MATHEMATICS AND COMPUTATIONAL SCIENCES

Vol 4(2), 2023, PP:33-45
DOI:10.30511/MCS.2023.2001580.1118
ISSN:2717-2708
http://mcs.qut.ac.ir/

Qom University of Technology

Improving augmented reality with the help of deep learning methods in the tourism industry

Mehdi Jabbari *[1], Maryam Amini [2], Hossein Malekinezhad [3], Zeynab Berahmand [4]

Abstract: From an economic point of view, the tourism industry has a special place. Especially in the -single product economy of Iran, it can be used as the best and most optimal alternative to oil. Augmented reality technology is one of the world's newest and most up-to-date applied technologies, highly regarded today. This research focuses on augmented reality and its patterns. This research aims to investigate and develop a practical pattern identification of augmented reality (ar) and its tracking in the tourism industry. Designs are provided by capturing the position and orientation of the device and its location using sensors and Computer vision with screen technology (augmented reality guide). A guide is designed, implemented, and evaluated as an augmented reality application on a mobile phone. The proposed solution has been using deep learning in marker identification. The deep learning architecture used is Yolo, and the proposed method's marker identification results have an accuracy of 68.73 maps.

Keywords: Deep learning, Augmented Reality, tourism industry, Deep learning network
2020 Mathematics Subject Classification: 6800X
Receive: 04 May 2023, **Accepted:** 31 May 2023

1 Introduction

From an economic point of view, the tourism industry has a special place. Especially in Iran's single-product economy, it can be used as the best and most optimal alternative to oil. At the same time, the tourism industry faces significant challenges in terms of welcoming and providing services to tourists entering the country. Currently, tourism in Iran is active by relying on traditional methods, human guides, and audio guides (in international languages). However, the lack of specialized and operational human resources on the one hand and the immense skill and training gap of these forces with international standards on the other hand in welcoming foreign tourists have created severe problems in the direction of the development of this industry in Iran. Ar technology is one of the world's newest and most up-to-date applied technologies, highly regarded today. This technology aims to create a diverse application system to guide tourists so that we can partially solve

[1]*Corresponding author: Department of computer science, Qom University of Technology, Qom. Iran, Email: jabbari@qut.ac.ir
[2]Department of computer science, Islamic Azad University, Naragh Branch, Iran
[3]Department of computer science, Islamic Azad University, Naragh Branch, Iran
[4]Department of Industrial Engineering, University of Qom, Qom, Iran

the gaps in this field. From a fundamental point of view, ar, abbreviated as AR, deals with the simple combination of natural and virtual (computer-based) worlds. Ar technology adds layers of digital information to an authentic subject, the image recorded on a video or camera. Ar integrates digital data and the user's physical environment in real-time. Unlike virtual reality, which creates a general artificial environment, ar works by using the physical environment around the user and overlaying new information on it. Ar is more social and easier to interact. For the first time, Boeing researcher Thomas Caudle proposed the term "augmented reality "in 1990. Ar has several applications: entertainment, education, medicine, travel, military, art and support, archeology, and navigation.
One of ar technology's first commercial applications was the yellow line displayed at the bottom of the TV screen when televised football games began in 1998. Nevertheless, this technology is used in many industries today, including health and treatment, public security, oil and gas, tourism, and marketing. The possibilities that are (AR) can add to the tourism industry are vast. Since humans perceive the world around them through their senses, the simultaneous combination of the natural world and virtual images forms the concept of ar. In ar, the original value of the existing reality is preserved, and its sensitive information is covered by computer production. Ar in the tourism industry can appear as a practical application.
The purpose of using technology is to create a tourism guide system. Ar in this industry starts much potential to enhance travelers' experiences by providing helpful travel information, mapping, and translation through an application on tourists' mobile phones. This research focuses on art and its patterns. This research aims to investigate and develop a practical pattern identification of ar (AR) and its tracking in the tourism industry. Designs are provided by capturing the position and orientation of the device and its location, using sensors and machine vision with screen technology (Ar Guide) or AR Guide in short. AR Guide is designed, implemented, and evaluated as an ar application on a mobile phone. In this research, the focus is on marker identification, and the proposed solution is marker identification using deep learning. The deep learning architecture used is Yolo, and the proposed method's marker identification results have an accuracy of 68.73 mAP.
The paper is organized as follows. In section 2, we review the literature on augmented reality. Section 3 includes the Architecture and Advantages of Convolutional Neural Networks (CNN). In Section 4, the Solution method is based on mathematics models. Section 5, Computational research, includes the mathematical programming methods, followed by knowledge. Conclusions are given in Section 6, followed by references.

