Person Tracking and Reidentification for Multicamera Indoor Video Surveillance Systems

S. Ye*, R. P. Bohush**, H. Chen***, I. Yu. Zakharava****, and S. V. Ablameyko c,d,*****

a Zhejiang Shuren University, Hangzhou, 310015 China
b Polotsk State University, Novopolotsk, 211440 Belarus
c Belarusian State University, Minsk, 220030 Belarus
d United Institute for Informatics Problems, National Academy of Sciences of Belarus, Minsk, 220012 Belarus
* e-mail: zjsruysp@163.com
** e-mail: bogushr@mail.ru
*** e-mail: eric.hf.chen@hotmail.com
**** e-mail: i.zakharova@psu.by
***** e-mail: ablameyko@bsu.by

Abstract—For practical use, the relevance of indoor surveillance from multiple cameras to track the movement of people and reidentify them in video sequences is constantly increasing. This is a complex task due to the effect of uneven illumination, background inhomogeneity, overlap, uncertainty of the trajectories of people, and the similarity of their visual features. The article presents an approach to track people by video sequences and reidentify them in multicamera video surveillance systems that are used indoors. At the first step, people are detected using a YOLO v4 convolution neural network (CNN) and described by a rectangular area. Further, the search for the face area and the calculation of its features are carried out, which in the developed method are used when accompanying a person in a video sequence and during his intercamera reidentification. This approach improves the accuracy of tracking with a complex movement trajectory and multiple intersections of people with similar characteristics. The search for faces is carried out on the detected areas based on the multitasking MTCNN, and the MobileFaceNetwork model is used to form the vector of the features of the face. Human features are generated using a modified CNN based on ResNet34 and an HSV color tone channel histogram. The correspondence between people on different frames is established based on the analysis of the spatial coordinates of faces and people, as well as their CNN features, using the Hungarian algorithm. To ensure the accuracy of intercamera tracking, reidentification is performed based on the facial features. Five test video sequences of different numbers of people captured indoors with a fixed video camera were used to test and compare different approaches. The obtained experimental results confirmed the strength of the characteristics of the proposed approach.

Keywords: tracking people, face recognition, internal video surveillance, convolution neural networks

DOI: 10.1134/S1054661820040136

INTRODUCTION

Currently, the use of video surveillance systems is steadily increasing, which is explained by the wide range of tasks they solve and the continuously developing algorithms and hardware. A similar trend is also valid for the future due to the continuous and rapid improvement of the hardware base, including the increased resolution of video cameras of 4K, the increase in the bandwidth of communication channels, the introduction of 5G technology, the development and application of artificial intelligence methods of processing information (neural networks, genetic algorithms, fuzzy logic), technologies for processing large amounts of data (big data), cloud solutions (cloud technology), Internet of things (IoT), and the blockchain [1]. The most effective systems are spatially distributed video surveillance systems that involve the use of geographically dispersed IP cameras and are organized using a multiagent or central architecture. In such systems, the tracking of many objects in video sequences formed by one camera, and their further reidentification, are especially relevant [2]. The reidentification (intercamera tracking, reidentification) of people can be defined as the task of assigning the same name or index to all images of the same person, obtained from spatially distributed cameras, based on the selection and analysis of the features of his images [3].

Received June 23, 2020; revised June 23, 2020; accepted July 3, 2020
A monitoring system that uses the reidentification of people consists of two or more video cameras, the field of view of which does not overlap with each other. At the same time, a complex algorithm for the reidentification of people detects objects of interest and compares the found areas on video from different cameras at different times. The use of reidentification in spatially distributed video surveillance systems makes it possible to collect statistics on the number of unique entries of a person over a large area covered by several CCTV cameras. One of the main problems in the implementation of reidentification is the choice of a descriptor describing a person, which largely determines the effectiveness of solving this problem.

Based on tracking and reidentifying people, it is possible to implement various practical tasks: monitoring the movement of people and other objects in the Smart Home and Smart City systems, analyzing the environment in automated driving systems, assessing the correctness of the movement in medicine and sports, tracking objects in technical vision systems in production, recognition of the type of human activity in monitoring and security systems, etc. [4, 5].

