Detecting Tagged People in Camera Images
Kamera Görüntülerinde Etiketlenen Kişilerin Tespit Edilmesi

Muhammed TELÇEKEN[1*], Yakup KUTLU[2]
[1]Department of Computer Engineering, Sakarya University of Applied Sciences, Sakarya, Turkey
[2]Department of Computer Engineering, Iskenderun Technical University, Hatay, Turkey
ORCIDs: 0000-0001-5223-2856, 0000-0002-9853-2878
E-mails: muhammedtelceken@subu.edu.tr, yakup.kutlu@iste.edu.tr
*Corresponding author.

***Abstract*—With the development of technology, cameras are used more widely. It is possible to evaluate the widespread use of cameras in various subjects in daily life. Especially face recognition systems are one of the most important areas of use of cameras. Facial recognition systems can be used in many areas such as cyber security, entertainment, security applications of daily used devices, and faster and easier transactions in financial areas. Although a lot of progress has been made in this regard, face recognition systems are still used widely enough because it is thought that they have weaknesses in terms of security. Many scientists are working on facial recognition. In this study, it is aimed to detect the faces of people determined from videos or live camera images in the best and safest way. Yolov4 object detection algorithm, a ready-made algorithm, was used for the detection of human faces on images. The faces of the people in the images were detected by training the data set we created in the Yolov4 algorithm. An accuracy of 99.1 has been achieved for detecting people's faces on images. The data set we created with pictures of certain people is trained in the CNN algorithm. The faces of the people detected on the images were classified on the model trained with the CNN algorithm for the identification of the people, and the accuracy value was examined for the detection of the identified people on the video recordings or live images from the cameras.**

***Keywords*—Face Recognition; Yolov4; CNN Algorithm**

***Özetçe*—Teknolojinin gelişmesi ile birlikte kameralar daha yaygın olarak kullanılmaktadır. Kameraların yaygın kullanımını günlük hayatın içerisinde çeşitli konularda değerlendirmek mümkündür. Özellikle yüz tanıma sistemleri kameraların kullanım alanlarının en önemlilerinden bir tanesidir. Yüz tanıma sistemleri siber güvenlik, eğlence, günlük kullanılan cihazların güvenlik uygulamalarında ve finansal alanlarda daha hızlı ve kolay işlem yapabilmek gibi birçok alanda kullanılabilmektedir. Bu konuda bayağı bir ilerleme kaydedilmiş olsa da henüz güvenlik açısından zayıf yönlerinin olduğunun düşünülmesinden dolayı yüz tanıma sistemleri henüz yeteri kadar yaygın şekilde kullanılmaktadır. Yüz tanıma üzerine birçok bilim insanı tarafından çalışmalar yapılmaktadır. Bu çalışmada videolardan veya canlı kamera görüntüleri üzerinden belirlenen kişilerin yüzlerini en iyi ve en güvenli şekilde tespit etmek amaçlanmıştır. İnsan yüzlerinin görüntüler üzerinde ki tespiti için hazır bir algoritma olan Yolov4 nesne algılama algoritması kullanılmıştır. Oluşturduğumuz veri seti Yolov4 algoritmasında eğitilerek görüntülerdeki insanların yüzleri tespit edilmiştir. Görüntüler üzerinde insanların yüzlerinin tespiti için yüzde 99,1'e ulaşan doğruluk elde edilmiştir. Belirli kişilerin resimleri ile oluşturduğumuz veri seti CNN algoritmasında eğitilmiştir. Görüntüler üzerinde tespit edilen insanların yüzleri kişilerin kimlik tespiti için CNN algoritması ile eğitilmiş model üzerinde sınıflandırılmış ve belirlenen kişilerin video kayıtlarında veya kameralardaki canlı görüntüler üzerinde tespiti için doğruluk değeri incelenmiştir.**

