Detection of Appearance and Behavior Anomalies in Stationary Camera Videos Using Convolutional Neural Networks

H. Chen[a,b,*], R. Bohush[c,], I. Kurnosov[d,***], G. Ma[e,****], Y. Weichen[e,*****], and S. Ablameyko[d,f,******]**

[a] *Zhejiang Shuren University, Hangzhou, 310015 China*
[b] *International Science and Technology Cooperation Base of Zhejiang Province: Remote Sensing Image Processing and Application, Hangzhou, 310000 China*
[c] *Polotsk State University, Novopolotsk, 211440 Republic of Belarus*
[d] *Belarusian State University, Minsk, 220030 Republic of Belarus*
[e] *EarthView Image Inc., Huzhou, 313200 China*
[f] *United Institute for Informatics Problems, National Academy of Sciences of Belarus, Minsk, 220012 Republic of Belarus*
**e-mail: eric.hf.chen@hotmail.com*
***e-mail: bogushr@mail.ru*
****e-mail: fpm.kurnosov@bsu.by*
*****e-mail: magd@ev-image.com*
******e-mail: yangwch@ev-image.com*
*******e-mail: ablameyko@bsu.by*

Abstract—The automatic detection and tracking of appearance and behavior anomalies in video surveillance systems is one of the promising areas for the development and implementation of artificial intelligence. In this paper, we present a formalization of these problems. Based on the proposed generalization, a detection and tracking algorithm that uses the tracking-by-detection paradigm and convolutional neural networks (CNNs) is developed. At the first stage, people are detected using the YOLOv5 CNN and are marked with bounding boxes. Then, their faces in the selected regions are detected and the presence or absence of face masks is determined. Our approach to face-mask detection also uses YOLOv5 as a detector and classifier. For this problem, we generate a training dataset by combining the Kaggle dataset and a modified Wider Face dataset, in which face masks were superimposed on half of the images. To ensure a high accuracy of tracking and trajectory construction, the CNN features of the images are included in a composite descriptor, which also contains geometric and color features, to describe each person detected in the current frame and compare this person with all people detected in the next frame. The results of the experiments are presented, including some examples of frames from processed video sequences with visualized trajectories for loitering and falls.

Keywords: video surveillance, face mask, tracking-by-detection, motion features, loitering
DOI: 10.1134/S1054661822020067

INTRODUCTION

People detection and behavior analysis using video sequences are important computer-vision problems that currently have an increasing number of applications, including activity recognition in surveillance and security systems. In contrast to images, video sequences contain a much larger amount of information that varies both in space and time. That is why their processing and analysis make it possible to extract not only static but also dynamic features of objects, which improves the overall effectiveness of automated video surveillance systems.

In recent times, so-called situational analytics has been intensively developed [6]. This methodology analyses nonstandard behavior patterns in crowds, as well as anomalous behavior of individuals or groups of individuals. The following behavior anomalies can be distinguished: loitering (remaining in a particular public place for a long time without any apparent purpose), fall, sudden acceleration (e.g., when someone starts running), sudden stop, and entering/exiting a certain area.

Loitering is considered an anomalous (suspicious) behavior. In some countries, it is prohibited by law as it may indirectly indicate illegal intentions; therefore, individuals that exhibit loitering need to be detected by video surveillance.

Received January 24, 2022; revised January 24, 2022; accepted January 24, 2022

Anomalies in people's appearance are detected based on whether their look differs from that generally accepted in a given region, season, or situation. Because of the coronavirus pandemic, it is especially important to detect violations of the mask regulation, which involves the detection of appearance anomalies (the presence or absence of a face mask) depending on effective restrictions.

Behavior anomalies of certain types can be recognized by tracking the movement of a person and analyzing its trajectory [9]. This makes it possible to prevent situations such as the installation of an explosive device, damage to property, and filming in restricted areas. Multiple-object tracking is key to the detection of appearance and behavior anomalies; however, it also complicates the process because it is required to detect and track all of the individuals that appear in the frame, as well as compute their features and trajectories.

Currently, algorithms based on deep learning, namely, convolutional neural networks (CNNs), are widely employed for object detection. These algorithms are robust to changes in lightning conditions and dynamic background and also enable detection of highly overlapped objects. In [20], a method that requires detection of moving objects and calculation of time gradients for them in an input video sequence was proposed. Points of interest in this method are determined based on calculated gradients and accelerated segmentation. Wavelet analysis is used to describe detected spatial-temporal blocks. At the final step of the method, classification based on a Gaussian mixture model (GMM) is carried out.

CNNs are widely employed to detect behavior anomalies in video sequences. In [7], a CNN autoencoder (AE) was used to restore original data when detecting behavior anomalies based on modeling. In [12], a fully convolutional neural network was proposed to detect fast movements that characterize anomalous behavior of people.

In [11], the efficiency of several deep neural networks for anomaly detection was evaluated and the possibility of using an adversarial autoencoder (AAE) network in a generated model to recognize normal patterns in images was investigated. In [3], a model for anomaly detection in video using a convolutional AE was constructed. Even though the input was a continuous set of frames, two-dimensional convolution was implemented. After the first convolution, temporal information was completely eliminated. In [16], an improvement of this method was proposed. Considering that the fully convolutional network did not take temporal information into account, an LSTM convolutional layer was added between the encoding and decoding layers in the Conv-AE network. However, with the temporal information of the input image being lost, the effect of the LSTM layer on the extraction of temporal information was very insignificant.

In [8], a new spatio-temporal U-Net for frame prediction and anomaly detection was proposed. This framework combines the benefits of spatial information modules with the capabilities of ConvLSTM for modeling temporal motion data. In addition, to further improve the accuracy of anomaly detection, a new regular score function was proposed, which consists of a prediction error for both the current frame and future frames. A number of approaches were considered in [2, 18, 19]. Anomaly detection using a moving camera was described in [4].