However, this task is complicated to implement and requires the precise localization of objects in frames and correct identification in the current frame relative to the previous ones. Tracking and reidentifying many people in a room is also complicated by a number of factors: low light levels; heterogeneous background, fragments of which can be similar in shape, texture, and color to images of people; the presence of shadows and, accordingly, a changing background; multiple overlaps of people with each other and with other objects in the room; and the similarity of the features of different tracked people and their quite fast movement, as well as—in some cases—the changing acceleration and nonlinear transformation of the trajectory of movement.

Methods based on comparing the features of images of people and faces are used to track people. Today, tracking-by-detection is currently the most effective method. This approach uses an ensemble of an object detector and an algorithm to combine the detection results on two frames. An effective solution to the problem of combining allows us to correctly correlate the results of detecting various objects and form stable trajectories of movement for each of them. At present, classification algorithms using convolutional neural networks (CNNs) [5–7] have been widely developed and applied to detect objects, which are resistant to changes in illumination and a dynamic background, and they allow detection even in the case of partial overlaps, which increases the quality of the tracking. However, with the similarity of the characteristics of people and complex trajectories of movement, the effectiveness of their work is significantly reduced.

Tracking is also carried out using a class of algorithms that use the selection of the key points of a person and the formation of the set of features based on this, taking into account the distances between them, the relationship between distances, etc. In [8], the PoseNet CNN is proposed to highlight such key points. The algorithm is based on the analysis of the displacement of all points in space and changes in the ratio of the distances between them. Such algorithms are characterized by greater stability with the partial overlap of objects; however, they require significant computational costs.

Facial features are used to match people in frames. This makes it possible to increase the efficiency of tracking when analyzing the trajectories of people’s movement, when they are hidden for a long time behind background objects, and when the external features of people are similar. In the article [9], at the first stage, a background extraction algorithm is used, designed for video cameras with the function of measuring the depth of the image, which allows us to select moving candidate areas. Further, the HOG features are formed for them, which are classified using SVM to decide on the presence of a person in this area. Then, based on the HOG and SVM, the face is detected in this area. For the face area, the feature vector is calculated containing 128 values using the FaceNet’s CNN [10]. After normalization of the obtained descriptor, the persons in the database are matched those present in the frame with the assignment of names to the latter. To assess the quality of the algorithm, two video sequences of 162 and 218 frames, respectively, were used, in each of which six people were present. The article notes that a greater number of errors occur when a person’s face is hidden, for example, with a beard or for a long time its features cannot be distinguished. In addition, the proposed algorithm can only be used in systems equipped with a depth sensor.

To solve the problem of the reidentification of people captured by two different cameras, the SiameseNet CNN is considered in [11]. The model uses two branches of the same architecture that exchange features during the comparison process. The end result is a binary classification that decides whether two images are similar in input. In [12], the PartNet architecture is presented, which preliminarily selects the most informative parts of a person, then extracts high-level features from them using the CNN. This approach, when tracking people in a video obtained from various surveillance cameras, shows satisfactory results but requires large computing resources. The work [13] considers the reidentification of people using the analysis of the characteristics of a person and his face. A modified AlexNet CNN is used to isolate human fea-
A modified VGG-16 CNN is used to highlight the features of a human face. Both models produce a vector of 4096 features as a descriptor for a person and a face, respectively. In the training process, the ImageNet database, which is not adapted for the task of reidentification, was used. Further, the position of the face and the person in the frame, together with the features selected from the detected areas, are fed to the combining algorithm using the Markov chains. The accuracy of this algorithm based on the DukeMTMC database is estimated by the AUC metric: 92.07% when using facial features and 99.99% when applying human traits. The combination of features was not evaluated in this work. However, in the database used, there are few intersections of the trajectories of people and there are practically no long-term overlaps; i.e., there are only a few instances of the most difficult situations.

Despite the presence of a fairly large number of existing algorithmic solutions that can be used to track and reidentify people, the reliability of their tracking on a video sequence when using several cameras leaves much to be desired.

This paper proposes a new approach for tracking and reidentifying people in multicamera video surveillance systems using facial features. At the first step, people are detected using the YOLO v4 CNN and are described by a rectangular area. Further, the search for the area of the face and the calculation of its features are carried out. The search for faces is carried out on the detected areas based on the multitasking MTCNN, and the MobileFaceNetwork model is used to form the vector of the features of a face. Human features are generated using a modified CNN based on ResNet34 and an HSV color tone channel histogram. The correspondence between people in different frames is carried out based on the analysis of the spatial coordinates of faces and people, as well as their CNN features, using the Hungarian algorithm. This approach will improve the accuracy of tracking with a complex trajectory of movement and multiple intersections of people with similar characteristics.

