Formalization of People and Crowd Detection and Tracking in Video

Rykhard Bohush
Computer systems and Networks Department
Polotsk State University
Novopolotsk, Belarus
ORCID 0000-0002-6609-5810

Huafeng Chen
College of Information Science and Technology
Zhejiang Shuren University
Hangzhou, China
ORCID 0000-0003-4229-4505

Sergey Ablameyko
United Institute of Informatics Problems of NAS of Belarus
Belarusian State University
Minsk, Belarus
ORCID 0000-0001-9404-1206

Abstract. **One of the promising areas of development and implementation of artificial intelligence is the automatic detection and tracking of moving objects in video sequence. The paper presents a formalization of the detection and tracking of people and crowd in video. The approach for tracking multiple people on video sequences for indoor and outdoor is described. The results of experiments for video sequences obtained using a stationary and moving video camera are presented.**

Keywords: **video surveillance, moving object, convolutional neural network, tracking by detection, motion trajectory**

I. INTRODUCTION

Detection and tracking of objects on video sequences are one of the main tasks in computer vision, which currently have a different number of technical applications and will be increasingly used for: analyzing the environment in automated systems of driving vehicles; assessing the movement of people in medicine and sports; tracking objects in industrial vision systems; recognizing the type of human activity in monitoring and security systems [1]. Unlike images, video sequences contain a much larger amount of information, which changes both in space and time. Therefore, processing and analyzing them allows to identify not only static, but also dynamic features of objects, which leads to an increase in the effectiveness of automated operation of video surveillance systems as a whole.

There are many object detection and tracking algorithms have been developed. In [2], an algorithm for tracking people on video based on the Monte Carlo method for Markov chains is proposed, in [3] – an algorithm for detecting and tracking people in intelligent emergency detectors based on the support vector machine method. Currently, algorithms based on convolutional neural networks, which are resistant to changes in illumination, dynamic background, and allow detection even in the case of significant overlaps of objects, are widely developed and used for object detection [4, 5].

Tracking a group of people is one of the most urgent tasks for video surveillance systems, but at present it is not fully solved. There are a number of approaches to solve this problem, however, due to these problems, the effectiveness of their work is insufficient. The stage of forming an effective set of features that will be used to detect and track objects in the video sequence is one of the most difficult, since there are restrictions for it: use features that can be obtained in advance to describe objects; determine a limited set of features that will allow to get the maximum effectiveness, i.e. it is necessary to exclude uninformative features; it is possible to apply algorithms that meet the computational requirements of the developed systems. We have developed a number of solutions that are designed to detect and track dynamic objects in video sequences [6], people [7], smoke and flames [8]. Accordingly, we can say that the set of features used is largely related to the detection and tracking algorithms used to solve the tasks set. Therefore, in order to develop effective methods, techniques, and algorithms for detecting and tracking objects on video sequences, it is necessary to clearly formalize these tasks. It is necessary to determine the objects that will be detected and tracked, to determine the main stages of this process, as well as criteria that allow to evaluate the quality of processing and show how this is implemented in practice.

In this paper, we propose a formalization of the tasks of detecting and tracking objects on video sequences. On the basis of the considered generalization, an algorithm for tracking a set of people and an algorithm for tracking crowd are developed. The results of experiments on the basis of the considered criteria, allowing to evaluate the quality of the algorithms, are presented.

II. FORMALIZATION OF PERSON DETECTION PROBLEM

A video sequence or video stream is a sequence of digital images (frames) $V = \{F_k\}$, k - the number of the image in the sequence. The object in the image (Ob) is a local area that differs from the surrounding

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

background and displays some of the features of the real-world object.

On each frame of the sequence obtained from a stationary video camera, as a rule, there are many objects: $OB_{F_k} = \{Ob_q^{F_k}\}, q = 1, \dots, Q$. According to the criterion of movement, each of them can be assigned to two main classes:

A stationary (stationary) object in a sequence of images is described by a set of features ($Ft_{Ob_q}^S$) and its coordinates $\left(x_{Ob_q}, y_{Ob_q}\right)$, which do not change during a time interval (t). Such an object can be represented by a formal model: $Ob_q^S = \left(Ft_{Ob_q}^S, x_{Ob_q}, y_{Ob_q}, Ns_{Ob_q}^{F_k}\right)$, where $\left(Ft_{Ob_q}, x_{Ob_q}, y_{Ob_q}\right) = const \forall F_k, k \in t$, $Ns_{Ob_q}^{F_k}$ – the set of possible noise effects on the object.

