Spatiotemporal Moving Crowd Tracking via Integral Optical Flow

Huafeng Chen[*1,2], Rui Tao[6], Hanjie Gu[*1], Shiping Ye[1,2], Rykhard Bohush[4], Ping Xu[1] and Sergey Ablameyko[3,5]
[1]Zhejiang Shuren University, Hangzhou 310015, China
[2]International Science and Technology Cooperation Base of Zhejiang Province: Remote Sensing Image Processing and Application, Hangzhou 310000, China
[3]Belarusian State University, Minsk, 220030, Republic of Belarus
[4]Polotsk State University, Novopolotsk, 211440, Republic of Belarus
[5]United Institute for Informatics Problems, National Academy of Sciences of Belarus, Minsk, 220012, Republic of Belarus
[6]Hangzhou Weizheng Intellectual Property Agency Co., Ltd, Hangzhou 310000, China
E-mail: eric.hf.chen@hotmail.com, tr0612@qq.com, guhanjie@zjsru.edu.cn, zjsruysp@163.com, bogushr@mail.ru, 66184750@qq.com, ablameyko@bsu.by
[*]Corresponding author

Keywords: moving crowd tracking, integral optical flow, crowd merging and splitting, ID management, motion pattern analysis

Received: September 9, 2025

This paper presents a novel approach for tracking moving crowds. Departing from conventional methods that focus on individual pedestrians, our method conceptualizes a moving crowd as a single, dynamically evolving entity. This entity can split into smaller sub-crowds or merge with others to form larger aggregations, making the approach particularly suitable for highly crowded scenarios. The proposed framework operates in two primary stages. First, moving crowds are detected using an integral optical flow technique, which accumulates optical flow vectors across consecutive video frames. Second, crowd identities are maintained via an ID management mechanism underpinned by a contribution matrix. This matrix records the contribution degree of detected crowds in the previous frame to those identified in the subsequent frame. The method is evaluated on manually annotated clips from three publicly available videos. The evaluation yields an average Multiple Object Tracking Accuracy (MOTA) of 0.361. Furthermore, the method demonstrates high performance in capturing crowd dynamics, with average precision and recall for crowd merging reaching 0.942 and 0.811, respectively, and for crowd splitting reaching 0.905 and 0.952, respectively. Additionally, the study defines internal motion patterns, referred to as "groups", within the moving crowds. These groups are identified based on local motion feature similarity and can be tracked in a manner analogous to the crowds themselves. Finally, several parameters are proposed, which hold potential for enabling more in-depth analysis of crowd movement behaviors.

Povzetek: Obravnavano je sledenje gibajočim se množicam z uporabo integralnega optičnega toka, kjer množico obravnava kot celoto. Predstavljena metoda omogoča zaznavanje združevanja, razdruževanja ter notranjih gibalnih vzorcev brez sledenja posameznikom.

1 Introduction

Crowd analysis represents a significant and challenging research domain within computer vision, encompassing critical tasks such as crowd density estimation, behavior recognition, and abnormal event detection. Several comprehensive reviews have effectively summarized the advancements in this field [1-3]. While some studies, like [4], attempt to detect and track individuals within a crowd, these methods often encounter limitations in high-density scenarios due to severe occlusion and substantial inter-object overlapping. Consequently, a holistic approach that treats the crowd as a single entity is frequently adopted to overcome these challenges.

Prevailing crowd analysis methods can be broadly categorized into optical flow-based techniques and deep learning approaches utilizing Convolutional Neural Networks (CNNs). For instance, Chen et al. [5] developed an end-to-end Crowd Attention Convolutional Neural Network (CAT-CNN) for accurate crowd counting. To enhance performance, Guo et al. [6] proposed a dual-CNN architecture, where one network generates density maps from crowd images, and another reconstructs the images from these maps, ensuring consistency. Similarly, Sharma et al. [7] introduced a unified CNN-based framework that integrates multi-scale information to simultaneously address crowd density estimation and behavior analysis, effectively handling scale variations.

As a classical and powerful tool for motion analysis, optical flow [8] has been widely applied to crowd dynamics. Nayan et al. [9] leveraged optical flow correlation analysis for anomaly detection in crowds. Drawing inspiration from fluid mechanics, Wang et al. [10] combined streaklines with a high-accuracy variational optical flow model for robust crowd behavior identification. Altalbi et al. [11] utilized optical flow to identify panic-induced distortions in crowd movements, while Zhang et al. [12] proposed a radar particle flow (RPF) method grounded in optical flow principles for crowd motion analysis. Furthermore, Bhuiyan et al. [13] demonstrated the effectiveness of fusing optical flow features with CNNs for abnormality detection in Hajj pilgrimage videos.

Building upon our previous work [14], which introduced the concept of integral optical flow and motion maps for categorizing three fundamental crowd behaviors, we further advanced a crowd tracking methodology [15]. This method tracks crowds frame-by-frame by calculating crowd centroids and measuring inter-frame correlations, visualizing overall crowd trajectories. In this paper, we extend our research by proposing a comprehensive framework for analyzing the evolution of moving crowds from a holistic perspective, based on motion information. The principal contributions of this work are threefold:

(1) We propose a novel method for detecting and tracking moving crowds as dynamic, evolving entities, eliminating the reliance on individual-level tracking. This approach offers a robust solution for analyzing crowd movements in high-density environments.

(2) We introduce definitions and detection mechanisms for distinct motion patterns within a crowd, providing a novel means to reveal and analyze internal crowd dynamics in detail.

(3) We define a set of quantitative crowd parameters to facilitate the future recognition and classification of crowd behavior patterns.

To evaluate the efficacy of our proposed crowd tracking method, we conducted experiments on three publicly available video clips, which were manually annotated to create evaluation datasets. Performance was assessed using established metrics, including precision, recall, MOTP, and MOTA, alongside four novel metrics specifically designed to measure the accuracy of crowd merging and splitting events. Experimental results indicate that the proposed method, despite being in its early developmental stages, demonstrates promising effectiveness.

2 Crowd detection and tracking

2.1 Crowd definition and detection

In the context of computer vision, a moving crowd specifically refers to such a gathering that exhibits collective motion, as opposed to a static crowd that remains stationary. For analytical purposes, moving crowds are often the primary focus due to their dynamic nature and associated challenges.

