Scientific Research Publishing

Open Journal of Applied Sciences, 2024, 14, 2610-2627
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

The Fusion of Temporal Sequence with Scene Priori Information in Deep Learning Object Recognition

Yongkang Cao, Fengjun Liu, Xian Wang, Wenyun Wang, Zhaoxin Peng

School of Mechanical Engineering, Hunan University of Science and Technology, Xiangtan, China
Email: liufengjun120@163.com

How to cite this paper: Cao, Y.K., Liu, F.J., Wang, X., Wang, W.Y. and Peng Z.X. (2024) The Fusion of Temporal Sequence with Scene Priori Information in Deep Learning Object Recognition. *Open Journal of Applied Sciences*, **14**, 2610-2627.
https://doi.org/10.4236/ojapps.2024.149172

Received: September 8, 2024
Accepted: September 24, 2024
Published: September 27, 2024

Abstract

For some important object recognition applications such as intelligent robots and unmanned driving, images are collected on a consecutive basis and associated among themselves, besides, the scenes have steady prior features. Yet existing technologies do not take full advantage of this information. In order to take object recognition further than existing algorithms in the above application, an object recognition method that fuses temporal sequence with scene priori information is proposed. This method first employs YOLOv3 as the basic algorithm to recognize objects in single-frame images, then the DeepSort algorithm to establish association among potential objects recognized in images of different moments, and finally the confidence fusion method and temporal boundary processing method designed herein to fuse, at the decision level, temporal sequence information with scene priori information. Experiments using public datasets and self-built industrial scene datasets show that due to the expansion of information sources, the quality of single-frame images has less impact on the recognition results, whereby the object recognition is greatly improved. It is presented herein as a widely applicable framework for the fusion of information under multiple classes. All the object recognition algorithms that output object class, location information and recognition confidence at the same time can be integrated into this information fusion framework to improve performance.

Keywords

Computer Vison, Object Recognition, Deep Learning, Consecutive Scene, Information Fusion

1. Introduction and Motivation

As an important branch of computer vision and artificial intelligence, object

recognition is widely used in intelligent robots [1] [2], industrial detection [3] [4], unmanned driving [5] [6], video surveillance [7] [8] and other fields and a research hotspot these days. The algorithms in the earliest days relied on basic vision, such as grayscale, color, texture, and shape, and traditional machine learning classification methods to recognize objects. In 2001, the VJ detector designed by P. Viola and M. Jones to extract Haar features for face recognition made a breakthrough in face recognition technology [9], marking that higher-order features began to be applied to object recognition. DPM (Deformable Parts Models) proposed by P. Felzenszwalb in 2008 played an important role for a long time back then [10], and then R. Girshick improved it, increasing the processing speed by more than 10 times without any compromise on accuracy [11].

The recognition rate and accuracy have been greatly improved when deep learning, which began to rise in 2012, smashed the limitation where traditional object recognition methods need manual efforts to extract visual features [12] [13]. RCNN (Region Convolutional Neural Network) [14] proposed by R. Girshick in 2014 is such a representative method of deep learning object recognition (DLOR). that works by generating candidate regions from image input, extracting the features of each candidate region via CNN, then sending the features to SVM (Support Vector Machine) classifier for the identification of object class, and finally correcting the position of candidate frame by the regressor, whereas SPPNet (Spatial Pyramid Pooliing Networks) [15] proposed by K. He *et al.* in the same year converts the corresponding part of the feature map and the candidate region into fixed-size features, reducing the loss of useful information. Many iconic methods have been proposed one after another in the year of 2015 that DLOR technologies mushroomed. The Fast RCNN [16] model proposed by R. Girshick based on RCNN has a simplified network structure due to ROI (Region of Interest) in lieu of the pyramid pooling layer of SPPNet and consumes less computing resources thanks to new multi-task loss function. The Faster RCNN [17] model proposed by S. Ren *et al.* replaces ROI with RPN (Region Proposal Network) to further increase its running speed. It is the first end-to-end framework for DLOR. The methods above are all two-stage ones where recognition process undergoes such two-stages as first extracting image features by regions, and then performing object classification and bounding box regression. But one-stage methods only extract features once before they recognize objects at an accuracy slightly lower but a speed greater than those of the two-stage ones. Typical one-stage methods include: SSD [18] (Single Shot MultiBox Detector) proposed by W. Liu *et al.* and YOLO [19] (You Only Look Once) proposed by R. Joseph *et al.* The way YOLO solves object recognition as a regression problem is advantageous in terms of fast speed and low background false positive rate, but disadvantaged in the aspects of precision as well as recall rate of small objects. YOLOv2 [20] and YOLOv3 [21] appeared in response to the shortcomings of YOLO after researchers improved the feature network structure and increased the prediction scale. After years of development, the YOLO series algorithms have gradually become one of the most popular DLOR algorithms and led to a large number of applied research on YOLO

series algorithms. W. Qin *et al.* propose a ship-detection method based on a deep convolutional neural network that is modified from YOLOv3. They added a squeeze-and-excitation (SE) structure to the backbone network to strengthen the ability to extract features. Through a large number of experiments prove this method improves the speed [22]. J. Jun Feng *et al.* propose a real-time performance fabric defect detection method based on the YOLOv3 model which combined with the high-level information and the low-level features. And YOLO detection layer was added to feature maps of different sizes to make it more suitable for defect detection of grey fabric and lattice fabric [23]. The YOLO algorithm is still under continuous development. In some recent studies, data processing, network structure, and loss function are further optimized.