2 Research literature

Ar has a rich history of 50 years of research and development in software. As shown by Gartner in the figure below, ar is currently in research stagnation; however, it is expected to progress in the next 5 to 10 years. In the coming years, according to research (Billinghurst, 2015), ar will develop in three primary technology areas: display, interaction, and tracking.
To understand and obtain a complete image, we should not only focus on classifying different images but also try to accurately estimate the concepts and locations of objects in each image. This work is called object detection [9], which usually includes various activities such as face detection [33], pedestrian detection [6], and skeleton detection [17]. One of the most fundamental problems of computer vision is whether object recognition can provide meaningful, valuable information for understanding images and videos that are related to many applications, including image classification [14], [19], human behavior analysis [1], face recognition [37] and self-driving [2], [4].
Meanwhile, it develops the field using neural networks, related learning systems, and advanced neural network algorithms. Also, it significantly impacts object recognition techniques, which can be considered a learning system [7]-[32]. However, due to the many differences in opinions and views, gestures, and lighting conditions,

it is challenging to implement object detection and positioning fully. In recent years, much attention has been paid to this field [10]-[36].

The definition problem is object recognition, positioning, and image classification. Therefore, the traditional object recognition model has three stages: information area selection, feature extraction, and variety. Hsu et al. [12] proposed defect inspection of indoor components in buildings using deep learning object detection and augmented reality. Liu et al. [22] developed an intelligent predictive maintenance approach with deep learning and augmented reality for machine tools in IoT-enabled manufacturing.

Selection of information area: It appears for different objects in each image position and is based on different angles and sizes. It is normal to generate multiple windows of the thing to scan the entire image. Although this comprehensive strategy can find objects in all possible situations, its deficiency is also apparent. The number of candidate windows is computationally expensive, and many are overproduced. However, if only a fixed number of sliding window templates are applied, areas of dissatisfaction may be produced. Li et al. [21] proposed integrated registration and occlusion handling based on deep learning for augmented-reality-assisted assembly instruction. Puri et al. [28] studied blockchain propels tourism industry—an attempt to explore topics and information in smart tourism management through machine learning. Kontogianni et al [18] recognise the already exploited AI approaches in this field, as well as ways of utilising technology to resume travel and reboot tourism worldwide safely. Murugan et al. [26] proposed Autonomous Vehicle Assisted by Heads up Display (HUD) with Augmented Reality Based on Machine Learning Techniques.

Feature extraction: To recognize different objects, we need to extract visual features that can provide meaningful and robust redevelopment. The parts are SIFT [30], HOG [24], and Haar-Like [5]. These features can generate complex cell-related drawings in the human brain [30]. However, due to the variety of appearances, lighting conditions, and backgrounds, it is not easy to manually design a robust descriptor to describe different objects fully.

Classification: The classifier must distinguish a target from all other varieties and build more hierarchical, meaningful, and informative redevelopment for visual recognition. Usually, Support Vector Machines (SVM), AdaBoost, and Deformable Part-Based Models (DPM) are good choices. DPM is a flexible model combining components with higher deformation costs to achieve more changes among these categories. DPM, with the help of a graphical model, carefully designed and inspired by systematic features and combined analysis, differential learning allows graphical models to be used to build high-accuracy-based models of various classifications.

Thanks to deep neural networks (DNNs), features of CNN, and range (R-CNN), it is more noticeable that DNNs or CNNs are utterly different from traditional approaches. They have deeper architectures that are capable of learning more complex features. Also, robust training algorithms allow learning manual parts without designing the training object redevelopment [29].

To a systematic review to summarize the model and its various features in several application areas, including general object recognition [10], [38], [36], salient object recognition [8], [20], face recognition [23, 35] and pedestrian detection [15, 3]. Whose relationships are shown in (Figure 2-8). Based on the original CNN methods, general object detection is achieved by bounding box regression, while salient object detection is performed by local contrast enhancement and pixel-level segmentation. Luo [25] proposed Question Text Classification Method of Tourism Based on Deep Learning Model.

CNN is the most redevelopment deep learning model [27]. A typical CNN architecture is named VGG16. Each CNN layer is known as a feature. The input layer feature is a 3D matrix of pixel intensities for different color channels (e.g., RGB).

Deep models can be called neural networks with deep structures. The history of neural networks dates back to the 1940s [27], the main goal of which was to simulate the human brain system in principle to solve general learning problems. It became popular in the 1980s and 1990s with Hinton, and his colleagues proposed the back-propagation algorithm (re-propagation). [31] Nevertheless, due to the overuse of training, the lack of extensive training data, limited computing power, and poor performance in comparison, Along with other machine learning tools, neural networks fell out of favor in the early 2000s. Deep learning has become popular since 2006 with the progress in speech recognition, and the improvement of deep understanding can be divided into the following factors.

- ✓ The emergence of large-scale training datasets such as ImageNet fully demonstrates the ability of massive learning.
- ✓ The rapid development of high-performance parallel computing systems, such as GPU clusters
- ✓ Significant advances in the design of network structures and educational strategies. Using Auto-Encoder (AE) or Restricted Boltzmann Machine (RBM) with unsupervised and layered prefetching is well-developed. By editing information, overload problems in training have been reduced [19], [11].