***Anahtar Kelimeler*—Yüz Tanıma; Yolov4; CNN Algoritması**

I. INTRODUCTION

With the rapid development of technology, cameras are used in many parts of our daily lives. Cameras are used in the mobile devices we use, in streets, in buildings, in the media, in medicine and in human life in many fields of science. Cameras are used in subjects such as security and entertainment. With its affordable cost, the interest in camera-based software has increased even more, and camera-based software has gained great importance, especially in issues such as social security and cyber security [1]. One of the software applications based on cameras is face recognition systems. Many studies have been done on this subject and many studies are still being carried out on this subject. Facial recognition systems are more preferred in terms of use, as they have better results than retinal recognition and fingerprint recognition systems in terms of ease of use and high success rates among biometric recognition systems [2]. Face recognition systems are the method of detecting the human face on video recordings or live images with certain algorithms over the images in the cameras. Very successful results have been obtained in studies on images to recognize the human face. One of the general purposes of face recognition systems from video is to detect human faces first through images with good image quality, where human faces can be easily distinguished, and then to determine who the detected face belongs to among the images registered by the system.

Some of the studies on facial recognition systems are as follows. Yang Zhiqi [3] preferred convolutional neural network (CNN) as the classification algorithm in his study in 2021 and suggested the MicroFace architecture, and achieved 96,26

accuracy in his study on 120000 images for 2000 people. Jun Wang et al. [4] preferred CNN as their classification algorithm in 2021, they proposed a FaceXzoo (open source) approach using the PyTtorch toolbox, and obtained 79,26 accuracy using a dataset with masked and unmasked images in the study. Ting Chen et al. [5] proposed a new face recognition method based on the fusion of LBP and HOG in their study in 2021, in their study on 3 different data sets, they obtained 82.00 accuracy for LBP and 94.20 for CS - NWALBP + HOG. Mingna Wu et al. [6] suggested JLSRCI based on LatLRR,LBR,SRC in their study in 2021,and they achieved 99.25 accuracy for EyalebB dataset 99.17 for AR dataset 99.17 for ORL dataset. Dongmei Shi and Hongyu Tang [7] preferred the CNN algorithm to classify LBP for feature extraction in their study in 2021, and they achieved 95.20 accuracy in their study on 3 different data sets. Min Hao, Guangyuan Liu [8] proposed a new approach based on LBP (local binary pattern) in their study published in 2021, examined CMU, UWA, PolyU datasets, and for each dataset 98.50, 96.60 and They achieved 94.0 accuracy. YU WANG et al. [9] obtained 89.70 and 87.30 accuracy in honda and youtube face datasets, respectively, using ST-VLAD and LBP algorithms in their study in 2021.

This study was examined in two stages. First, the detection of people's faces in camera images was examined. Yolov4 object detection algorithm, which is a machine learning algorithm based on deep learning, has been examined to detect people's faces. The dataset obtained by tagging people's faces on various images was trained on the Yolov4 algorithm. With the trained model, the detection of human faces on images with crowded groups of people was examined. CNN algorithm has been examined to determine the identities of individuals. In order to determine the identities of the people, the data set we created from the pictures of 10 different people was trained in the CNN algorithm. A new image was obtained by combining certain sections from the images obtained in different times and environments for 10 people. On this video, firstly, the faces of the people in the images were determined by using the model trained with Yolov4. The detected facial images were classified in the model trained with the CNN algorithm and the identification of 10 people was examined. This study was carried out on the Google Colab application.

II. Materials & Methods

This study was carried out in two stages. First, two different data sets were created and the first data set was used in the training of the YoloV4 algorithm for detecting faces from videos in the YoloV4 algorithm. The second data set was created to be used in the training of the CNN algorithm. After completing the training stages of the algorithms, the detection of faces from video images with the YoloV4 algorithm was examined with the recorded models. Then, the classified results were examined in the CNN algorithm model, where the detected faces were trained. The flow chart of the study is shown in Fig. 1.

A. Obtaining Data

In our study, 2 different data groups were examined. In order to detect people's faces in the Yolov4 object detection algorithm, the first data set was obtained by tagging 425 people's faces on 125 different pictures. Then, the second data set was obtained by tagging the faces of 10 Hollywood stars on 750 pictures to be used in the training to determine the identities of the people in the CNN algorithm. In order to determine the identities of the people on the images, our third data group was obtained by combining 2 minutes of images from 3 different recordings for 10 different people.