A moving object in a sequence of images is characterized by a change in one or more basic parameters: shape, size, and coordinates over a time interval (t). The transformation of the shape and / or size of an object leads to a change in its features in the frames $\left(ft_{Ob_q}^{F_k}\right)$. Such an object can be represented by a 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 $x_{Ob_q}^{F_k}, y_{Ob_q}^{F_k}$ - object coordinates; $Ft_{Ob_q}^D$ – a set of features of moving object, $Ft_{Ob_q}^D \supseteq ft_{Ob_q}^{F_k}, \forall k \in t$. Then $ft_{Ob_q}^{F_k} \cap \cap ft_{Ob_q}^{F_{k+i}}$, that is, for the same moving object on a sequence of frames, a change in its features is characteristic.

Object detection is the determination of the location of a given object Ob^e in the image F, while its size is smaller than the size of the image, and the number of objects in the image is obviously unknown. In general, the object Ob^e detection process is implemented by comparing the features of the reference (Ft^e) and all possible fragments on the image plane using the rule-based method:

$$S\left(Ft^e, Ft_{Ob_q}^F\right) \overset{M,Z}{-} \max,$$

where S - detection accuracy; Z - a set of restrictions.

When detecting a stationary object on a sequence of images, it is necessary to consider the variability of a dynamic scene, since in addition to static ones, there are moving objects on it, and their number can change. Objects $\left(Ob_q^D\right)$ can overlap a stationary object, which will cause its features to change on the sequence of frames, i.e.: $Ft_{Ob_q}^{S,F_k} \neq Ft_{Ob_q}^{S,F_{k+1}}$. Therefore, to detect a stationary object on a sequence of images displaying a dynamic scene, it is necessary to use a method M_{STV}, that takes into account the change in features over time:

$$S\left(Ft^e, Ft_{Ob_q}^{S,F_k}\right) \overset{M_{STV},Z_{STV}}{-} \max,$$

where Z_{STV} - a set of constraints when detecting a stationary object on a sequence of images.

Detection of moving object means to determine a location of the object Ob^D on the current frame F_k of the video sequence based on the specified F_e:

$$SD\left(Ob_{F_k}^D, Ob_{F_e}^D\right) \overset{MD,ZD}{-} \max,$$

where SD - the accuracy of detecting a moving object; MD - the method used; ZD - a set of restrictions.

III. Formalization of person and crowd tracking problem

Object tracking can be divided into three types:

- tracking of a single object (Visual object tracking, VOT);

- tracking of multiple objects (Multiple object tracking, MOT);

- tracking of crowds.

The first case of tracking is characterized by the fact that the object is detected and localized in the first frame, other objects are not detected.

A. Single person tracking

Tracking a moving object - determining the location of the same object on each frame of the video sequence during a time interval. This makes it possible to plot the trajectory of an object, determine its speed and acceleration. When solving practical problems, in some cases, an analysis of the trajectory of movement is required.

To perform maintenance, detection and localization procedures are required. Different ways of describing the objects of observation are used:

- a single point that characterizes the center of mass of the object or the center of the minimum possible rectangle described around the object;

- a set of key points by which the object can be uniquely identified in subsequent frames;

- a geometric primitive described around the object (most often a rectangle, less often an ellipse);

- the external contour of the object;

- a set of areas that are as stable as possible when moving, or the entire area of the object;

- invariant characteristics of the object (for example, texture, color scheme, etc.).

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

The trajectory of the object is a sequential display of the movement of this object on the video sequence: $Tr(Ob^D) = \left(Ob^D_{F_k}\right), \forall k \in t$.

In physics, the trajectory of motion is called the line that a particle describes when it moves. It is obvious that for the problem to be solved, in the end, it is necessary to determine a line on the required frame of the video sequence, which will show how the observed object moved over a certain period of time.

There may be different ways to determine the coordinates of an object on the frame, but the most commonly used approach involves finding its center (one pixel per frame) with coordinates $\left(x^{F_k}_{Ob_q}, y^{F_k}_{Ob_q}\right)$. As a rule, the movement is considered in the frame coordinate system. Then the trajectory of the object on the video sequence is described by a sequence in the form of a set of coordinates of the center of the object on each frame:

$$Tr(Ob^D) = \left(Ob^D_{F_k}\right) = \left(\left(x^{F_1}_{Ob_q}, y^{F_1}_{Ob_q}\right), \cdots, \left(x^{F_n}_{Ob_q}, y^{F_n}_{Ob_q}\right)\right).$$

The trajectory Tr' of its movement can be found using method MTS and constraints ZTS:

$$STS(Tr') \xrightarrow{MTS,ZTS} \max,$$

where STS - the accuracy of tracking a moving single object.