The accurate detection and tracking of individuals within a dense moving crowd present significant difficulty. Severe occlusions, where individuals are blocked from view by others, make it frequently impractical or even impossible to reliably detect and track every single person. Furthermore, a moving crowd is not a static entity; it evolves over time. Individuals may leave the crowd, others may join it, and the relative positions of people within the crowd can change constantly. These factors, among others, justify adopting a holistic perspective that treats the moving crowd as a single, evolving entity, rather than attempting to track its constituent individuals.

From a holistic viewpoint, a moving crowd can be conceptually decomposed into smaller groups of people and individual persons. A small group within a crowd can be characterized as a subset of individuals who are spatially proximate and may exhibit coordinated behavior or share a common goal.

Since our approach does not rely on detecting or tracking individual persons, we define the basic analytical unit based on distinguishable motion patterns. These basic units can be pixels, blocks of pixels, or other small image regions.

(1) Conditional connectivity

To formally define the spatial coherence of a crowd based on motion, we introduce the concept of conditional connectivity. For any two basic units P_a and P_b, if there exists a path $P_1, P_2, \dots, P_n$ connecting them (where $P_1 = P_a$ and $P_2 = P_b$), such that for every i $(1 \le i < n)$, units P_i and P_{i+1} are adjacent, and each unit $P_i (1 \le i \le n)$ satisfies a specific condition Con, then P_a and P_b are considered connected under the constraint Con, denoted as $P_a \overset{Con}{\longleftrightarrow} P_b$.

(2) Moving crowd detection

A moving crowd can be detected by analyzing the motion features of the basic units. In a two-dimensional image domain, the motion feature of a basic unit P_i can be represented as a vector $MF_i = (u, v)$, where u and v denote the horizontal and vertical components of displacement, respectively. A moving crowd is thus defined as a set of basic units that are conditionally connected based on a motion magnitude threshold. Formally, a crowd C is defined as:

$$C = \left\{ P_i \middle| P_k \overset{\|MF\|_2 \ge T_D}{\longleftrightarrow} P_l, 1 \le k, l \le N_P, k \ne l \right\}, i = 1, \dots, N_P, \quad (1)$$

where T_D is a predefined threshold for the magnitude of the motion vector (e.g., the L2 norm $\|MF\|_2$), and N_P is the total number of basic units. This set comprises all units interconnected through paths where each unit's motion magnitude meets or exceeds the threshold T_D.

2.2 Group determination

Within a moving crowd, individuals do not necessarily exhibit uniform motion. It is common to observe spatially connected groups of people moving in different directions, which may soon separate from one another. This observation indicates that multiple distinct motion patterns can coexist within a same moving crowd. From this perspective, a moving crowd can be conceptualized as

being composed of several groups. Within each group, individuals share a similar motion pattern. Formally, a moving crowd C can be represented as a collection of these groups:

$$C = \{G_i\}, i = 1, \cdots, N_G, \quad (2)$$

where N_G denotes the number of spatially separated motion patterns, which corresponds to the number of groups.

Mathematically, a group within a moving crowd is defined as a subset of the crowd, comprising basic units that are connected under a specific condition. To quantify the dissimilarity between the movements of two basic units, a distance function is defined based on their motion features (MF):

$$dis\left(MF_{i_1}, MF_{i_2}\right) = \left\|MF_{i_1} - MF_{i_2}\right\|_2, \quad (3)$$

where $MF = (u, v)$ represents the motion vector. If the distance $dis\left(MF_{i_1}, MF_{i_2}\right)$ is less than or equal to a predefined threshold T_{MF}, the two basic units P_{i_1} and P_{i_2} are considered to have similar motion patterns.

In certain scenarios, such as when two units are located far apart on a curved path, their instantaneous motion vectors may differ significantly even if they belong to the same group. To prevent the over-segmentation of groups in such cases, membership is determined not only by the direct similarity between units but also by the consistency with their local neighborhood. Specifically, a basic unit is assigned to a group if its motion feature is similar to the average motion feature of its neighboring units. Therefore, a group G is formally defined as:

$$G = \left\{P_i \middle| P_k \overset{Con}{\longleftrightarrow} P_l,\ 1 \le k, l \le N_P, k \ne l, P_i \in C\right\}, i = 1, \cdots, N_P, \quad (4)$$

where the connectivity condition Con requires that $\|MF_k - \overline{MF_k}\|_2 \le T_{MF}$ and $\|MF_l - \overline{MF_l}\|_2 \le T_{MF}$. Here, N_P is the number of basic units, C is the overarching crowd, and $\overline{MF_k}$ and $\overline{MF_l}$ represent the average motion features of the neighbors of units P_k and P_l, respectively.

2.3 Identity management

A moving object, whether it is a crowd as a whole or an internal small group, undergoes continuous evolution over time in a video sequence. In surveillance videos, this evolution is often manifested through events such as individuals splitting from the principal group or multiple subgroups gathering together. Effectively tracking a crowd or a group thus necessitates addressing three key aspects:

(1) Does the object continue to exist in the subsequent frame?

(2) Does any part of the object separate from its principal region (i.e., the largest contiguous area)?

(3) Do any separated parts of the object, or the principal part itself, merge with parts from other distinct objects?

To manage the identities of these evolving objects, we assign a unique identifier (ID) to each detected moving object. The core of the tracking process lies in the inheritance of these ID numbers across frames, which is determined based on the spatial correlation between objects in consecutive frames.

Suppose $\left\{Ob_1^i, Ob_2^i \ldots, Ob_{N_i}^i\right\}$ represents the set of all N_i moving objects detected in frame i. As the scene progresses to a later frame j, these objects may fragment and regroup. New objects may also appear. The resulting set of objects in frame j is $\left\{Ob_1^j, Ob_2^j \ldots, Ob_{N_j}^j\right\}$. To quantify the relationship between objects across these frames, a contribution matrix $C_{N_i \times N_j}$ is computed. Each element C_{kl} in this matrix denotes the unit number that fragment originating from object Ob_k^i in frame i contribute to the formation of object Ob_l^j in frame j.