By using Bunch Normalization (BN), the training of intense neural networks becomes very effective. [13] Meanwhile, various network structures, such as AlexNet, Overfeat, GoogLeNet, VGG, and ResNet, have been studied to improve performance.

Continuous efforts have shown that deep learning has brought revolutionary progress on enormous challenges rather than dramatic improvements on small datasets. Its success comes from training a prominent CNN on 1.2 million images with several techniques. (For example, ReLU operations that are illegal to set).

3 Architecture and Advantages of CNN

CNN is the most redevelopment deep learning model [20]. A typical CNN architecture is named VGG16. Each CNN layer is known as a feature. The input layer feature is a 3D matrix of pixel intensities for different color channels (e.g., RGB).

The feature of each internal layer of an image is a few inductive channels whose pixels can be viewed as a unique feature. Each neuron is connected to a small part of the previous layer's neighboring neurons (receptive field). A variety of changes can occur in features such as filtering and aggregation. The filtering operation (convolution) connects a filter matrix (statistical weight) with the values of neighboring neurons and takes a non-linear function such as sigmoid, ReLU to reach the final responses (pooling operations, such as maximum sum, average sum, L2-pooling, and local contrast normalization [34]), summarizes the responses of a field into a single value to produce more full descriptions.

An initial feature hierarchy is established by the interplay between convolution and summation, which can be super visibly tuned by adding multiple fully connected (FC) layers to adapt to different visual tasks. The final layer with other active functions is added to provide a specific probability for each output neuron according to the charges. Moreover, the whole network can be optimized based on an objective function (for example, squared error or cross-entropy loss) using stochastic gradient descent (SGD). A typical VGG16 has 13 convolutions (conv) layers, three fully connected layers, three max-pooling layers, and a softmax classification layer. Conv feature maps are generated by applying 3 × 3 filter windows and downscaling with two max-pooling layers.

An arbitrary test image of the same size as the trained network can process the training samples. If different sizes are developed, rescaling or cropping operations may be required [19].

The advantages of CNN over traditional methods can be summarized as follows:

- ✓ Hierarchical feature redevelopment, which is multi-level redevelopment from pixel to high-level semantic features learned by a multi-stage, hierarchical structure [10], [16], can be learned from the data automatically, and the hidden factors of the input data can be Separated through several non-linear levels.
- ✓ The Deep architecture shows a higher expressive capacity than traditional shallow models.
- ✓ The CNN architecture provides a shared opportunity to optimize several related tasks together (for example, RCNN combines classification and bounded box regression in a fast multi-objective method).
- ✓ By taking advantage of the extensive learning capacity of deep CNNs, some classical computer vision challenges can be solved as information transmission problems with new dimensions and from different perspectives.
- ✓ Due to these advantages, CNN is widely used in many research fields, such as high-resolution image reconstruction, image classification, face recognition image retrieval, pedestrian detection, and video analysis.

4 Solution method

This work aims to recognize the object in the video using the texture information from the adjacent video frames. This work was done in two stages.

First step: We train a dummy label, that is, a convolutional neural network, for object recognition. Video training is taught on video frames separately. The exact YOLO network finder setup is shown in Figure 1, which was first trained on the PASCAL VOC 20 class dataset, then trained on the YouTube video dataset. When fine-tuning to 10 subsets in video datasets, the goal is to minimize the weighted squared detection, similar to the value set in YOLO. If the tuning is okay, here we only learn more parameters. We keep fully connected layers, 24 convolution layers, and four pooling layers unchanged. This tutorial generates approximately 50 cycles for synchronization, using the RMSProp optimizer with a speed of 0.9 and a small batch size of 128. With YOLO, tuning is fine. The dummy tag takes 224×224 frames as input and returns the types of categories and possible locations of objects on each non-overlapping S×S grid cell. The model provides a conditional probability output class, such as bounded box B, with communication certainty for each grid cell. As in YOLO, we consider a corresponding secret box for a grid cell as one of the B-boxes to predict the actual shared area and terrain through the IoU Association. During training, we simultaneously optimize the classification error. We minimize the localization error for each grid cell for the corresponding secret box concerning the natural ground only when an object appears. In the second step, we train a neural network (RNN) with recurrent units (GRU). This network serves as the input to the dummy tag sequences, a target optimization that encourages both accuracies on the target frame and consistency across successive frames. Given sets of dummy labels $X^{(1)}. \ldots . X^{(T)}$, we train the RNN to generate advanced predictions. $\hat{y}^{(1)}. \ldots . \hat{y}^{(T)}$ Concerning the natural ground $y^{(T)}$is available only at the final stage in each sequence. Here, t redevelopes the steps of the series, and T redevelopes the length of the line. We use a fully connected layer with a linear activation function as output since it is a regression problem. In the final experiments, we use a two-layer GRU with 150 nodes in each layer. The above parameters are based on the validation data.

