Detect weapons in crowds of people using artificial intelligence

Dr. Qasim Jaleel
A teacher in the Karbala Education Directorate
qasim.jaleel1984@itnet.uobabylon.edu.iq

Abstract

Security cameras and video surveillance systems are crucial infrastructures for ensuring the safety and security of the general population. Nevertheless, the identification of high-risk scenarios using these methods is still predominantly carried out manually in several places, particularly in Karbala. Insufficient personnel in the security industry and restricted human capabilities might result in unnoticed dangers or delays in identifying potential threats, so endangering the public during crucial events. As a reaction, many entities have created robotic and automated methods to detect potential dangers by analyzing CCTV footage. The objective of this project is to create an affordable and efficient artificial intelligence (AI) system that can accurately recognize and identify weapons in real-time surveillance films, even in various situations. The system consists of a training phase for the Convolutional Neural Networks (CNNs) using a data set containing images of weapons. The second stage is testing the system, which demonstrated its high ability to detect weapons with high accuracy compared to previous research, the accuracy of the proposed method reached 93%.

Keywords: artificial intelligence, CNN, deep learning, machine learning.

Introduction

Artificial intellect) AI (is the replication of human intellect in computers, which are programmed to imitate cognitive abilities including learning, problem-solving, perception, reasoning, and decision-making(Saxena et al., 2020). Artificial intelligence empowers robots to carry out jobs that have historically relied on human intelligence, including

anything from basic activities such as voice and picture recognition to intricate endeavors like operating self-driving cars, engaging in strategic games, and producing innovative material(Prongnuch & Sitjongsataporn, 2021).

AI comprises a wide array of approaches, algorithms, and procedures with the goal of developing intelligent systems that can comprehend and engage with the environment in a way that resembles human behavior(Ni et al., 2023). AI encompasses several essential elements and specialized areas, such as:

Machine Learning, Deep Learning: Deep learning is a distinct subfield of machine learning that use artificial neural networks with several layers (known as deep neural networks) to acquire intricate patterns from vast quantities of data(Teuwen et al., 2023)since facial landmarks can provide precise AU locations to facilitate the extraction of meaningful local features for AU detection. However, most existing AU detection works handle the two tasks independently by treating face alignment as a preprocessing, and often use landmarks to predefine a fixed region or attention for each AU. In this paper, we propose a novel end-to-end deep learning framework for joint AU detection and face alignment, which has not been explored before. In particular, multi-scale shared feature is learned firstly, and high-level feature of face alignment is fed into AU detection. Moreover, to extract precise local features, we propose an adaptive attention learning module to refine the attention map of each AU adaptively. Finally, the assembled local features are integrated with face alignment feature and global feature for AU detection. Extensive experiments demonstrate that our framework (i.

Deep learning has demonstrated exceptional accomplishments in tasks like as picture and audio recognition, natural language processing, and game playing(Jaleel & Ali, 2022b). Natural Language Processing (NLP) is a specialized branch of Artificial Intelligence (AI) that specifically aims to facilitate computers in comprehending, deciphering, and producing human language(Jaleel & Ali, 2022a). Natural Language Processing (NLP) approaches find use in several domains, including language translation, sentiment analysis, chatbots, and information retrieval(Jaleel & Hadi, 2022). Computer Vision refers to the field of study that focuses on enabling machines to analyze and comprehend visual data, such as photos and videos, by interpreting the information they contain(Wang et al., 2020). Computer vision techniques are employed in several applications, including object identification, facial recognition, medical image analysis, and autonomous driving(Sahu & Dash, 2021).

