Performance Evaluation for the ISBI'2012 Multiple Particle Tracking Challenge

The Challenge Committee

We describe hereafter the methods that are used for evaluating the results of the particle tracking challenge. Their is yet no well-accepted method to measure the global tracking performance, in particular when tracking a varying number of particles in a cluttered environment. We have therefore chosen to use a set of complementary criteria that will hopefully give a complete and intuitive characterization of the tracking results. We have implemented these performance measures and they are made available to the participants through an open-source software platform.

I. Track definition

For this challenge, we consider that a track is a temporal series of subsequent spatial positions. In the case of a 3D image sequence, the spatial position at a given time point $t \geq 0$ is a vector $(x(t), y(t), z(t))$, with $x(t)$, $y(t)$, and $z(t)$ the coordinates at this time along the x-axis, y-axis, and z-axis, respectively. A track θ that exists from time $t_{\text{init}} \geq 0$ to time $t_{\text{end}} \geq t_{\text{init}}$ is therefore defined as the set $\theta = \{\theta(t) = (x(t), y(t), z(t)),\ \text{with}\ t = t_{\text{init}} \ldots t_{\text{end}}\}$. Note that we do not allow for missing positions in the interval $(t_{\text{init}} \ldots t_{\text{end}})$. It is the responsibility of the tracking method to provide a full sequence of subsequent positions. Missing positions are marked as non-matching and are penalized as described below.

II. Distance between two tracks

For the purpose of track distance measurement, we use the following gated Euclidian distance between two positions $\theta_1(t)$ and $\theta_2(t)$:

$$\|\theta_1(t) - \theta_2(t)\|_{2,\epsilon} \triangleq \min(\|\theta_1(t) - \theta_2(t)\|_2, \epsilon),$$

with $\|\cdot\|_2$ the standard ℓ_2-norm of $\mathbb{R}_3$, and $\epsilon \in \mathbb{R}_+$. The rationale behind the use of the gate ϵ is to limit the penalization of tracks that separate. When two tracks are more than ϵ apart, it is indeed considered that their positions do not *match*. In this case, it is not relevant to measure the distance between their positions and a penalty ϵ is considered instead. In the context of the particle tracking challenge, the value of ϵ is set to 5 pixels.

It may happen that two tracks have different temporal supports. For instance, θ_1 may exist at a given time t for which θ_2 does not. In this case, we consider that tracks do not match at this time. It leads us to apply a penalty ϵ and to use the abuse of notation $\|\theta_1(t) - \theta_2(t)\|_{2,\epsilon} = \epsilon$. Along the same line, if none of the two tracks are defined at time t, we write: $\|\theta_1(t) - \theta_2(t)\|_{2,\epsilon} = 0$. This allows for a compact formulation of the distance d between tracks that is used for the challenge:

$$d(\theta_1, \theta_2) = \sum_{t=0}^{T-1} \|\theta_1(t) - \theta_2(t)\|_{2,\epsilon},$$

with T the length of the image sequence.

III. Distance between track sets and optimal pairing

Let us denote $\mathcal{X} = \{\theta^k_{\mathcal{X}},\ \text{for}\ k = 1 \ldots |\mathcal{X}|\}$ an ordered set of ground-truth tracks that is provided by the challenge organizers, and $\mathcal{Y}$ a set of candidate tracks whose quality needs to be assessed with respect to $\mathcal{X}$. Because some tracks in $\mathcal{X}$ may not match a track in $\mathcal{Y}$, we extend $\mathcal{Y}$ with $|\mathcal{X}|$ *dummy* tracks that are empty. We denote $\tilde{\mathcal{Y}}$ this extended set of tracks. We also denote $\Omega_{\mathcal{Y}}$ the ensemble of ordered sets of tracks that can be obtained by taking $|\mathcal{X}|$ elements from $\tilde{\mathcal{Y}}$. We compute the distance between any element $\mathcal{Z} \in \Omega_{\mathcal{Y}}$ and $\mathcal{X}$ as the sum of the distances between the $|\mathcal{X}|$ pairs of tracks defined by the ordering of the sets. This allows us to define the distance between $\mathcal{X}$ and $\mathcal{Y}$ as the minimum distance between $\mathcal{X}$ and an element of $\Omega_{\mathcal{Y}}$:

$$d(\mathcal{X}, \mathcal{Y}) = \min_{\mathcal{Z} \in \Omega_{\mathcal{Y}}} \sum_{k=1}^{|\mathcal{X}|} d(\theta^k_{\mathcal{X}}, \theta^k_{\mathcal{Z}})$$

Building the set of tracks $\mathcal{Z}^\star \in \Omega_{\mathcal{Y}}$ that gives minimum distance to $\mathcal{X}$ involves reordering $\tilde{\mathcal{Y}}$ and taking a subset of $|\mathcal{X}|$ elements from it. This task can be viewed as a rectangular assignment problem between the set of tracks $\mathcal{X}$ and $\tilde{\mathcal{Y}}$. Because of the additive and positive nature of the costs of association between tracks, this problem can be solved efficiently. In practice, we use the Munkres algorithm that gives the optimal solution in polynomial time.

