Particle Tracking Accuracy Measurement Based on Comparison of Linear Oriented Forests

Martin Maška and Pavel Matula
Centre for Biomedical Image Analysis, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{xmaska,pam}@fi.muni.cz

Abstract

Particle tracking is of fundamental importance in diverse quantitative analyses of dynamic intracellular processes using time-lapse microscopy. Due to frequent impracticability of tracking particles manually, a number of fully automated algorithms have been developed over past decades, carrying out the tracking task in two subsequent phases: (1) particle detection and (2) particle linking. An objective benchmark for assessing the performance of such algorithms was recently established by the Particle Tracking Challenge. Because its performance evaluation protocol finds correspondences between a reference and algorithm-generated tracking result at the level of individual tracks, the performance assessment strongly depends on the algorithm linking capabilities. In this paper, we propose a novel performance evaluation protocol based on a simplified version of the tracking accuracy measure employed in the Cell Tracking Challenge, which establishes the correspondences at the level of individual particle detections, thus allowing one to evaluate the performance of each of the two phases in an isolated, unbiased manner. By analyzing the tracking results of all 14 algorithms competing in the Particle Tracking Challenge using the proposed evaluation protocol, we reveal substantial changes in their detection and linking performance, yielding rankings different from those reported previously.

1. Introduction

Particle tracking plays a key role in many biomedical applications focusing on dynamic intracellular processes. The particle can be anything from a single molecule to a macromolecular complex, organelle, virus, or microsphere manifesting itself as a small dot in the image data. The problem of particle tracking can be formulated as having a recorded time-lapse sequence of moving dot-like objects, one is interested in spatiotemporal positions of individual objects.

There are a dozen of software tools and fully automated algorithms for particle tracking [8, 10]. They typically work in two phases: (1) particle detection and (2) particle linking. First, all particles are detected separately in every frame of a given time-lapse sequence. Second, the detected particles are linked into tracks, a set of which forms a linear oriented forest (LOF) in the graph theory terminology. Knowing the performance of individual phases is of great importance for potential users when composing robust application-oriented image analysis pipelines as well as for algorithm developers when aiming at further algorithmic improvements.

An objective comparison of 14 particle trackers, using a completely annotated repository of computer-generated image data and a diverse set of performance evaluation criteria, was performed recently within the Particle Tracking Challenge (PTC) [1]. Having a reference LOF and an algorithm-generated LOF to be evaluated, the PTC evaluation protocol establishes particle correspondences at the level of individual tracks, yielding a possibly inconsistent scoring for identical configurations of tracking errors with different temporal contexts, as shown in Figure 1 and listed in Table 1. Furthermore, it provides neither users nor algorithm developers with a direct information about individual linking errors committed by the algorithm, which may complicate its parameter fine-tuning and further algorithmic developments.

In this paper, we propose a new evaluation protocol that establishes correspondences between a reference LOF and an algorithm-generated LOF at the level of individual particle detections. It allows one to assess detection and linking performance of the algorithm independently of each other, thus consistently penalizing identical, time-varying configurations of tracking errors. After having particle detections paired, the detection and linking performance is evaluated using a simplified version of the Acyclic Oriented Graphs Matching (AOGM) measure [6], which exploits only a limited subset of allowed graph operations available in LOFs. The adoption of the AOGM concept allows one to directly identify allowed graph operations needed when transforming the algorithm-generated LOF to the reference one, thus recognizing individual tracking errors committed by the al-

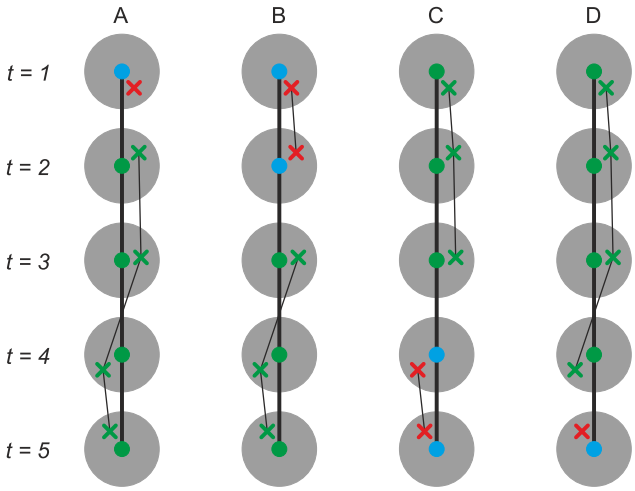

Figure 1. Tracking results of four hypothetical algorithms that use the same detection routine and four different linking routines, each committing one linking error at a different time point. The tiny circles correspond to reference detections, the crosses to algorithm-generated detections with localization errors of three pixels, and the gray disks indicate the gating areas of five pixels within which two detections are considered matching. The colors encode categories assigned by the PTC evaluation protocol to each detection: matching (in green), missing (in blue), and spurious (in red), yielding inconsistent detection and linking scores, as listed in Table 1.