Although great success has been achieved for DLOR algorithms represented by the YOLO series, when some traditional difficulties, such as poor light, misshapen objects, and large-scale changes in object scale, come up in object recognition, guaranteed accuracy and stability are still a challenge to results. The above problems will persist as long as we use only single sources of information and single models of recognition. Therefore, efforts are being made to further improve object recognition by expanding the information sources used for object recognition, as a result, information fusion technology is widely used in object recognition. H. Yang *et al.* propose a method for low-dimensional, strongly robust, and fast space target ISAR image recognition based on local and global structural feature fusion. The method makes up for the missing structural information of the trace feature and ensures the integrity of the ISAR image feature information [24]. Chen *et al.* propose a target-level fusion method for intelligent vehicle target detection based on information collected millimeter-wave (MMW) radar and camera. The experimental results indicate that the proposed algorithm can complete a tracks association between the MMW radar and camera, and the method has an excellent performance [25]. W. Chang *et al.* propose a method of fusing the spectral feature difference image (DI) and textural feature (grey level co-occurrence matrix) DI obtained by change vector analysis (CVA) to improve the accuracy of multi-source remote sensing image building change detection. Experiments results show that this method can significantly improve the change detection performance of multi-source remote sensing image building [26]. These fruitful studies have played an important role in practical applications.

In the existing information fusion research on object recognition, most of the information used therein comes from one or more sensors at the same moment, and most of the methods used therefor belong to information fusion on data level and information fusion on feature level. For some important object recognition applications such as intelligent robots and unmanned driving, images are collected on a consecutive basis and associated among themselves, besides, the scenes have steady prior features. This means that more information is available from these scenes to improve object recognition. Efforts are expected to further expand the research of information fusion in the field of object recognition.

An object recognition method that fuses temporal sequence with scene priori

information is proposed herein for consecutive scenes with steady features. This method first employs YOLOv3 as the basic algorithm to recognize objects in single-frame images, then the DeepSort [27] [28] (simple online and real-time tracking with a deep association metric) algorithm to establish association among potential objects recognized in images of different moments, and finally the confidence fusion method and temporal boundary processing method designed herein to fuse, at the decision level, temporal sequence information with scene priori information. Experimental studies show that the method introduced herein better suppresses the influence of interference factors such as light conditions and object scale changes, and significantly improves the comprehensive performance of object recognition.

The remainder of this paper is organized as follows: Part 2 describes the principle of YOLOv3; Part 3 introduces the information fusion object recognition method proposed herein; Part 4 is devoted to the analysis and discussion of the results of the proposed method applied to the instances. Finally, this paper is concluded in Part 5.

2. YOLO Algorithm for Object Recognition

Although some newer DLOR algorithms have come out recently, the YOLOv3 algorithm with excellent synthetic performance in extensive testing is still one of the mainstream choices. The overall structure of its network is shown in **Figure 1** where Res represents the residual network (ResNet) structure designed to solve the degradation caused by the deepening of the network structure; the CBL module consists of three parts: convolution level, Batch Normalization and Leaky ReLU activation function, effectively solving the problems such as gradient vanishing, gradient explosion, and overfitting that are adverse to training; and Concat operation is used for the fusion of deep and shallow feature maps. When the algorithm works, first, the detection images are resized to 416 pixel × 416 pixel and input to the Darknet-53 backbone network for feature extraction. And, the feature pyramid structure capable of multi-scale detection introduced drawing on the idea of FPN (Feature Pyramid Networks, FPN) extracts 8-time, 16-time and 32-time feature maps from the backbone network respectively, and anchor boxes and non-maximum value suppression are used to output, in the form of detection box, the recognition results of large-, medium- and small-scale objects and the confidence information thereof.

3. Proposed Method

The method proposed herein fuses the object recognition results obtained from multi-frame images consecutively collected by the deep learning model with scene priori information at the decision level, thereby improving the comprehensive performance of object recognition. The new method has an overall technical framework as shown in **Figure 2**, mainly including: single-frame image object recognition, inter-frame object registration, and multi-class information fusion-based correction of recognition results.

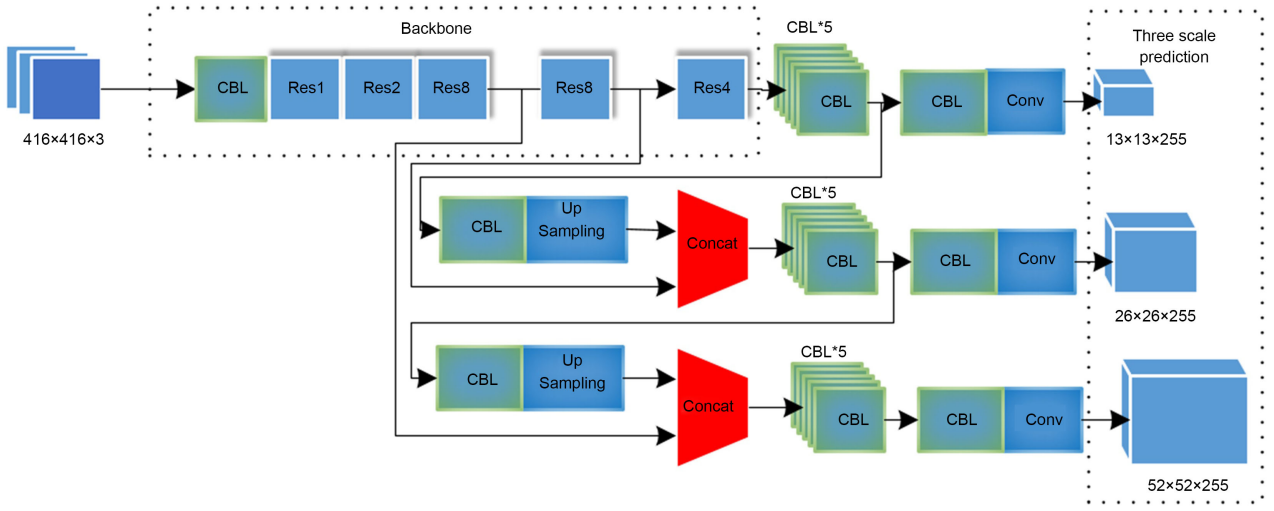


Figure 1. Overall structure of YOLOv3.