Recently, a number of works devoted to monitoring the compliance with mask-wearing requirements have been published. In [10], two deep neural networks were trained. The first one—single-shot multibox detector—was used to detect faces in images. Then, the MobileNetV2 network was used for face classification. The authors collected a composite dataset based on datasets used in other studies. In [15], Faster-RCNN and YOLOv3 were compared. Faster-RCNN showed high quality; however, the runtime was not adequate for real-time applications. In [13], a two-step approach to this problem was described, with the library from OpenCV as a detector and a classifier based on MobileNetV2. In [14], an approach was proposed whereby the complexity of an image is first estimated and then, depending on it, the faster MobileNet or the more accurate ResNet50 is used. Since the authors used a cascade of models for detection, complexity estimation, and classification, the overall runtime made it difficult to use this approach in real-time scenarios.

In this paper, we formalize the problem of detecting appearance and behavior anomalies in video sequences. To recognize anomalous behavior, it is required to analyze the features that characterize the movement of a person. For this purpose, we use tracking-by-detection, i.e., the first step is to detect the person. In the detected region, the face of the person is found and analyzed for the presence of a face mask. In this case, the main difficulties are the lack of a training dataset and the high-speed requirements. That is why we generate a large class-balanced dataset and train YOLOv5 to detect masked faces. The results of the experiments, which confirm the high efficiency of the proposed approach, are presented. To recognize falls, local trajectory analysis is carried out, while global trajectory analysis is used to recognize loitering.

1. FORMALIZATION OF ANOMALY DETECTION

A *video sequence* or *video stream* is a sequence of digital images (frames) $V = \{F_k\}$, where k is the number of images in the sequence. An *image object* (*Ob*) is a local region that differs from the surrounding background and reflects some of the features of a real-world object. Each frame of the sequence, which is

obtained using a stationary camera, generally contains many objects:

$$OB_{F_k} = \left\{Ob_q^{F_k}\right\}, \quad q = 1,\ldots,Q.$$

In the general case, the *detection of an object* is the localization of object Ob^e in image F, with its size being less than the size of the image and the number of objects in the image being a priori known.

Based on the movement criterion, each of them can be attributed to either of the two following classes.

A stationary object in a sequence of images is described by a set of features ($Ft_{Ob_q}^S$) and its coordinates $\left(x_{Ob_q}, y_{Ob_q}\right)$ that do not change over a period of time (t). This object can be represented by the following formal model:

$$Ob_q^S = \left(Ft_{Ob_q}^S, x_{Ob_q}, y_{Ob_q}, Ns_{Ob_q}^{F_k}\right),$$

where $\left(Ft_{Ob_q}, x_{Ob_q}, y_{Ob_q}\right) = \text{const}\ \forall F_k,\ k \in t$, and $Ns_{Ob_q}^{F_k}$ is a set of noise influences on the object.

A dynamic object in a sequence of images is characterized by a variation in one or more of its basic parameters (shape, size, or coordinates) on a certain time interval (t). A transformation of the shape and/or size of the object affects its features in image frames ($ft_{Ob_q}^{F_k}$). Hence, this object can be described by the following formal model:

$$Ob_q^D = \left(ft_{Ob_q}^{F_k}, x_{Ob_q}^{F_k}, y_{Ob_q}^{F_k}, Ns_{Ob_q}^{F_k}\right),$$

where $x_{Ob_q}^{F_k}$, $y_{Ob_q}^{F_k}$ are the coordinates of the dynamic object; $Ft_{Ob_q}^D \supseteq ft_{Ob_q}^{F_k},\ \forall k \in t$; $Ft_{Ob_q}^D$ is a set of its features; and $\left(ft_{Ob_q}^{F_k} \cap ft_{Ob_q}^{F_{k+i}}\right) \in Ft_{Ob_q}^D$.

Anomaly detection in surveillance video is carried out based on the following assumption: to solve the problem, it is sufficient to run the search procedure in the current frame without taking into account individual movement histories, i.e., without analyzing the dynamic features of objects. Therefore, formal model (2) should take into account certain features based on which the appearance or behavior of a person is classified as an anomaly. To detect appearance anomalies, it is required to determine features that can indicate anomalies of this type. For this purpose, we can use features associated with physical parts of the human body, e.g., the head, the position of the hands or the whole body, etc. Hence, the process of anomaly detection is implemented by matching a specified set of static features (Ft_e^P) with features of all possible fragments in a frame by using method M based on the rule

$$S\left(Ft_e^P, Ft_{P_q}\right) \xrightarrow[Z]{M} \max,$$

where S is the accuracy of detection and Z is a set of constraints.

To detect behavior anomalies, it is required to analyze dynamic features, i.e., the history of movement in a spatial region recorded by the camera. It is reasonable to use the following features that characterize individual movement in a video sequence: the magnitude and angle of shift between neighbor frames, trajectory, speed, acceleration, and time.

The trajectory represents the movement of a person in video frames as a thin line, or a line of unit-width cross section. This line can be constructed based on the coordinates of the centers of images of the person in the previous frames. This can be done in different ways; the most common one is to find the center (one pixel per frame) with coordinates $\left(x_{P_q}^{F_k}, y_{P_q}^{F_k}\right)$. Then, an individual trajectory in a video recorded by a stationary camera is described by a set of coordinates of the object's center in each frame:

$$Tr\left(P^D\right) = \left(P_{F_k}^D\right)$$
$$= \left\{\left(x_{P_q}^{F_1}, y_{P_q}^{F_1}\right), \left(x_{P_q}^{F_2}, y_{P_q}^{F_2}\right), \left(x_{P_q}^{F_3}, y_{P_q}^{F_3}\right), \ldots, \left(x_{P_q}^{F_n}, y_{P_q}^{F_n}\right)\right\}.$$

It should be taken into account that different people can appear in a video sequence starting from frames other than the first one, as well as disappear at some instant from the nth frame other than the last one. Then, the trajectory can be constructed starting from frame m, in which a person is detected, to frame n. Thus, the trajectory is generated for each person who appears in the frame and is deleted when this person disappears from the frame. When a moving person is temporarily overlapped by other objects or individuals, or when it is impossible to detect the person, the trajectory is interrupted and then restored after a few frames, i.e., the trajectory is fragmented. Thus, to recognize the type of movement, it is necessary to take into account the specific characteristics of human motion. This can be done by using some a priori information about individual movement from previous frames, as well as its dynamic characteristics, on the basis of which a trajectory model can be preliminarily constructed (this model can be linear or nonlinear, periodic or nonperiodic). Based on this trajectory, it is possible to calculate individual movement in a video over a certain time interval, the boundaries of which are determined by the numbers of the frames in which the coordinates of the person's position are determined and the shortest distance between them is found.