Algorithm	TP	FN	FP	JSC	α
A	4	1	1	0.67	0.32
B	3	2	2	0.43	0.24
C	3	2	2	0.43	0.24
D	4	1	1	0.67	0.32

Table 1. The performance scores for the tracking results of four hypothetical algorithms from Figure 1, obtained using the PTC evaluation protocol. Although the four algorithms generated the identical sets of particle detections and committed a single linking error only, their detection and linking performance, in terms of the Jaccard similarity coefficient (JSC) and normalized track-wise pairing distance (α), respectively, was scored inconsistently due to differences in the number of matching (TP), missing (FN), and spurious detections (FP).

gorithm. By applying the proposed evaluation scheme to the tracking results of all 14 algorithms that participated in PTC and compiling their rankings in terms of particle detection, localization, and linking, we show how much these rankings correlate to those previously reported in [1].

2. Methods

A track is a temporal series of subsequent spatial positions. The spatial position of a particle at a given time point $t \geq 1$ is a vector $\theta(t) = (x(t), y(t), z(t))$, where $x(t)$, $y(t)$, and $z(t)$ are the particle coordinates along the particular image axes. In the case of 2D sequences, the third coordinate, $z(t)$, is constant and usually set to zero. A track θ that spans the range of time points $[t_{init} \geq 1, t_{end} \geq t_{init}]$ is therefore defined as a set $\theta = \{\theta(t) : t = t_{init}, \ldots, t_{end}\}$. If a particle position is unknown for a particular time point, its coordinates are not listed in such a set. Clearly, every track can be converted to an oriented path where vertices correspond to individual spatial positions and edges to their temporal relationships. Therefore, any tracking result composed of a set of tracks can directly be represented as a LOF.

Given a reference and algorithm-generated set of tracks, Θ_1 and Θ_2, respectively, our aim is to measure how similar these two sets are. For this purpose, it is essential to establish correspondences between individual spatial positions included in Θ_1 and Θ_2 at corresponding time points, which in turn allows one to straightforwardly assess the similarity of these two sets of tracks at the level of detection, localization, and linking. At the level of detection, one is interested in whether or not each spatial position included in one set of tracks simultaneously appears in the other set of tracks, up to a given gating distance, $\varepsilon \in \mathbb{R}, \varepsilon \geq 0$. At the level of localization, by contrast, one considers only spatial positions paired with respect to the given gating distance, evaluating errors in the localization of particle positions for individual pairs. Finally, at the level of linking, the temporal aspect of each set of tracks is mutually compared.

As the correspondences between spatial positions are established based on their distances, the gated distance of two spatial positions, $\theta_1(t)$ and $\theta_2(t)$, is defined as [1]:

$$\|\theta_1(t) - \theta_2(t)\|_{2,\varepsilon} = \min(\varepsilon, \|\theta_1(t) - \theta_2(t)\|_2) \quad (1)$$

where $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^3$. The gated distance of two spatial positions farther than ε or of an unknown spatial position and a known one is set to ε, whereas that of two unknown spatial positions is defined to be zero.

2.1. The PTC Evaluation Protocol

The PTC evaluation protocol [1] establishes correspondences between the spatial positions in Θ_1 and Θ_2 by finding an optimal pairing at the level of whole tracks in Θ_1 and Θ_2 with respect to their distances. A distance between two tracks, θ_1 and θ_2, is defined as:

$$d(\theta_1, \theta_2) = \sum_{t=1}^{T} \|\theta_1(t) - \theta_2(t)\|_{2,\varepsilon} \quad (2)$$

where T is the total number of time points in the sequence. An optimal pairing of tracks is established by extending Θ_2 with $|\Theta_1|$ dummy tracks (denoted by $\varnothing$), each formally being an empty set of spatial positions, and by solving an optimal subpattern assignment using the Munkres algorithm. This yields a globally best possible pairing, in terms of the minimum total distance over all tracks in Θ_1, of each track $\theta_i \in \Theta_1$ with a track $\hat{\theta}_i \in \Theta_2 \cup \varnothing$, being either an original

track in Θ_2 (if available) or a dummy track (in the absence of a suitable track in Θ_2).