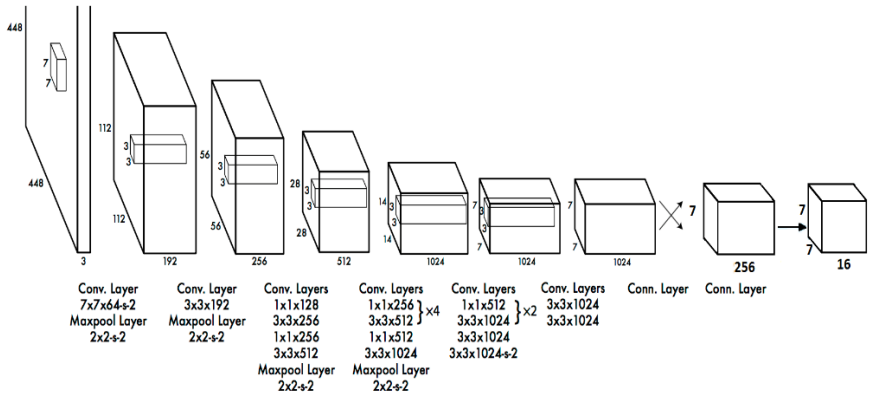

Figure 1: Versus the proposed network. The value of output channel 16 corresponds to 1 class-conditional and 1 + 4 votes and coordinates for each box B = 3.

The following equations define the default through a GRU layer where $h_l^{(t)}$redevelopes the layer's output at the current time step. $h_{l-1}^{(t)}$shows the work of the previous layer at the same stage of the sequence. Equations 1 shows the defining the default through the GRU layer.

$$r_l^{(t)} = \sigma\left(h_{l-1}^{(t)} w_i^{xr} + h_l^{(t-1)} w_l^{hr} + b_l^r\right) \tag{4-1}$$

$$u_l^{(t)} = \sigma\left(h_{l-1}^{(t)} w_i^{xu} + h_l^{(t-1)} w_l^{hu} + b_l^u\right) \tag{4-2}$$

$$c_l^{(t)} = \sigma\left(h_{l-1}^{(t)} w_i^{xc} + r_t \odot (h_l^{(t-1)} w_l^{hc}) + b_l^c\right) \tag{4-3}$$

$$h_l^{(t)} = \sigma\left(1 - u_l^{(t)}\right) \odot h_t^{(t-1)} + u_l^{(t)} \odot c_l^{(t)} \tag{4-4}$$

Here σ denotes a matrix logistic function, and ⊙ is the (matrix) product. The candidates' reset gate, update gate, and hidden state are redeveloped by r, u, and c, respectively. For S=7 and B=2, the dummy label $X^{(T)}$and the prediction $\hat{y}^{(t)}$are both at R^{1470}.

We design an objective function (Relation 2) that is calculated for each precision in the target frame and consistency of predictions in the time steps of the sequence by the following methods. Equation 2 shows the objective function

$$loss = d\_loss + \alpha.s\_loss + \beta.c\_loss + \gamma.pc\_loss \tag{4-5}$$

d_loss, s_loss, c_loss, pc_loss related to detection loss, similarity loss, category loss and prediction_consistency_loss are explained in the following sections. Values of the above parameters α=0.2, β=0.2, and γ=0.1 are selected based on the recognition performance in the validation set. The training is

synchronized in 80 cycles to update the parameters using RMSProp and a driving force of 0.9. During training, we use a small batch of 128 and a sequence of 30.

In the final output, where the actual terrain classification and localization are available, we describe a multipart object detection_loss according to YOLO. Equation 3 the shows function $\boldsymbol{detection_{loss}}$:

$$
\begin{aligned}
detection_{loss} = \lambda_{coord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left(X_i^{(T)} - \hat{X}_i^{(T)}\right)^2 + \left(y_i^{(T)} - \hat{y}_i^{(T)}\right)^2 \\
+ \lambda_{coord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left(\sqrt{w}_i^{(T)} - \sqrt{\widehat{w}}_i^{(T)}\right)^2 + \left(\sqrt{h}_i^{(T)} - \sqrt{\hat{h}}_i^{(T)}\right)^2 \\
+ \sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left(C_i^{(T)} - \hat{C}_i^{(T)}\right)^2 \\
+ \lambda_{noobj} \sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{ij}^{noobj} \left(C_i^{(T)} - \hat{C}_i^{(T)}\right)^2 \\
+ \sum_{i=0}^{s^2} 1_i^{obj} \sum_{c \epsilon classes} \left(p_i^{(T)}(c) - \hat{p}_i^{(T)}(c)\right)^2
\end{aligned}
\tag{4-6}
$$

1_{ij}^{obj} indicates if the object appears in the cell I, and 1_{ij}^{obj} suggests that this prediction is the j-box predictor in cell i. The loss function distinguishes the classification and localization error from an object in the grid cell based on the presence or absence of the error.