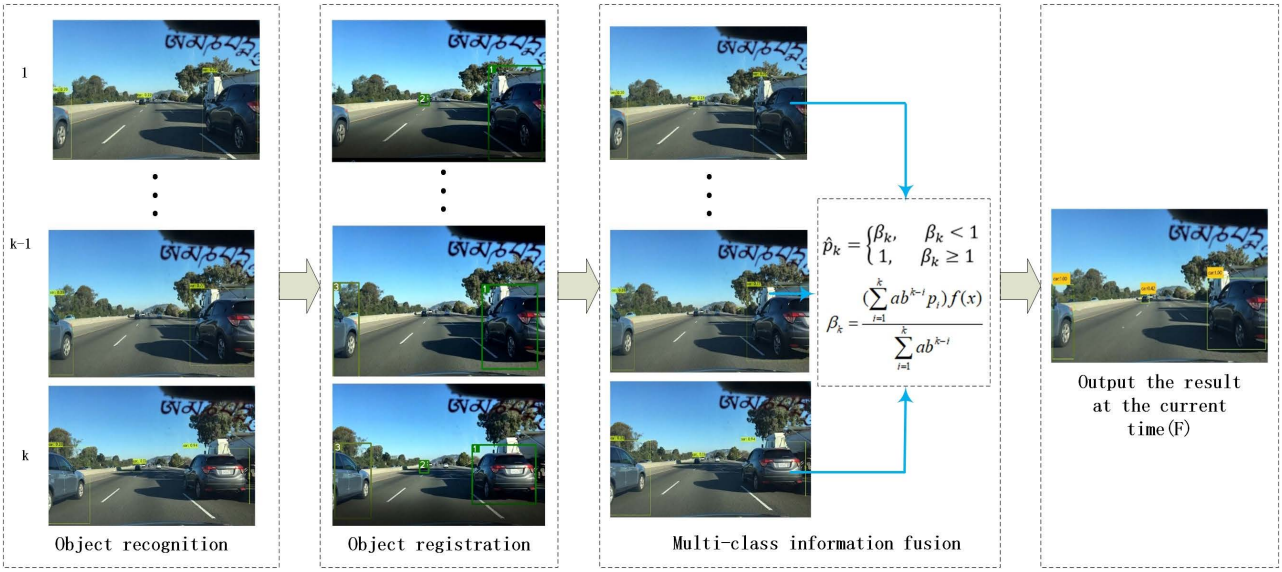


Figure 2. Overall technical framework.

3.1. Single-Frame Image Object Recognition and Inter-Frame Object Registration

The detailed flow of single-frame object recognition and inter-frame registration is shown in Figure 3. Single-frame image object recognition is used to obtain the class, location and confidence information of potential objects. In the technical framework proposed herein, all object recognition algorithms that provide the above information can be used to recognize objects from single-frame images. Considering that the great success achieved by YOLOv3 algorithm in the field of object recognition in recent years, YOLOv3 is used herein to accomplish this step.

The step of inter-frame object registration is taken to establish the association among potential objects recognized at different moments. DeepSort, a technical framework commonly used for object tracking, is employed herein to accomplish

this step. By first resorting to recursive Kalman filtering to predict the positions of all potential objects that have appeared before in the current frame (frame F), then matching the inter-frame objects by considering both motion information and appearance information. The formula for calculating the matching degree of motion information $d^{(1)}(u,v)$ is:

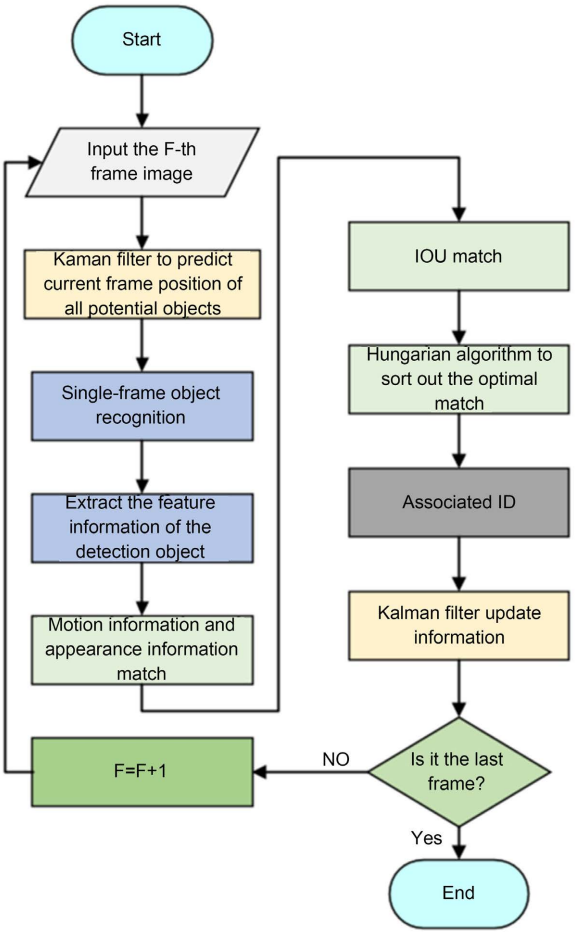


Figure 3. Single-frame object recognition and inter-frame object registration process.

$$d^{(1)}(u,v)=\left(d_u-y_v\right)^T S_{uv}^{-1}\left(d_u-y_v\right) \tag{1}$$

Wherein d_u is the position of detection box of the u-th object identified by YOLOv3 in the image of the F-th frame, and y_v is the very place that the object whose ID is I_v and whose frames have appeared is in the image of the F-th frame as predicted by the recursive Kalman filter algorithm, and S_{uv} is the covariance matrix between d_u and y_v.