Convolutional Neural Networks (CNNs) are a type of advanced neural networks specifically developed to handle and examine visual input, such as photos and movies(Teuwen et al., 2023)since facial landmarks can provide precise AU locations to facilitate the extraction of meaningful local features for AU detection. However, most existing AU detection works handle the two tasks independently by treating face alignment as a preprocessing, and often use landmarks to predefine a fixed region or attention for each AU. In this paper, we propose a novel end-to-end deep learning framework for joint AU detection and face alignment, which has not been explored before. In particular, multi-scale shared feature is learned firstly, and high-level feature of face alignment is fed into AU detection. Moreover, to extract precise local features, we propose an adaptive attention learning module to refine the attention map of each

AU adaptively. Finally, the assembled local features are integrated with face alignment feature and global feature for AU detection. Extensive experiments demonstrate that our framework (i. Convolutional Neural Networks (CNNs) have emerged as the most advanced method for a range of computer vision applications, such as picture classification, object identification, segmentation, and others(Milosevic, 2020).

Robotics is the integration of artificial intelligence and mechanical engineering to create and construct intelligent devices, commonly known as robots, that can carry out activities independently or with minimal human intervention. Robotics is used in several fields, including industrial automation, manufacturing, service robots, and autonomous drones(Kim et al., 2023)to defeat their opponents, players need to choose and implement the correct sequential actions. Because RTS games like StarCraft II are real-time, players have a very limited time to choose how to develop their strategy. In addition, players can only partially observe the parts of the map that they have explored. Therefore, unlike Chess or Go, players do not know what their opponents are doing. For these reasons, applying generally used artificial intelligence models to forecast sequential actions in RTS games is a challenge. To address this, we propose depthwise separable convolution-based multimodal deep learning (DESEM.

Video surveillance systems employ cameras and software to observe and document activity inside a specified region. These systems serve several functions, such as ensuring security, promoting safety, and facilitating monitoring. Here is a concise explanation of the functioning of video surveillance systems, Cameras: Video surveillance systems

include of one or more strategically positioned cameras to gather footage of the monitored area(Bhatti et al., 2021)it must ensure a safe and secure environment for investors and tourists. Having said that, Closed Circuit Television (CCTV. Cameras can differ in their kind, such as being fixed or pan-tilt-zoom, as well as in their resolution and capabilities, such as having night vision or infrared imaging(Q. Yang et al., 2020).

Video Recording: Cameras collect video footage of the monitored area, which is then saved on a digital video recorder (DVR), network video recorder (NVR), or cloud storage system. The captured video can be retrieved and examined at a later time for the purpose of analysis or as proof(Dolhansky et al., 2020).

Real-Time Monitoring: Video surveillance systems not only record but also offer real-time monitoring features, enabling security personnel or operators to observe live video feeds from the cameras. Real-time monitoring allows for prompt action to be taken in response to security risks or incidents as they happen(Albiero et al., 2021).

Motion Detection and Analytics: Numerous video surveillance systems are furnished with motion detection technology, which activates recording or notifications when movement is detected inside the camera's visual range. Advanced systems can employ video analytics algorithms to automatically analyze film and identify certain events or behaviors, such as object detection, facial recognition, and aberrant behavior detection(T. Y. Yang et al., 2019)we employ the soft stagewise regression scheme. Existing feature aggregation methods treat inputs as a bag of features and thus ignore their spatial relationship in a feature map. We propose to learn a fine-grained structure mapping for spatially grouping features before

aggregation. The fine-grained structure provides part-based information and pooled values. By utilizing learnable and non-learnable importance over the spatial location, different model variants can be generated and form a complementary ensemble. Experiments show that our method outperforms the state-of-the-art methods including both the landmark-free ones and the ones based on landmark or depth estimation. With only a single RGB frame as input, our method even outperforms methods utilizing multi-modality information (RGB-D, RGB-Time.