1. SYSTEM MODEL

A spatially distributed indoor video surveillance system consists of geographically dispersed IP cameras that are installed in different rooms, and is organized, as a rule, based on a single data processing center. The use of surveillance cameras, which are mounted on the wall, provides a fairly large observation area. One camera can be used to monitor a room that is sufficiently large or a long corridor. Thus, in addition to algorithms for reidentification, which allows determining the movement of a person around the premises, it is necessary to use methods of tracking people on a video sequence formed by one camera.

Figure 1 shows a simplified structure of a video system with tracking and reidentification of people for three IP cameras. The object \( P_{C1}^{f} \) leaves the surveillance area of IP camera \( I \) (C1), while the moment of its exit is fixed at frame \( k \) of video sequence \( F \). This object goes into the surveillance area of IP camera 3 (C3); therefore, a reidentification procedure is necessary for it. For another person, \( P_{C2}^{f} \), it is necessary to track him during a period of time \( t \) in the video sequence formed by IP camera 2 (C2) and reidentify him when moving to the surveillance area of IP camera 3.

The proposed algorithm for tracking and reidentifying people in video sequences consists of the following main stages: people detection; formation of a vector of features for each person; detection and recognition of a face in the image area detected at the previous stage; identification of a person by face; in the first computation of the features of a face in the processed video sequence based on the features of a face, inter-camera reidentification is performed; establishing the correspondence between people in the frame; indexing people; determining their visibility in the frame; and highlighting a person with a frame when he is present in the frame.

During tracking, in order to establish a correspondence between people in the input frame and those being tracked, it is proposed to form a composite descriptor when describing each person (\( P_{ID} \)), which is represented as a vector of features. At the same time, the spatial and CNN features of a person’s face have been introduced into this descriptor, which are calculated and used for reidentification, and will also ensure more efficient tracking in the video sequence formed by one camera. The presence in the descriptor of spatial, CNN, and histogram features of a person’s image will provide correct tracking if it is impossible to identify him by face.

A composite descriptor of the form is proposed to describe a person during tracking and reidentification:

\[
P_{ID} = (f_{rec}^{FCNN}, Nm^p, Np_C^p, Np_f^p, dF, x_{det}^F, y_{det}^F, w_{det}^F, h_{det}^F, N_{det}^F, f_{PCNN}, f_{TopPCNN}, H_{det}^p, x_{det}^p, y_{det}^p, w_{det}^p, h_{det}^p),
\]

where \( f_{rec}^{FCNN} \) are the CNN features for the last correct face recognition of the tracked object; \( Nm^p \) is the name of the person; \( Np_C^p \) is the video camera’s index; \( Np_f^p \) is the index of the person in the video sequence; \( dF \) is the distance between the calculated features and the features of the image of the face from the database; \( x_{det}^F \) and \( y_{det}^F \) are the coordinates of the center of the face area in the frame during its previous detection;
and \( h_{\text{det}}^F \) are the width and height of the face area in the frame during its previous detection; \( N_{\text{det}}^F \) is the number of continuous face detection results in the absence of recognition; \( f_{\text{det}}^{PCNN} \) are the CNN’s features of the entire human figure at the last correct detection; \( f_{\text{det}}^{\text{TopPCNN}} \) are the CNN’s features of the upper half of the human figure at the last correct detection; \( H_{\text{det}}^P \) are the histogram’s features of a person at the last correct detection; \( x_{\text{det}}^P \) and \( y_{\text{det}}^P \) are the coordinates of the center of the area of a person in the frame during its previous detection; and \( w_{\text{det}}^P \) and \( h_{\text{det}}^P \) are the width and height of the person’s area in the frame during the previous detection.

To ensure the real-time mode of the complex task of detecting and tracking people, it is necessary for the first stage to use an accurate and fast CNN. Among the existing CNNs, the YOLO model is designed to reduce the computational processing costs, and its fourth version presented in 2020 [14] is also characterized by improved accuracy and speed in comparison with the previous version of this CNN. YOLO v4 achieves 94.8% accuracy in the top-5 metric. At the same time, the model supports a processing speed of about 65 frames per second on a Tesla V100 video card.