Despite the development of numerous algorithms over the past decade, due to the possibility of significant visual changes in the object and illumination, background noise, occlusions, the task of tracking a single object (VOT) is not fully completed.

B. Multiple person tracking

When tracking multiple objects, one need to determine the set of trajectories $TR' = \{Tr'_q\}$ of the objects in the frames, and then compare them to each other to determine the movement of all objects between frames:

$$STM(TR, TR') \xrightarrow{MTM,ZTM} \max.$$

Several objects can be present in the frame at the same time. Moreover, objects can have almost identical visual features. Thus, it is possible to lose an object that occurs due to its intersection with a similar one, or overlap with a background element. Tracking of multiple objects is performed at long time intervals, and it is possible to predict the location on subsequent frames.

C. Crowd tracking

Crowd is a large group of people with severe occlusions. Crowd tracking is to monitor state changing of crowds. Usually abnormally sudden change indicates emergency, for example, a crow splits into small groups could mean people run away from danger, two large crowds merge into one along with intensive motion could mean clash.

There may exist several crowds in one frame: $CR_{F_k} = \{Cr^{F_k}_q\}, q = 1, \dots, Q$, where $Cr^{F_k}_q = \{Ob^{F_k}_{q_i}\}, i = 1, \dots, n_q$, n_q is the number of people that compose crowd $Cr^{F_k}_q$. One thing deserves to be mentioned is that crowd does not keep its composition through time, it can split, or join together with other crowds.

Because of severe occlusions, single person $Ob^{F_k}_{q_i}$ in the crowd can hardly be detected or tracked. One common way is to treat crowd as a single entity and consider imaginary particles occupy the crowd area. Along with particles moving, crowd will reshape or regroup. It is possible to track for one crowd in a certain frame where its sub-groups go in next frames ($Cr^{F_k}_q = \{Sb^{F_{k+t}}_{q_i}\}, q_i = 1, \dots, l_q$, where $Sb^{F_{k+t}}_{q_i}$ is a separated sub-group of $Cr^{F_k}_q$ in F_{k+t}), or where its sub-groups came from ($Cr^{F_k}_q = \{Sb^{F_{k-t}}_{q_j}\}, q_j = 1, \dots, m_q$, where $Sb^{F_{k-t}}_{q_j}$ is a separated sub-group of $Cr^{F_k}_q$ in F_{k-t}).

Once sub-groups of a crowd is located in a previous frame or a posterior frame, further analysis of the crowd can be performed to determine whether certain crowd behavior happens. Three types of behavior are considered:

- crowd directional movement: many people move in the same direction;

- crowd aggregation: many people move toward a certain region from different directions;

- crowd dispersion: many people move away from a certain region in different directions.

One of the main problems for the practical use of the tracking algorithm is to ensure high accuracy with limited hardware resources and input data. In general, accurate tracking can be achieved by solving the global optimization problem, which requires the entire sequence of frames at once, which is impossible in real video surveillance systems. In the existing methods, the object tracking problem is often formulated as an optimization problem using graph algorithms [9]. Each detected object is represented as a vertex, and the transition from one vertex to another is determined by the similarity function used. Establishing an association on graphs can be solved by the method of finding the path with the minimum cost, which is most effectively solved by global optimization. Tracking algorithms based on the selection and analysis of special points require the presence of corners in the image contours. With a small number of them, the effectiveness of