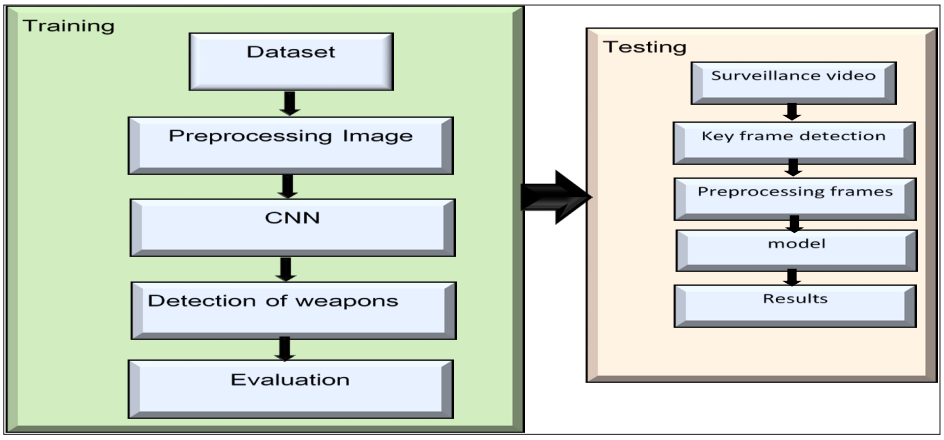
Remote Access: Contemporary video surveillance systems frequently have remote access functionality, enabling authorized users to remotely examine live or recorded video feeds from any location with an internet connection. Remote access allows for the surveillance systems to be monitored and managed from cellphones, tablets, or PCs(Truong et al., 2023). In this paper, modify CNN for detection weapons. Then the CNN network is trained by dataset. And then test the videos by extracting the frames and inserting it into the network for detecting weapons.

methodology

The system consists of two stages, as shown in the figure (1). The first stage is the stage of building the model and training it on various types of weapons. In the first stage, the image containing various types of weapons is read. The image data is then pre-processed to suit the requirements of the CNN network. After that, the results obtained during the training process are evaluated. The second stage is the testing stage, where the accuracy of the system in the process of detecting weapons is tested. In the second stage, which is the testing stage, the system is tested with new image data that is not included in the training process. This

stage is important because it demonstrates the ability of the proposed system to detect weapons in crowds of people. In addition, the system provides high accuracy in detecting weapons compared to previous methods.

Figure (1): general structure of the detection model weapons.