### 2. USE OF FACIAL FEATURES FOR REIDENTIFICATION AND TRACKING

The input data for these stages are the results of applying the detector based on YOLO v4 to the frames of the video sequence. Due to the fact that occasional false detections are sometimes possible, the detected objects are filtered by size.

The features of a person’s face in the approach under consideration are used for tracking and reidentification; therefore, the search of the face area is determined and scaled, the face area is detected and localized, the features are calculated, and the features of the face are compared. The selection of the face area is performed based on the analysis of the size of the

---

**Fig. 1.** Examples of assistance and reidentification of people in a distributed video surveillance system.
detected fragment. If its width is less than one-third of its height, then only the upper part of this fragment is analyzed, otherwise the entire area describing a person is analyzed. In this case, the search area is scaled to the dimensions of the input layer of 300 × 300.

In the proposed approach, to ensure the accuracy of the intercamera tracking, reidentification is performed only by the facial features, and consequently the accuracy of this stage is estimated by the parameters of the used CNN. The accurate calculation of the facial features ensures a high probability of a correct face recognition. One of the effective approaches to increase the effectiveness of face recognition is the use of the Additive Angular Margin Loss for Deep Face Recognition (ArcFace) function [15], which can be applied in training with almost any CNN architecture. Based on the data presented in [16], it can be seen that the highest accuracy is achieved for the LResNet100E-IR model, which is trained using ArcFace but it requires substantial computational costs, which imposes significant restrictions on processing video sequences. The MobileFaceNetwork model is characterized by significantly lower computational costs, while ensuring accuracy; for example, based on the LFW database, the accuracy is 99.5%; and for the LResNet100E-IR CNN, it is 99.77%. With this in mind, for face recognition, the MobileFaceNetwork architecture is used, which generates a vector of 128 features for a face and an input layer of [112 × 112] pixels.

To detect areas containing faces, a multitasking three-stage MTCNN model is used [17]: at the first stage, the candidate regions are detected; at the second stage, the false regions are suppressed; and at the third stage, the candidate regions are detected; at the second stage, the coordinates are refined and five key points of the face are detected. For this model based on the Wider Face database [18], an accuracy of 82% of correct detections is achieved, and at the same time, the computational costs are reduced by more than 50% compared to the RetinaFace model [19], which uses the ResNet100 architecture to achieve 92% accuracy.

For face identification the distances between the selected features in the current frame (\(f_{\text{curr}}^{\text{FCNN}}\)) and the facial features in the database (\(f_{\text{db}}^{\text{FCNN}}\)) are calculated:

\[
d^f_i = \frac{1}{128} \sum_{m=0}^{128} (f_{\text{db}}^{\text{FCNN}}(m) - f_{\text{curr}}^{\text{FCNN}}(m))^2.
\]

Among the calculated distances, the minimal distance is selected, and if it is greater than the predetermined threshold value, the recognition result is accepted as correct. With the correct face recognition from the database, the facial features (\(f_{\text{rec}}^{\text{FCNN}}\)), distance \(d^f\), and the person’s name (\(N^m\)) are updated in the composite descriptor. For reidentification, the features are compared with a database that contains the features of the faces of people from the neighboring cameras. If the obtained value is greater than the threshold value, a decision is made to move the person into the field of visibility of a neighboring video camera, and the value is updated in the composite descriptor (\(N_r^f\)).

In the case of the intersection of people in the treated area, several faces, which belong to different tracked objects, can be detected in one video sequence. Therefore, if the value of the intersection over the union (IoU) for faces is more than 0.8, then the similarity between the characteristics of the faces calculated by the CNN (\(f_{\text{curr}}^{\text{FCNN}}\)) and features (\(f_{\text{rec}}^{\text{FCNN}}\)) from the composite descriptors of the people being tracked is assessed. The maximum similarity among the calculated ones determines the correspondence of the face to the tracked object.

If a face is not recognized using the facial features from the database, then (\(f_{\text{curr}}^{\text{FCNN}}\)) and (\(f_{\text{rec}}^{\text{FCNN}}\)) are compared. If the threshold value for the obtained comparison result is exceeded, the recognition is considered positive and (\(f_{\text{rec}}^{\text{FCNN}}\)) is updated; otherwise, only (\(f_{\text{curr}}^{\text{FCNN}}\)) is updated.