Finally, the detection, localization, and linking similarity between Θ_1 and Θ_2 is calculated as follows [1]:

- *Detection*: $\mathrm{JSC} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}+\mathrm{FP})$. This defines the Jaccard similarity coefficient for spatial positions. It always falls in the $[0, 1]$ interval, with higher values corresponding to higher detection similarity. TP (true positives) denotes the number of matching positions (i.e., known spatial positions within the given gating distance) in the optimally paired tracks; FN (false negatives) denotes the number of dummy positions in the optimally paired tracks; and FP (false positives) denotes the number of nonmatching positions in Θ_2.

- *Localization*: RMSE. This is the root-mean-square error of all matching positions in the optimally paired tracks. It always falls in the $[0, \varepsilon]$ interval, with lower values corresponding to higher localization similarity.

- *Linking*: $\alpha(\Theta_1, \Theta_2) = 1 - d(\Theta_1, \Theta_2)/d(\Theta_1, \varnothing)$. α always falls in the $[0, 1]$ interval, with higher values corresponding to higher linking similarity. $d(\Theta_1, \varnothing)$ is the maximum possible total distance from a reference set of tracks, being considered a normalization factor.

2.2. The Proposed Evaluation Protocol

Although JSC is easy to compute and interpret, the values of TP, and FN and FP, calculated using whole-track-paired spatial positions, are often underestimated and overestimated, respectively, not always leading to a natural picture of the detection similarity between two tracking results, as shown in Figure 1 and listed in Table 1. Bearing this in mind, we establish correspondences between the spatial positions in Θ_1 and Θ_2 in a manner to simultaneously maximize TP and minimize FN and FP. To this end, a complete bipartite graph, with the vertices corresponding to the spatial positions in Θ_1 for one vertex subset and to those in Θ_2 extended by N dummy vertices for the other vertex subset, where N is the number of spatial positions in Θ_1, and with the edge weights set to the reciprocal gated distances between the respective spatial positions, is constructed and an optimal subpattern assignment is solved using the Munkres algorithm. This yields a globally best possible pairing at the level of individual spatial positions and with respect to the minimum total gated distance between them.

The proposed alternative to finding correspondences between the spatial positions precludes evaluation of the linking similarity using α. Furthemore, this measure provides neither users nor algorithm developers with any clues about dissimilar linking decisions, ruling out their identification and possible curation. Nevertheless, both limitations can be overcome by adapting the AOGM measure [6] to LOFs.

Being developed specifically for the Cell Tracking Challenge [7], the AOGM measure evaluates how difficult it is to transform an algorithm-generated acyclic oriented graph to a reference graph. The difficulty is measured as a weighted sum of the lowest number of allowed graph operations that make both graphs identical. When working with LOFs, the graph operations allowed can be reduced to *add/delete vertex* and *add/delete edge*, ignoring the graph operations, *split vertex* and *change edge semantics*, related to undersegmentation and branching events, which do not occur in LOFs. Such a simplification also leads to the fact that any nonnegative configuration of the weights used for the reduced set of graph operations allowed always satisfies the minimality condition, which guarantees the measured difficulty to be not only the weighted sum of the lowest number of allowed graph operations but also the minimum weighted sum [6].

Let $G_1 = (V_1, E_1)$ be a reference LOF that corresponds to Θ_1, and $G_2 = (V_2, E_2)$ be an algorithm-generated LOF that corresponds to Θ_2. Let the vertex correspondences between the sets V_1 and V_2 be defined using two pairing functions, $p_1 : V_1 \to V_2 \cup \{\perp\}$ and $p_2 : V_2 \to V_1 \cup \{\perp\}$, where $\perp \notin V_1 \cup V_2$ is a dummy vertex, and where $p_j(p_i(v_i)) = v_i$, $i, j \in \{1, 2\}$, $i \neq j$, for every $v_i \in V_i$ such that $p_i(v_i) \in V_j$. These pairing functions allow one to classify the vertices as matching (i.e., true positives), missing (i.e., false negatives), and spurious (i.e., false positives) as follows:

$$V_1^{\mathrm{TP}} = \{v_1 \in V_1 : p_1(v_1) \neq \perp\}\,, \tag{3}$$

$$V_1^{\mathrm{FN}} = \{v_1 \in V_1 : p_1(v_1) = \perp\}\,, \tag{4}$$

$$V_2^{\mathrm{TP}} = \{v_2 \in V_2 : p_2(v_2) \neq \perp\}\,, \tag{5}$$

$$V_2^{\mathrm{FP}} = \{v_2 \in V_2 : p_2(v_2) = \perp\}\,, \tag{6}$$

where $V_1 = V_1^{\mathrm{TP}} \cup V_1^{\mathrm{FN}}$ and $V_2 = V_2^{\mathrm{TP}} \cup V_2^{\mathrm{FP}}$, and to count the cardinality of each class: $\mathrm{TP} = |V_1^{\mathrm{TP}}| = |V_2^{\mathrm{TP}}|$, $\mathrm{FN} = |V_1^{\mathrm{FN}}|$, and $\mathrm{FP} = |V_2^{\mathrm{FP}}|$. Next, the numbers of missing and redundant edges, EA and ED, are deduced from the induced subgraphs, $\hat{G}_1 = (\hat{V}_1, \hat{E}_1)$ and $\hat{G}_2 = (\hat{V}_2, \hat{E}_2)$, by the vertex sets V_1^{TP} and V_2^{TP}, being formed of the matching vertices and their incident edges solely, as follows:

$$\mathrm{EA} = |\{(u, v) \in \hat{E}_1 : (p_1(u), p_1(v)) \notin \hat{E}_2\}|\,, \tag{7}$$

$$\mathrm{ED} = |\{(u, v) \in \hat{E}_2 : (p_2(u), p_2(v)) \notin \hat{E}_1\}|\,. \tag{8}$$

Finally, the detection and linking similarity between G_1 and G_2 is calculated using a normalized, user-weighted sum of those quantities that relate to the respective similarity type, being referred to as $\mathrm{LOFM_D}$ and $\mathrm{LOFM_L}$ as the abbreviations for the Linear Oriented Forests Matching measure for particle detection and linking, respectively, as follows:

$$\mathrm{LOFM_D} = 1 - \frac{\min(w_{\mathrm{FN}} \cdot \mathrm{FN} + w_{\mathrm{FP}} \cdot \mathrm{FP}, e_D)}{e_D}\,, \tag{9}$$

$$\mathrm{LOFM_L} = 1 - \frac{\min(w_{\mathrm{EA}} \cdot \mathrm{EA} + w_{\mathrm{ED}} \cdot \mathrm{ED}, e_L)}{e_L}\,, \quad (10)$$

where $(w_{\mathrm{FN}}, w_{\mathrm{FP}}, w_{\mathrm{EA}}, w_{\mathrm{ED}})$ is a user-defined configuration of nonnegative weights for the individual graph operations allowed, and $e_D = w_{\mathrm{FN}} \cdot |V_1|$ and $e_L = w_{\mathrm{EA}} \cdot |\hat{E}_1|$ are the costs of creating the respective reference output from scratch (i.e., when $V_2 = \emptyset$ and $\hat{E}_2 = \emptyset$, respectively). The minimum operator in the numerators prevents from having final negative values when it is cheaper to create the respective reference outputs from scratch rather than to transform the algorithm-generated output to them. The normalization ensures that $\mathrm{LOFM_D}$ and $\mathrm{LOFM_L}$ always fall in the $[0, 1]$ interval, with higher values corresponding to higher detection and linking similarity, respectively. Because e_D and e_L could theoretically be zero, $\mathrm{LOFM_D}$ and $\mathrm{LOFM_L}$, respectively, are defined to be zero if this is the case.

For the sake of completeness, the localization similarity between G_1 and G_2 is calculated again using RMSE, but this time over the matching vertices in the optimally paired detections (i.e., over the vertex sets V_1^{TP} and V_2^{TP}).

3. Results and Discussion

To validate the proposed evaluation protocol, we applied it on the hypothetical tracking results shown in Figure 1 and on the tracking results of all 14 algorithms that participated in PTC, analyzing all four PTC biological scenarios (Microtubules, Receptors, Vesicles, and Viruses) of three levels of particle density (low, mid, and high) and of two levels of signal-to-noise ratio (4 and 7). The four scenarios displayed realistic renderings of simulated, green-fluorescent-protein-labeled particles of scenario-variable dynamics (randomly oriented directed motion, switching between random-walk and randomly oriented directed motion, random-walk motion, and switching between random-walk and orientation-restricted directed motion) and appearance attributed to applying diverse point-spread-function models for virtual fluorescence microscopy (2D anisotropic Gaussian, 2D confocal, 2D wide-field, and 3D confocal). In total, we evaluated 264 PTC tracking results because not all 14 algorithms provided tracking results for all analyzed scenarios. The gating distance, ε, was set to 5, as used in PTC. The weight configuration, $(w_{\mathrm{FN}}, w_{\mathrm{FP}}, w_{\mathrm{EA}}, w_{\mathrm{ED}})$, of $\mathrm{LOFM_D}$ and $\mathrm{LOFM_L}$ was fixed at $(1, 1, 1.5, 1)$, adopting the edge-related weights from [6] but changing the vertex-related weights to 1, thus better reflecting a point-like nature of individual particles.