The localization of an object in each video frame over a time interval (t) is called tracking. Tracking is required to construct individual trajectories, as well as determine speed and acceleration. Considering that many people can simultaneously appear in a video frame, for their correct tracking, it is required to determine the set of their trajectories $TR' = \{Tr_q'\}$ with their

subsequent matching to determine all individual movements between frames:

$$STM\left(TR,TR'\right) \xrightarrow[ZTM]{MTM} \max,$$

The speed (*SP*) and acceleration (*AC*) of a person moving in the coordinates of a video frame constitute the next group of important features, which can be evaluated by computing the first and second derivatives.

The fourth feature required in the analysis is the movement time (*TM*), which is a temporal feature. The main parameters of this feature are the times of the beginning and end of object movement, while its secondary parameters are the time intervals that characterize the continuous movement of the object and its stationarity for the entire period of observation.

Thus, taking into account the features considered above, the model of individual movement in the general case can be represented as follows:

$$M(P_{\text{mov}}) = \left(Tr, SP, AC, TM\right).$$

To detect behavior anomalies in a video sequence, it is required to estimate the correspondence between a generated model of individual movement and model $M(P_{\text{mov}}^{\text{abn}})$, which characterizes anomalous behavior:

$$Sm\left(M(P_{\text{mov}}), M(P_{\text{mov}}^{\text{abn}})\right) = X.$$

Based on it, a decision about the type of individual movement is made.

Depending on the problem, model $M(P_{\text{mov}})$ can be incomplete, i.e., some features may not be used to describe human movement.

2. GENERAL STRUCTURE OF THE ANOMALY-DETECTION ALGORITHM

To detect behavior anomalies, it is required to construct individual trajectories of each person in a frame. For this purpose, we propose an algorithm that uses tracking-by-detection as in [1] and, upon detecting a person, evaluates the corresponding features. At the first stage, the procedure of people detection in the current frame is carried out. For this purpose, we need a high-speed CNN, which should also have high accuracy because the more accurate the detector, the better the tracking. Among the existing CNNs, the YOLOv5 model meets our requirements, which is why it is used at this stage. As a result of the detection stage, a bounding box is constructed around the detected person. Then, the features of the image of this person are determined. For the detected regions, the features used for tracking and trajectory-construction are formed. The analysis of the movement features is carried out to determine the type of movement.

Thus, the proposed algorithm consists of the following steps (see Fig. 1).

1. Detect all people in the tth frame P_t^{det} by using the CNN. The people detected in frame $t-1$ are tracked P_{t-1}^{tr}.

2. Determine the features of the image of the detected person.

3. Form an array of two sets of CNN features $V_{P_t^{\text{det}}} = \left(V_{P_t^{\text{det}}}^1, V_{P_t^{\text{det}}}^2\right)$ taking into account the height (h) and width (w) of the regions detected at the previous step: $V_{P_t^{\text{det}}}^1 = (f_{\text{det}}^{\text{FullPCNN}})$ for the entire image of the person and/or $V_{P_t^{\text{det}}}^2 = (f_{\text{det}}^{\text{TopPCNN}})$ for its top part.

3.1. If $(w/h) < 1$ for the image of the detected person, then CNN features are formed for the entire image ($f_{\text{det}}^{\text{FullPCNN}}$) and for its top part ($f_{\text{det}}^{\text{TopPCNN}}$).

3.2. Otherwise, only the upper part of the human body is considered detected, and CNN features are formed only for the second part of the array, i.e., $V_{P_t^{\text{det}}} = \left(V_{P_t^{\text{det}}}^2\right)$.

3.3. If the tracked person is correctly detected three or more times, then the set of CNN features is formed by averaging: $V_{P^{\text{tr}}} = \frac{1}{n}\sum_{i=1}^{n} V_{Ob_i^{\text{det}}}$, $n = 3$–5.

4. Filter the objects to determine their similarity by comparing the distances between the centers of the objects, their width, and their height in the previous and current frames:

$$(l_{\text{curr}} - l_{\text{prev}}) > t_l \quad \text{or} \quad (w_{\text{curr}} - w_{\text{prev}}) > t_w$$
$$\text{or} \quad (h_{\text{curr}} - h_{\text{prev}}) > t_h.$$

If these conditions hold, then the similarity estimation for these objects is not carried out.

5. Generate a similarity matrix for the individuals detected in the current and previous frames.

5.1. If $(w/h) < 1$ for the image of the detected person, then ($f_{\text{det}}^{\text{FullPCNN}}$) are used as CNN features of this image.

5.2. Otherwise, ($f_{\text{det}}^{\text{TopPCNN}}$) are used as CNN features.

6. Apply the Hungarian algorithm to match the objects in the frames and perform the following substeps.

6.1. If there is a match for the detected person in the current frame, then assign this person the name (index) from the previous frame.

6.2. Check the presence of the person in the frame in accordance with [1].

6.3. If a new person is detected, then assign this person a unique name (index) and consider this person a new object to be tracked.