The ID management procedure for each new frame (after the first) involves the following three steps:

Step 1: Primary inheritance.

Identify all pairs (k, l) for which the contribution C_{kl} is simultaneously the maximum value in its column $\max(\{C_{\cdot l}\})$ AND the maximum value in its row $\max(\{C_{k\cdot}\})$. For each such pair, object Ob_l^j inherits the ID number from object Ob_k^i. Each ID number from frame i can be inherited at most once.

Step 2: Secondary assignment.

For any Ob_l^j in frame j that has not yet received an ID number via Step 1, find the object Ob_k^i in frame i for which C_{kl} is the maximum value in column l (i.e., $\max(\{C_{\cdot l}\})$). If the ID number of Ob_k^i has not yet been inherited by any object in frame j, then assign this ID to Ob_l^j.

Step 3: New ID assignment.

Assign a new unique ID number for each remaining object. Any object Ob_l^j in frame j that remains without an ID number after Steps 1 and 2 is assigned a brand new, unique ID number.

An object from frame i is considered to have ceased existence if its ID number is not inherited by any object in frame j.

3 Crowd parameters

Once moving crowds and their internal groups are detected, a set of quantitative parameters can be derived to characterize their dynamic properties. This section defines several parameters proposed for the subsequent analysis of crowd behavior. These parameters are designed to provide theoretical insights into crowd motion patterns.

(1) Centroid and trajectory

The centroid of a crowd or a group represents its geometric center, calculated as the average coordinates of all its constituent basic units. The centroid coordinates $(\bar{x}, \bar{y})$ are computed as follows:

$$Ct = (\bar{x}, \bar{y}) = \left(\frac{1}{N}\sum_{j=1}^{N} x_j, \frac{1}{N}\sum_{j=1}^{N} y_j\right), \quad (5)$$

where N is the total number of basic units within the crowd or group, and (x_j, y_j) are the coordinates of the j-th unit.

The trajectory of a crowd or group over its lifetime is defined as the temporal sequence of its centroid positions across consecutive frames. This trajectory is denoted as

$(Ct_1, Ct_2, \ldots, Ct_n)$, where n is the number of frames during which the entity exists.

(2) Motion intensity

Motion intensity quantifies the overall magnitude of movement within a crowd or group between two consecutive frames. It is defined as the average magnitude of the displacement vectors of all its basic units:

$$Mi = \frac{1}{N}\sum_{j=1}^{N}\left\|d_j\right\|_2, \quad (6)$$

where N is the number of basic units, d_j is the displacement vector of the j-th unit, and $\|\cdot\|_2$ denotes the L2 norm (magnitude) of the vector.

(3) Crowd movement homogeneity

This parameter measures the degree to which the motion within a crowd is uniform. A crowd comprising a single group typically exhibits high homogeneity, as all units share a similar motion pattern. In contrast, a crowd containing multiple groups with divergent motion patterns displays lower homogeneity and appears more chaotic. Homogeneity Mh is calculated as the ratio of the size of the principal group (the largest group within the crowd) to the total size of the crowd:

$$Mh = \frac{N_p}{N}, \quad (7)$$

where N_p is the number of basic units in the principal group, N is the total number of units in the crowd. The value of Mh lies in the interval $(0,1]$. A higher value indicates a more homogeneous crowd movement.

(4) Group movement consistency

While the motion within a group may be homogeneous (spatially similar), it is not necessarily consistent in terms of direction. Movement consistency characterizes the alignment of motion vectors within a group. For instance, units moving in a straight line at similar speeds exhibit high consistency, whereas units following a curved path or moving in a circular formation may have opposing directional vectors, leading to low consistency. Consistency Mc is defined as the ratio of the magnitude of the average displacement vector to the motion intensity:

$$Mc = \frac{\left\|\frac{1}{N}\sum_{j=1}^{N} d_j\right\|_2}{Mi}, \quad (8)$$

where N is the number of units, d_j is the displacement vector of j-th unit, and Mi is the motion intensity as defined in Equation (6). The value of Mc falls within $[0,1]$. A value closer to 1 indicates highly consistent (directional) movement, while a value closer to 0 suggests that individual motions cancel each other out due to a lack of directional alignment.

4 Algorithm for crowd analysis

4.1 Optical flow

Optical flow provides a fundamental technique for analyzing the apparent motion of pixels between consecutive frames in a video sequence, forming the basis for motion analysis in dynamic scenes. Numerous methods have been developed for computing optical flow, which can be broadly categorized into different classes, such as gradient-based, matching-based, energy-based, and phase-based methods, depending on their underlying theoretical foundations and mathematical formulations. In this work, we employ the method detailed in [16] to compute dense optical flow, which estimates the motion vector for every pixel in the frame, as opposed to sparse methods that only track a limited set of feature points.

A standard optical flow field, denoted as OF_t for frame t, captures the displacement vector of each pixel between two consecutive frames. However, due to the extremely short time interval (Δt) involved, the displacement magnitude of moving foreground objects (e.g., people) is often comparable to the inherent, random motion of the background (e.g., slight camera jitter or environmental noise). This makes it challenging to robustly distinguish foreground from background based on a single two-frame optical flow calculation. While background motion often appears random (e.g., small back-and-forth or circular movements) over short periods, this characteristic is not discernible instantaneously. Over a sufficiently long duration, however, the random nature of background motion causes its accumulated displacement vectors to remain small, whereas the consistent motion of foreground objects leads to steadily growing displacement vectors.

To leverage this temporal characteristic, we introduce the concept of Integral Optical Flow (IOF). The core idea is intuitive: instead of relying on the optical flow between two frames, we accumulate the optical flow vectors over a series of consecutive frames. This accumulation amplifies the motion signals of consistently moving foreground objects while suppressing the random noise associated with background motion.

