Measures for Ranking Cell Trackers without Manual Validation

Andrey Kan[a,*], Christopher Leckie[a], James Bailey[a], John Markham[b], Rajib Chakravorty[b]

[a] *Victoria Research Laboratory, National ICT Australia (NICTA), Department of Computing and Information Systems, University of Melbourne, VIC, Australia*
[b] *Victoria Research Laboratory, National ICT Australia (NICTA), Department of Electrical and Electronic Engineering, University of Melbourne, VIC, Australia*

Abstract

Cell tracking is often implemented as cell detection and data association steps. For a particular detection output it is a challenge to automatically select the best association algorithm. We approach this challenge by developing novel measures for ranking the association algorithms according to their performance without the need for a ground truth. We formulate tracking as a binary classification task and develop our principal measure (ED-score) based on the definitions of precision and recall. On a range of real cell videos tested, ED-score has a strong correlation (-0.87) with F-score. However, ED-score does not require a ground truth for computation.

Keywords: Cell tracking, Data association, Tracking quality, Performance measures, Tracker selection, Bayes theorem

1. Introduction

Automated cell tracking has become an important tool in a wide range of biological studies, including anti-cancer drug screening, measuring the prolif-

*Corresponding author. Address: 3.25b, ICT Building, 111 Barry St, Carlton, VIC 3053, Australia. Tel.: +61 3 8344 1423; fax: +61 3 9348 1184. E-mail addresses: akan@csse.unimelb.edu.au (A. Kan), caleckie@csse.unimelb.edu.au (C. Leckie), baileyj@unimelb.edu.au (J. Bailey), jmarkham@wehi.edu.au (J. Markham), Rajib.Chakravorty@nicta.com.au (R. Chakravorty).

eration of immune system cells, and conducting wound healing assays [1, 2]. Systems for tracking cells often comprise two separate steps: cell detection and association of detected cells across frames [2, 3]. In this paper, we focus on the association step, and by cell tracker we mean an algorithm for data association.

Many cell trackers have parameters that need to be specified, for example, the process noise covariance for the Kalman filter used in the method of Li et al. [4], and the weights for association costs in the algorithm by Padfield et al. [3]. Given a cell video processed by a cell detector, users naturally want to select a tracking algorithm with appropriate parameter values that leads to the most accurate reconstruction of cell tracks. We refer to the accuracy of track reconstruction as tracking performance.

Existing measures of tracking performance [5, 6, 4] require the availability of ground truth (GT) information, for example, identities of cells throughout the video. A common approach is to generate a GT by manually annotating a subset of video frames. However, the best choice of tracking algorithm and its associated parameter settings can vary between videos, and it can be very time consuming to manually annotate every new video. Therefore, an important and open research problem in automated cell tracking is how to select algorithms and their associated parameter settings without access to the GT for videos.

To address this problem, our aim is to develop a measure for ranking trackers according to their performance without the need for a GT. The challenge in developing such measures is the absence of labeled training data. We propose a method for estimating the performance of a tracker based solely on the information available from the tracking results, such as the lengths of links made by a tracker. Our method relies on an assumption that the distribution of the lengths of wrong associations can be estimated from the sample of distances between locations within frames (Section 4.3.1). Our evaluation on real and

synthetic videos shows that the assumption holds in practical scenarios. The main application for our measures is selecting the most accurate tracking algorithm (and its parameters) for a given video and a fixed detection step. This problem can arise, for example, when cells of the same type are recorded under different treatments, so that the visual appearance of the cells is the same, but the motility varies across videos.

Although, in this paper, we consider cell tracking as our main application, our method is sufficiently general to be applicable to other domains, such as particle tracking or vehicle tracking. This is due to the fact that we develop our formulations upon a general points association problem. Furthermore, we approach the tracking problem from the perspective of pattern recognition by breaking tracks down into inter-frame links, and categorizing the links into positive and negative classes. The tracking problem then essentially becomes recognizing positive links from the set of all possible links on a given detection. This allows us to apply concepts of precision and recall in our solution.