The formula for calculating the matching degree $d^{(2)}(u,v)$ of appearance information is:

$$d^{(2)}(u,v)=\min\left\{1-r_u^T r_k^{(v)} \middle| r_k^{(v)} \in R_v\right\} \tag{2}$$

Wherein r_u is the appearance feature vector extracted by d_u, and r_v is the

set of those q times appearance feature vectors of the object whose ID is I_v that have appeared (q ≤ 100; 30 is assigned to it in this paper). DeepSort sets the linear weighted result $C_{u,v}$ from $d^{(1)}(u,v)$ and $d^{(2)}(u,v)$ to be the measure of how well the motion information matches the appearance information:

$$C_{u,v} = \lambda d^{(1)}(u,v) + (1-\lambda) d^{(2)}(u,v) \tag{3}$$

Wherein λ is a hyperparameter that controls the influence of the two metrics on the results (0.01 is assigned to it in this paper). If $C_{u,v}$ falls within the threshold range (0.32 is assigned to it in this paper)where the two match each other, it is considered that the *u*-th potential object identified in the F-th frame is the very potential object with ID I_v that has appeared before. Motion information and appearance information match successfully. The *u*-th potential object is identified after ID I_v is assigned to the F-th frame. Otherwise, a further step should be taken to judge whether the object matches by the indicator Intersection Over Union (IOU) between d_u and y_v.

The object detection and registration process of the method proposed herein is shown in Figure 3. It should be pointed out that yet there is a combinatorial optimization problem in the one-to-one registration of the *v* potential objects that have appeared and the u objects identified in the F-th frame. The technical framework of DeepSort works by the Hungarian algorithm to recursively sort out the optimal match between the v potential objects that have appeared and the u object identified in the F-th frame.

3.2. Fusion of Temporal Sequence Information

Real objects usually appear simultaneously in multi-frame images captured consecutively. The method proposed herein fuses the single-frame detection confidences of the same potential object at multiple moments after object recognition in a single-frame image is registered with inter-frame object, and outputs the corrected new confidences. If after detection and registration, the confidences of object recognition for a potential object inconsecutive *k* frame images are p_1, p_2, …, p_k, respectively, then the confidence $\hat{p}_k$ obtained after information fusion and planning are:

$$\hat{p}_k = \begin{cases} \beta_k, \beta_k < 1 \\ 1, \beta_k \geq 1 \end{cases} \tag{4}$$

Equation (4) ensures that the resulting confidence $\hat{p}_k$ is a constant between [0, 1]. And in the equation, β_k is the calculated value for the multi-frame confidence fusion:

$$\beta_k = \frac{\left(\sum_{i=1}^{k} ab^{k-i} p_i\right) f(x)}{\sum_{i=1}^{k} ab^{k-i}} \tag{5}$$

Wherein a is the single-frame basic contribution; and *b* is the forgetting factor ($b \in (0, 1)$), whose function is to gradually reduce over time the weight of the

DLOR result of a certain frame in the confidence fusion for the subsequent moments. $f(x)$ is a multiplication function, whose calculation formula is:

$$f(x)=\begin{cases}1, & x<z\\ \dfrac{c}{1+2e^{-d(x-z)}}, & x\geq z\end{cases} \tag{6}$$

Wherein the variable x is the proportion of the images where YOLOv3 can detect an object.to a consecutive sequence of images captured since the first appearance of the object. If x is greater than or equal to z, the calculated value β_k of the multi-frame confidence fusion will be greatly increased, otherwise it will stay as it is. c is the multiplication factor, and d, the multiplication adjustment factor. Parameters z, c, and d are valued as appropriate.

3.3. Fusion of Scene Prior Information

It is often that new objects appear in the process of object recognition in consecutive scenes. Let m be the minimum length of temporal sequence for the fusion of temporal sequence information under the method proposed herein (m is valued as appropriate). If the number of image frames collected for a new object from the time-series images in which it appears is less than m, it is obviously impossible to perform time-series information fusion by the method in Section 3.2. This problem happens to provide an interface for the method proposed herein to fuse scene priori information. The specific process to fuse scene priori information is shown in Figure 4, where area A is the reasonable area delineated according to scene priori knowledge where a new object might appear for the first time, and area B is the unreasonable area where it might appear for the first time. The object recognition confidence of (m-k) moments missing from the fusion of temporal sequence information is set to be either the object recognition confidence of the current frame image if the new object appears in area A for the first time, or 0 and if it appears in area B for the first time. After the scene priori information is fused, the confidence for the new object can also be corrected by the method described in Section 3.2. And if the area where a new object appears for the first time does not match the scene priori information, the recognition confidence of the object will be significantly reduced.

For ease of understanding, scenes used in part 4 are taken as examples to illustrate how to the delineate areas based on the scene priori knowledge. As shown in Figure 5, according to common sense, the central area in front of the observer and the areas on the left and right sides are reasonable areas where new object might appear for the first time in this scene and thus should be each delineated as area A, while the other area is the unreasonable area where new objects might appear for the first time and thus should be delineated as area B.