x_i ، y_i ،w_i ،h_i correspond to the coordinates of the center of the earth, width, and height for objects in the grid cell (exists) and $\hat{x}_i$ ،$\hat{y}_i$ ،$\widehat{w}_i$ ،$\hat{h}_i$ correspond to the corresponding projections. c_i و $\hat{c}_i$ redevelopes confidence in cell network I for the natural and predicted terrain. $p_i(c)$ و $\hat{p}_i(c)$ Corresponds to the conditional probability for object class c at cell index I for real and predicted terrain, respectively. We use similar settings to reduce $detection_{loss}$of the YOLO object and $\lambda_{coord} = 5$, $\lambda_{noobj} = 0.5$.

The objective function includes an adjuster that monitors the difference between the dummy and predicted labels at each frame. Equation 4 shows the function $\boldsymbol{similarity_{loss}}$.

$$
similarity_{loss} = \sum_{t=0}^{T} \sum_{i=0}^{s^2} \hat{C}_i^{(t)} \left(X_i^{(t)} - \hat{y}_i^{(t)}\right)^2 \tag{4-7}
$$

Where $x_i^{(t)}$ و $\hat{y}_i^{(t)}$denote the dummy label and the corresponding predictions of the i cell network at the t time step, respectively, we minimize the unweighted square with confidence in the prediction made in the corresponding cell.

5 Computational research

In this section, the model on the objects-YouTube dataset is evaluated. Quantitative results (measuring the average accuracy) and subjective evaluation of the model's performance is provided concerning success predictions and failure cases.

The YouTube-Objects dataset consists of videos collected from YouTube by searching for the names of ten object classes from the PASCAL VOC challenge. It contains 155 movies, between 9 and 24 for each category. The duration of each video varies between 30 seconds and 3 minutes. However, only 6087 frames contain 6975 box samples. A division of training and testing is provided.

We implement the matching of the YOLO domain and the proposed RNN model using MATLAB software. The best RNN model using two GRU layers of 150 hidden units each and skipping a probability of 0.5 between layers is significantly more significant than YOLO domain matching alone. While we can only check the prediction quality on labeled frames, subjective evaluations are provided on sequences.

We compare the method with other evaluated methods in the collection of objects-YouTube. As shown in Table 1 and Table 2, the detector reports based on nondeformable models (DPM) (Felzenszwalb. P et al., 2008) average accuracy is less than 30, abysmal performance in some categories such as cat, method (VOP) (Tripathi. S et al., 2016) uses an adaptive video object proposal that is classified by Alexnet (5 convolutional layers, three fully connected) in an R-CNN and obtains an mAP of 37.41.

We also compare with YOLO (24 convolutional layers, two fully connected layers), which unifies the classification and localization tasks and achieves an average accuracy above 55.

Table 1. Accuracy on ten categories of objects

Train	motorcycle	Horse	Dog	cow	Cat	Machine	Boat	Bird	Airplane	Methods
39.58	31.61	35.10	15.84	19.24	1.69	48.99	25.50	48.14	28.42	DPM
29.23	29.77	54.52	34.42	57.56	33.7	41.00	35.34	28.82	29.77	VOP
62.03	24.62	36.96	55.81	23.48	43.03	65.52	57.66	89.51	76.67	YOLO
67.09	32.31	42.53	53.49	56.78	46.67	81.95	59.91	91.98	83.89	DA YOLO
58.23	41.54	81.77	58.72	78.02	62.42	80.69	62.16	87.65	76.11	RNN- PS

We adapt all video frames to generate dummy labels on the YOLO method. They are used as inputs of RNN. We choose YOLO as the dummy label because it is the most accurate fast image surface detector. Improves the consistency and performance of the YOLO domain; we get an mAP of 61.66 accuracies. n categories of objects.

This model uses RNN-based prediction; mAP generally achieves all the main bases. RNN model uses both input/output similarity. RNN model has the best input/output similarity performance, weak levels such as category and prediction accuracy, and mAP has reached 68.73 accuracies. This means the relative improvement compared to the best scores is 11.5%. In addition, RNN improves recognition accuracy in many categories. (Table 1).

While comparing Hidden Markov Models (HMMs) with Conditional Random Fields (CRFs), recent works have shown that superior prediction accuracy is achieved through RNNs. In various tasks, from image analysis to speech recognition, when much information is available, the performance of RNNs is technical. It shows that RNNs stimulate CRF-based independent methods for structural prediction on image segmentation.

Table 2. The overall recognition results are in the object dataset - YouTube. Our best model (DA YOLO) provides improvement compared to (RNN – Propose Method).