7. Draw a bounding box around the detected person.

8. Construct and analyze the person's trajectory.

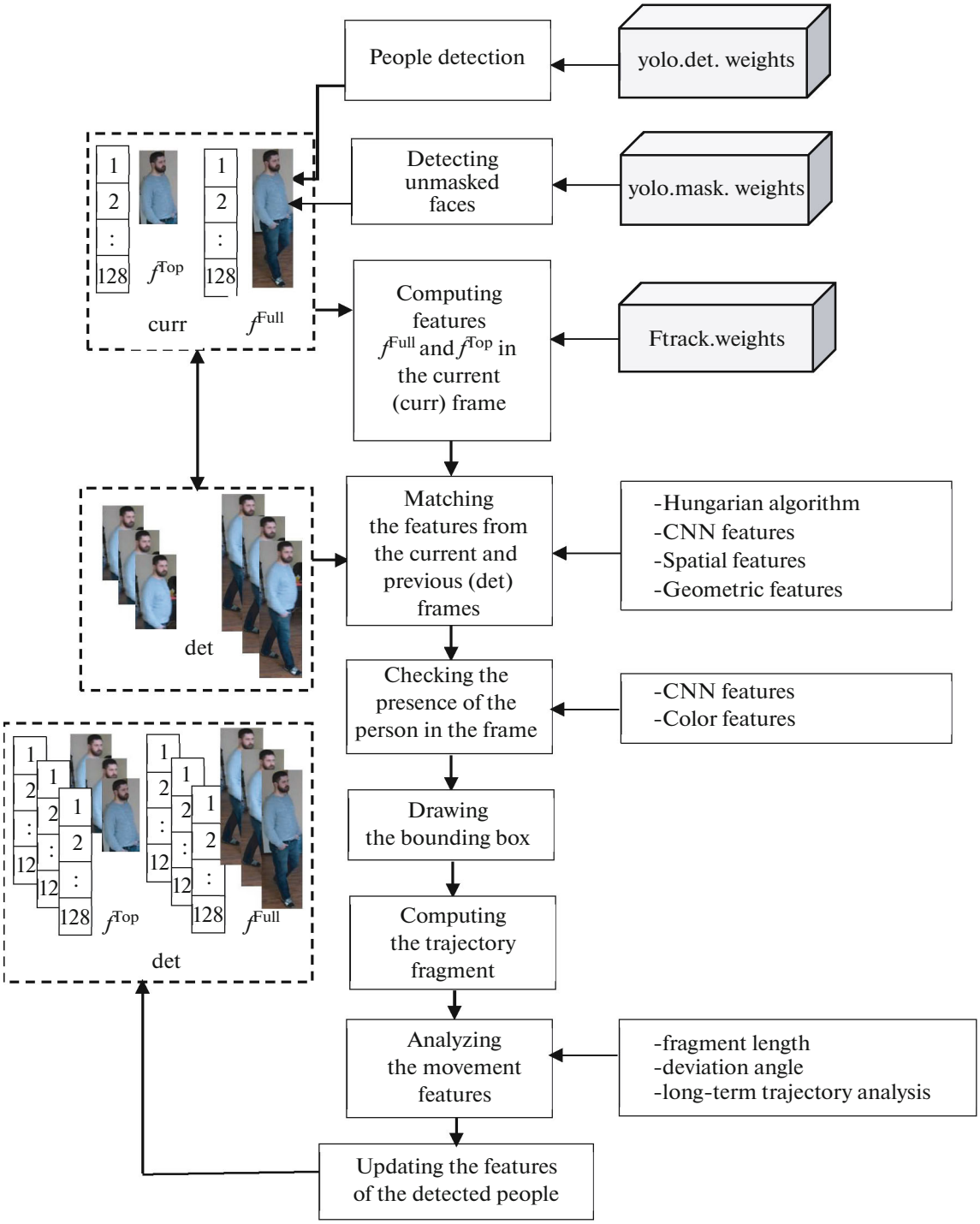


Fig. 1. General scheme of the algorithm: yolo.det. weights and yolo.mask. weights are YOLOv5 weight coefficients for people detection and face-mask detection, respectively, while Ftrack.weights are weight coefficients for the CNN that computes the features of the tracked person.

9. Update the features for the tracked person in the current frame.

9.1. Write new values of the width, height, and coordinates of the image centers to array $V_{P_t^{tr}}$.

9.2. Write new values of the CNN features and histogram descriptors to array $V_{P_t^{tr}}$ if the IoU overlap parameter is below 0.8.

10. Receive the next input frame, go to step 1.

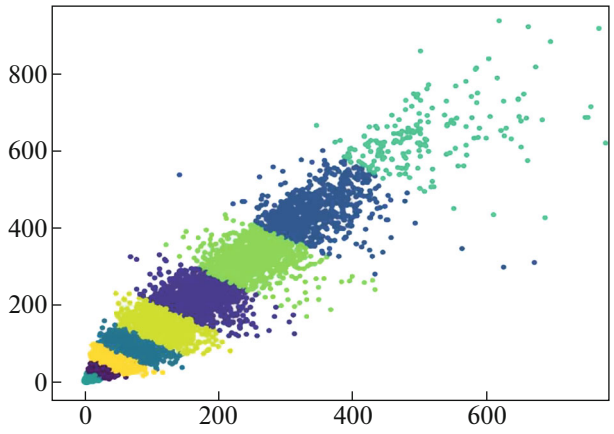


Fig. 2. Applying the *k*-means algorithm to the training dataset.

3. DETECTION OF APPEARANCE ANOMALIES

In this work, by the appearance anomaly, we mean the absence of a face mask, which has become especially important in recent years. Before the COVID-19 pandemic, wearing a face mask in European countries was considered an atypical appearance aimed at hiding the person's face. However, during the pandemic and lockdowns, the absence of a face mask is generally considered an anomaly.

In this case, the detection problem is complicated by the lack of a proper training dataset, as well as the high-speed requirements for the algorithm. The key features of our approach are the use of the YOLOv5 one-step CNN simultaneously as a detector and classifier, as well as the generation of our own dataset, various augmentations, and fine-tuning of the CNN parameters by using a clustering algorithm.

3.1. Generating the Training Dataset

The problem of face-mask detection is relatively new, which is why high-quality datasets are not yet publicly available. There are two datasets on the Kaggle platform [5], which consist of 800 and 1000 images (respectively) and are similar in terms of labeling quality and type of photographs. We combine these datasets into a single set, which is referred to as Kaggle. This dataset is well-labeled, even though some small faces were skipped. Its disadvantages are as follows:

– Uniformity of viewing angles and lighting: most of the images were taken under uniform sunlight with the camera at eye level. Frontal angles dominate and there are few faces shot from the side.