To minimize the cases of a false identification of a person (\(P\)) after the overlapping of a set of people, the analysis of features (\(N_{\text{det}}^F\)) and IoU is used according to the following rule: if \((N_{\text{det}}^F < 5)\) and \(\text{IOU}(P_{\text{curr}}', P_{\text{det}}') > 0.3\) and \(\text{IOU}(P_{\text{det}}, P_{\text{curr}}) < 0.6\), then the person cannot be identified.

### 3. TRACKING BASED ON THE ANALYSIS OF IMAGES OF PEOPLE

In cases where faces cannot always be detected or recognized, tracking is performed based on an algorithm that includes an assessment of the presence of the entire human figure, the formation of CNN features for the entire area and for its upper part and their accumulation, the formation of spatial features and filtering by distance and size, calculating the similarity between all objects that were tracked and detected in the current frame, establishing a correspondence between them, indexing and naming people, determining their visibility in the frame, and highlighting a person with a frame when he is present in the frame. For the formation of human features, a CNN of 29 convolution and one fully connected layers is used, which forms a vector of 128 feature values for the input image [20].

When moving indoors, a person can go behind the background object; accordingly, the features will be calculated for the upper part of his figure. Therefore, the CNN features are calculated for the entire human
If a \((w_{\text{curr}}^p < h_{\text{curr}}^p)\), the similarity between the tracked people \((P_t)\) and the people found in the current frame \((P_{\text{curr}})\) is assessed based on the expression taking into account the CNN of the features for the last five detector results:

\[
d(P_t, P_{\text{curr}}) = \frac{\alpha}{N_{\text{det}}} \sum_{i=1}^{N_{\text{det}}} \left( \sum_{j=1}^{128} (f_{\text{curr}}^\text{PCNN} - f_{\text{det}}^\text{PCNN})^2 \right) + \beta \left( \sqrt{(x_{\text{curr}}^p - x_{\text{prev}}^p)^2} + \sqrt{(y_{\text{curr}}^p - y_{\text{prev}}^p)^2} + \sqrt{(w_{\text{curr}}^p - w_{\text{prev}}^p)^2} + \sqrt{(h_{\text{curr}}^p - h_{\text{prev}}^p)^2} \right),
\]

where \((f_{\text{curr}}^\text{PCNN}), (x_{\text{curr}}^p, y_{\text{curr}}^p), (w_{\text{curr}}^p),\) and \((h_{\text{curr}}^p)\) are the CNN features and coordinates of the center, height, and width of the object, respectively, in the input frame; \(\alpha\) and \(\beta\) are the correction factors.

Otherwise, the similarity score is calculated for the CNN of the features of the upper half of the human figure and the recalculating of the coordinates of the center and height.

Filtering by the distance and size of objects is provided to exclude errors due to the influence of people who may be similar according to the CNN features but are far from the tracked object.

As a result of calculating the distances \(d(P_t, P_{\text{curr}})\) for all the tracked people in the previous frames and detected objects, a similarity matrix is formed in the input frame, to which the Hungarian algorithm is applied, which solves the assignment problem [21]. As a result, the detected person in the current frame is assigned the name or index of the person being tracked. Moreover, the need for processing \(P_t\) not only from the previous frame but also from the earlier frames is a feature, since a short-term loss of optical communication between the camera and the person is possible. This is due to the fact that inside buildings the trajectories of the movement of people quite often intersect and the objects of interest overlap relative to the CCTV camera. This happens, for example, when people are talking or moving together. In addition, a person can have several points of entry and exit in the frame, partially or completely overlapped by other static objects. To reduce the likelihood of false indexing changes after complex cases of motion of objects with multiple overlaps, in a composite descriptor \(P_{ID}\), based on which continuous tracking is performed, only the coordinates of the object, its width, and its height are updated.

The correct operation of the detector is important for tracking-by-detection methods, because if a person is not found in one or more frames due to a false skip, then the features in the composite descriptor will not be updated. Accordingly, this can lead to a significant difference between the stored features and those calculated in subsequent frames and, accordingly, to indexing errors during tracking. In addition to this problem, the moment a person leaves the frame or is completely hidden behind another object to prevent being tracked has to be detected; and correct indexing is also relevant when a person reenters a frame after some time.