RNN Propose Method	DA YOLO	YOLO	VOP	DPM	Methods
68.73	61.66	56.53	37.41	29.41	mAP

We objectively evaluate the proposed RNN model in Figure 2. The top and bottom rows in each pair of sequences correspond to the pseudo-tags from the proposed method. While only the last frame in each sequence contains the natural terrain, we can see that the RNN produces a more accurate and consistent prediction across time frames. Predictions are compatible according to classification scores, localization, and confidence coefficient.
In the first example, the RNN consistently detects the dog throughout the sequence, even if the dummy label is wrong for the first two frames (bird). In the second example, the pseudo-labels did not have motorcycles, persons, bicycles, or other steps at different times. However, our approach consistently predicted the motor.
The third example shows that the RNN predicts both cars while the dummy tag detects only the smaller car in two frames in the sequence.
The last two examples show how the RNN increases its confidence coefficient scores and shows positive detections for cats and cars, respectively, each below the dummy label detection threshold.

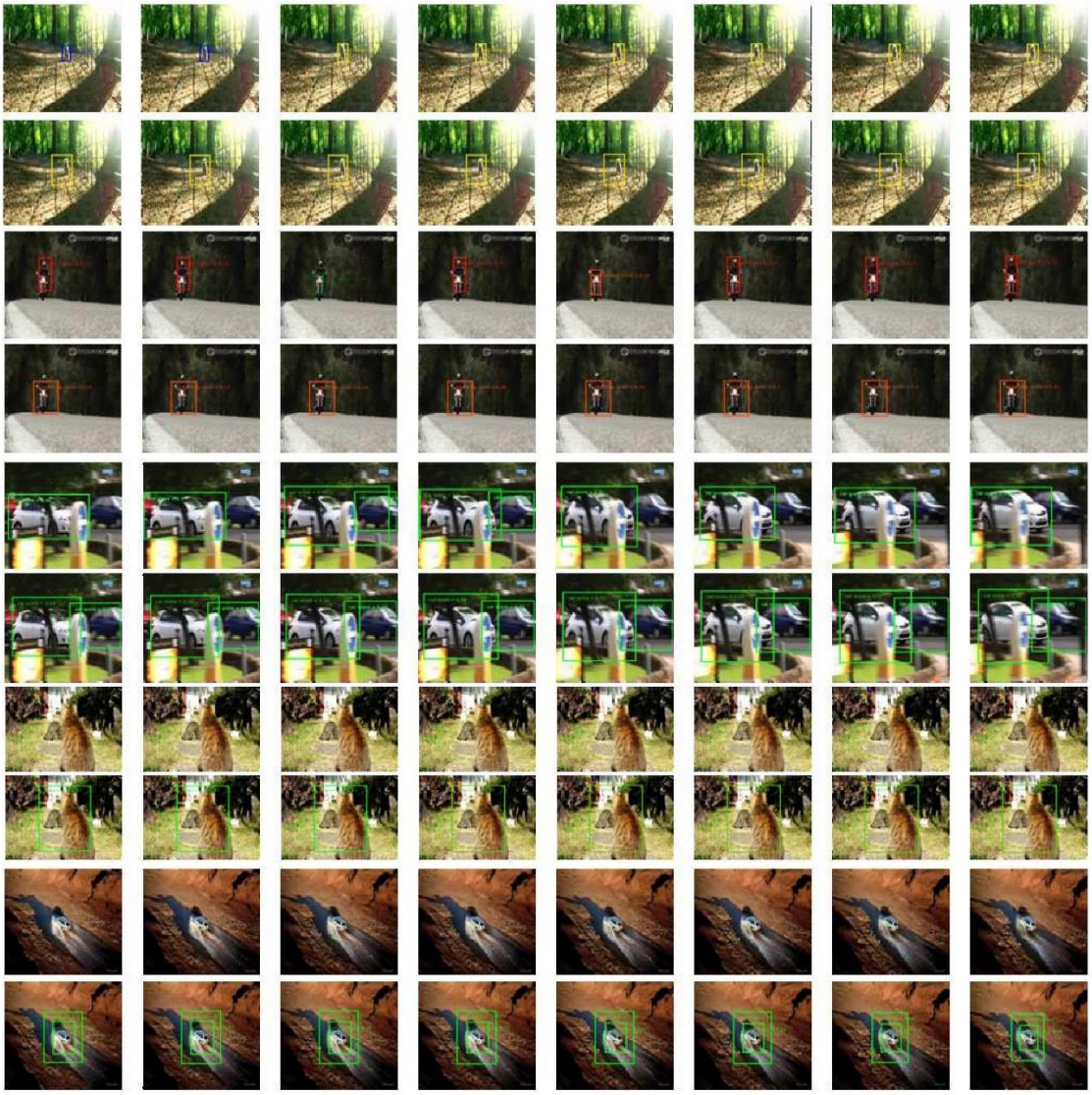

Figure 2: The results of the final eight frames of five different test set sequences. Each pair's top and bottom rows show the dummy label and RNN, respectively. RNN predicts correct values and multiple samples and detects missing objects by increasing the confidence factor

This limits the set of possible predictions, which may be undesirable in cases where many objects are nearby. In addition, the robustness of the YOLO model may have problems lighting up the RNN, which makes the predictions transparent across frames.