4. Experiments and Analysis

4.1. Test Platform

The configuration of the hardware platform for the object recognition algorithm

in the experiment is: dual Nvidia GTX3080 graphics cards, and dual Intel(R) Xeon(R) Gold 6139M CPUs. The deep learning and information fusion algorithms are implemented using Pytorch (1.8.1) and CUDA (11.1) under the integrated development environment Pycharm (2020.2.3) in the Windows 10 operating system.

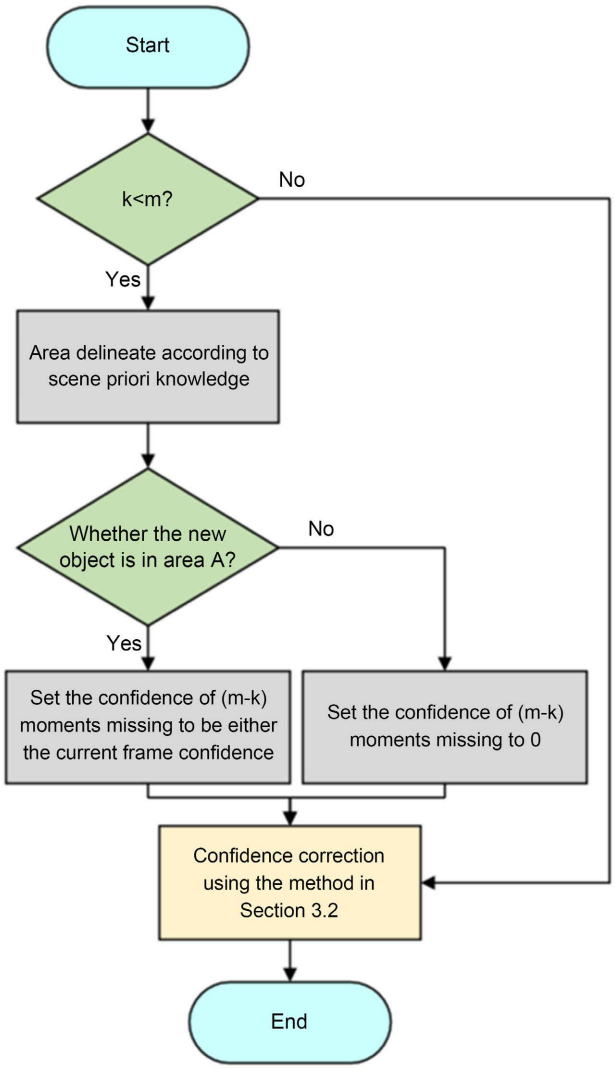


Figure 4. Flow chart of the fusion of scene priori information.

(a) (b)

Figure 5. Area delineation according to priori knowledge. (a) Public dataset scenes; (b) Self-built dataset scene.

4.2. Experiments Using Public Dataset

The information fusion object recognition method proposed herein this paper was first tested on public datasets. During the experiment, 80% of the samples were extracted from the "Urban object detection" dataset [29] for the training and testing of basic DLOR (the samples were divided into training set, validation set and test set in the proportion of 3:1:1). The method proposed herein is applicable to consecutive scenes only, but the "Urban object detection" dataset contains few consecutive scenes. Therefore, in the comparison of recognition accuracy between this method and basic DLOR, in addition to the two consecutive scenes taken from the "Urban object detection" data, four consecutive scenes were also extracted from the "BDD100K" dataset [30]. The scenes used are shown in **Figure 6**, and the number of images, objects and types included in each scene, in **Table 1**. The Areas A and B in each scene are divided according to the example in Section 3.3, and the parameter values for information fusion to be conducted by the method proposed herein are shown in **Table 2**. Given 0.05 and 0.3 as the confidence thresholds to judge whether an object exists, the basic DLOR (YOLOv3) and the method proposed herein are compared and analyzed by the false positive rate and the miss rate, leading to the results shown in **Table 3** and **Table 4**.

(a) (b) (c) (d) (e) (f)

Figure 6. Public dataset scenes used for algorithm performance evaluation (a) Scene 1; (b) Scene 2; (c) Scene 3; (d) Scene 4; (e) Scene 5; (f) Scene 6.

Table 1. Basic information about six consecutive scenes.

Scenes	Number of images	Total objects	Object type
Scene 1	13	43	person/traffic signal
Scene 2	15	58	car/traffic light
Scene 3	20	83	car
Scene 4	20	107	car/person
Scene 5	102	646	car/person/traffic light
Scene 6	202	858	car/person/traffic light

Table 2. Parameter valuing to run the algorithm.

Parameter	a	b	c	d	z	m	L (pixel)	W (pixel)	R (pixel)
Value	1/3	2/3	3	8	0.5	3	475	460	135
	1/3	2/3	5	5	0.5	3	475	460	135

Table 3. False positive rate and miss rate under the two methods with the threshold of 0.05.

Scene	Object Type	Basic DLOR		Method herein	
		False positive rate	Miss rate	False positive rate	Miss rate
Scene 1	person	0	27.2%	0	8.0%
	traffic signal	4.5%	18.8%	0	4.5%
Scene 2	car	5.4%	17.6%	4.1%	6.7%
	traffic light	5.3%	39.5%	2.6%	23.7%
Scene 3	car	2.5%	4.8%	1.2%	0
Scene 4	car	7.7%	12.3%	3.2%	5.9%
	person	12.0%	23.0%	4.0%	7.0%
Scene 5	car	10.1%	27.7%	9.8%	22.8%
	person	4.1%	16.3%	4.1%	0
	traffic light	3.4%	18.8%	2.9%	15.6%
Scene 6	car	3.4%	3.8%	3.0%	1.3%
	traffic light	4.7%	2.6%	4.7%	2.0%
	person	15.9%	7.5%	13.6%	3.0%

It can be seen from Tables 1-4 that when the confidence threshold is set to 0.05, the performance of the method proposed herein is significantly better than that of basic DLOR in all the six scenes, to be specific, the false positive rate of various objects in each scene under the former algorithm is 1.98% lower than that under basic DLOR on average, and the miss rate, 9.18% lower than that under the latter. The object recognition of person in scene 4 by the method proposed herein shows the most significant performance improvement, as suggested by the miss rate and false positive rate reduced by 16% and 8%, respectively.