By establishing the particle correspondences at the level of individual detections, instead of at the level of individual tracks, the proposed evaluation protocol consistently scored the identical sets of particle detections (Table 2), not considering temporal relationships between the particles, as the PTC evaluation protocol did (Table 1). Furthermore, we can observe an increase in the number of matching detections, approximately from 11 % to 48 %, along with a decrease in the number of missing and spurious detections (Figure 2), approximately from 6 % to 23 % and from 6 % to 26 %, respectively. This has led to an increase in the average detection score over all detection scores for each algorithm and analyzed scenario, from $\mathrm{JSC} = 0.58 \pm 0.24$, when applying the PTC evaluation protocol, to $\mathrm{LOFM_D} = 0.82 \pm 0.26$, when applying the proposed evaluation protocol. Note that having the particles paired at the level of individual detections gives $\mathrm{JSC} = 0.84 \pm 0.21$.

Algorithm	TP	FN	FP	$\mathrm{LOFM_D}$	$\mathrm{LOFM_L}$
A–D	5	0	0	1.0	0.75

Table 2. The performance scores for the tracking results of four hypothetical algorithms from Figure 1, obtained using the proposed evaluation protocol.

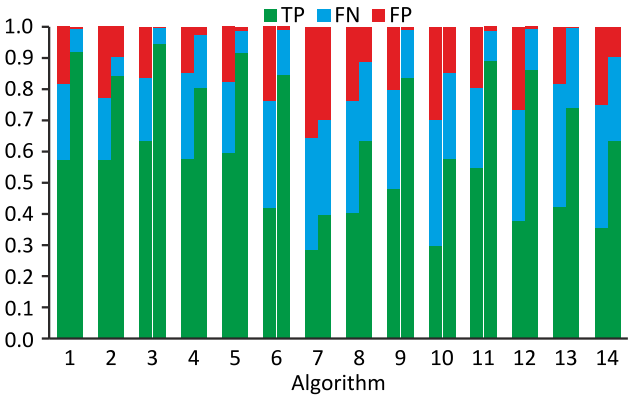

Figure 2. The overall distribution of particle detections, identified by individual algorithms participating in PTC, obtained using the PTC and proposed evaluation protocols (the left and right horizontal bars, respectively, in each pair) across all analyzed scenarios.

The isolated evaluation of the detection and linking parts of tracking results also led to the consistent linking scores in the case of one missing link at different time points (Table 2). More importantly, it allows one to evaluate tracking results at the level of individual tracking errors (Figure 3), thus providing users and algorithm developers with relevant practical clues on the algorithm behavior, which might help in fine-tuning its parameters and improving its components focused on individual phases of the particle tracking task.

The total rankings of all algorithms participating in PTC, being compiled using the PTC and proposed evaluation protocols for each scenario, strongly correlate, in terms of the Kendall rank correlation coefficient τ, for particle localization, but not for particle detection and even less for particle linking (Figure 4). Such an observation is not surprising because the proposed evaluation protocol adopts the localization measure, RMSE, used in PTC, while involving conceptually different detection and linking measures over differently paired tracking results. The proposed evaluation protocol calculates RMSE over the set of matching detections

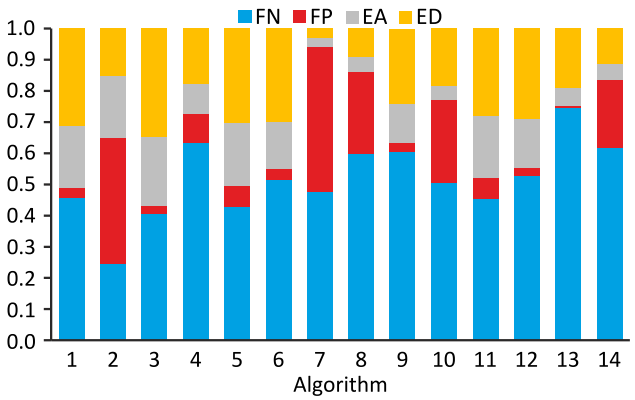

Figure 3. The overall distribution of errors, committed by individual algorithms participating in PTC, obtained using the proposed evaluation protocol across all analyzed scenarios.