At the first step, to clarify the presence of a person in the frame, the algorithm uses search and face recognition for the area in which the detector had a positive result in the previous frame. If a face is found, then \((j_{\text{curr}}^\text{PCNN})\) and \((j_{\text{rec}}^\text{PCNN})\) are compared, and if the obtained value exceeds the threshold value, the person is considered to be present in the frame, and the area is highlighted with a frame.

If the face is not found, then the CNN features and the features of the color histograms of the areas of the person in the adjacent frames are analyzed. To reduce the impact of changes in the light, the images are converted from the color space RGB to HSV and only the data of the color tone is used to assess the similarity. A person is considered to be present in the frame if the following condition is met:

\[
d_{\text{PCNN}} \geq \varepsilon \quad \text{if} \quad R \leq \eta,
\]

where \(d_{\text{PCNN}} = \frac{1}{N_{\text{det}}} \sum_{i=1}^{N_{\text{det}}} \left( \sum_{j=1}^{128} (f_{\text{curr}}^\text{PCNN} - f_{\text{det}}^\text{PCNN})^2 \right)\); \(R\) is a measure of the similarity of the histogram features of the color tone for human images at the last correct detection and the current frame, which is calculated based on the Euclidean distance; and \(\varepsilon\) and \(\eta\) are the threshold levels, the values of which are the same as in [20].

4. EXPERIMENTS AND RESULTS

For testing, we used video sequences with a total number of frames 10250, shown in Fig. 2, which were obtained from a stationary video camera in rooms with different lighting, nonlinear trajectory of movement, full and partial overlap of people with similar external characteristics, people leaving the room followed by returning to the frame, etc.

The first video sequence (Fig. 2a) includes 2320 frames with a low image quality, uneven illumination, and the presence of a high level of shadows. The number of people in the frame varies from one to
Moreover, two of them have similar characteristics as their clothes are almost identical in color, and their height, physique, and hair color are also similar. The people in the shooting scene disperse and then intersect again with a multiple overlap; i.e., the trajectory of the movement is complex, and the location of faces relative to the camera and the poor quality does not allow identifying people in many of the frames.

Figure 2b shows the frames of the second video sequence of 1350 frames, in which there are two people whose image features are quite similar. This video is characterized by lower lighting, as well as nonuniform lighting and position of their faces relative to the camera. Thus, the low quality does not allow identifying people in many frames, as tracking is possible only based on the features of the people. In the third test video (Fig. 2c), there are two people who move around
the room along a complex trajectory, partially hiding behind the background objects, i.e., behind the board and tables, significantly moving away from the video camera. This test video consists of 1280 frames. The fourth video sequence (Fig. 2d) consists of 3450 frames, in which from one to three people are present. They have the following distinctive features: when moving, the lower half of a person’s figure is often hidden by a table, there is an intersection of trajectories, and there are multiple entries into and exits out of the premises of people in different order. Figure 2e shows examples of frames of the fifth video of 1850 frames, in which there are two or three people, and the features of the images of two of them are quite similar. The movement is carried out along a complex trajectory with a significant distance from the video camera and multiple overlap, when two people are hidden by a third person. Therefore, the location of faces at a considerable distance from the video camera with a low quality and a large angle of deviation from the facial profile does not allow identifying people in many of the frames.

The labelImg tool [22] was used to label the frames. In this case, the video sequence is divided into separate frames, after which objects are selected in them, and indices corresponding to the order in which people appear in the video are set as the object classes. Then, the resulting text files are combined to evaluate the tracking algorithm.

The developed approach for tracking and reidentification of people in spatially distributed video surveillance systems is implemented in C++ using the OpenCV 3.4 and dlib computer vision libraries. All processing procedures for detecting, tracking, and identifying people based on the CNN were carried out on a graphics processor using CUDA parallel processing technology.

The effectiveness of the proposed algorithm was compared with others (Table 1) using the main parameters that are widely used to assess the effectiveness of tracking objects [23]: IDF shows the percentage of correct identification of tracked objects; MOTA takes into account the number of false positive (FP), false negative (FN), and IDF results and characterizes the accuracy of tracking objects in time, taking into account the restoration of the trajectory in the short-term absence of the object; MOTP shows how accurately the object was localized in the frame during tracking without regard to detection.