Table 4. False positive rate and miss rate under the two methods with the threshold of 0.3.

Scene	Object Type	Basic DLOR		Method herein	
		False positive rate	Miss rate	False positive rate	Miss rate
Scene 1	person	0	35.6%	0	24.7%
	traffic signal	0	40.9%	0	27.7%
Scene 2	car	0	27.0%	0	18.9%
	traffic light	0	42.1%	0	29.4%
Scene 3	car	0	29.5 %	0	21.8%
Scene 4	car	5.1%	23.0%	3.2%	17.6%
	person	0	45.0%	0	37.5%
Scene 5	car	0	49.5%	0	42.1%
	person	0	51.0%	0	39.9%
	traffic light	0	25.0%	0	20.6%
Scene 6	car	0.6%	19.0%	0.2%	13.1%
	person	0	26.2%	0	18.8%
	traffic light	0	51.5%	0	40.8%

When the confidence threshold is set to 0.3, the missed detection rate performance of the method proposed herein is also significantly better than that of basic DLOR in all the six scenes, to be specific, the miss rate of various objects in each scene under the former algorithm is 8.6% lower than that under basic DLOR on average. The object recognition of traffic signal in scene 1 by the method proposed herein shows the most significant performance improvement, as suggested by the miss rate reduced by 13.2%. Regardless of whether the confidence threshold is 0.05 or 0.3, the two performance evaluation indicators in relation with the method proposed herein are not lower than those in relation with basic DLOR in all the scenes. The method proposed is a great improvement on the object recognition performance of the basic DLOR.

Figure 7(a) shows the object recognition result obtained by basic DLOR under the condition of threshold 0.05 in the seventh frame image in scene 2, and **Figure 7(b)**, the recognition result obtained by the method proposed herein under the same condition. It can be seen from the figures that due to the poor light conditions of the images, the Basic DLOR mistakenly recognizes a one-person object on the top of the car in the right of center of the image, which false positive recognition can be avoided by the method proposed herein.

Figure 8(a) shows the object recognition result obtained by Basic DLOR under the condition of threshold 0.3 in the fourth frame image in scene 4, and **Figure 8(b)**, the recognition result obtained by the method proposed herein under the same condition. It can be seen from the figures that due to the far distance, the Basic DLOR fails to recognize the car and the traffic signal in the middle of the image, which miss can be avoided by the method proposed herein.

(a) (b)

Figure 7. Recognition results obtained by Basic DLOR and the method proposed herein in the seventh frame of scene 2. (a) Basic DLOR; (b) The method proposed herein.

(a) (b)

Figure 8. Recognition results obtained by Basic DLOR and the method proposed herein in the fourth frame of scene 4. (a) Basic DLOR; (b) The method proposed herein.

4.3. Experiments Using Self-Built Dataset

Intelligent robot environment perception is one of the important applications of object recognition. In order to further verify the practicability of the method proposed, 1961 images from two production workshops were collected for this paper by reference to the working environment of industrial robots to build an industrial scene dataset for experiments. The dataset includes such two types of objects as machine tool and person. During the experiment, 90% of the samples were taken from the self-built industrial dataset for the training of the DLOR model and validation (the samples are split into the training set, the validation set and the test set in the proportion of 8:1:1), and two consecutive scenes with a large number of images were selected from the remaining images for the purpose of algorithm performance evaluation. The two scenes are shown in **Figure 9**, among which, the first one includes 45 images and a total of 202 objects, and the second one, 45 images and a total of 189 objects.

Considering that in the images, some of the objects in this dataset were larger than those in the public dataset used in Section 4.2, the parameters W and R of the information fusion process in this experiment were increased to 540 and 155 respectively, while other parameters were the same as those in Section 4.2. After many attempts, it was found that both algorithms performed well when the confidence threshold was set to 0.3 on the dataset, hence this confidence threshold. The performance of Basic DLOR and our method is shown in **Table 5**.

It can be seen from the table that there was no false positive in the two methods. This is because, limited by the research conditions, the appearance of the objects from the self-built dataset is less diversified than that from public dataset, and the

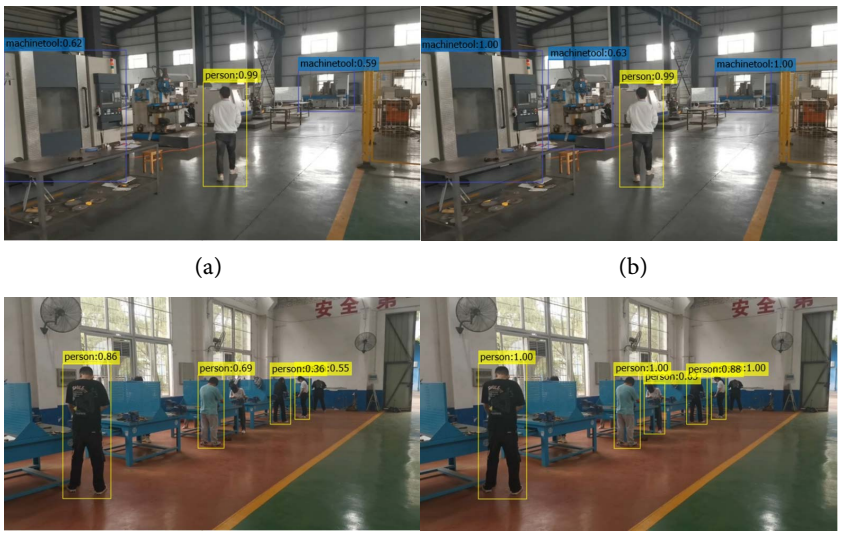
difference between the objects in the training samples and the objects in the performance evaluation images is relatively small. The miss rate of our method is 7.1% lower on average than that of Basic DLOR. Figure 10(a) and Figure 10(b) show an example of reduced object miss rate obtained by our method in scene one, and Figure 10(c) and Figure 10(d), scene two. Based on the experimental results in Section 4.2 and this Section, it is concluded that it is because of the expansion of information sources that the method proposed herein greatly improves the performance of object recognition algorithm, and that the quality of a single frame image becomes less influential on the recognition result.