For formal description, let I_t denote the t-th frame of a video sequence I , and $I_t(p)$ denote the pixel at coordinates p in that frame. Let OF_t represent the basic optical flow field computed between frame t and $t+1$. The Integral Optical Flow for frame t over an interval of itv frames is denoted as IOF_t^{itv}. This IOF_t^{itv} is a vector field that records the accumulated displacement information for all pixels in I_t over a period of itv frames. For any pixel $I_t(p)$, its integral optical flow vector $IOF_t^{itv}(p)$ is computed as follows:

$$IOF_t^{itv}(p) = \sum_{i=0}^{itv-1} OF_{t+i}(p_{t+i}), \quad (9)$$

where the path of the pixel is tracked recursively:

$$\begin{cases} p_{t+0} = p, \\ p_{t+i} = p_{t+(i-1)} + OF_{t+(i-1)}\left(p_{t+(i-1)}\right), i > 0. \end{cases}$$

This formulation ensures that the integral optical flow at a starting point p in frame t is the vector sum of the displacements along the trajectory that the pixel (or the scene point it represents) follows through the subsequent itv frames. This accumulated vector provides a more robust and significant motion measure for detecting coherently moving regions like crowds.

4.2 Moving crowd tracking flow

The overall procedure for tracking moving crowds is executed according to the workflow illustrated in Figure 1. The process involves distinct computational steps for different types of frames within a video sequence. For every frame except the final one in the sequence, the basic

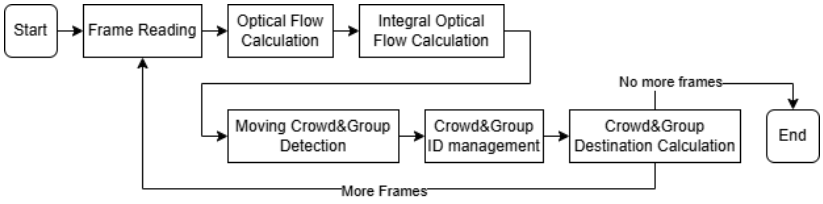

Figure 1: Moving crowd tracking flow.

optical flow between it and its subsequent frame is computed. Additionally, for frames at positions defined by $1 + n \times itv$ (where $n \geq 0$), the integral optical flow is calculated over an interval of itv frames, as defined in Equation (9) of Section 4.1. The parameter itv is a key frame interval parameter that determines the temporal scope for motion accumulation.

The tracking pipeline initializes with the first frame of the sequence ($n = 0$). Once its integral optical flow is computed, the algorithm proceeds to detect all moving crowds and the distinct motion patterns (groups) within them. Each detected crowd and internal group is then assigned a unique identifier (ID number), starting from 1 and sequentially incremented. This establishes the initial state of the tracking system.

For all subsequent frames where integral optical flow is calculated, specifically at frames $1 + n \times itv$ (with $n \geq 1$), the tracking system updates the identities of the evolving entities. This step involves managing the inheritance of existing ID numbers by crowds and groups that persist from the previous integral optical flow frame. Simultaneously, newly emerged moving crowds and motion patterns that were not present before are assigned new, unique ID numbers. The specific rules governing this ID inheritance and assignment are detailed in Section 2.3 (Identity Management), which ensures consistent tracking across temporal evolution.

A critical aspect handled at each integral optical flow frame is the prediction of fragment destinations. As crowds and groups evolve, they may split into spatially separated fragments. Based on the computed integral optical flow vectors, the algorithm predicts the future positions of these fragments. This prediction is essential for correctly associating fragments in the current frame with the reconstituted or merged entities they will form in subsequent frames, thereby maintaining holistic tracking of dynamically evolving objects.

5 Experimental results

This section presents a comprehensive evaluation of the proposed crowd analysis framework using manually annotated real-world video sequences. To visually demonstrate the performance of our method, the experimental results are illustrated with figures that delineate the detected moving crowds and their internal motion patterns. Specifically, the outer boundaries of moving crowds are explicitly outlined, while distinct motion patterns (groups) within each crowd are highlighted using transparent masks. For clear identification in these visualizations, a unique identifier is assigned to each entity: moving crowds are prefixed with the symbol #, and internal motion patterns are prefixed with the symbol *.

5.1 In-door moving crowd tracking

The performance of the proposed tracking algorithm in an indoor environment is demonstrated in Figure 2. The test sequence is a surveillance video capturing a large hall where numerous pedestrians walk towards various destinations, with a minority making brief stops. The video comprises 201 frames, each with a resolution of 856×568 pixels. The key parameters for the algorithm were set as follows: the frame interval (itv) for integral optical flow calculation was 10 frames; the displacement magnitude threshold (T_D) for crowd detection was 8 pixels; and the motion feature difference threshold (T_{MF}) for group determination was 1.5 pixels. Furthermore, to filter out noise and very small moving objects, any detected entity with an area of less than 500 pixels was disregarded in the analysis.

Figures 2a, 2b, and 2c visually present the detection and tracking results at frames 11, 21, and 31, respectively. These subfigures illustrate the original video frames overlaid with the detected crowd boundaries, internal motion patterns (groups) highlighted with masks, and their corresponding unique identifiers (crowds prefixed with #, groups with *). For enhanced clarity in observing the spatiotemporal evolution of the crowds, Figures 2d, 2e, and 2f show the corresponding simplified representations at the same frames, displaying only the outer boundaries and the ID numbers of the tracked crowds.

A detailed analysis of the sequence reveals the dynamic nature of crowd movement and interaction:

- **At Frame 11 (Fig. 2a & 2d)**: In the bottom-left corner, two distinct crowds are observed: crowd #22 is moving predominantly upwards, while crowd #17 is moving downwards. Crowd #13 exhibits more complex internal dynamics; individuals in the center of this crowd are moving rightwards, while those on the periphery are moving upwards. This heterogeneity in motion vectors within #13 leads to the detection of six distinct internal motion patterns (*1 to *6). The segmentation within crowd #10 provides an intuitive example of the group determination logic. Although all four individuals are moving in a generally upward direction, slight variations in their precise