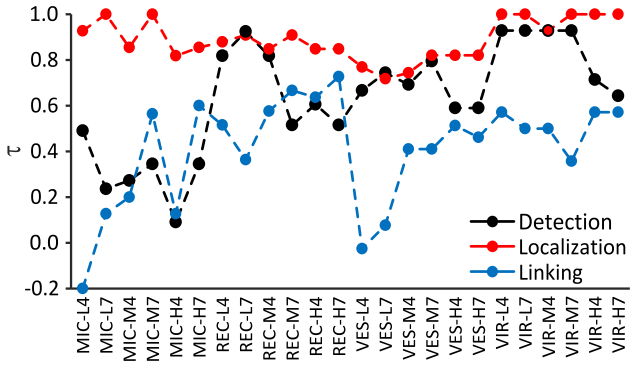

Figure 4. The Kendall rank correlation coefficient, τ, between the detection, localization, and linking rankings of all algorithms participating in PTC, being compiled using the PTC and proposed evaluation protocols for each analyzed scenario.

established at the level of individual detections, which can be considered a superset of the set consisting of the matching detections established at the level of individual tracks, yielding the values of τ not being lower than 0.71 and being 0.89 ± 0.09 on average across all analyzed scenarios. On the contrary, the values of τ for particle detection and particle linking were 0.63 ± 0.24 and 0.41 ± 0.23, respectively, on average across all analyzed scenarios, not exceeding the level of 0.93 and 0.73, respectively.

Limiting the focus on the top three best-performing algorithms only, we can generally conclude that the algorithms were ranked as top-3 performers with comparable frequencies using the PTC and proposed evaluation protocols, with the exception of Algorithms 5, 8, and 12 (Figures 5 and 6). Indeed, the localization scores of Algorithm 8 and the detection scores of Algorithm 12 were underrated using the PTC evaluation protocol due to their lower whole-track-oriented linking performance as compared to the best-performing algorithms, which can be attributted to not always incorporating an appropriate particle motion model in the case of Algorithm 8 [1, 5] and to making the linking decisions based on a limited multi-frame information solely [1, 3]. Finally, Algorithm 5 was ranked less frequent as a top-3 performer