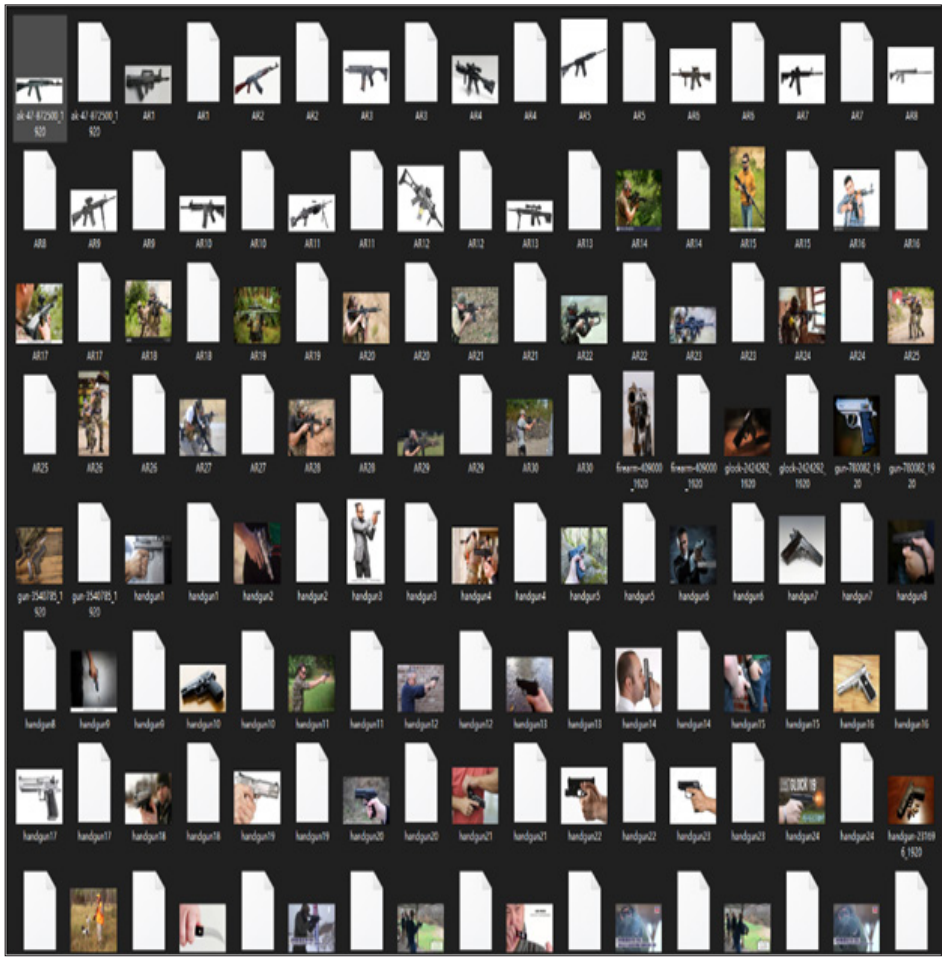
Dataset

The COCO Dataset, also known as Common Objects in Context, is a widely used dataset in computer vision research. The COCO dataset is a highly popular dataset that is extensively utilized for applications like as object detection, segmentation, and captioning. Its main emphasis is on commonplace things and settings, as shown in figure (2).

Preprocessing Image

This preprocessing procedure effectively prepares the image dataset for training CNN models. Preprocessing is essential for ensuring that the input data is correctly prepared and standardized, hence enhancing the training process and optimizing the performance of the CNN model.

Figure (2) part of dataset.



Convolutional Neural Network (CNN) model

To detect weapons, Model CNN was used. This model has the ability to deal with images and videos. In addition to the ability to detect objects such as people, cars, weapons, etc. The number of layers in a Convolutional Neural Network (CNN) for weapon identification can be flexible and influenced by factors like as the complexity of the

detection task, the quantity and variety of the dataset, and the specific architecture being employed.

In the proposed system, the CNN network consists of the following layers: The term “Feature Extraction Layers” refers to a set of layers in a neural network that are responsible for extracting relevant features from input data. The earliest layers of the CNN architecture usually comprise convolutional and pooling layers that are responsible for extracting features from the input pictures.

The phrase “Intermediate Layers” refers to the layers that exist between the input and output layers of a neural network. The intermediate layers in the CNN architecture serve to enhance the retrieved features and capture more intricate representations of the input data. These layers frequently include of supplementary convolutional and pooling layers.