– Lack of small faces: in most of the images, there are approximately five medium-sized faces. The number of labeled small faces (several pixels in size) is extremely small.

The Kaggle dataset is significantly inferior in quality to publicly available datasets that contain face images, which have larger volume, higher face-size diversity, larger number of people present per image, better lighting and viewing angles, more diverse emotions and accessories (hats), etc. Wider Face is one of these datasets [17], which consists of 32000 images with 400000 faces. However, only half of this dataset is publicly available.

3.2. Initializing the Model

YOLOv5 uses anchors for high-quality detection of faces of different sizes in a single run. A grid is superimposed on the input image, and the model searches for the objects inside each grid cell. The anchor is the estimated size of the object. It determines the width and height of the region to be selected inside each cell. To make predictions more accurate, several anchors of different sizes can be used (in this work, we use nine). The sizes of the anchors correspond to expected sizes of faces, which is why it is reasonable to initialize them based on the training dataset. We cluster the sizes of all faces in the training dataset into nine groups by using the *k*-means algorithm (see Fig. 2). The anchors are initialized by the centers of these clusters: [6.9419, 8.7715], [12.886, 16.466], [20.605, 26.093], [31.871, 40.189], [48.509, 60.48], [75.842, 97.367], [125.57, 161.59], [219.84, 283.9], and [381.06, 482.96].

3.3. Training the Model

First, the model is pretrained on the Wider Face dataset to achieve high detection quality. Then, we train it on the Kaggle dataset to provide the required classification functionality. The final results of training during 30 epochs are shown in Fig. 3. It can be seen that mAP with IoU = 0.5 (hereinafter, it is referred to as mAP@0.5) and mAP@0.95 still tend to grow; however, the slowdown is sufficient to conclude that mAP@0.5 is 0.94, which means that the model performs well as a detector.

3.4. Classification

In the process of detecting and filtering the image regions that possibly contain faces, the model extracts a sufficient number of features for image classification. In this work, faces are classified into two classes: "mask" and "no mask." In addition, the model estimates confidence for each classification decision. To filter out the results with insufficient confidence, a confidence threshold is set. For this purpose, we use the F_β metric, as it reflects a balance between precision and recall:

$$F_\beta = \left(1+\beta^2\right)\cdot\frac{precision\cdot recall}{\left(\beta^2\cdot precision\right)+recall}.$$

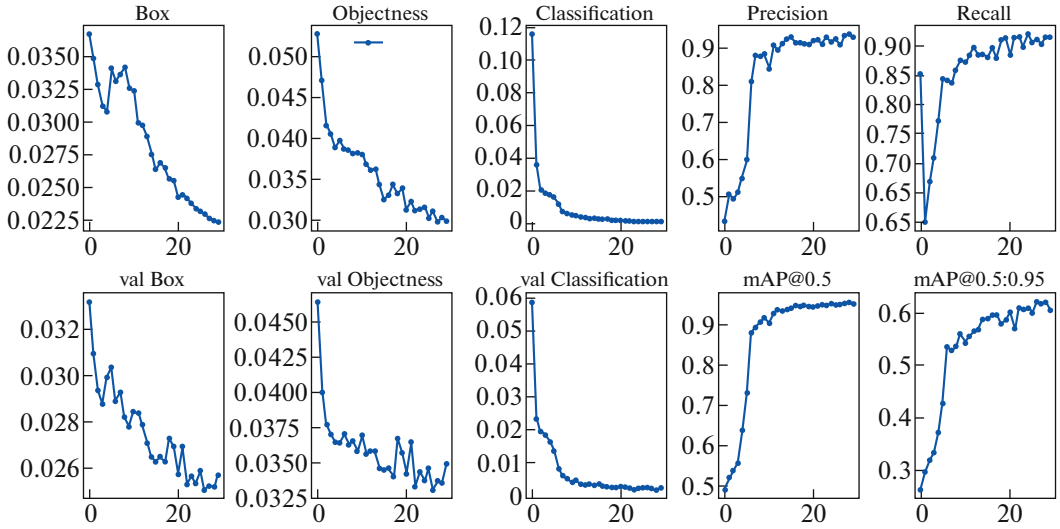


Fig. 3. Result of 30 training epochs for the pretrained model on the Kaggle dataset.

For the resulting model, the precision and recall are determined by the confidence threshold. Figure 4 plots $F_{\beta=1}$ versus the confidence threshold.

The optimal confidence level of 0.428 maximizes $F_{\beta=1}$. Using this value, we can evaluate the accuracy of the model on the test data: 0.9411.

3.5. Modifying the Dataset for Training and Classification

The analysis of the graphs that represent the training results and the visual assessment of the processed images suggest that we have reached a certain quality limit on the Kaggle dataset. Moreover, the composition of this dataset does not correspond to that of Wider Face (or similar datasets). That is why it needs to be extended. For this purpose, face masks were superimposed on a random half of the faces from the Wider Face images, which have a relatively frontal position (see Fig. 5), and then this modified Wider Face dataset was combined with Kaggle. The resulting dataset has the following advantages: racial and national diversity, different lighting conditions and viewing angles, class balance, and large size. However, there are certain disadvantages: more than half of the face masks are artificially generated and are located frontally, so the model is trained for detection and classification on these artificial faces.

3.6. Experimental Results of Face Mask Detection

The training on the modified dataset was performed for ten epochs. The quality metrics are shown in Fig. 6. It can be concluded that the classification accuracy is high and reaches a plateau. The values of mAP@0.5 and mAP@0.95 grow almost linearly. This means that the number of training epochs can be increased. The optimal level of confidence is determined based on the graph of $F_{\beta=1}$ (Fig. 7), and the new model is tested on the same test data. The accuracy is 0.9611, which is higher than in previous works.