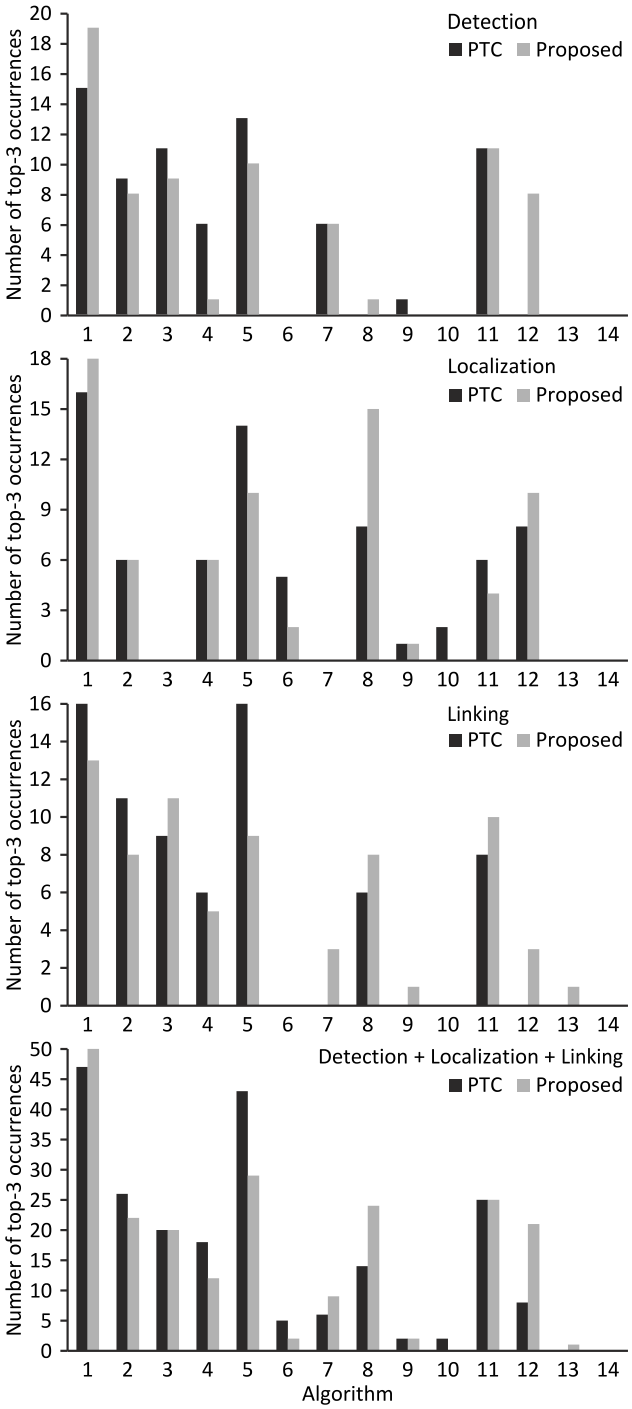

Figure 5. The total number of occurrences of individual algorithms participating in PTC, being ranked among the top three detection, localization, and linking performers according to the PTC and proposed evaluation protocols across all analyzed scenarios. The bottom chart was obtained by aggregating the three charts above it.

Scenario		Microtubules						Receptors						Vesicles						Viruses					
Density		Low		Mid		High		Low		Mid		High		Low		Mid		High		Low		Mid		High	
SNR		4	7	4	7	4	7	4	7	4	7	4	7	4	7	4	7	4	7	4	7	4	7	4	7
Detection JSC	#1	5	2	4	2	4	2	3	3	3	3	3	3	5	7	7	7	7	7	1	1	5	1	2	2
	#2	11	4	11	4	11	4	11	11	11	11	11	11	7	3	3	3	3	1	5	5	2	2	5	5
	#3	4	5	5	11	5	11	1	1	1	9	1	1	3	5	1	1	1	5	2	2	1	5	1	1
Localization RMSE	#1	8	8	4	4	4	4	1	1	1	11	1	1	8	1	1	1	1	1	2	2	2	2	2	2
	#2	4	4	8	8	8	9	11	6	11	12	11	11	10	5	10	12	8	12	5	5	5	5	5	5
	#3	5	5	5	5	5	5	6	11	12	6	12	12	5	6	8	6	12	8	1	1	1	1	1	1
Linking α	#1	4	4	4	4	4	4	11	11	3	3	3	3	5	1	1	1	1	1	1	1	5	2	2	2
	#2	8	5	2	2	2	2	1	1	11	11	11	11	11	8	8	8	3	5	5	5	2	1	5	5
	#3	5	2	5	5	11	5	3	8	1	5	1	1	3	5	3	3	5	8	2	2	1	5	1	1

Scenario		Microtubules						Receptors						Vesicles						Viruses					
Density		Low		Mid		High		Low		Mid		High		Low		Mid		High		Low		Mid		High	
SNR		4	7	4	7	4	7	4	7	4	7	4	7	4	7	4	7	4	7	4	7	4	7	4	7
Detection $\mathrm{LOFM_D}$	#1	1	1	1	1	12	1	1	11	3	3	3	3	7	7	7	7	7	7	5	1	2	1	2	2
	#2	8	12	12	12	1	12	11	3	11	11	11	1	5	11	3	3	1	1	1	5	5	2	1	1
	#3	11	4	5	2	5	2	3	1	12	1	12	5	11	11	3	1	11	11	5	2	2	1	5	5	12
Localization RMSE	#1	8	8	4	4	4	4	1	8	1	11	1	1	8	8	8	8	8	1	2	2	2	2	2	2
	#2	4	4	8	8	8	1	11	1	11	12	12	12	12	1	12	1	12	8	5	5	5	5	5	5
	#3	5	5	5	5	1	9	8	6	12	6	8	8	11	12	1	12	1	12	1	1	1	1	1	1
Linking $\mathrm{LOFM_L}$	#1	5	4	4	4	4	4	3	3	3	3	11	11	3	3	1	1	8	7	1	1	5	5	5	2
	#2	11	2	5	2	8	2	11	11	11	11	3	3	1	1	3	3	3	1	12	5	2	1	2	5
	#3	13	5	8	11	11	11	1	9	1	8	2	1	12	12	7	7	1	2	5	8	1	8	8	8

Figure 6. The top three best-performing algorithms in terms of particle detection, localization, and linking according to the PTC (the upper panel) and proposed (the lower panel) evaluation protocols for each analyzed scenario.

using the proposed evaluation protocol than using the PTC evaluation protocol namely because of evaluating the linking performance on the induced subgraphs of LOFs by the sets of matching detections, not taking into account the localization aspect as α did. Nevertheless, it is important to note that in four out of seven situations when Algorithm 5 was not ranked as a top-3 linking performer, its scores were lower than those of the third-ranked algorithms by less than 0.01. In the remaining three cases, its scores were lower by 0.02 to 0.05 than those of the third-ranked algorithms.