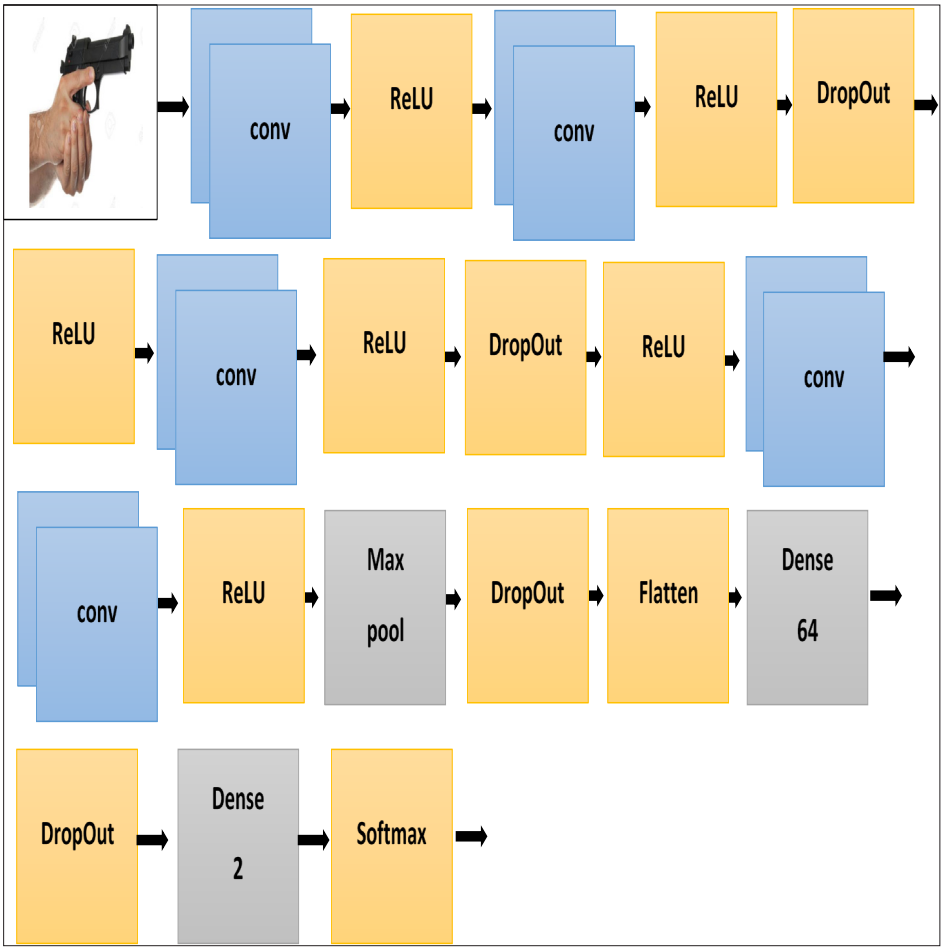
The Localization and Classification Layers are. At the later stages of the CNN architecture, there are usually layers that have the task of identifying the location and category of weapon in the images. The architecture typically consists of convolutional layers, which are then followed by fully connected layers. This design is used to make predictions about the bounding box coordinates and class probabilities for identified weapon.

The output layer is the final layer of a neural network that produces the network’s output or prediction. The output layer of the Convolutional Neural Network (CNN) architecture generates the ultimate predictions for tasks related to weapon detection. The output layer for weapon detection may have neurons that correspond to several categories of weapons, such as firearms and knives, as well as the background.

Figure (3) below illustrates the architecture that it employed for

modify CNN model. Some of the model's hyperparameters include the filter size, which should be 256 for the first convolution layer, 128 for the second, 64 for the third, 32 for the fourth layer, and 16 for final the layer. also using in model, the operation of the pooling layer, whose stride is always set to 1 for both the pooling and convolution layers. It also explained how used ReLU in CNN after each convolution layer. It is using decided to use Max Pooling with a filter size of 2 × 2.

Figure (3) CNN model.



Detection of weapons

In the weapons detection phase, the network is trained on various types of weapons. The data set is divided into 70% training data and 30% test data. In the training phase, the data set designated for training is used, which contains various positions, sizes, and shapes of weapons. This stage is considered very important because it is essential for learning about weapons, and the more pictures of weapons there are, the more accurate the system will be.