The analysis of Table 1 shows that when processing test video sequences, the developed approach to tracking people using face identification allows improving the two main parameters: the accuracy of tracking objects in time, taking into account the restoration of the trajectory in the short-term absence of the object (MOTA = 91.6) and the accuracy of localizing the object in the frame during tracking without taking into account detection (MOTP = 83) compared to the
algorithms from [7, 20], which do not use the calculation and analysis of the features of human faces. The speed of the proposed algorithm on the computer used for testing with the characteristics of Intel Core i5-8600, 3.6 GHz, 16 Gb RAM, GPU Nvidia GTX 1060 with three people in the frame is 12 frames/s, which allows us to achieve real-time processing when analyzing each second frame of the input video.

For comparison, Fig. 3 shows the results of tracking objects in identical frames of a video sequence, in which there are three people and their trajectories of movement have multiple intersections, with one object completely overlapping the other. The color of the clothes of two persons is almost identical in color, and their physique and height are almost the same.

Figures 3a and 3d show a video frame containing three people with the correct indexing for the tracking algorithm without identification by faces (Fig. 3a) and with the correct indexing and identification (Fig. 3d). When the trajectories of their movement intersect for the proposed approach, the improvement in the processing result is obvious, since at the same time, two objects with indices 2 and three 3 are tracked (Fig. 3e), in contrast to the algorithm from [7], which tracks only one person (moreover, with the wrong index 3, Fig. 3b). After the further movement of people, this tracking algorithm assigns the wrong index 6 to the second object (Fig. 3c). The proposed algorithm allows us to accompany people with the correct indexes (Fig. 3f). The object with index 2, which had been hidden for a long time behind object 3 in Fig. 3f has no name, because their face was turned at an angle of 90° relative to the camera and it was not recognized.

CONCLUSIONS

An approach is proposed for tracking and reidentification by video sequences using the analysis of the features of human faces for spatially distributed indoor video surveillance systems. For this, people are detected; a vector of the features for each person is formed; detection and extraction of the facial features, as well as their recognition in the image area detected at the previous stage, are carried out; a person is reidentified by face; the correspondence between people in the frame is established; people are indexed; the tracking is made more precise; and the outline of a person present in the frame is highlighted. At the same time, the facial features are used when tracking a person in a video sequence and in the intercamera reidentification. To test the algorithm, five video sequences of different quality and with a different number of people were used. Based on them the characteristics of the developed algorithm for tracking people by video sequences are compared with the existing ones, which showed that the developed algorithm can improve the following points:

- the accuracy of tracking objects in time, taking into account the restoration of the trajectory in the short-term absence of the object;
- the accuracy of the localization of the object in the frame during tracking without taking into account the detection.

The proposed algorithm has a fast speed of operation, which makes it possible to achieve real-time processing when analyzing every second frame of the input video.

FUNDING

This work is supported by Public Welfare Technology Applied Research Program of Zhejiang Province (LGF19F020016, LGJ18F020001 and LGJ19F020002) and the National High-end Foreign Experts Program (GDW20183300463).

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1. P. W. Khan, Y.-Ch. Byun, and N. Park, “A data verification system for CCTV surveillance cameras using blockchain technology in smart cities,” Electronics 9 (3), 484 (2020).
2. L. Duan, Y. Lou, Sh. Wang, W. Gao, and Y. Ru, “AI oriented large-scale video management for smart city: Technologies, standards, and beyond,” IEEE MultiMedia 26 (2), 8–20 (2019).
3. R. Vezzani, D. Baltieri, and R. Cucchiara, “People reidentification in surveillance and forensics: A survey,” ACM Comput. Surv. 46 (2), 1–37 (2013).
4. X. Wang, “Intelligent multi-camera video surveillance: A review,” Pattern Recognit. Lett. 34 (1), 3–19 (2013).
5. R. P. Bohush and I. Y. Zakharava, “Person tracking algorithm based on convolutional neural network for indoor video surveillance,” Comput. Opt. 40 (1), 109–116 (2020).
6. D. Chahyati, M. I. Fanany, and A. M. Arymurthy, “Tracking people by detection using CNN features,” in Proc. of the 4th Information Systems International Conference (ISICO 2017) (Bali, 2017), pp. 167–172.
7. N. Wojke, A. Bewley, and D. Paulus, “Simple online and realtime tracking with a deep association metric,” in Proc. of the IEEE International Conference on Image Processing (ICIP2017) (Beijing, 2017), pp. 3645–3649.
8. U. Iqbal, A. Milan, and J. Gall, “PoseTrack: Joint multi-person pose estimation and tracking,” in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2017) (Honolulu, HI, 2017), pp. 4654–4663.
9. K. Koide, E. Menegatti, M. Carraro, M. Munaro, and J. Miura, “People tracking and re-identification by face recognition for RGB-D camera networks,” in Proc. of the European Conference on Mobile Robots (ECMR) (Paris, 2017), pp. 1–7.