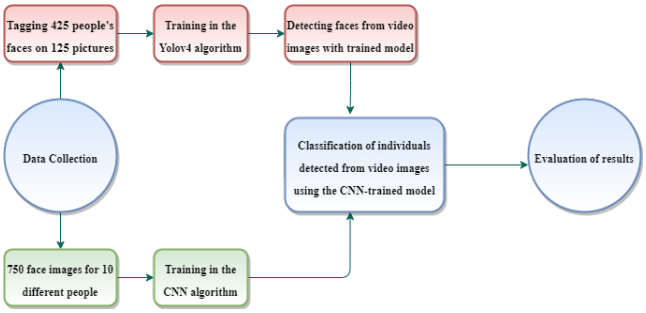

Figure 1: Flow Chart

B. Google Colab

The Colab notebook is an interactive environment that allows us to write and execute code. Colab notebook (https://colab.research.google.com/) allows us to combine code execution, rich text possibilities, images, HTML, Latex and other elements into a single document. It allows us to analyze and visualize data by taking full advantage of the popular Python libraries. Colab notebook executes code on Google's cloud server. In other words, it allows you to perform your training regardless of the power of your machine. To run deep learning applications, features such as high GPU and RAM are needed. Due to the high costs of a computer with a high GPU, it is an application that allows deep learning models to be run comfortably as a GPU and RAM supported application provided by Google.

C. Yolov4 Algorithm

The most common algorithm used for object detection in recent years is the YOLO (Yolo Only Look Once) algorithm. The reason why the Yolo algorithm is so popular is that it can detect objects much faster than previous algorithms. The most important feature that distinguishes YOLO from other algorithms is that it is an algorithm that can detect objects in real time. Although there were algorithms for real-time object detection before YOLO, the overall average precision values were not sufficient. The basis of the speed of the YOLO algorithm is that it predicts the class and coordinates of all objects in the picture by passing a picture through the neural network at once [10]–[12]. The basis of this estimation process is that it treats object detection as a single regression problem. To do this, it first divides the input picture into SxS grids. These grids are 3x3, 5x5, 19x19 etc. it could be. The structure

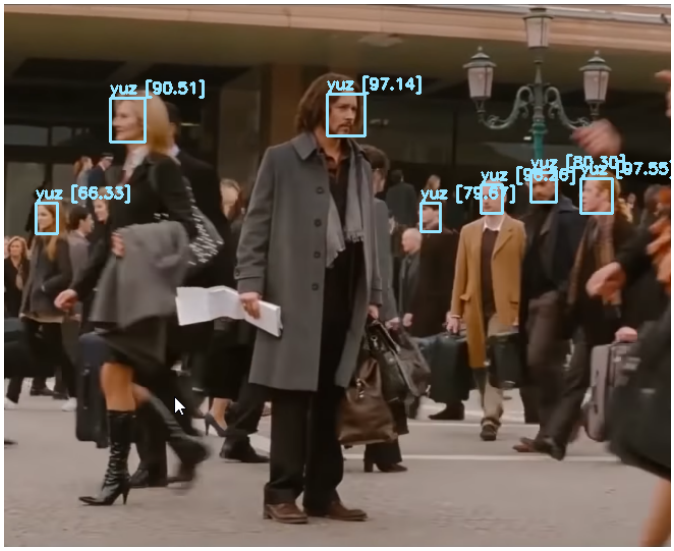

Figure 2: Output of an image from the YOLO Algorithm

of a picture obtained after the picture passes through the neural network is shown in Fig. 2.

Each grid finds within itself whether the object is in the area or not, if its midpoint is in it, and if its midpoint is in it, it finds its length, height, and what class it is. In the picture in Fig. 2, each of the grids corresponding to the midpoints of the faces is responsible for detecting the face it corresponds to and drawing a box around it. For this, YOLO creates a separate prediction vector for each grid.

