given. Finally, 10 typical remote sensing picture scenes are used to verify the method. The number of scene types to be recognized in the test was 10, and the visual vocabulary extraction method remained the same. The fixed visual words were 600, and the latent semantic topics varied from 8 to 50. Compared to the average recognition rate, latent semantic topics were 20. Inappropriate latent semantic topics will lead to a decline in recognition rates.

2. Research Methods

Sports target detection was an important basic technology of intelligent video monitoring system, and was also the basis of behavior recognition, target tracking, and other intelligent analysis technology in national and social public safety, aerospace, and other important fields, and many civil fields have a pivotal role in target tracking, human-machine interaction, traffic control, video retrieval, and other fields have practical value. The application prospect of motion target detection in video systems is very great and has a wide market demand, so it has attracted high attention from researchers in relevant fields around the world, and many researchers have put forward their own algorithms. However, various algorithms have their own characteristics and applicable objects. Under different backgrounds, it was a more reasonable idea to choose different algorithms to achieve the detection of motion goals.

2.1. Concepts and Methods of Picture Scene Semantic Recognition. The methods of video picture acquisition and scene semantic recognition can be classified into 3 categories: (1) a semantic object method was constructed to show the whole scene by checking or distinguishing the semantic objects in the picture. [7]. (2) the scene Gwast model, which prevents the division of a single target or region and uses a low-dimensional spatial envelope to describe the structure of the scene, in which the five sensory attributes of naturalness, openness, roughness, extensibility, and roughness correspond to one dimension in the spatial envelope space, respectively, and each dimension corresponds to a meaningful spatial attribute in the scene as the basis for scene semantic division. (3) Establish the local semantic concept of the picture. Firstly, the points of interest are automatically detected in the picture, and local descriptors are used to describe these points. Then, the mapping from the local descriptor to a local semantic concept was established, and the distribution of local semantic concepts in the picture was used to realize the picture scene recognition. This method was mainly used for scene recognition in remote sensing pictures.

2.2. Visual Feature Packet Description of Remote Sensing Pictures. To accurately identify remote sensing picture scenes, discriminating features must be extracted from remote sensing pictures either by low-level or middle-level semantic modeling like block characteristics, regional characteristics, local invariant characteristics, etc. Different distinguishing features reflect different types of information, so they have their own benefits for specific categories. In addition, it is necessary for picture content analysis to be combined with different features in many cases, so the fusion of multiple features was conducive to improving the effect of

picture scene recognition. The Bag of Words (BOW) is the most commonly used simplified text description model in text processing, which expresses text into disordered word combinations without regard to syntax and word order. In text classification applications, the BOW model is often combined with the SVM classifier and the naive Bayesian classifier to obtain excellent classification results. Application of the model in computer vision was generalized as the feature packet (Bag of Features, BOF) method. The basic principle was to quantify the quantification of various local visual features by vectors and describe a picture or set of pictures by generating visual words or vocabulary. The CSIFT features of remote sensing pictures (or regions) to be recognized are usually extracted using the same method as training pictures. The visual vocabulary category of each CSIFT was determined according to the nearest neighbor rules, and the frequency of each visual vocabulary occurred in remote sensing pictures (regions) to be sorted. That was, the visual feature packet description of the remote sensing picture to be recognized. The visual feature package description of a remote sensing picture can transform the target segmentation and detection in the scene into the learning of visual vocabulary distribution and realize the connection between low-level picture feature representation and high-level semantics.

2.3. Picture Scene Semantic Recognition based on PLSA. When distinguishing different scenes, the visual vocabulary frequency can be used as a major basis, but in complex remote sensing picture scenes, there will be polysemy and similarity between visual words, because the same target may appear in different scene categories. Under the condition of insufficient training samples, the recognition method of directly linking scene category with an extracted feature vector was unable to approximate the actual scene semantics, leading to a decrease in the accuracy of the scene recognition. The main thinking method of this paper was to extract the latent semantics in the picture by applying the probabilistic latent semantic analysis (PLSA) model to the common training picture and completing the scene type judgment of the pictures to be identified based on the probability distribution of the underlying semantics.

The algorithm process was as follows:

(1) We extract the features of all the pictures. Some pictures were randomly selected from each training picture set. The feature vectors of CSIFT of these pictures were extracted, and M visual words were generated by the K-mean clustering algorithm. A similarity metric was carried out between every visual word and the feature vector of every training picture. The co-occurrence frequency matrix $n_0(\mathrm{d}_i, w_j)$ of $N \times M$ dimension "picture-terms" was obtained. Among them, $i \in (1, N)$, $j \in (1, M)$ represents the frequency of the visual word w_j in the picture d_i.

(2) The EM algorithm was applied to obtain the similar maximum likelihood plan of the PLSA model and the distribution rule $p(w_j|z_{\mathrm{Train}})$ of visual words when the potential semantics found in the pictures were obtained.