For example, an object that moves slightly but passes from one grid cell to another. Here the correctness of predictions is unfavorable.

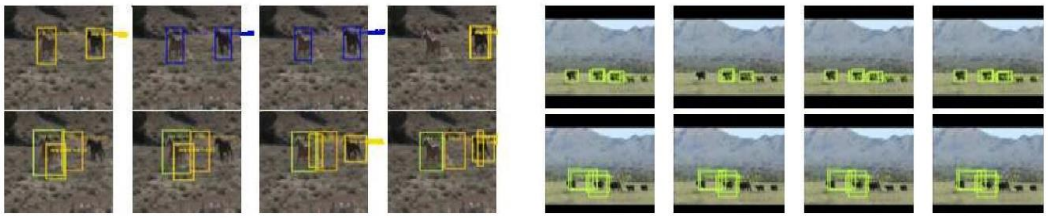

Figure 3: Failure cases for the proposed model. Left: RNN cannot recover from false pseudo-labels. Right: RNN localization is worse than pseudo-labels, probably due to multiple instances of the same object class

Figure 3 shows some failure cases. In the first case, the dummy label classifies the samples as dogs and even as birds in two frames, while the actual ground samples are horses. RNN cannot recover from false brands. Strangely enough, this model significantly increases the confidence factor for another wrong cow.

In the second case, the RNN predicts the correct values but fails to localize, probably due to the movement and re-approach of several samples of the same category.

Weak supervision in level classification in the current scheme assumes the presence of all objects in adjacent frames. While this assumption usually holds for short video clips, it may be violated in the event of occlusion or sudden entry or exit from objects. Furthermore, beliefs about the desirability and correctness of prediction can be broken for fast-moving objects.

Conclusion

In this research, we focused on marker identification, and the proposed solution was marker identification using deep learning. The deep learning model used the supervised learning method, which includes Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The deep learning architecture used is YOLO, and the marker identification results using this method have an accuracy of 61.33 mAP. We implement the proposed YOLO model using MATLAB software. The feature extraction network is usually a pre-trained CNN (we used Pretrained Deep Neural Networks), which uses RNN for feature extraction. We develop a method to extend the integration of object detection. Our process transfers learning from the image domain to image frames, which applies it. In addition, we developed a new Recurrent Neural Networks (RNN) method, which predicts using texture information in adjacent structures. To summarize, we propose a method for correcting video-based object detection that consists of two parts: 1- a dummy label, which assigns temporary labels to all available video frames, and 2- a recurrent neural network, in which A sequence of labeled frames is read, defined using texture information for output predictions. The following is a practical training strategy: 1- Supervised level classification at each time 2- Strong localization level supervision at the last time step 3- Correct prediction at successive times 4- Similarity constraints between dummy labels and output prediction at each step. Finally, we have conducted extensive experimental research showing YouTube datasets in our framework, achieving an average accuracy of 68.73 (mAP) on the test data, which compares to the best-published results of 37.41 and 61.66 for a domain compatible with the YOLO network.

Future research is expected to be conducted on multi-modal samples (i.e., speech and gestures) and intelligent interfaces that recognize its type in the long term.

Most importantly, it is expected that the investigation of 3D models and tracking of unstable objects will be done to track the environment using vertical sensors. Also, make it indoor search outdoors and unique indoors in the future. In the same case, other studies will also focus on social aspects.

References

[1] Z. Cao, T. Simon, S. E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 7291-7299.

[2] C. Chen, A. Seff, A. Kornhauser, J. Xiao, Deepdriving: Learning affordance for direct perception in autonomous driving. In Proceedings of the IEEE international conference on computer vision, 2015, 2722-2730.

[3] D. Chen, S. Ren, Y. Wei, X. Cao, J.Sun. Joint cascade face detection and alignment. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, Proceedings, Part VI 13, 2014, 109-122. Springer International Publishing.

[4] X. Chen, H. Ma, J. Wan, B. Li, T. Xia, Multi-view 3d object detection network for autonomous driving. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, 1907-1915.

[5] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), 1, 2005, 886-893.

[6] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: An evaluation of the state of the art. IEEE transactions on pattern analysis and machine intelligence, 34(4) 2011, 743-761.

[7] A. Dundar, J. Jin, B. Martini, E. Culurciello, Embedded streaming deep neural networks accelerator with applications. IEEE transactions on neural networks and learning systems, 28(7) 2016, 1572-1583.

[8] M. Everingham, The PASCAL visual object classes challenge 2009 (VOC2009) results, 2007.

[9] PF. Felzenszwalb, RB. Girshick, D. McAllester, D. Ramanan. Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence, 32(9) 2009, 1627-1645.

[10] R. Girshick, J.Donahue, T.Darrell, J.Malik, Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, 580-587.