For each vector:

- Confidence score: this score shows how certain the model is whether there are objects in the valid grid (if 0 definitely not, 1 definitely exists). If it thinks it is an object, it shows how sure it is whether it is really that object and the coordinates of the box around it.
- Bx: The x coordinate of the object's midpoint
- By: The y-coordinate of the object's midpoint
- Bw: width of the object
- Bh: Height of the object
- Associated Class Probability: As many predictive values as there are different classes in our model.
- Confidence Score: Box Confidence Score * Affiliated Class probability
- Box Confidence Score: P(object)*IoU
- P(object): The probability that its box contains the object
- IoU: IoU value between ground truth and predicted box

In grids with no objects, the confidence score will be 0 because the connected class probability must be 0. According to the output vector in Fig. 2, only one object can be defined for each grid. If only a 3x3 grid was used, 9 objects could be estimated. Anchor Boxes are mounted on the YOLO algorithm in order to eliminate the problem that may arise against the middle point of 2 different objects in a grid. With the arrival of Anchor Boxes, the output vector will be calculated by the equation in (1).

$$SxS(\#AX(5+\#C)) \quad (1)$$

That is, since S x S indicates the total number of grids, the probability for Confidence,x,y,w,h and other classes will be calculated as much as the number of anchor boxes for each grid. Tx,Ty,Tw andTh are estimated by the network for each tile. At the same time, we know which grid is the processed grid, so we can find the distance of the grid to the upper left corner. If we call these distances Cx and Cy and the width and heights of the anchor box we have determined before, Pw and Ph, the setup of the system is shown in Fig. 3.

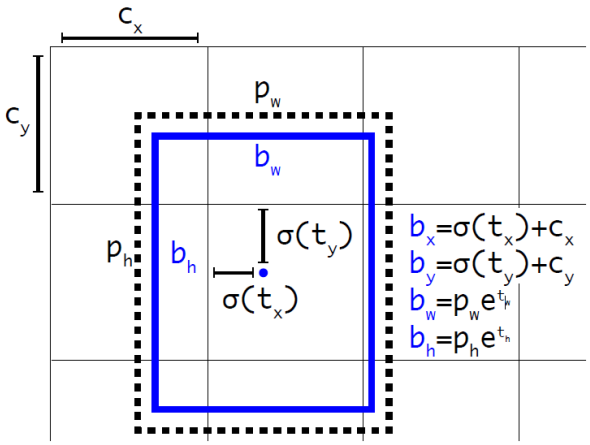

Figure 3: A relative system for showing estimates in practice

With this apparent system, the network will be more stable as we normalize the parameters between 0 and 1. However, while the algorithm is running, too many unnecessary boxes will appear, even a few different boxes for just one object. It will be easy to throw out the unnecessary boxes, if we already have an object in that grid and if more than one grid thinks that the object is the middle point for the same object, then the Non max Suppression algorithm will be activated. The following is what the Non max Suppression algorithm will do.

- It will remove all boxes with a confidence score below a certain level.
- It will select the box with the highest confidence score and output it. If we call this box A;
- A and IoU will remove all other boxes with a value greater than 0.5.

At the end of these processes, only one box will remain for each object. Compared to Yolov3, YOLOv4 introduced mosaic data enhancement in data processing. Additionally, backbone, network training, activation function and loss function were optimized, making YOLOv4 faster, providing the best balance between accuracy and real time [13]. An open source neural network framework, CSPDarknet53, was used as the main backbone network to train and extract the YOLOv4 network image features [13]. They then used PANet as a neck net to better combine the features extracted from the images [14]. They used YOLOv3 to perform object detection [15].

D. Convolutional Neural Network (CNN)

It is a type of Multi-Layer Perceptron (MLP). It was suggested that it was inspired by the visual centers of animals. Mathematical convolution operations can be thought of as a response to stimuli [16]. A CNN consists of one or more convolution layers followed by one or more fully connected layers such as a standard multilayer neural network [17]. CNN network is the architecture named LeNet, which was first proposed by Yann LeCun in 1988 and continued until 1998 [18]. CNN algorithm can be applied in many different fields such as image and sound processing, natural language processing (NLP) biomedicine. Especially in the field of image processing, the best results have been obtained [19]–[21].