(a) (b)

Figure 9. Self-built dataset scenes used for algorithm performance evaluation. (a) Scene 1; (b) Scene 2.

Table 5. False positive rate and miss rate under the two methods.

Scene	Object Type	YOLOv3		Method herein	
		False positive rate	Miss rate	False positive rate	Miss rate
Scene 1	machine tool	0	18.2%	0	11.2%
Scene 2	person	0	21.7%	0	14.5%



(a) (b)

(c) (d)

Figure 10. Recognition results obtained by Basic DLOR and the method proposed herein in industrial scenes. (a) The recognition results of 23rd frame in scene 1 under Basic DLOR; (b) The recognition results of 23rd frame in scene 1 under the method proposed herein; (c) The recognition results of 17th frame in scene 2 under Basic DLOR; (d) The recognition results of 17th frame in scene 2 under the method proposed herein.

5. Conclusions

As an important branch of computer vision and artificial intelligence, object recognition is widely applied in engineering practice. In recent years, DLOR algorithms have achieved great success. However, recognition results obtained in reliance on single information source and recognition model only are yet unsatisfactory when some traditional difficulties come up in object recognition, such as poor light conditions, misshapen objects, and large-scale changes in object scale. Although information fusion technology has been widely used to further improve the performance of object recognition algorithms, most of the information used in information fusion research on object recognition comes from one or more sensors at the same moment, and most of the methods used belong to information fusion on such data level and information fusion on such feature level that useful information are not fully utilized, therefore, the research on information fusion technology in the field of object recognition needs to be further expanded.

For some important object recognition applications such as intelligent robots and unmanned driving, images are collected on a consecutive basis and associated among themselves, besides, the scenes have steady prior features. Existing technologies does not take full advantage of this information. In order to take object recognition further than existing algorithms in the above application, an object recognition method that fuses temporal sequence with scene priori information is proposed herein. This method first employs YOLOv3 as the basic algorithm to recognize objects in single-frame images, then the DeepSort algorithm to establish association among potential objects recognized in images of different moments, and finally the confidence fusion method and temporal boundary processing method designed herein to fuse, at the decision level, temporal sequence information with scene priori information. Experiments using public datasets and self-built industrial scene datasets show that due to the expansion of information sources, the quality of single-frame image has a less impact on the recognition results, whereby the object recognition is greatly improved.

It should be noted that it is a widely applicable framework that is put forward herein for multi-class information fusion. In addition to YOLOv3 used herein, the algorithms that output object class, location information and recognition confidence at the same time can all be integrated into this information fusion framework to improve performance.

Acknowledgements

This work was supported the National Natural Science Foundation of Hunan Province (Grant No. 2018JJ3167).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Hatano, M. and Fujii, T. (2019) 3-D Shape Recognitions of Target Objects for Stacked Rubble Withdrawal Works Performed by Rescue Robots. *Artificial Life and Robotics*, **25**, 94-99. https://doi.org/10.1007/s10015-019-00566-6

[2] Cai, H. and Mostofi, Y. (2021) Exploiting Object Similarity for Robotic Visual Recognition. *IEEE Transactions on Robotics*, **37**, 16-33. https://doi.org/10.1109/tro.2020.3005531

[3] Tan, D., Chen, L., Jiang, C., Zhong, W., Du, W., Qian, F., *et al.* (2021) A Circular Target Feature Detection Framework Based on DCNN for Industrial Applications. *IEEE Transactions on Industrial Informatics*, **17**, 3303-3313. https://doi.org/10.1109/tii.2020.3024578

[4] Yang, J., Xi, M., Jiang, B., Man, J., Meng, Q. and Li, B. (2021) FADN: Fully Connected Attitude Detection Network Based on Industrial Video. *IEEE Transactions on Industrial Informatics*, **17**, 2011-2020. https://doi.org/10.1109/tii.2020.2984370

[5] Lang, Y. and Yuan, B. (2021) RETRACTED: Algorithm Application Based on the Infrared Image in Unmanned Ship Target Image Recognition. *Microprocessors and Microsystems*, **80**, Article ID: 103554. https://doi.org/10.1016/j.micpro.2020.103554

[6] Topple, J.M. and Fawcett, J.A. (2021) MiNet: Efficient Deep Learning Automatic Target Recognition for Small Autonomous Vehicles. *IEEE Geoscience and Remote Sensing Letters*, **18**, 1014-1018. https://doi.org/10.1109/lgrs.2020.2993652

[7] Felici-Castell, S., García-Pineda, M., Segura-Garcia, J., Fayos-Jordan, R. and Lopez-Ballester, J. (2021) Adaptive Live Video Streaming on Low-Cost Wireless Multihop Networks for Road Traffic Surveillance in Smart Cities. *Future Generation Computer Systems*, **115**, 741-755. https://doi.org/10.1016/j.future.2020.10.010