Visual comparison of mAPs for the two models described above confirms that the second model is a better detector. This is due to the fact that the previous model did not "see" the Wider Face dataset on the last training epochs and began to "forget" how to detect small faces. Figure 8 shows an example that illustrates the operation of the model.

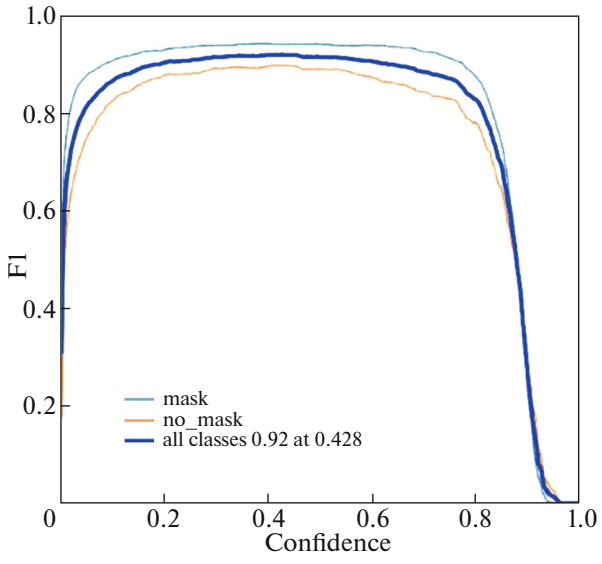


Fig. 4. $F_{\beta=1}$ versus confidence threshold on the Kaggle dataset.

Fig. 5. Example of superimposing face masks on images from the Wider Face dataset.

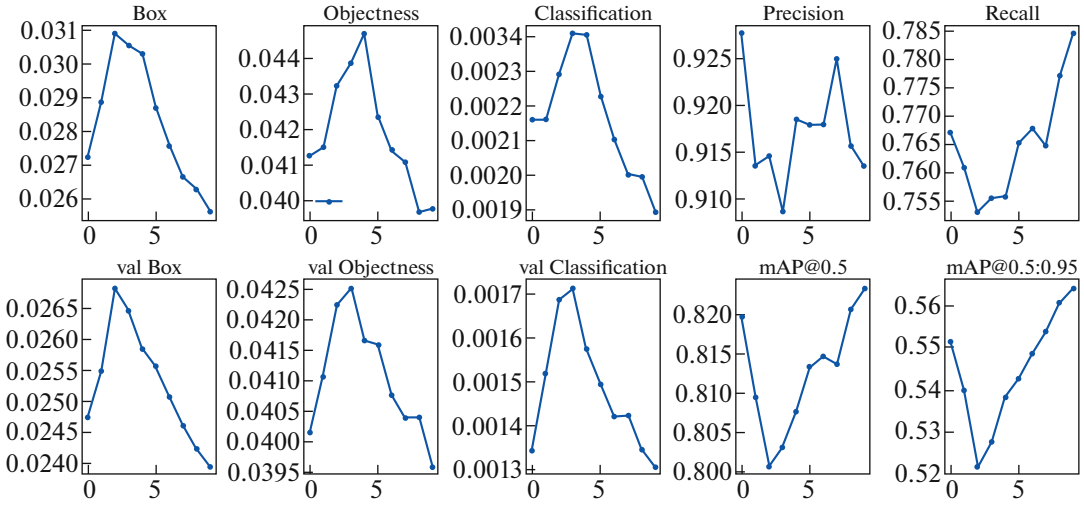


Fig. 6. Result of 10 training epochs on the modified Kaggle dataset.

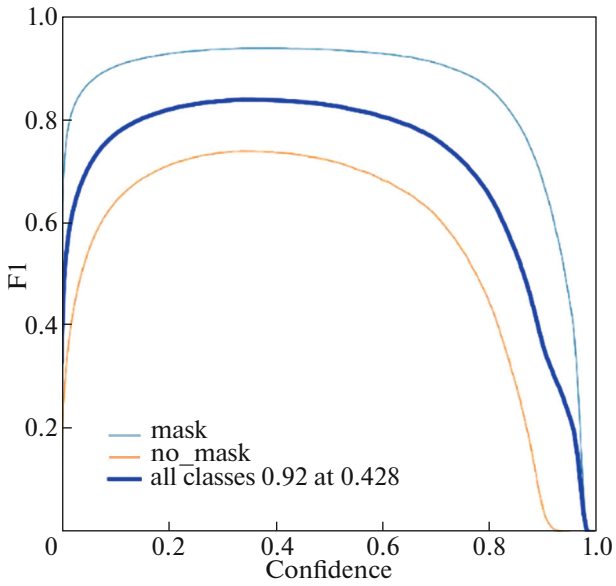


Fig. 7. $F_{\beta=1}$ versus confidence threshold on the modified Kaggle dataset.

4. DETECTION OF BEHAVIOR ANOMALIES

4.1. Motion Tracking and Analysis

To track people and construct individual motion trajectories, we propose a composite descriptor that includes the geometric and CNN features of the entire image of a person and its top portion, which are obtained using the CNN architecture of 29 convolutional layers and one fully connected layer [1]. The composite descriptor also includes a person index, which remains constant in subsequent frames if tracking is carried out properly. Indoors, a moving person can be overlapped by background objects; in this case, the features are computed for the upper body. Thus, the CNN features are computed for the entire body and for its upper half if the width of the detected object is less than its height; otherwise, the obtained CNN features characterize the upper body. Then, the composite descriptors of the tracked objects are matched with those detected in the current frame.

To improve tracking, the features for the last n correct detections are compared; then, the similarity

Fig. 8. Operation of the algorithm under different photography conditions.

matrix is constructed. This matrix is fed to the Hungarian algorithm, which is used to solve the assignment problem. For this purpose, a weight matrix is generated; its elements characterize the similarity between the descriptors of all objects detected in the current frame (the number of which corresponds to the number of columns) and those of the objects tracked from the previous frames (the number of which corresponds to the number of rows). The Hungarian algorithm subtracts the maximum similarity value from each element of the weight matrix so as to transform it into a matrix of optimal assignments, each row and each column of which contain only one zero. Their coordinates determine the correspondence between the objects in the frames. This allows us to correctly match the objects detected in the current frame with those tracked from the previous frames, as well as construct correct trajectories for them.