The CNN algorithm processes the image in various layers. If it is necessary to write down the layers in the CNN algorithm and the operations they do superficially;

- Convolution Layer: It is used to detect properties.
- Non-Linearity Layer: Introducing nonlinearities to the system
- Pooling (Downsampling) Layer: It provides control of the fit by reducing the number of weights.
- Flattening Layer: It is the data preparation layer for Classical Neural Network.
- Fully-Connected Layer: Standard neural network used in classification

On the basis of the CNN algorithm, a standard neural network is used as a classification solution. The layers we specified are applied to determine the information and features before classification.

It is the main building block of the CNN algorithm. It is responsible for determining the properties of images. It applies some filters to the image to determine the high and low level features of the images. Filters are usually multidimensional and contain pixel values (5x5x3). 5 represents the height and width of the matrix, and 3 represents the depth of the matrix.The visuals of the application of the filter are shown in fig. 4.

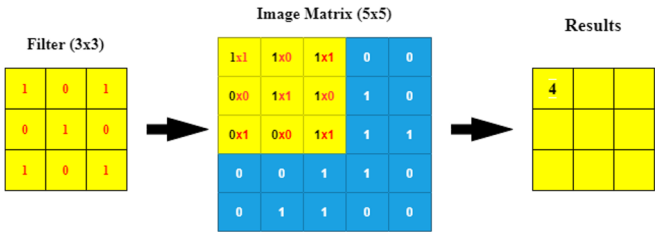

Figure 4: Applying the filter on the image

First, the filter is moved over the image by sliding it starting from the upper left corner of the image. While the filter is moving over the image, the indices between the two matrices (image and filter) are multiplied with each other and all results are collected and stored in the output matrix. The scrolling process continues 1 column to the right after each addition operation. After the end of the line, the next line is passed and the same operations are applied again. After all the operations are completed, a 3x3 output matrix is obtained as a result of a 3x3 filter applied on an image with a 5x5 matrix. The output matrix is called Feature Map. Indicates the location of the image in the feature represented by the filter. Multiple filters are applied to detect more than one feature. More complex filters can be applied if complex features are desired to be extracted.

Stride is often used with the term padding. Stride controls how the filter evolves around the image. In the example shown in Fig. 3, Stride is set to 1 pixel, but this can be set to a larger pixel if desired. However, setting Stride large affects the output size. When the first filter is applied in the CNN algorithm, it is necessary to preserve as much information as possible for the other convolution layers. That's why padding is used. As seen in the example in Fig. 3, the output matrix is smaller than the input image. For this reason, padding will add 0 values around the image matrix to preserve the size of the image. Fig. 5 shows this structure.

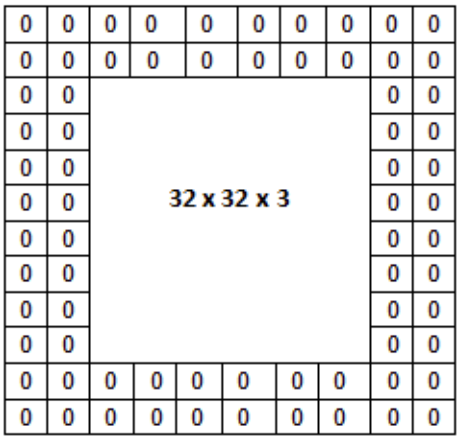

Figure 5: Padding image matrix

After the convolution layers, the non-linearity layer usually comes. Since all layers can be a linear function, the Neural network behaves like a single perception. Result outputs can be calculated as a linear combination. This layer can also be called (Activation layer). In the past, this 'sigmoid', 'tanh' was used more. However, since the best results on the speed of training of neural networks are obtained in the Rectifier (ReLu) function, the use of this function is generally preferred. It is calculated using the Relu function (2).

$$ReLu : f(x) = \max(0, x) \tag{2}$$

It is a frequently used layer between consecutive convolution layers in CovNet. It is used to reduce the shift size of the representation and the number of computations of parameters within the network. Thus, incompatibility on the network is checked. It has many operations such as MaxPooling, AveragePooling, L2-normPooling. The most popular Pooling app is MaxPooling. The implementation of the MaxPooling process is shown in Fig. 6.