[8] Yin, L. and He, R. (2021) RETRACTED: Target State Recognition of Basketball Players Based on Video Image Detection and FPGA. *Microprocessors and Microsystems*, **80**, Article ID: 103340. https://doi.org/10.1016/j.micpro.2020.103340

[9] Viola, P. and Jones, M.J. (2004) Robust Real-Time Face Detection. *International Journal of Computer Vision*, **57**, 137-154. https://doi.org/10.1023/b:visi.0000013087.49260.fb

[10] Felzenszwalb, P., McAllester, D. and Ramanan, D. (2008) A Discriminatively Trained, Multiscale, Deformable Part Model. 2008 *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, 23-28 June 2008, 1-8. https://doi.org/10.1109/cvpr.2008.4587597

[11] Felzenszwalb, P.F., Girshick, R.B. and McAllester, D. (2010) Cascade Object Detection with Deformable Part Models. 2010 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, 13-18 June 2010, 2241-2248. https://doi.org/10.1109/cvpr.2010.5539906

[12] Lee, S., Lee, H., Back, J., An, K., Yoon, Y., Yum, K., *et al.* (2021) Prediction of Tire Pattern Noise in Early Design Stage Based on Convolutional Neural Network. *Applied Acoustics*, **172**, Article ID: 107617. https://doi.org/10.1016/j.apacoust.2020.107617

[13] D'Angelo, G. and Palmieri, F. (2021) Network Traffic Classification Using Deep Convolutional Recurrent Autoencoder Neural Networks for Spatial-Temporal Features Extraction. *Journal of Network and Computer Applications*, **173**, Article ID: 102890. https://doi.org/10.1016/j.jnca.2020.102890

[14] Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2016) Region-based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, **38**, 142-158.
https://doi.org/10.1109/tpami.2015.2437384

[15] He, K., Zhang, X., Ren, S. and Sun, J. (2015) Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**, 1904-1916.
https://doi.org/10.1109/tpami.2015.2389824

[16] Girshick, R. (2015) Fast R-CNN. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 1440-1448.
https://doi.org/10.1109/iccv.2015.169

[17] Ren, S., He, K., Girshick, R. and Sun, J. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149.
https://doi.org/10.1109/tpami.2016.2577031

[18] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., *et al.* (2016) SSD: Single Shot Multibox Detector. In: Leibe, B., Matas, J., Sebe, N., andWelling, M., Eds., *Computer Vision—ECCV* 2016, Springer, 21-37.
https://doi.org/10.1007/978-3-319-46448-0_2

[19] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 779-788.
https://doi.org/10.1109/cvpr.2016.91

[20] Redmon, J. and Farhadi, A. (2017) YOLO9000: Better, Faster, Stronger. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6517-6525. https://doi.org/10.1109/cvpr.2017.690

[21] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement. arXiv: 1804.02767.

[22] Wang, Q., Shen, F., Cheng, L., Jiang, J., He, G., Sheng, W., *et al.* (2020) Ship Detection Based on Fused Features and Rebuilt YOLOv3 Networks in Optical Remote-Sensing Images. *International Journal of Remote Sensing*, **42**, 520-536.
https://doi.org/10.1080/01431161.2020.1811422

[23] Jing, J., Zhuo, D., Zhang, H., Liang, Y. and Zheng, M. (2020) Fabric Defect Detection Using the Improved YOLOv3 Model. *Journal of Engineered Fibers and Fabrics*, **15**.
https://doi.org/10.1177/1558925020908268

[24] Yang, H., Zhang, Y. and Ding, W. (2020) A Fast Recognition Method for Space Targets in ISAR Images Based on Local and Global Structural Fusion Features with Lower Dimensions. *International Journal of Aerospace Engineering*, **2020**, Article ID: 3412582. https://doi.org/10.1155/2020/3412582

[25] Chen, B., Pei, X. and Chen, Z. (2019) Research on Target Detection Based on Distributed Track Fusion for Intelligent Vehicles. *Sensors*, **20**, Article 56.
https://doi.org/10.3390/s20010056

[26] Wang, C. and Wang, X. (2020) Building Change Detection from Multi-Source Remote Sensing Images Based on Multi-Feature Fusion and Extreme Learning Machine. *International Journal of Remote Sensing*, **42**, 2246-2257.
https://doi.org/10.1080/2150704x.2020.1805134

[27] Veeramani, B., Raymond, J.W. and Chanda, P. (2018) DeepSort: Deep Convolutional Networks for Sorting Haploid Maize Seeds. *BMC Bioinformatics*, **19**, Article No. 289.
https://doi.org/10.1186/s12859-018-2267-2

[28] Doan, T. and Truong, M. (2020) Real-Time Vehicle Detection and Counting Based

on YOLO and DeepSort. 2020 12*th International Conference on Knowledge and Systems Engineering* (*KSE*), Can Tho, 12-14 November 2020, 67-72. https://doi.org/10.1109/kse50997.2020.9287483

[29] Dominguez-Sanchez, A., Orts-Escolano, S., Garcia-Rodriguez, J. and Cazorla, M. (2018) A New Dataset and Performance Evaluation of a Region-Based CNN for Urban Object Detection. 2018 *International Joint Conference on Neural Networks* (*IJCNN*), Rio de Janeiro, 8-13 July 2018, 1-8. https://doi.org/10.1109/ijcnn.2018.8489478

[30] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., *et al.* (2020) BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 2633-2642. https://doi.org/10.1109/cvpr42600.2020.00271