(3) The feature vectors of the test pictures were extracted, and the similarity measurement was carried out with M visual words obtained in step (1). The co-occurrence frequency matrix $n_T(d_T, w_j)$ of "picture-term" of the test picture was obtained. The co-occurrence frequency matrix $n_T(d_T, w_j)$ of $p(w_j|z_{\mathrm{Train}})$ and the test picture were used as the inputs of the PLSA model. By keeping $p(w_j|z_{\mathrm{Train}})$ the same, the potential semantic distribution $p(z|d_T)$ of the test picture can be obtained, which forms the K-dimension meaning vector of the test picture.

(4) The scene recognition of the picture was completed by using the KNN classifier to classify the latent semantic vector of the test picture.

2.4. Moving Object Detection Algorithm Integrating Scene Semantics. Since background objects are relatively stationary or slow moving, while the foreground objects' movements are related to the background, target detection was considered as a classification problem. That is, judging whether pixels belong to the foreground or background, a sample set at the position of each pixel was established by the background model, and it determines if the current pixel belongs to the background by comparing the pixel value at the corresponding position of the new frame with the sample set [8]. Setting up the background model requires the first frame to initialize. The filling of the background sample set was to fully contain the spatial and temporal distribution information in the first frame picture and use the similar spatial and temporal distribution characteristics of similar pixels. That is, for a pixel point, several neighborhood points are randomly selected as the sample set. The background pixel $N_G(x)$ of the initial model was the pixel value of the neighbor points, as presented in the following formula:

$$M(x) = \{v(y|y \in N_G(x))\}. \tag{1}$$

After adding the semantic prior information of the scene, the marked area in the scene was set as D_1. As shown in formula (2), if the current pixel value does not belong to D_1 and the number of pixel difference between this pixel, and the corresponding sample set less than a certain threshold was less than min, the current pixel was considered as the foreground point. If the pixel belongs to D_1 or the number of pixels whose difference between pixel and corresponding sample set was less than a certain threshold was greater than min, the current pixel was considered as the background point.

$$v(x) = \begin{cases} 1, & (v(x) \notin D_1) \wedge M(x)\left[A_{\vee(x)} < T\right] < \min, \\ 0, & (v(x) \in D_1) \wedge M(x)\left[A_{\vee(x)} < T\right] > \min, \end{cases} \tag{2}$$

where $v(x)$ was the pixel value of pixel point x, $M(x) = \{v_1, v_2, \ldots, v_n\}$ was the size of the background sample set (n

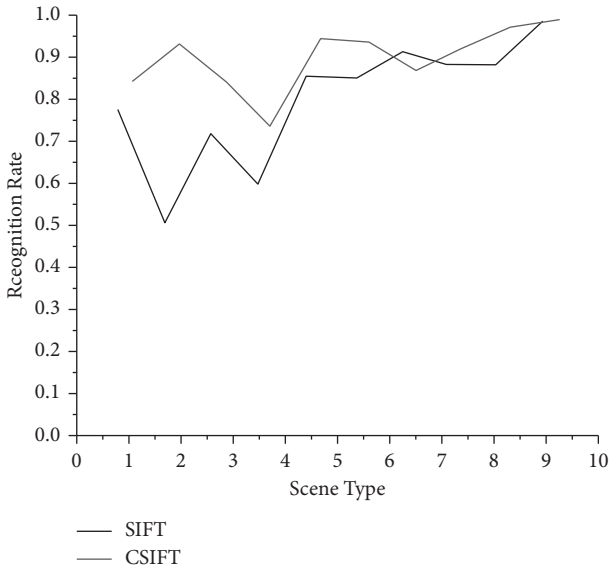

Figure 1: Scene recognition results with different low-level feature extraction methods.

was the size of sample set), $A_{v(x)}$ was the pixel point, $v(x)$ was the pixel difference of the corresponding sample set, and 1 and 0 represent the foreground and background, respectively.

3. Result Analysis

Google Earth captured 1794 picture clips of 10 different scenes. The quality of the picture was not limited, and the scene type was decided by the target in the slice. Among the 10 types of pictures, 50 pictures are stochastically selected as training pictures and the rest as test pictures [9]. The effects of different feature extraction methods on the recognition results are analyzed. The recognition effects of BOF description based on CSIFT feature and nearest neighbor classification directly by introducing the PLSA model are compared, and the recognition effects under different latent semantic topics and different visual words are compared to verify the effect of the algorithm.

3.1. Comparison of Different Low-Level Feature Extraction Methods. The design of the visual vocabulary generation method was based on CSIFT features. The most commonly used SIFT features are mainly for gray-scale pictures. When extracting features, it was first necessary to convert color pictures into gray-scale pictures. In the test, the dense grid's sampling interval was set to 8×8. The visual words were 600 and the potential semantic topics were 20. Test results with different underlying feature description methods are shown in Figure 1.