In summary, *the contributions of our paper* are as follows: (a) we present several novel measures for ranking cell trackers according to their performance. These measures do not require a GT (Section 4); and (b) we evaluate our proposed measures using both real and synthetic videos for different trackers, and show that our measures correlate with tracking performance in practical cell tracking scenarios (Section 5).

2. Related Work

Over recent decades, there have been many proposed cell tracking methods [7, 8]. Each of these methods has a number of parameter values that need to be specified by the user. Rittscher (2010) remarks: "it is often not clear which particular ... tracking algorithm is well suited for the given data type. It would

be helpful if the decision on what type of algorithm should be used, or what particular parameter setting should be used, could be made automatically" [8]. We address this challenge by developing measures for ranking trackers according to their performance without the need for a ground truth.

We only consider the data association step, hence our method in its present form is only useful for cell tracking systems that perform cell detection and association steps separately. We note that there has been a growing number of such systems [9, 10, 6, 3], and that such systems tend to be more resilient to abrupt cell movements [3]. Abrupt movements may arise because it is often desirable to keep the time between frames large due to the known effect of phototoxicity [11]. Furthermore, we discuss a way to extend our present method to ranking arbitrary tracking systems (Section 6).

Object tracking is a fundamental task, which can be divided into two subtasks: segmentation or appearance modeling [12], and tracking locations over time through matching [13]. The matching problem arises when tracking multiple targets [14], or even for one target in the presence of noisy measurements [15]. Often the two sub-tasks can be addressed simultaneously [16].

As object appearance can vary greatly across videos and application domains, in this paper, we consider only the second sub-task to isolate the matching part. Furthermore, we do not develop a tracking algorithm, but instead focus on estimating tracking performance without manual validation. This problem has been previously addressed in the context of video surveillance systems (see the survey of SanMiguel et al. [17] and the references therein). However, this previous work is not directly applicable in the cell tracking context. In the surveillance domain, such methods tend to rely on visual keys such as color and shape, or on directed object movements. In cell videos, motions tend to be more stochastic, and color and shape tend to have less discriminative power compared

to objects (e.g., people or cars) in surveillance applications.

Finally, there has been work on automated parameter tuning or quality estimation in other domains. Abdul-Karim et al. [18] suggest to use the minimum description length principle for an automated selection of optimal parameter settings for vessel/neurite segmentation algorithms. Warfield et al. [19] use a modification of the expectation-maximization algorithm in order to automatically estimate a ground truth for medical image segmentation, and simultaneously estimate the performance of given segmentation algorithms. These previous methods focus on image segmentation and cannot be readily applied in the context of data association.

3. Problem Statement

3.1. Cell Tracking Preliminaries

Consider a video that has been processed by a *cell detector* (e.g., presented in [20]). The cell detector identifies cells in each frame of the video. A *location* is a k-dimensional vector $\vec{x} = \{x_1, \ldots, x_k\}$, where $x_i \in \mathbb{R}$, $i = 1, \ldots, k$. For example, the location can be a vector comprising the centroid position, fluorescence and size of a cell.[1] The output of the cell detector is called a *detection*. The detection is a set of discovered cell locations in a video. Note that this set can contain errors, such as spurious locations and missing true cells.

A *link* is a tuple $\{t, \vec{x}, \vec{y}\}$, such that $\vec{x}$ is a location in frame t and $\vec{y}$ is a location in frame $t + 1$. We denote the set of all possible links on detection $\mathbb{D}$ as $\mathbb{L}_{all}$. If $\vec{x}$ and $\vec{y}$ relate to the same cell then the link $\{t, \vec{x}, \vec{y}\}$ is called a *true link*. If a cell divides, the mother and each of the daughter cells are considered

[1] Note that individual components of the vector (e.g., centroid, size and fluorescence) are properties of a region of a frame that presumably corresponds to a cell. That is, the location is a point in the space of properties, and this point summarizes the region of a frame that presumably corresponds to a cell.