Evaluation

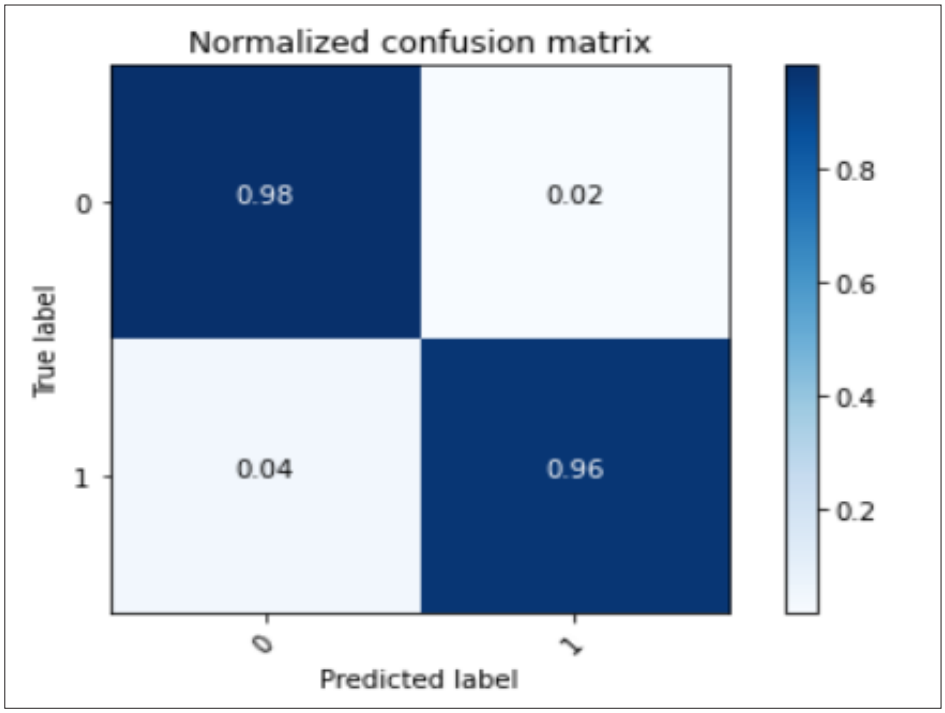
Evaluating the system is an important stage, whether it is in the training stage or the testing stage. The measure used in the proposed system is accuracy, which is considered one of the most important measures for the Neural network. The figure (4) shows the connection between lost training data and evaluation data, as well as the connection between model loss and accuracy. The figure (5) depicts the confusion matrix for the discrimination model, which shows the distribution of number ratios between weapons and not weapons. The proposed system proved an accuracy of 97% in the training phase, while the accuracy reached 93% in the testing phase.

In the testing phase, the mechanism will be different because it will depend on new data that the proposed system was not trained on. The first step is to take the monitored videos, where the videos are divided into frames. Not all frames that enter into the testing process are useful, so a processing is used to extract the frames, which is the Key frame detection. In video processing and analysis, key frame identification is a technique used to pinpoint representative frames within a video stream. These frames are important or representative because they provide important details about the video's content. Applications such as video

summarizing, video indexing, and video browsing can benefit from key frame detection. The calculated similarity scores are used to determine which key frames to use. Key frames are sometimes defined as frames that show notable variances or changes in comparison to their nearby frames. As an alternative, key frames might be chosen on a regular basis to guarantee that the whole video sequence is covered.

Figure (4)The relationship between the loss, accuracy between train, and validation.

Figure (5) Predicted label of the proposal model.



Conclusion

The proposed model has proven its ability to detect various types of weapons in a crowd of people. The model was trained on the Coco data set, which contains various types of weapons and is divided into training data and test data. When compared with previous research, we find that the proposed model has the highest detection rate, as in table (1). When using the Model Yolo 4, the accuracy of the system is 91%. Faster R-CNN achieves an average Accuracy of 88%. UZI Model achieves an average Accuracy of 88%. The accuracy of the proposed model in the training phase reached 97%, while in the testing phase it was 93%.

The accuracy of the system depends greatly on the accuracy of the weapons images on which the proposed model is trained. In addition

to the types of weapons presents in the training data set. It is possible to develop the proposed model using other deep learning networks that may increase the accuracy of the proposed model

Table 1Comparison Between the Results of The Traditional Methods and The Proposed Method.

Method type	Accuracy
Yolov4(Bhatti et al., 2021)it must ensure a safe and secure environment for investors and tourists. Having said that, Closed Circuit Television (CCTV	91%
Faster RCNN(Jain et al., 2020)	91%
UZI Model(Jain et al., 2020)	88%
The propose method	93%

References

1. Albiero, V., Chen, X., Yin, X., Pang, G., & Hassner, T. (2021). img-2pose: Face Alignment and Detection via 6DoF, Face Pose Estimation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 7613–7623. https://doi.org/10.1109/CVPR46437.2021.00753

2. Bhatti, M. T.,Khan, M. G., Aslam, M., & Fiaz, M. J. (2021). Weapon Detection in Real-Time CCTV Videos Using Deep Learning. IEEE Access, 9, 34366–34382. https://doi.org/10.1109/ACCESS.2021.3059170

3. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) Dataset. http://arxiv.org/abs/2006.07397

4. Jain, H., Vikram, A., Mohana, Kashyap, A., & Jain, A. (2020). Weapon Detection using Artificial Intelligence and Deep Learning for Security Applications. Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020, November, 193–198. https://doi.org/10.1109/ICESC48915.2020.9155832

5. Jaleel, Q., & Ali, I. H. (2022a). Facial behavior analysis-based deepfake video detection using gan discriminator. 2022 International Conference on Data Science and Intelligent Computing (ICDSIC), 36–40.