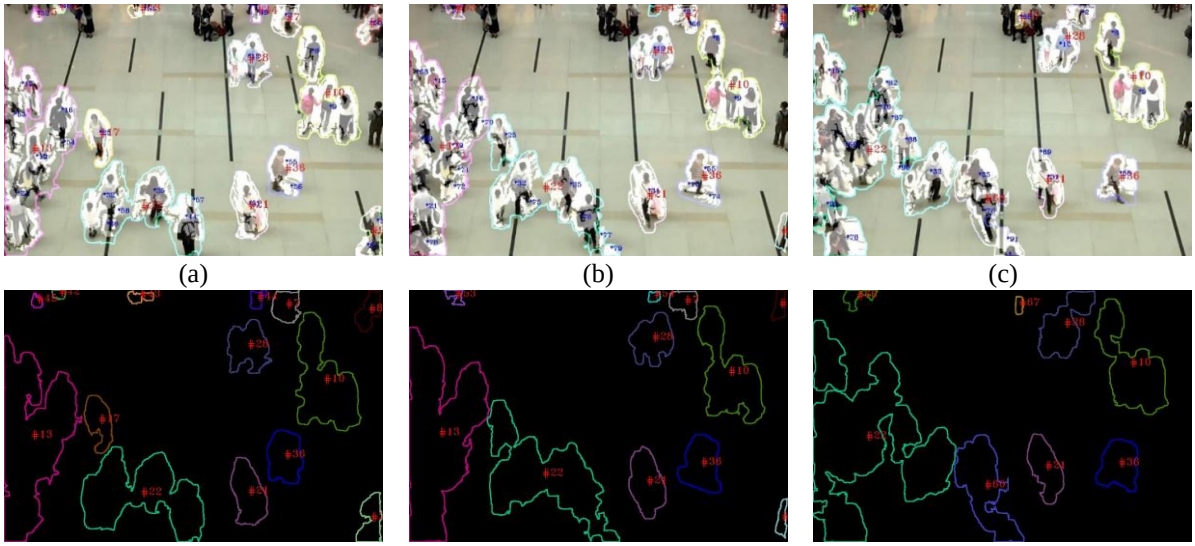

Figure 3: In-door moving crowd tracking results.

motion vectors cause the algorithm to distinguish the person positioned slightly higher (assigned to group *6) from the cluster of three below (assigned to group *9).

- **At Frame 21 (Fig. 2b & 2e)**: A significant merging event occurs between crowds #17 and #22. Following the ID inheritance rules defined in Section 2.3, the merged entity retains the ID #22 of the larger original crowd (#22) which contributed the greater area to the new formation.
- **At Frame 31 (Fig. 2c & 2f)**: The previously merged crowd #22 undergoes a splitting event. A portion of its constituents breaks away to form a new, independent crowd assigned a new ID (#60). The remainder of the original #22 merges with the existing crowd #13 and some newly appeared individuals. This combined entity forms a single, larger crowd. Consistent with the ID management protocol, this new crowd inherits the ID #22 from its largest constituent part. Furthermore, within the evolving scene, the motion pattern *69 (located towards the middle-left) is noteworthy. The individuals within this group are moving to the right, creating a clear motion contrast with the surrounding groups which are exhibiting different movement patterns.

5.2 Out-door moving crowd tracking

The proposed method is further evaluated using an outdoor video sequence depicting a dynamic riot control scenario, as illustrated in Figure 3. The video captures a simulated confrontation: South Korean police forces are lined up on the right side, while a group of rioter's charges from the left. This sequence, comprising 168 frames with a resolution of 480×360 pixels, presents a challenging environment with rapid movements and complex interactions between two opposing groups. The algorithm parameters were configured as follows: the frame interval (itv) for integral optical flow was set to 10 frames; the displacement threshold (T_D) was 10 pixels; and the displacement difference threshold (T_{MF}) was 1. To filter out noise, any detected moving object with an area smaller than 300 pixels was ignored.

5.2.1 Initial Deployment and police maneuvers (Frames 11 & 21)

Figures 3a and 3b present the detailed tracking results for frames 11 and 21, respectively, while Figures 3g and 3h provide the corresponding simplified views showing only crowd boundaries and their IDs. In the initial phase, the police line is the primary moving entity. The analysis shows that police officers at both ends of the line advance more quickly. As they move, their motion patterns become increasingly complex and are segmented into distinct groups by the algorithm, demonstrating its sensitivity to variations in velocity and direction within a seemingly cohesive line.

5.2.2 Engagement: rioters' charge and police response (frames 81 & 91)

The scenario intensifies at frames 81 and 91, shown in Figures 3c/3d (detailed results) and 3i/3j (simplified boundaries). The rioters on the left emerge and begin a rapid charge towards the police line. Meanwhile, the police response evolves dynamically; some officers who had previously moved out from the right flank have already established new positions, while others are still in transit. Concurrently, parts of the original police queue begin to advance forward, creating a multi-directional and

multi-group movement landscape that the algorithm successfully tracks.

5.2.3 Climax and ID inheritance: merger and identity transition (frames 101 & 141)

The climax of the interaction is captured in frames 101 and 141, displayed in Figures 3e/3f and 3k/3l. The rioters continue their advance, eventually meeting the police in the middle of the scene. A critical observation from Figure 3e is the merging of police subgroups. Specifically, the police officers originally marked as motion patterns *24 and *35 in the previous frame (Fig. 3d) merge into a single pattern, *24, and subsequently combine with other still-moving police to form the larger crowd identified as #1.

The most significant event regarding identity management occurs towards the end of the sequence. The crowd of rioters (initially with ID #20) merges with the primary police crowd (ID #1) during the confrontation. Following the merger, the police cease moving. According to the ID inheritance rules defined in Section 2.3, and because the rioters constitute the principal moving component after the merger, the entire resulting conglomerate inherits the ID #1. This outcome correctly reflects the final state of the scene, where the only coherently moving entity is the single crowd of rioters, which has effectively absorbed the identity of the larger group it merged with.

5.3 Method evaluation

The task of moving crowd tracking shares similarities with Multiple Object Tracking (MOT) but is fundamentally distinct due to the dynamic and collective nature of crowds.