It is important to note that the optimal pairing of particle detections at the level of individual vertices in two given LOFs does not have to be always unique in terms of established pairs. Indeed, two possible configurations of colliding vertices exist, which in turn may lead to multiple optimal pairings of the same minimum total gated distance between the vertex sets. One is defined by multiple algorithm-generated detections within the gating area of a single reference detection, being also in the same minimum distance from it, whereas the other is formed of a single algorithm-generated detection within the gating areas of multiple reference detections, being also in the same minimum distance from each of them. In spite of not impacting $\mathrm{LOFM_D}$, these colliding configurations can possibly influence $\mathrm{LOFM_L}$ by wrongly increasing ED by no more than two for each colliding configuration. Nevertheless, their presence seems to be very rare in the PTC tracking results, affecting less than $0.56\,\%$ of all particle detections analyzed in this study. It is, therefore, feasible to assess all optimal pairings and select among them that with the minimum $\mathrm{LOFM_L}$ value.

4. Conclusion

In this paper, a new protocol for evaluating the detection and linking performance of particle tracking algorithms has been proposed. Treating particle tracking results as LOFs, the proposed protocol evaluates how difficult it is to transform an algorithm-generated LOF to a reference LOF. Such a difficulty is measured by normalizing a weighted sum of the lowest number of graph operations needed to make both LOFs identical, after having their vertices optimally paired at the level of individual detections. By analyzing the tracking results of all 14 algorithms that competed in PTC using the proposed protocol, we have shown that it has compiled substantially different rankings from those reported previously using the PTC protocol. In addition to performance-oriented evaluation of particle tracking algorithms, the sym-

metric nature of established vertex correspondences allows one to use the proposed protocol also for determining differences between multiple manual annotations of point-like targets, thus finding its application in annotation fusing too.

In future work, we intend to make use of collision-related information directly when establishing particle correspondences at the level of individual detections, hence ensuring the pairing optimality not only with respect to $\mathrm{LOFM_D}$, but also with respect to $\mathrm{LOFM_L}$, without exhaustively evaluating all possible optimal pairings. We also intend to focus on relationships between the proposed evaluation protocol and relevant approaches, used by the computer vision community, to evaluating performance of multi-object trackers [4], with the primary aim of interconnecting performance evaluation protocols developed separately by the bioimage analysis and computer vision communities for conceptually similar, but domain-specific point-like trackers [9].

The Java-based implementation of the proposed evaluation protocol is made publicly available as an Icy plugin [2]. The complete list of scores of all 14 algorithms that participated in PTC, obtained using the proposed evaluation protocol, can be found at http://cbia.fi.muni.cz/projects/lofm.

Acknowledgments.

This work was supported by the Czech Science Foundation under the grant number GJ16-03909Y.

References

[1] N. Chenouard et al. Objective comparison of particle tracking methods. *Nature Methods*, 11(3):281–289, 2014.

[2] F. de Chaumont et al. Icy: An open bioimage informatics platform for extended reproducible research. *Nature Methods*, 9(7):690–696, 2012.

[3] K. Jaqaman et al. Robust single-particle tracking in live-cell time-lapse sequences. *Nature Methods*, 5(8):695–702, 2008.

[4] R. Kasturi et al. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):319–336, 2009.

[5] K. E. G. Magnusson and J. Jaldén. Tracking of non-Brownian particles using the Viterbi algorithm. In *IEEE International Symposium on Biomedical Imaging*, pages 380–384, 2015.

[6] P. Matula et al. Cell tracking accuracy measurement based on comparison of acyclic oriented graphs. *PLoS One*, 10(12):e0144959, 2015.

[7] M. Maška et al. A benchmark for comparison of cell tracking algorithms. *Bioinformatics*, 30(11):1609–1617, 2014.

[8] E. Meijering et al. Methods for cell and particle tracking. *Methods in Enzymology*, 504(2):183–200, 2012.

[9] E. Meijering et al. Imaging the future of bioimage analysis. *Nature Biotechnology*, 34(12):1250–1255, 2016.

[10] H. Shen et al. Single particle tracking: From theory to biophysical applications. *Chemical Reviews*, 117(11):7331–7376, 2017.