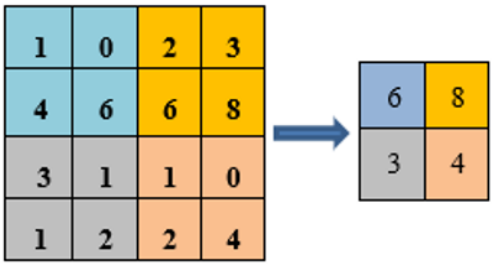

Figure 6: MaxPooling process

Finally, and most importantly, in this layer, the data in the Fully Connected Layer inputs are prepared. Neural networks are a one-dimensional array of input data. In this neural network, the matrices are not one-dimensional since the data comes from the Convolution and Pooling layer. In this layer, matrices are converted to one-dimensional array. This conversion process is shown in Fig. 7.

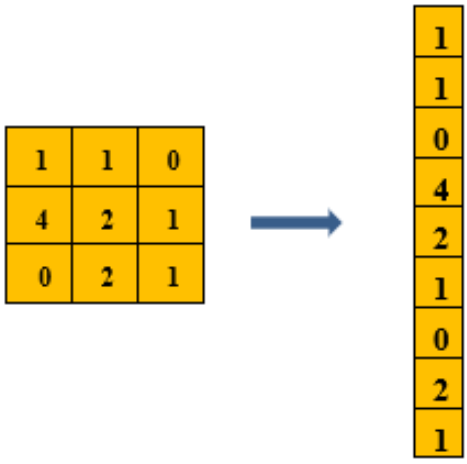

Figure 7: Flatten layer

Fully connected layers connect each neuron in each layer to each neuron in the next layer. It is the same as a multilayer neural network (MLP). Flattened matrices pass through fully connected layers to classify images.

As a result of the classification of the face images detected by the yolov4 algorithm from the video recordings with the trained CNN algorithm model, the identities of the people on the images were determined as in Fig. 8.

E. Performance Metric

In this study, 3 different data groups were studied. First, a new data set was created to be used in the Yolov4 algorithm to detect faces in the images. Secondly, the data set to be used for training in the CNN algorithm was created to determine the identities of the faces to be detected. Finally, a data set obtained from videos was created to detect faces with Yolov4.

Figure 8: Identification in video

Accuracy was examined for the faces to be detected from the video images with Yolov4 and then for the performance criterion for the classification of these detected faces in the CNN algorithm. The accuracy criterion was calculated using (3).

$$Accuracy = \frac{TP + TN}{TN + FN + TP + FP} \tag{3}$$

Here TP indicates the number of correctly classified data, TF indicates classified negatives, FP misclassified positives, FN misclassified negatives.

III. Results & Discussion

This study was carried out on Google Colab, which provides GPU and RAM support. Firstly, the identification of faces through images was examined. Yolov4 algorithm was used to detect faces on images. For training on Yolov4, a new data set was created by labeling 425 face images on 125 images. The dataset in which the created faces are labeled is bent on Google Colab in the Yolov4 algorithm. Training of faces in the Yolov4 algorithm took 6 hours and 22 minutes. At the end of the training, the final weights of the objects, the best weights of the objects and the weights of the objects at the end of 1000 iterations were recorded in a folder, and it was aimed to use the trained model in detecting the faces in the images in the next stage. One of the most important points for detecting faces is the distance of the human face from the camera in the image, the amount of appearance of the human face. During the test phase, 99.1 accuracy was obtained with the model trained with the Yolov4 algorithm in the images in which 225 human faces were determined.