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

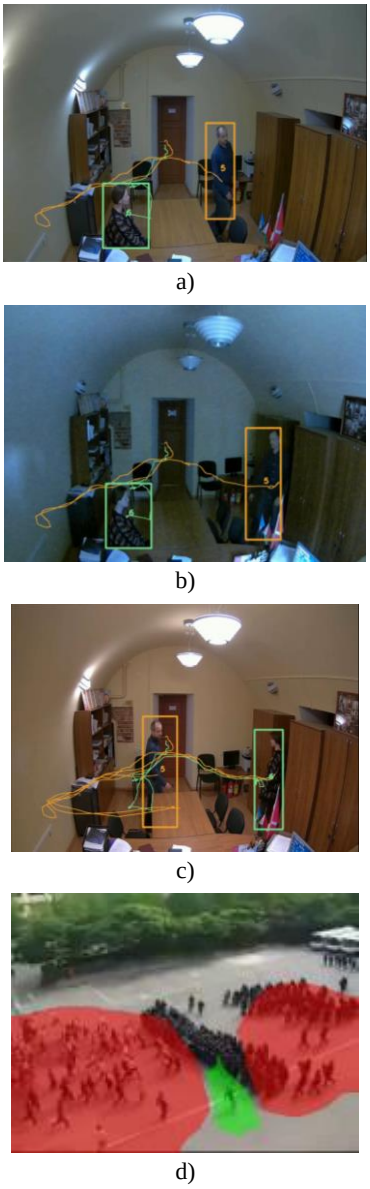
tracking will be low. The use of colour characteristics of objects for tracking is considered in the method from [10]. Probabilistic approaches use the statement that a moving object has a certain state, which is measured on each frame, and to estimate its position on the next one, it is necessary to generalize the values from the previous ones. For this purpose, methods based on the Kalman filter [11] or the particle filter [12] are used. However, objects can have a pronounced nonlinear trajectory of movement, and in this case, the assessment of the new state based on the previous ones will be determined with a high error. Therefore, different approaches are used to solve different applied problems.

IV. Experimental results

The proposed mathematical background have been tested in many applied tasks that have been described in our papers [6-8]. Tracking by detection method is effective for people tracking. In this case, the detection stage is one of the key ones. The quality of its work largely determines the accuracy of people tracking in video. Therefore, for what follows, we will use a more accurate CNN YOLOv4, the advantages of which are indicated in [13]. After person detection in the frame, features of the selected fragment in the spatial area in the frame and in time domain in video sequence are calculated. We use features such as CNN and histogram features of H channel in HSV space when this person was last correctly detected in frame, centre coordinates for selected area of a person in frame, displacement in the current frame relative to the previous one, width and height of the area in previous frame, motion trajectory, motion time. The values of the similarity function are calculated for all accompanied and detected people in the current frame. Based on these values, a correspondence is established between the detected and tracked objects using Hungarian algorithm. The trajectory is created when a person is first detected. The trajectory is deleted if this person is not detected for a certain number of consecutive frames and there is no comparison for him with previous frames. thus, in this case, we consider that he left the scene, which is recorded by a video camera.

An important task is to determine the effectiveness of tracking, taking into account the joint work of the detection and tracking stages to assess the possibility of practical use of the algorithm. Therefore, testing of the tracking algorithm for indoor surveillance and a modified algorithm for outdoor, which uses the Kalman filter, was carried out taking into account the results of human detection by CNN YOLOv4. In this case, errors in the operation of this CNN lead to a deterioration in the criteria for tracking. However, experiments reflect the real effectiveness for tracking by detection algorithms, which is very important for making a decision about their application in practice.

For evaluation of proposed approach MOT16 metric [14] is used. Experiments for indoor video sequences were carried out on six videos from a stationary surveillance camera. The total number of frames is 11890. Video are characterized by change in lighting, a nonlinear trajectory for people, overlap by background objects or the intersection of people trajectories, similar characteristics. The experiments performed have confirmed that our proposed approach improves the tracking accuracy on test video sequences obtained indoors and outdoors. To implement the algorithm from [15], MOTA = 0.288306 is provided for all videos from [14], and for our proposed MOTA = 0.300860. An improvement is also provided for video sequences from a fixed indoor surveillance camera: for the algorithm from [15] MOTA = 0.8793, for the proposed algorithm MOTA = 0.9266.



a) b) c) d)

Fig. 1. Results of mutiple people tracking and crowd behavior detection

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

For crowd tracking, algorithm based on integral optical flow and motion maps [16] is used. The key step of this algorithm is to analyze geometric structure formed by movements of crowds and their sub-groups, which involves considering speed and incline of the movements and density of crowds. Based on geometric structure analysis, certain crowd behaviors can be detected.

Fig. 1 show examples of detecting and tracking people and crowd. Example for indoor video surveillance with a changing lighting level and building trajectories of people movement is shown in Fig. 1a-c. Fig. 1d shows crowd behavior detection during a riot control exercise. In Fig. 1d, both red areas show crowd directional movements are happening. Furthermore, these two moving crowds from the left and the right are about to meet in the middle, the core area of the collision zone is painted green.