[11] GE. Hinton, N.Srivastava, A. Krizhevsky, I. Sutskever, RR. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580, 2012.

[12] SH. Hsu, HT. Hung, YQ. Lin, CM. Chang, Defect inspection of indoor components in buildings using deep learning object detection and augmented reality. Earthquake Engineering and Engineering Vibration, 2023, 1-14.

[13] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning, 2015, 448-456.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, T. Darrell, Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia, 2014, 675-678.

[15] H. Jiang, E. Learned-Miller, Face detection with the faster R-CNN. In 2017 12th IEEE international conference on automatic face & gesture recognition, 2017, 650-657.

[16] K. Kavukcuoglu, P. Sermanet, YL. Boureau, K. Gregor, M. Mathieu, Y. Cun, Learning convolutional feature hierarchies for visual recognition. Advances in neural information processing systems, 2010, 23.

[17] H. Kobatake, Y. Yoshinaga, Detection of spicules on mammogram based on skeleton analysis. IEEE Transactions on Medical Imaging, 15(3) 1996, 235-245.

[18] A. Kontogianni, E. Alepis, C. Patsakis. Smart tourism and artificial intelligence: Paving the way to the post-covid-19 era. Advances in Artificial Intelligence-based Technologies: Selected Papers in Honour of Professor Nikolaos G. Bourbakis, 1 2022, 93-109.

[19] A. K rizhevsky, I. Sutskever, G.E. Hinton, Image Net classification with deep convolutional neural networks. Advances in neural information processing systems, 25(2) 2012, 1097-1105.

[20] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature, 521(7553) 2015, 436-444.

[21] W. Li, J. Wang, M. Liu, S. Zhao, X. Ding, Integrated Registration and Occlusion Handling Based on Deep Learning for Augmented Reality Assisted Assembly Instruction. IEEE Transactions on Industrial Informatics, 2022.

[22] C. Liu, H. Zhu, D. Tang, Q. Nie, T. Zhou, L. Wang, Y. Song, Probing an intelligent predictive maintenance approach with deep learning and augmented reality for machine tools in IoT-enabled manufacturing, Robotics and Computer-Integrated Manufacturing, 77 2022, 102357.

[23] N. Liu, J. Han, D. Zhang, S. Wen, T. Liu, Predicting eye fixations using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, 362-370.

[24] DG. Lowe, Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60, 2004, 91-110.

[25] W. Luo, L. Zhang, Question text classification method of tourism based on deep learning model. Wireless Communications and Mobile Computing, 2022, 1-9.

[26] S. Murugan, A. Sampathkumar, S. Kanaga Suba Raja, S. Ramesh, R. Manikandan, D. Gupta, Autonomous Vehicle Assisted by Heads up Display (HUD) with Augmented Reality Based on Machine Learning Techniques. In Virtual and Augmented Reality for Automobile Industry: Innovation Vision and Applications, 2022,45-64.

[27] W. Pitts, WS. McCulloch, How we know universals the perception of auditory and visual forms. The Bulletin of mathematical biophysics, 9 1947, 127-147.
[28] V. Puri, S. Mondal, S. Das, VG. Vrana, Blockchain Propels Tourism Industry-An Attempt to Explore Topics and Information in Smart Tourism Management through Text Mining and Machine Learning. In Informatics, 10(1) 2023.
[29] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 779-788.
[30] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 2015, 28.
[31] DE. Rumelhart, GE. Hinton, RJ. Williams, Learning representations by back-propagating errors. nature, 323(6088) 1986, 533-536.
[32] A. Stuhlsatz, J. Lippel, T. Zielke, Feature extraction with deep neural networks by a generalized discriminant analysis. IEEE transactions on neural networks and learning systems, 23(4) 2012, 596-608.
[33] KK. Sung, T. Poggio, Example-based learning for view-based human face detection. IEEE Transactions on pattern analysis and machine intelligence, 20(1) 1998, 39-51.
[34] FM. Wadley, Probit analysis: a statistical treatment of the sigmoidresponse curve, Annals of the Entomological Soc. Of America, 67(4) 1974, 549–553.
[35] R. Wang, F. Bunyak, G. Seetharaman, K. Palaniappan, Static and moving object detection using flux tensor with split Gaussian models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, 414-418.
[36] X. Wang, H. Ma, X. Chen, Salient object detection via fast R-CNN and low-level cues. In 2016 IEEE International Conference on Image Processing, 2016, 1042-1046.
[37] Z. Yang, R. Nevatia, A multi-scale cascade fully convolutional network face detector. In 2016 23rd International Conference on Pattern Recognition, 2016, 633-638.
[38] X. Zhao, W. Li, Y. Zhang, TA. Gulliver, S. Chang, Z. Feng, A faster RCNN-based pedestrian detection system. In 2016 IEEE 84th Vehicular Technology Conference (VTC-Fall) 2016, 1-5.