Each fragment of the trajectory is visualized by computing the shortest distance between the coordinates of a detected person $\left(x_{P_q}^{F_k}, y_{P_q}^{F_k}\right)$ in the current frame and the coordinates $\left(x_{P_q}^{F_{k-l}}, y_{P_q}^{F_{k-l}}\right)$ of this person in the $(k-l)$th frame for the last correct detection:

$$\mathrm{tr}_k = \sqrt{\left(x_{P_q}^{F_k} - x_{P_q}^{F_{k-l}}\right) + \left(y_{P_q}^{F_k} - y_{P_q}^{F_{k-l}}\right)}.$$

When a person is detected in the previous frame, $l = 1$. The use of this expression for each frame, starting from the one that follows the detection, allows us to construct and visualize the entire trajectory of this person.

The following characteristics of the trajectories can indicate behavior anomalies: the length of the trajectory can indicate that the person moves in a certain area for a long period of time (loitering), self-intersections of the trajectory also indicate loitering, a sudden change in the curvature indicates a possible fall, and an interruption (completion) of the trajectory indicates that the person stops.

It should be noted that, to detect anomalies of this type (loitering), long-term tracking is required to accumulate information about the movement of the object. Obviously, the proposed approach enables long-term tracking and trajectory computation; then, it is required to analyze the type of the trajectory.

A short-term analysis of the trajectory makes it possible to determine anomalies of another type, namely, falls. For this purpose, sudden changes in the speed and angle of movement in a local area are estimated. If a certain threshold is exceeded, then a decision is made to classify this event as a fall.

4.2. Tracking Experiments and Results

For testing, we used video sequences obtained from a stationary camera in rooms with different lighting conditions, nonlinear movement trajectories, complete and partial overlaps of objects, similar physical characteristics of people, appearance and disappearance of people within frames, etc.

The movement of the individuals with indices 1 and 2 in Fig. 9b is described by the trajectories that indicate the intersection of the same spot from different directions (short-term zigzags). This can be qualified as loitering. For the individual with index 3 (see Fig. 9b), it is difficult to determine the type of movement and further tracking is required. An example of a sudden change in the trajectory that indicates a fall is shown in Fig. 9d (see the person with index 7).

CONCLUSIONS

In this paper, we have proposed a formalization of the people detection and tracking problem, including the detection of appearance and behavior anomalies, in video sequences. Based on it, we have developed an

Fig. 9. People tracking examples with the visualization of trajectories to analyze the type of movement.

algorithm for detecting anomalous appearance and behavior of people that is based on tracking-by-detection and uses convolutional neural networks (CNNs) for people detection, face detection, and classification based on the mask-wearing criterion, and feature extraction in the process of tracking. To solve the first two problems, the YOLOv5 CNN with different weight coefficients has been employed. The classification of faces based on the mask-wearing criterion is significantly complicated by the lack of proper training datasets. In this connection, we have described a procedure for forming a large class-balanced dataset that can be used for CNN training. We have also described the process of training YOLOv5 to detect masked faces and have presented the corresponding test results, which confirm the high efficiency of the proposed approach.

In addition, we have proposed an algorithm for people tracking, as well as extraction and analysis of movement features. For people tracking, CNN features are included in the composite descriptor, which also contains geometric and color features, to describe each person detected in the frame.

We have also presented some examples of frames from processed video sequences with the results of face classification based on the mask-wearing criterion and visualization of individual trajectories, which make it possible to detect behavior anomalies based on long-term and short-term movements.

FUNDING

The work was supported in part by the Public Welfare Technology Applied Research Program of Zhejiang Province (LGF19F020016), the National High-End Foreign Experts Program (G2021016028L, G2021016002L, and G2021016001L) and Zhejiang Shuren University Basic Scientific Research Special Funds.

COMPLIANCE WITH ETHICAL STANDARDS

This article is a completely original work of its authors; it has not been published before and will not be sent to other publications until the PRIA Editorial Board decides not to accept it for publication.

Conflict of Interest

The authors declare that they have no conflicts of interest.

Ethical Approval

All procedures performed in the studies involving human participants were in accordance with the ethical standards of the institutional and/or national research com-

mittee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed Consent

Informed consent was obtained from all individual participants involved in the study.