IV. Performance measures

Several indices to assess the quality of a candidate set of tracks $\mathcal{Y}$ are computed based on the optimal pairing $(\mathcal{X}, \mathcal{Z}^\star)$ as defined above:

1) The distance $d(\mathcal{X}, \mathcal{Y})$, and the normalized score:

$$\alpha(\mathcal{X}, \mathcal{Y}) \triangleq 1 - d(\mathcal{X}, \mathcal{Y})/d(\mathcal{X}, \emptyset),$$

where $\emptyset$ stands for the set of $|\mathcal{X}|$ dummy tracks. By definition of $\|\cdot\|_{2,\epsilon}$, the minimal value of $d(\mathcal{X}, \mathcal{Y})$ is 0. Still by definition, $d(\mathcal{X}, \emptyset)$ is an upper bound of $d(\mathcal{X}, \mathcal{Y})$. Indeed, a pair of tracks $(\theta_\mathcal{X}, \theta_\mathcal{Y})$ is guaranteed not to be selected by the optimization process if $d(\theta_\mathcal{X}, \theta_\mathcal{Y})$ is larger than the distance between $\theta_\mathcal{X}$ and a dummy track. The normalized score $\alpha(\mathcal{X}, \mathcal{Y})$ therefore lies in the interval $[0, 1]$. It takes value 1 if the pairs of tracks in $\mathcal{X}$ and $\mathcal{Z}^\star$ exactly match. It takes value 0 if no valid match has been found: $\mathcal{Z}^\star = \emptyset$. It quantifies the quality of the best pairing of tracks, regardless of the tracks in $\mathcal{Y}$ that are not selected in $\mathcal{Z}^\star$.

2) The criterion

$$\beta(\mathcal{X}, \mathcal{Y}) \triangleq \frac{d(\mathcal{X}, \emptyset) - d(\mathcal{X}, \mathcal{Y})}{d(\mathcal{X}, \emptyset) + d(\bar{\mathcal{Y}}, \emptyset_{\bar{\mathcal{Y}}})}$$

where $\bar{\mathcal{Y}} = (\tilde{\mathcal{Y}} \setminus \mathcal{Z}^\star) \cap \mathcal{Y}$ stands for the set of tracks in $\mathcal{Y}$ that have not been selected in $\mathcal{Z}^\star$ and $\emptyset_{\bar{\mathcal{Y}}}$ denotes the set of dummy tracks of size $|\bar{\mathcal{Y}}|$. A track in $\mathcal{Y}$ may not be selected in $\mathcal{Z}^\star$ because association with other candidate tracks or dummy tracks have been preferred over it. This kind of spurious track typically consists of a combination of positions corresponding to different reference tracks or to erroneous positions originating from clutter. We thus term such a track *spurious*. The criterion $\beta(\mathcal{X}, \mathcal{Y})$ lies in $[0, \alpha(\mathcal{X}, \mathcal{Y})]$. It take values $\alpha(\mathcal{X}, \mathcal{Y})$ if there is no spurious track in $\mathcal{Y}$, that is if $\bar{\mathcal{Y}}$ is empty.

The two criteria α and β account for both association and localization error. For a more detailed characterization of tracking algorithm performance, it is useful to separate these two factors. We exclude the localization error by quantifying the number of matching positions between the pairs of tracks $(\mathcal{X}, \mathcal{Z}^\star)$. Two positions at a given time t are counted as a matching pair if they are both non-dummy and if $\|\theta_\mathcal{X}(t) - \theta_{\mathcal{Z}^\star}(t)\|_{2,\epsilon} < \epsilon$. Otherwise, the pair is counted as non-matching (with the exception of two dummy positions, which are not counted). We therefore compute:

3) The number of matching pairs of positions in $(\mathcal{X}, \mathcal{Z}^\star)$. They are called *true positive* pairs of positions and are denoted TP.
4) The number of non-matching pairs of positions in $(\mathcal{X}, \mathcal{Z}^\star)$. They are called *false negative* pairs of detections because they correspond to dummy positions or non-matching positions that are associated with track positions in the reference set $\mathcal{X}$. They are denoted FN.
5) The number of positions in $\bar{\mathcal{Y}}$. They are called *false positive* positions and are denoted FP.
6) The Jaccard similarity index for positions:

$$JSC \triangleq \frac{TP}{TP + FN + FP}$$

which lies in the interval $[0, 1]$. It takes value 1 only if all positions in $(\mathcal{X}, \mathcal{Z}^\star)$ are matching and if $\bar{\mathcal{Y}}$ is empty.

We introduce similar criteria, but that quantify the tracking quality at the track level:

7) The number of tracks in $\mathcal{Z}^\star$ that are not dummy tracks. These tracks are selected because they contain a majority of matching detections with a single track in $\mathcal{X}$. They are called *true positive* tracks and their number is denoted TP_θ.
8) The number of tracks in $\mathcal{Z}^\star$ that are dummy tracks. These *false negative* tracks are selected because there exist tracks in $\mathcal{X}$ for which a single track in $\mathcal{Y}$ with a majority of matching positions cannot be found. We denote FN_θ their number.
9) The number of tracks in $\bar{\mathcal{Y}}$. Their number is denoted FP_θ, for *false positive* tracks.
10) The Jaccard similarity index for tracks:

$$JSC_\theta \triangleq \frac{TP_\theta}{TP_\theta + FN_\theta + FP_\theta}$$

Finally, we consider criteria that exclude the association error, and quantify only the localization error. Since non-matching positions are already penalized by the above scores, we limit the assessment of localization error to matching positions:

11) The root mean-square error (RMSE) of true positive pairs (TP as above).
12) The minimum error in TP pairs.
13) The maximum error in TP pairs.
14) The standard deviation of error in TP pairs.

We believe that these 14 criteria provide a complete characterization of tracking performance.