V. Conclusion

The paper presents a formalization of the problem of detecting stationary and moving objects on video sequences, taking into account their features. Three cases of tracking are considered and described: single object, multiple objects and crowds on video sequences. The approach for tracking multiple people on video sequences for indoor and outdoor video surveillance is described. The first stage requires detecting person in the input frames by YOLOv4 convolutional neural network. For solving assignment problem of person, we store information about individual object in spatial domain of frames and in the time domain on a video sequence. For person description, the following feature set is used: neural network and histogram features, centre coordinates of a person in the frame, offset in the current frame relative to the previous one, person width and height in the previous frame, trajectory and time of movement.

References

[1] Cavallaro and E. Maggio, Video tracking: theory and practice. Wiley, 2011.

[2] D. Kuplyakov, E. Shalnov and A. Konushin, “Markov chain Monte Carlo based video tracking algorithm,” Programming and Computer Software, vol. 43, pp. 224-229, July 2017. doi: 10.1134/s0361768817040053.

[3] A. P. Kirpichnikov, S. A. Lyasheva and M. P. Shleymovich, “Detection and tracking of people in intelligent detectors of emergency situations,” Kazan Technological University Bulletin, vol. 17, pp.351-356, 2014. (in Russian).

[4] Mohana and HV Ravish Aradhya, “Object Detection and Tracking using Deep Learning and Artificial Intelligence for Video Surveillance Applications,” Int. Journal of Advanced Computer Science and Applications, vol. 10, pp. 517-530, 2019. doi: 10.14569/IJACSA.2019.0101269.

[5] S. Mane and S. Mangale, “Moving Object Detection and Tracking Using Convolutional Neural Networks,” 2018 Second Int. Conf. on Intelligent Computing and Control Systems (ICICCS), Madurai, India, pp. 1809-1813, 2018. doi: 10.1109/ICCONS.2018.8662921.

[6] R. Bogush, S.Maltsev, A. Kastryuk, N. Brovko and D. Gluhau, “Motion Detection and Tracking Algorithms in Video Streams,” VNU Journal of Science, Mathematics - Physics, vol. 25, pp. 143-151, 2009.

[7] Bogush R. P., Zakharova I. Yu. and Ablameyko S. V., “Algorithm for person tracking on video sequences using face identification for indoor surveillance,” Herald of computer and information technologies, pp. 3-14, 2020. doi: 10.14489/vkit.2020.07.pp.003-014.

[8] S. Ye, B. Zhican, C. Chen, R. Bohush and S. Ablameyko, “An Effective Algorithm to Detect Both Smoke and Flame Using Color and Wavelet Analysis,” Pattern Recognition and Image Analysis, 2017, vol. 27, pp. 131-138. doi:10.1134/S1054661817010138.

[9] B.A. Zalesky and A.I. Kravchonok, “Tracking dynamical objects and their recognition by graph algorithms,” Informatics, vol. 2, pp. 7-26, 2006.

[10] B. A. Zalesky, “Object tracking algorithm by moving video camera,” Doklady of the National Academy of Sciences of Belarus, vol. 64, pp. 144-149, 2020. doi: 10.29235/1561-8323-2020-64-2-144-149.

[11] V.Yu. Agafonov, V.L. Rozaliev and A.V. Zaboleeva-Zotova, “Using the Kalman filter in object tracking tasks,” Intelligent systems. Theory and applications, vol. 20, pp. 13-17, 2016.

[12] F. Gustafsson, F. Gunnarsson and N. Bergman, “Particle Filters for Positioning, Navigation and Tracking,” IEEE Transactions on Signal Processing, vol. 50, pp. 425-437, Feb. 2002. doi: 10.1109/78.978396.

[13] A. Bochkovskiy, Ch.-Y. Wang and H.-Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” https://arxiv.org/abs/2004.10934. last accessed June 12, 2020.

[14] MOTChallenge: The Multiple Object Tracking Benchmark, https://motchallenge.net, last accessed April 14, 2021.

[15] Real-time Multi-person tracker using YOLO v3 and deep_sort with tensorflow, https://github.com/Qidian213/deep_sort_yolov3,last accessed Feb. 22, 2021.

[16] H. Chen, S. Ye, O. Nedzvedz, S. Ablameyko and Zhican Bai, “Motion Maps and Their Applications for Dynamic Object Monitoring,” Pattern Recognition and Image Analysis, vol. 29, pp. 131-143, April 2019.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.