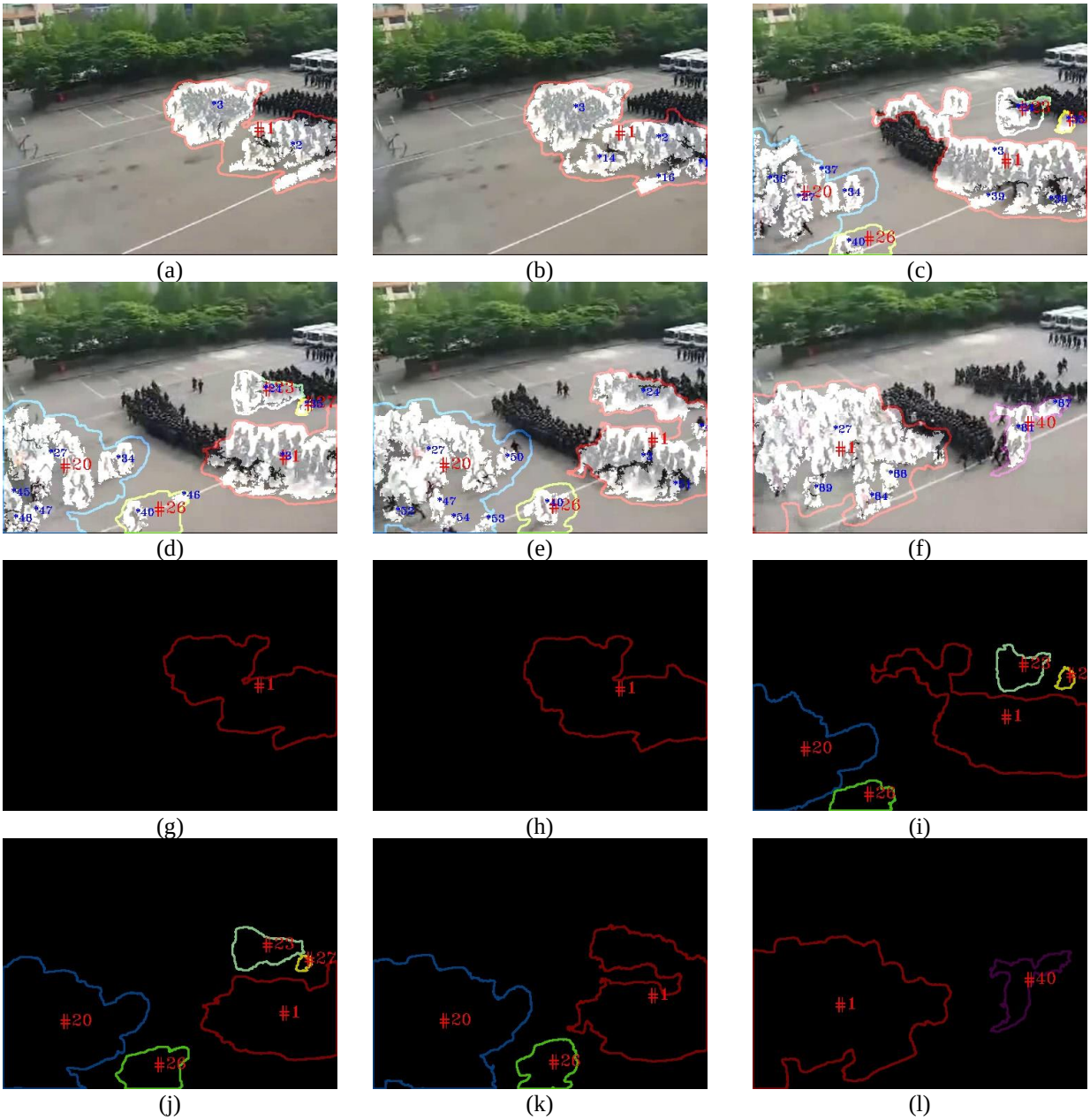

Figure 4: Out-door moving crowd tracking results.

The key differentiator lies in the phenomena of splitting and merging, where crowds can dissolve into smaller groups or coalesce into larger ones. In contrast, objects in standard MOT are typically treated as independent entities that do not merge or split. To bridge this gap and provide a comprehensive evaluation, we adopt established MOT metrics from [17]—including precision, recall, MOTA, and MOTP—while also introducing two novel metric pairs specifically designed to assess the algorithm's performance in handling crowd merging and splitting events.

5.3.1 Standard MOT metrics

The following standard metrics are utilized:

(1) True Positives (TP): A predicted object mask is considered a true positive if its Intersection over Union (IoU) with a ground-truth mask exceeds a predefined threshold (e.g., IoU > 0.5).

(2) False Positives (FP): This refers to predicted object masks that do not correspond to any real object in the ground truth. FP represents the total count of such erroneous detections across the entire video sequence.

(3) False Negatives (FN): This denotes real objects in the ground truth that the algorithm fails to detect. FN represents the total count of these missed detections in the sequence.

(4) Precision: This metric quantifies the accuracy of the detections, answering the question: "Of all the crowds detected, what proportion are genuine?" It is calculated to minimize false detections.

$$Precison = \frac{TP}{TP+FP}. \quad (10)$$

(5) Recall: This metric measures the algorithm's ability to find all genuine crowds, answering the question: "Of all the actual crowds present, what proportion did we successfully detect?" It is calculated to minimize missed cases.

$$Recall = \frac{TP}{TP+FN}. \quad (11)$$

(6) ID Switches (IDSW): An identity switch occurs when the tracking identity of a crowd is incorrectly changed. IDSW counts the total number of such identity changes.

(7) Multiple Object Tracking Accuracy (MOTA): This metric provides a holistic measure of tracking performance by combining errors from false positives, false negatives, and identity switches. It assesses the tracker's effectiveness in detecting targets and maintaining consistent trajectories, independent of localization precision.

$$MOTA = 1 - \frac{FN+FP+IDSW}{GT}, \quad (12)$$

where GT represents the total number of ground-truth objects in the entire video sequence.

(8) Multiple Object Tracking Precision (MOTP): This metric evaluates the average accuracy of the spatial localization for all correctly tracked targets (TPs). It is defined as the average IoU between the predicted masks and their corresponding ground-truth masks.

$$MOTP = \frac{\sum IoU_{TP}}{TP}, \quad (13)$$

where IoU_{TP} is the Intersection over Union between a True Positive detection and its matched ground-truth mask.