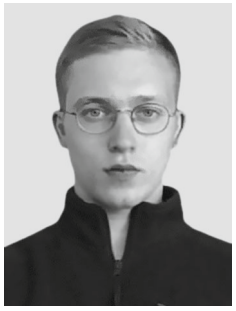
REFERENCES

1. R. Bohush and I. Zakharava, "Person tracking algorithm based on convolutional neural network for indoor video surveillance," Comput. Opt. **40**, 109–116 (2020). https://doi.org/10.18287/2412-6179-CO-565
2. T. Ganokratanaa, S. Aramvith, and N. Sebe, "Unsupervised anomaly detection and localization based on deep spatiotemporal translation network," IEEE Access **8**, 50312–50329 (2020). https://doi.org/10.1109/ACCESS.2020.2979869
3. M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016* (IEEE, 2016). https://doi.org/10.1109/CVPR.2016.86
4. E. Jardim, L. A. Thomaz, E. A. B. da Silva, and S. L. Netto, "Domain-transformable sparse representation for anomaly detection in moving-camera videos," IEEE Trans. Image Process. **29**, 1329–1343 (2020). https://doi.org/10.1109/TIP.2019.2940686
5. Kaggle: Data Science Platform. https://www.kaggle.com.
6. K. Kardas and N. K. Cicekli, "SVAS: Surveillance video analysis system," Expert Syst. Appl. **89**, 343–361 (2017). https://doi.org/10.1016/j.eswa.2017.07.051
7. D.-H. Lee, "Cascading denoising auto-encoder as a deep directed generative model" (2015). arXiv:1511.07118 [cs.LG]
8. Y. Li, Y. Cai, J. Liu, S. Lang, and X. Zhang, Spatio-temporal unity networking for video anomaly detection, IEEE Access **7**, 172425–172432 (2019). https://doi.org/10.1109/ACCESS.2019.2954540
9. W. Li, D. Zhang, M. Sun, Y. Yin, and Y. Shen, "Loitering detection based on trajectory analysis," in *8th Int. Conf. on Intelligent Computation Technology and Automation (ICICTA), Nanchang, China, 2015* (IEEE, 2015), pp. 530–533. https://doi.org/10.1109/ICICTA.2015.136
10. P. Nagrath, R. Jain, A.Madan, R. Arora, P. Kataria, and J. Hemanth, "SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2," Sustainable Cities Soc. **66**, 102692 (2021). https://doi.org/10.1016/j.scs.2020.102692
11. M. Ribeiro, A. E. Lazzaretti, and H.S. Lopes, "A study of deep convolutional auto-encoders for anomaly detection in videos," Pattern Recognit. Lett. **105**, 13–22 (2017). https://doi.org/10.1016/j.patrec.2017.07.016
12. M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayedd, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," Comput. Vision Image Understanding **172**, 88–97 (2018). https://doi.org/10.1016/j.cviu.2018.02.006
13. S. Sen and K. Sawant, "Face mask detection for covid_19 pandemic using pytorch in deep learning," IOP Conf. Ser.: Mater. Sci. Eng. **1070**, 012061 (2021). https://doi.org/10.1088/1757-899X/1070/1/012061
14. S. Sethia, M. Kathuria, and T. Kaushik, "Face mask detection using deep learning: An approach to reduce risk of Coronavirus spread," J. Biomed. Inf. **120**, 103848 (2021). https://doi.org/10.1016/j.jbi.2021.103848
15. S. Singh, U. Ahuja, M. Kumar, K. Kumar, and M. Sachdeva, "Face mask detection using YOLOv3 and faster R-CNN models: COVID-19 environment," Multimedia Tools Appl. **80**, 9753–19768 (2021). https://doi.org/10.1007/s11042-021-10711-8
16. L. Wang, F. Zhou, Z. Li, W. Zuo, and H. Tan, "Abnormal event detection in videos using hybrid spatio-temporal autoencoder," in *25th IEEE Int. Conf. on Image Processing (ICIP), Athens, 2018* (IEEE, 2018), pp. 2276–2280. https://doi.org/10.1109/ICIP.2018.8451070
17. Wider Face: A Face Detection Benchmark. https://shuoyang1213.me/WIDERFACE/.
18. M. Z. Zaheer, A. Mahmood, H. Shin, and S.-I. Lee, A self-reasoning framework for anomaly detection using video-level labels, IEEE Signal Process. Lett. **27**, 1705–1709 (2020). https://doi.org/10.1109/LSP.2020.3025688
19. Y. Zhang, X. Nie, R. He, M. Chen and Y. Yin, "Normality learning in multispace for video anomaly detection," IEEE Trans. Circuits Syst. Video Technol. **31**, 3694–3706 (2020). https://doi.org/10.1109/TCSVT.2020.3039798
20. W.K. Zhu, H.F. Li, J. S. Mu, G. H. Xue, and Z. Xu Dai, "An anomaly detection and correction method based on measurement representation model," Appl. Mech. Mater. **333–335**, 51–57 (2013). https://doi.org/10.4028/www.scientific.net/AMM.333-335.51

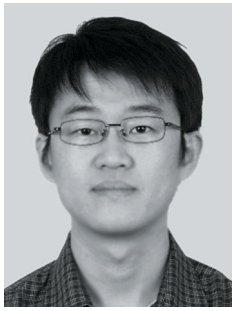
Translated by Yu. Kornienko

Huafeng Chen. Born in 1982. Associate Professor at Zhejiang Shuren University. Graduated from Zhejiang University in 2003. Received PhD degree in the field of earth exploration and information technology at the Institute of Space Information and Technique (Zhejiang University) in 2009. Scientific interests: remote sensing image processing, GIS application, image and video processing, and deep learning. Author of more than 20 papers.

Rykhard Bohush. Graduated from Polotsk State University in 1997. Received PhD degree in the field of information processing at the Institute of Engineering Cybernetics of the National Academy of Sciences of Belarus in 2002. Head of the Computer Systems and Networks Department of Polotsk State University. Scientific interests: image and video processing, object representation and recognition, intelligent systems, and machine learning.



Ivan Kurnosov. Born in 2001. Graduated from Belarusian State University in 2022. Software engineer, Leader of the Artificial Intelligence and Data Science Community at Exadel Inc. Bronze medal winner of the March Machine-Learning Mania 2021 – NCAWW competition on the Kaggle platform. Scientific interests: image classification and recognition, natural language processing, and segmentation.



Guangdi Ma. Born in 1985. Graduated from the Chinese Academy of Surveying and Mapping in 2011. Chief Engineer at EarthView Image Inc. Scientific interests: image analysis, photogrammetry, point cloud and oblique-photography-aided real 3D reconstruction.

Yang Weichen. Born in 1979. Graduated from Jilin University (China) in 2001. General manager at EarthView Image Inc. Scientific interests: image analysis, photogrammetry, and geographical information systems. Pioneered the business service mode of remote-sensing target-recognition to assist refined social governance in China.

Sergey Ablameyko. Born in 1956. Received Diploma in Mathematics in 1978, Candidate's degree in 1984, Doctoral degree in 1990, and Professor rank in 1992. Professor at Belarusian State University. Scientific interests: image analysis, pattern recognition, digital geometry, knowledge-based systems, geographical information systems, and medical imaging. Member of the Editorial Board of *Pattern Recognition and Image Analysis*, *Nonlinear Phenomena in Complex Systems*, and many other national and international journals. IAPR Fellow, AAIA Fellow, and Academician of the National Academy of Sciences of Belarus, European Academy, and many other academies. Honorary Professor at Moscow State University (Russia), Dalian University of Technology (China), and many other universities. Vice-President of the Asia-Pacific Artificial Intelligence Association.