As can be seen from Figure 1, the use of CSIFT features as low-level features was overall superior to the traditional SIFT features based on the gray-scale, and the recognition performance of SIFT features based on the gray-scale was slightly better only for the "oil-fuel-pot" scene. This was mainly because the goal of dominant position in such scenarios was some cylindrical storage tanks whose shape characteristics are the most effective identification features, and large differences in tone in different regions, so the benefits of CSIFT are not obvious for this type of scenario. In terms of the average recognition rate of category 10 targets, CSIFT was 90.2% and SIFT was 79.67%, the former being significantly dominant.

3.2. Improvement of Identification Results by Introducing PLSA. The algorithm was realized by introducing the PLSA model to train the KNN classifier on the basis of remote sensing picture BOF description, which was denoting PLSA + BOF-KNN, the obtained remote sensing pictures of BOF description can also be directly trained into the KNN classifier for scene recognition, which was denoised as BOF-KNN. The dense grid's sampling interval was still set to 8×8 and the visual words were set to 600. Figure 2 shows the identification results, which are respectively given in the form of a classification confusion matrix. BOF was directly applied for recognition. Some scenes share a large number of visual words, resulting in large ambiguity in the recognition results. After PLSA was introduced, the polysemy phenomenon can be effectively eliminated and the scene recognition performance can be improved.

3.3. The Influence of Different Visual Words on Recognition Results. In the test, the scene types to be recognized were 10, the extraction method of visual words was the same, fixed visual words were set to 600, and potential semantic topics varied between 8 and 50. Figure 3 shows the average recognition rate comparison result. The highest average recognition rate was presented with a potential semantic topic of 20, and inappropriate latent semantic topics will lead to a decline in the recognition rate. [10]. The optimal number of visual words and the number of potential semantic topics exist in theory, but it was very difficult to accurately solve the problem in practice. At present, a great number of tests are used to determine an empirical value. Figure 4 shows the influence of different visual terms on recognition results.

3.4. The Influence of Different Number of Potential Semantic Topics on Recognition Results. In the test, the scene types to be recognized were 10, the extraction method of visual words was the same, the fixed visual words were set to 600, and the potential semantic topics varied between 8 and 50. The average recognition rate was compared, and Figure 3 showed the result. The highest average recognition rate was presented with a potential semantic topic of 20, and inappropriate latent semantic topics will lead to a decline in the recognition rate [10]. The most suitable visual words and potential semantic topics exist in theory, but it was very difficult to accurately solve the problem in practice. At present, a great number of tests are used to determine an empirical value.

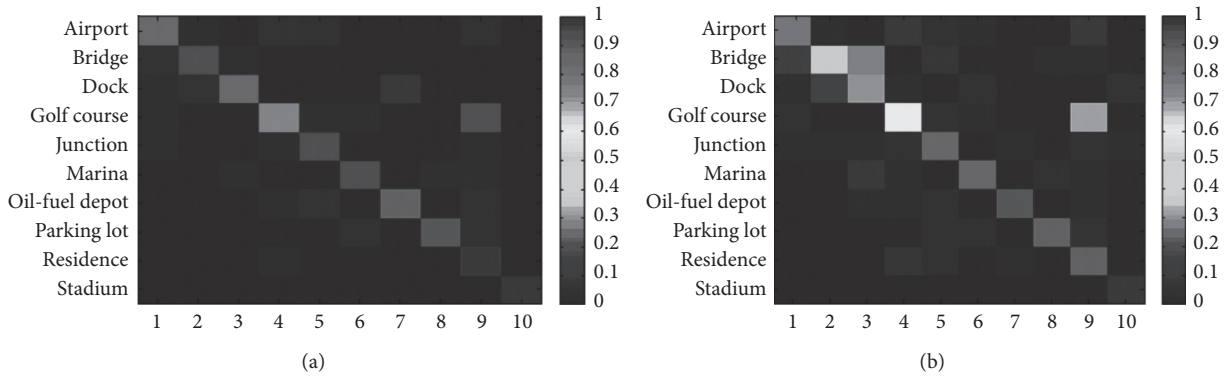

Figure 2: Scene recognition results obtained by using different recognition methods: (a) the classification mixed silk matrix obtained by ToF-KNN and (b) the classification mixed silk matrix obtained by PLSA + ToF-KNN.