5.3.2 Proposed metrics for crowd dynamics

Given the specific challenges of crowd tracking, it is insufficient to rely solely on standard MOT metrics. Our method is designed to track moving crowds; therefore, static crowds are intentionally not detected, which can inherently lower recall and MOTA scores in scenes containing stationary groups. Furthermore, the algorithm's identity management strategy—where a smaller crowd merging into a larger one loses its ID and is assigned a new one upon splitting—can lead to ID switches during transient interactions like crossing without genuine merger. To address these nuances and provide a fair evaluation, we propose two new pairs of metrics focused on the core crowd behaviors of merging and splitting:

- Merging Precision (MPrecision) and Merging Recall (MRecall): These metrics evaluate the algorithm's accuracy in detecting genuine merging events and its ability to identify all actual merges, respectively.

$$MPrecision = \frac{TPM}{TPM+FPM}, \quad (14)$$

$$MRecall = \frac{TPM}{TPM+FNM}. \quad (15)$$

- Splitting Precision (SPrecision) and Splitting Recall (SRecall): These metrics evaluate the algorithm's accuracy in detecting genuine splitting events and its ability to identify all actual splits, respectively.

$$SPrecision = \frac{TPS}{TPS+FPS}, \quad (16)$$

$$SRecall = \frac{TPS}{TPS+FNS}. \quad (17)$$

Here, TPM, FPM, FNM represent the true positives, false positives, and false negatives for merging events. Similarly, TPS, FPS, FNS represent the corresponding quantities for splitting events.

5.3.3 Datasets and evaluation setup

The evaluation was conducted on three video sequences:

- Hall: The indoor surveillance video from Section 5.1.
- Riot Control Exercise: The outdoor scenario from Section 5.2.
- Mall: A new video sequence depicting people shopping in a mall (resolution: 640×480, 200 frames).

The integral optical flow was calculated with a frame interval (itv) of 10 for all sequences. The comprehensive evaluation results are presented in Table 1.

5.4 Discussion

This section provides a comprehensive discussion on the performance, advantages, and limitations of the proposed moving crowd detection and tracking framework, based on the experimental results presented in the previous sections.

5.4.1 Advantages and application potential

The proposed method successfully enables the detection and tracking of moving crowds in video sequences. This capability forms a critical foundation for subsequent high-

Table 1: Evaluation results. TPM, FNM and FPM are true positives, false negatives, and false positives of merging, respectively; TPS, FNS and FPS are true positives, false negatives, and false positives of splitting, respectively; MPrecision, MRecall, SPresion and SRecall are precision, and recall of detection, merging, and splitting, respectively.

Video		Hall	Riot Control Exercise	Mall
Detection	TP	216	29	56
	FN	54	18	18
	FP	48	8	15
	Precision	0.818	0.784	0.789
	Recall	0.800	0.617	0.757
ID switches		49	6	17
Evolution	TPM	19	3	7
	FNM	2	1	2
	FPM	1	0	1
	MPrecision	0.950	1.000	0.875
	MRecall	0.905	0.750	0.778
	TPS	28	2	6
	FNS	0	0	1
	FPS	1	0	2
	SPrecion	0.966	1.000	0.750
	SRecall	1.000	1.000	0.857
MOTP		0.926	0.783	0.816
MOTA		0.441	0.319	0.324

level analysis of crowded scenarios, such as crowd behavior recognition and abnormal event detection. A significant advantage of our approach is that it is training-free, relying on optical flow and spatial-temporal constraints rather than large annotated datasets, which contributes to its computational efficiency and ease of implementation in various environments.

The method's strength is further highlighted by comparing it with our prior work [14]. While the method in [14] could recognize typical crowd behaviors within specific regions at a given time, it lacked the capability to track crowd evolution. In contrast, the method proposed herein monitors the entire lifecycle of crowds—including their emergence, merging, splitting, and dissipation. This ability to capture long-term motion trajectories and structural changes makes the current method a far more suitable candidate for understanding complex crowd behaviors that unfold over extended periods.

5.4.2 Performance analysis and influencing factors

The experimental evaluation on both indoor and outdoor videos reveals several key factors influencing the method's performance:

(1) Impact of Crowd Structure: The accuracy of detecting merging and splitting events is highly dependent on the spatial coherence of the crowds. The method performs best when crowd movement patterns are structured and cohesive. Conversely, performance decreases when crowds are scattered or exhibit highly irregular motion, as the spatial constraints for defining a unified crowd become less reliable.

(2) Sources of Error in Standard MOT Metrics: The analysis of standard tracking metrics (Section 5.3) identifies primary sources of error.

- Precision is primarily affected by noise in the optical flow computation. Inaccurate flow vectors can lead to false positive detections.
- Recall is mainly impacted by the presence of static crowds. Since the method is designed to detect moving entities, completely stationary groups are intentionally ignored, leading to false negatives in scenes with a mix of moving and non-moving people.
- MOTA (Multiple Object Tracking Accuracy) is influenced by Identity Switches (IDSW). As demonstrated in Sections 5.1 and 5.2, these switches frequently occur during complex crowd interactions like merging and splitting, which is an inherent challenge in crowd (as opposed to individual object) tracking.

(3) Explanation of MOTP Performance: The MOTP (Multiple Object Tracking Precision) value, which measures localization accuracy, is generally higher in scenarios where moving crowds dominate the scene. This is because the integral optical flow calculation provides a robust representation of collective motion, leading to more precise spatial delineation of the moving aggregates.

5.4.3 Limitations and future work

Despite its advantages, the method has limitations. The dependence on optical flow makes it susceptible to illumination changes and rapid motion. Furthermore, the heuristic rules for ID management during merge/split events, while effective, could be refined. Future work will

focus on integrating more robust optical flow techniques and exploring data-driven approaches to improve ID management without sacrificing the method's efficiency. Validation on larger and more complex datasets will also be a priority.

6 Conclusion

Crowd motion analysis represents a pivotal yet challenging frontier in computer vision. While significant hurdles exist in the accurate detection of crowds, the tasks of tracking and analyzing their dynamic movements present even greater complexities. This paper has built upon our previous research by introducing a novel approach centered on integral optical flow for the detection and tracking of moving crowds. Our method effectively monitors the entire evolutionary process of crowds, including their emergence, merging, splitting, and dissipation, thereby providing a robust foundation for understanding complex crowd behaviors.