10. F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Boston, Massachusetts, 2015), pp. 815–823.

11. R. Tao, E. Gavves, and A. W. M. Smeulders, “Siamese instance search for tracking,” in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Las Vegas, NV, 2016), pp. 1420–1429.

12. L. Zhao, X. Li, Y. Zhuang, and J. Wang, “Deeply-learned part-aligned representations for person re-identification,” in Proc. of the IEEE International Conference on Computer Vision (ICCV) (Venice, 2017), pp. 3239–3248.

13. N. Narayan, N. Sankaran, D. Arpit, K. Dantu, S. Setlur, and V. Govindaraju, “Person re-identification for improved multi-person multi-camera tracking by continuous entity association,” in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (Honolulu, HI, 2017), pp. 64–70.

14. A. Bochkovskiy, Ch.-Y. Wang, and H.-Y. M. Liao, YOLOv4: Optimal Speed and Accuracy of Object Detection (2020). https://arxiv.org/abs/2004.10934. Accessed June 12, 2020.

15. J. Deng, J. Guo, N. Xue, and S. Zafeiriou, ArcFace: Additive Angular Margin Loss for Deep Face Recognition (2019). https://arxiv.org/abs/1801.07698.pdf. Accessed May 20, 2020.

16. InsightFace Model Zoo. https://github.com/tzutalin/labelImg. Accessed June 16, 2019.

17. M. H. Ma and J. Wang, “Multi-view face detection and landmark localization based on MTCNN,” in Proc. of the Chinese Automation Congress (CAC) (Xi’an, 2018), pp. 4200–4205.

18. Sh. Yang, P. Luo, Ch. Change Loy, and X. Tang, WIDER FACE: A Face Detection Benchmark (2015). https://arxiv.org/pdf/1511.06523.pdf. Accessed May 20, 2020.

19. J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, RetinaFace: Single-stage Dense Face Localisation in the Wild (2019). https://arxiv.org/pdf/1905.00641.pdf. Accessed May 20, 2020.

20. R. Bohush and I. Zakharava, “Robust person tracking algorithm based on Convolutional Neural Network for indoor video surveillance,” in Pattern Recognition and Information Processing, PRIP 2019, Ed. by S. V. Ablameyko, V. V. Krasnoprosin, and M. M. Lukashevich; Commun. Comput. Inf. Sci. 1055, 289–300 (2019).

21. H. W. Kuhn, “The Hungarian method for the assignment problem,” Naval Res. Logist. Q. 2, 83–97 (1955).

22. LabelImg Is a Graphical Image Annotation Tool and Label Object Bounding Boxes in Images. https://github.com/tzutalin/labelImg. Accessed June 16, 2019.

23. B. Keni and R. Stiefelhagen, “Evaluating multiple object tracking performance: The CLEAR MOT metrics,” EURASIP J. Image Video Process. 1, 1–10 (2008).
Zakharava Iryna. Born 1995. Graduated from Polotsk State University in 2017. Received master of technical science degree in Mathematical Modeling, Numerical Methods, and Complexes of Programs in 2019. Studying for doctoral degree and working part-time as assistant in computer systems and networks department. Her scientific research is focused on image and video processing, scene understanding, and machine learning.

Sergey Ablameyko. Born in 1956. Received DipMath in 1978, doctoral degree in 1984, and doctor of science in 1990. Appointed professor in 1992 at Belarusian State University. Scientific interests: image analysis, pattern recognition, digital geometry, knowledge-based systems, geographical information systems, medical imaging. He is on the Editorial Board of Pattern Recognition Letters, Pattern Recognition and Image Analysis, and many other international and national journals. He is a senior member of IEEE, Fellow of IAPR, Fellow of Belarusian Engineering Academy, Academician of National Academy of Sciences of Belarus, Academician of the European Academy, and others. He was a First Vice-President of the International Association for Pattern Recognition (IAPR) (2006–2008) and President of the Belarusian Association for Image Analysis and Recognition.