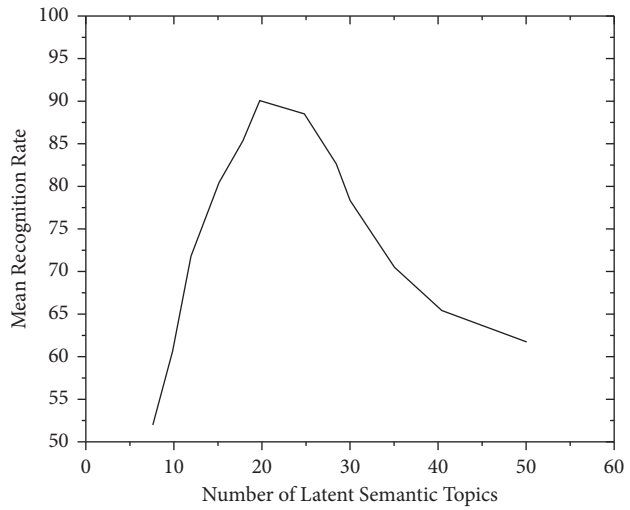

Figure 3: The effect of different number of potential semantic topics on recognition results.

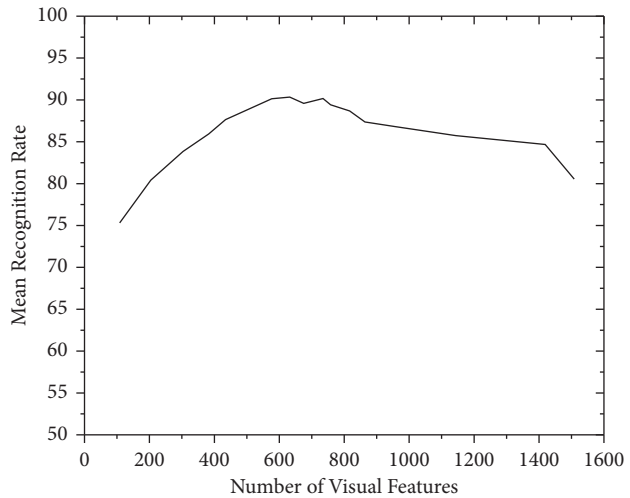

Figure 4: The influence of different visual terms on recognition results.

4. Conclusions

The detection of moving objects was easily affected by the environment, and the main problems it faces are as follows: light change, local occlusion, target scale change, picture jitter, noise interference, light change, shadow, reflection inside the region, moving target moving slowly, etc. Firstly, on the basis of combing the concepts and methods of picture recognition, the semantic information of the scene was fused to eliminate the interference factors in the unnecessary detection area. The qualitative analysis and quantitative analysis in the test part validate the proposed moving target detection algorithm.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] A. Yz, B. Zc, Z. A. Yue, C. A. Jie, A. Jy, and A. Yx, "Hybrid integration method for highly maneuvering radar target detection based on a Markov motion model-sciencedirect," *Chinese Journal of Aeronautics*, vol. 33, no. 6, pp. 1717–1730, 2020.

[2] R. Kruk and Z. Rempała, "Monitor for anti-aircraft guidance and observation systems," *Problems of Mechatronics Armament Aviation Safety Engineering*, vol. 10, no. 2, pp. 143–150, 2019.

[3] X. Wang, X. Feng, and Z. Xia, "Scene video text tracking based on hybrid deep text detection and layout constraint," *Neurocomputing*, vol. 363, no. 21, pp. 223–235, 2019.

[4] Q. Xu, "Using sensor network in motion detection based on deep full convolutional network model," *Complexity*, vol. 2021, no. 3, 11 pages, Article ID 909522, 2021.

[5] E. Dong, B. Han, H. Jian, J. Tong, and Z. Wang, "Moving target detection based on improved Gaussian mixture model considering camera motion," *Multimedia Tools and Applications*, vol. 79, no. 11, pp. 7005–7020, 2020.

[6] Y. Geng, J. Du, and M. Liang, "Abnormal event detection in tourism video based on salient spatio-temporal features and sparse combination learning," *World Wide Web*, vol. 22, no. 2, pp. 689–715, 2019.

[7] Z. Zhang, W. Li, and Y. Zhang, "Automatic construction and extraction of sports moment feature variables using artificial intelligence," *Complexity*, vol. 2021, no. 2, 13 pages, Article ID 5515357, 2021.

[8] M. B. Azizkhani, J. Kadkhodapour, S. Rastgordani, A. P. Anaraki, and B. Shirkavand Hadavand, "Highly sensitive, stretchable chopped carbon fiber/silicon rubber based sensors for human joint motion detection," *Fibers and Polymers*, vol. 20, no. 1, pp. 35–44, 2019.

[9] R. Liang, H. Zhi, and M. M. Kamruzzaman, "Methods of moving target detection and behavior recognition in intelligent vision monitoring," *Acta Microscopica*, vol. 28, no. 4, pp. 750–759, 2019.

[10] X. Fan, H. Guo, Z. Xu, and B. Li, "Dim and small targets detection in sequence pictures based on spatiotemporal motion characteristics," *Mathematical Problems in Engineering*, vol. 2020, no. 1, 19 pages, Article ID 7164859, 2020.