The experimental studies conducted demonstrate that moving crowds can be effectively detected and tracked across various scenarios. The results visually articulate the status and dynamic evolution of crowds, validating the practicality of our approach. A key advantage of the proposed method is its training-free nature, which allows for efficient application without the need for large annotated datasets. However, the performance is influenced by factors such as the spatial coherence of crowds and the inherent noise in optical flow computation.

Future research will focus on several promising directions. We plan to delve deeper into the analysis of crowd/group movement information and conduct more detailed experimental evaluations of the proposed crowd parameters. The objective is to develop more precise descriptors for the evolution process of crowds, which could significantly enhance the understanding of crowded scenes. Furthermore, integrating advanced techniques for handling complex interactions and exploring applications in abnormal behavior detection will be a priority. Ultimately, these efforts aim to provide richer information to support a higher-level, semantic understanding of crowded scenarios, paving the way for more intelligent video surveillance systems.

Acknowledgement

This work is supported by the National High-End Foreign Experts' Program (Grant No. G2023016002L, G2021016028L), Zhejiang Province Basic Public Welfare Research Program Project (Grant No. LGG21F020007), Zhejiang Shuren University Basic Scientific Research Special Funds and Major Technological Innovation Project of Hangzhou (Grant No. 2022AIZD0128).

References

[1] Elbishlawi, S., Abdelpakey, M.H., Eltantawy, A., Shehata, M.S. and Mohamed, M.M., Deep learning-based crowd scene analysis survey, J Imaging, 2020, vol. 6, no. 9, 95, DOI: 10.3390/jimaging6090095.

[2] Bendali-Braham, M., et al. (2021), Recent trends in crowd analysis: a review, machine learning with applications, Elsevier ltd., vol. 4(October 2020), 100023, DOI: 10.1016/j.mlwa.2021.100023.

[3] Chaudhary, D., Kumar, S. and Dhaka, V.S., Video based human crowd analysis using machine learning: a survey, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2022, vol. 10, p. 2, 113-131, DOI: 10.1080/21681163.2021.1986859.

[4] Li, M., Dong, H., Zhang, F. and Liu, X., Method for Top View Pedestrian Flow Detection Based on Small Target Tracking, Informatica, 2024, vol. 48, p. 59-70, DOI: 10.31449/inf.v48i11.6033.

[5] Chen, J.W., Su, W. and Wang, Z.F., Crowd counting with crowd attention convolutional neural network, Neurocomputing, 2020, vol. 382, p. 210-220, DOI: 10.1016/j.neucom.2019.11.064.

[6] Guo, H.P., Wang, R. and Sun, Y.E., Dual convolutional neural network for crowd counting, Multimedia Tools & Applications, 2024, vol. 83, no. 9, p. 26687-26709, DOI: 10.1007/s11042-023-16442-2.

[7] Sharma, V., Mir, R. N. and Singh, C., Scale-aware CNN for crowd density estimation and crowd behavior analysis, Computers & Electrical Engineering, 2023, vol. 106, 108569, DOI: 10.1016/j.compeleceng.2022.108569.

[8] Chen, Ch.X., Ye, Sh.P., Chen, H.F., Nedzvedz, A., Ablameyko, S. and Nedzvedz, O., Determination of blood flow characteristics in eye vessels in video sequence, Informatica, 2019, vol. 43, no. 4, p. 515-525, DOI: 10.31449/inf.v43i4.2598.

[9] Nayan, N., Sahu, S.S. and Kumar, S., Detecting anomalous crowd behavior using correlation analysis of optical flow, Signal Image and Video Processing, 2019, vol. 13, no. 6, p. 1233-1241, DOI: 10.1007/s11760-019-01474-9.

[10] Wang, X.F., He, Z.S., Sun, R., You, L., Hu, J. and Zhang, J., A Crowd Behavior Identification Method Combining the Streakline with the High-Accurate Variational Optical Flow Model, IEEE Access, 2019, vol. 7, p. 114572-114581, DOI: 10.1109/ACCESS.2019.2929200.

[11] Altalbi, A.A.H., Shaker, S.H. and Ali, A.E., Localization of Strangeness for Real Time Video in Crowd Activity Using Optical Flow and Entropy, International Journal of Online and Biomedical Engineering, 2023, vol. 19, no. 7, p. 52-68, DOI: 10.3991/ijoe. v19i07.38869.

[12] Zhang, L., Cao, L., Zhao, Z.M., Wang, D.F. and Fu, C., A Crowd Movement Analysis Method Based on Radar Particle Flow[J]. Sensors, 2024, vol. 24, no. 6, 1899, DOI: doi.org/10.3390/s24061899.

[13] Bhuiyan, M.R., Abdullah, J., Hashim, N., Farid, F.A. and Uddin, J., Hajj pilgrimage abnormal crowd movement monitoring using optical flow and FCNN, Journal of Big Data, 2023, vol. 10, 86, DOI: 10.1186/s40537-023-00779-4.

[14] Chen, H.F., Nedzvedz, O., Ye, Sh.P. and Ablameyko, S., Crowd Abnormal Behaviour Identification Based on Integral Optical Flow in Video Surveillance

Systems, Informatica, 2018, vol. 29, no. 2, p. 211-232, DOI: 10.5555/ios.INF1178.

[15] Chen, H.F., Pashkevich, A., Ye, Sh.P., Bohush, R. and Ablameyko, S., Crowd Movement Type Estimation in Video by Integral Optical Flow and Convolution Neural Network, Pattern Recognition and Image Analysis, 2024, vol. 34, no. 2, p. 266-274, DOI: 10.1134/S1054661824700068.

[16] Farnebäck, G., Two-frame motion estimation based on polynomial expansion, Proc. 13th Scandinavian Conference on Image Analysis, Halmstad, 2003, p. 363-370, DOI: 10.1007/3-540-45103-X_50.

[17] Luo, W.H., Xing, J.L. and Milan, A., et al. Multiple object tracking: A literature review, Artificial Intelligence, 2021, vol. 293, 103448. DOI: 10.1016/j.artint.2020.103448.