6. Jaleel,Q.,& Ali, I. H. (2022b). MesoNet3: A Deepfakes Facial Video Detection Network Based on Object Behavior Analysis. International Conference on New Trends in Information and Communications Technology Applications, 38–49.

7. Jaleel, Q.,& Hadi, I. (2022). Facial Action Unit-Based Deepfake Video Detection Using Deep Learning. 2022 4th International Conference on Current Research in Engineering and Science Applications (ICCRESA), 228–233.

8. Kim, C., Bae, J., Baek, I., Jeong, J., Lee, Y. J., Park, K., Shim, S. H., & Kim, S. B. (2023). DESEM: Depthwise Separable Convolution-Based Multimodal Deep Learning for In-Game Action Anticipation. IEEE Access, 11(May), 46504–46512. https://doi.org/10.1109/ACCESS.2023.3271282

9. Milosevic,N.(2020). Introduction to Convolutional Neural Networks. Introduction to Convolutional Neural Networks, April. https://doi.org/10.1007/978-1-4842-5648-0

10. Ni, J., Young, T., Pandelea, V., Xue, F., & Cambria, E. (2023). Recent advances in deep learning based dialogue systems: a systematic

survey. Artificial Intelligence Review, 56(4), 3055–3155. https://doi.org/10.1007/s10462-022-10248-8

11. Prongnuch, S.,& Sitjongsataporn, S. (2021). Differential Drive Analysis of Spherical Magnetic Robot Using Multi-Single Board Computer. International Journal of Intelligent Engineering and Systems, 14(4), 264–275. https://doi.org/10.22266/ijies2021.0831.24

12. Sahu, M., & Dash, R. (2021). A survey on deep learning: Convolution neural network (cnn). In Smart Innovation, Systems and Technologies (Vol. 153, Issue January). Springer Singapore. https://doi.org/10.1007/978-981-15-6202-0_32

13. Saxena, A., Khanna, A., & Gupta, D. (2020). Emotion Recognition and Detection Methods : A Comprehensive Survey. 53–79. https://doi.org/10.33969/AIS.2020.21005

14. Teuwen, J., Moriakov, N., Liu, M., Zhang, X., Hou, Y., Tran, V. N., Lee, S. H., Le, H. S., Kwon, K. R., Jentzen, A., Riekert, A., Yan, Z., Zhu, X. X., Wang, X., Ye, Z., Guo, F., Xie, L., Zhang, G., Patil, K., … Celik, T. (2023). Deepfakes: temporal sequential analysis to detect face-swapped video clips using convolutional long short-term memory. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13(3), 1–20. https://doi.org/10.1007/s11263-020-01378-z

15. Truong, T. X., Nhu, V.-H., Phuong, D. T. N., Nghi, L. T., Hung, N. N., Hoa, P. V., & Bui, D. T. (2023). A New Approach Based on TensorFlow Deep Neural Networks with ADAM Optimizer and GIS for Spatial Prediction of Forest Fire Danger in Tropical Areas. Remote Sensing, 15(14), 3458.

16. Wang, S.-Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot... for now. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8695–8704.

17. Yang, Q., Zhang, Y., Dai, W., & Pan, S. J. (2020). Transfer learning. Cambridge University Press.

18. Yang, T. Y., Chen, Y. T., Lin, Y. Y., & Chuang, Y. Y. (2019). Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019-June, 1087–1096. https://doi.org/10.1109/CVPR.2019.00118