Secondly, in order to determine the identities of the detected faces, training was conducted with 750 pictures of 10 different people in the CNN algorithm. The model trained in the CNN algorithm was then used to identify the faces detected on the images. In the CNN algorithm, the model was recorded according to the lowest 'val-loss' value during training. As a result of the classification performed on 223 face images, an accuracy of 87.44 was obtained. It is thought that better results will be obtained by improving the CNN algorithm architecture.

REFERENCES

[1] Sakaci B, Yildirim T. Smart suit design for military. Journal of Intelligent Systems with Applications 2019; 2(1): 66-71.

[2] Aydin Y, Akar F. Using local features in face recognition systems. Journal of Intelligent Systems with Applications 2018; 1(2): 131-134.

[3] Zhiqi Y. Face recognition based on improved VGGNET convolutional neural network. In: 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2021. p. 2530-2533.

[4] Wang J, Liu Y, Hu Y, Shi H, Mei T. FaceX-Zoo: A PyTorch toolbox for face recognition. Proceedings of the 29th ACM International Conference on Multimedia, October 2021, pp. 3779–3782.

[5] Chen T, Gao T, Li S, Zhang X, Cao J, Yao D, Li Y. A novel face recognition method based on fusion of LBP and HOG. IET Image Processing 2021; 15(6).

[6] Wu M, Wang S, Li Z, Zhang L, Wang L, Ren Z. Joint latent low-rank and non-negative induced sparse representation for face recognition. Applied Intelligence 2021; 51(11): 5349-8364.

[7] Shi D, Tang H. Face recognition algorithm based on self-adaptive blocking local binary pattern. Multimedia Tools and Applications 2021; 80: 23899-23921.

[8] Hao M, Liu G, Xie D. Hyperspectral face recognition with a spatial information fusion for local dynamic texture patterns and collaborative representation classifier. IET Image Processing 2021; 15(8): 1617-1628.

[9] Wamg Y, Huang YP, Shen XJ. ST-VLAD: Video face recognition based on aggregated local spatial-temporal descriptors. IEEE Access 2021; 9: 31170-31178.

[10] Wu D, Lv S, Jiang M, Song H. Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. Computers and Electronics in Agriculture 2020; 178: 105742.

[11] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016, June 27-30, Las Vegas, NV, USA, pp. 779-788.

[12] Silva G, Monteiro R, Ferreira A, Carvalho P, Corte-Real L. 2019. In: October. Face Detection in thermal images with YOLOv3. International Symposium on Visual Computing (ISVC 2019): Advances in Visual Computing 2019, pp. 89–99.

[13] Bochkovskiy A, Wang CY, Liao HYM. Yolov4: Optimal speed and accuracy of object detection. ArXiv 2020; https://arxiv.org/abs/2004.10934.

[14] Wang D, He D. Recognition of apple targets before fruits thinning by robot based on R-FCN deep convolution neural network. Transactions of the Chinese Society of Agricultural Engineering 2019; 35(3): 156-163.

[15] Redmon J, Farhadi A. Yolov3: An incremental improvement. ArXiv 2018; https://arxiv.org/abs/1804.02767.

[16] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics 1980; 36(4): 193–202.

[17] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; 521: 436-444.

[18] Le Cun Y, Jackel LD, Boser B, Denker JS, Graf HP, Guyon I, Henderson D, Howard RE, Hubbard W. Handwritten digit recognition: applications of neural network chips and automatic learning. IEEE Communications Magazine 1989; 27(11): 41–46.

[19] Yildirim O, Ucar A, Baloglu UB. Recognition of real-world texture images under challenging conditions with deep learning. Journal of Intelligent Systems with Applications 2018; 1(2): 122-126.

[20] Narin A, Pamuk Z. Effect of different batch size parameters on predicting of COVID19 cases. Journal of Intelligent Systems with Applications 2020; 3(2): 69-72.

[21] Balli O, Kutlu Y. Regional signal recognition of body sounds. Journal of Intelligent Systems with Applications 2021; 4(2): 157-160.