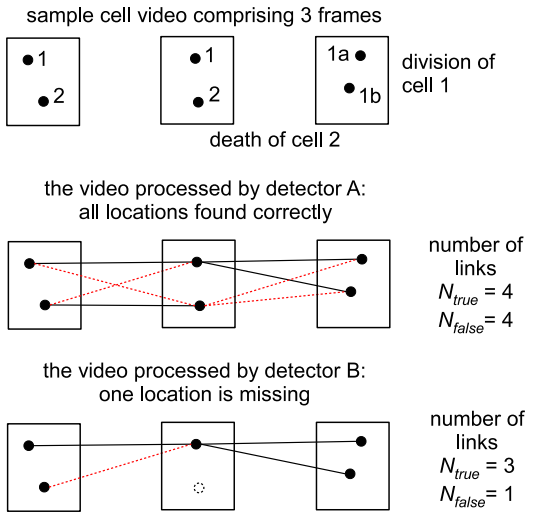

Figure 1: An illustration of the concepts of all possible links, true links, and false links. Links that connect locations of the same cell are called true links (solid black lines). Links that connect unrelated locations are called false links (dotted red lines).

to be different cells. However, a link between the mother and a daughter cell is a true link. All links that are not true links are called *false links* (Figure 1).

$\mathbb{L}_{true}$ (respectively $\mathbb{L}_{false}$) denotes the set of all true (respectively false) links in $\mathbb{L}_{all}$, so that $\mathbb{L}_{true} \cup \mathbb{L}_{false} = \mathbb{L}_{all}$ and $\mathbb{L}_{true} \cap \mathbb{L}_{false} = \emptyset$. For a given detection $\mathbb{D}$, *the ground truth (GT)* is a mapping from $\mathbb{L}_{all}$ to $\{true, false\}$.

A *cell tracker* $\mathbb{CT}$ is an algorithm that takes a cell detection $\mathbb{D}$ as input and produces a set of links $\mathbb{L}$ as output: $\mathbb{CT}(\mathbb{D}) \rightarrow \mathbb{L}_{trk} \subseteq \mathbb{L}_{all}$. The set of links $\mathbb{L}_{trk}$ produced by the tracker is a subset of all possible links $\mathbb{L}_{all}$ on the given detection $\mathbb{D}$. In other words, detection $\mathbb{D}$ implies a set of links $\mathbb{L}_{all}$ and this set is defined regardless of tracking. Cell tracking can be considered as the process of selecting links from $\mathbb{L}_{all}$.

By the cell tracker, we mean a tracking method with all its parameters set to certain values. For example, a nearest neighbor linking tracker with its gating distance set to 5, and a nearest neighbor linking tracker with its gating distance set to 7 are considered as two different trackers.

3.2. Cell Tracking Performance

In this paper, we consider tracking performance to mean a measure that reflects the accuracy of maintaining cell identities across a video. The running time of the tracker is usually less important, as long as the algorithm terminates in a reasonable amount of time. In this section, we review several previously proposed measures for tracking performance. Note that a GT is required in order to compute these previous measures.

On the level of inter-frame associations, the accuracy of maintaining cell identities can be measured by precision and recall, commonly used in the field of information retrieval. These measures capture, respectively, the proportion of spurious links made by the tracker and correct links overlooked by the tracker. Precision is defined as $prec = |\mathbb{L}_{true} \cap \mathbb{L}_{trk}|/|\mathbb{L}_{trk}|$, and recall is defined as $recl = |\mathbb{L}_{true} \cap \mathbb{L}_{trk}|/|\mathbb{L}_{true}|$ (we assume $|\mathbb{L}_{trk}| > 0$ and $|\mathbb{L}_{true}| > 0$.). Precision and recall are combined into a single measure called the F-score (also denoted as F_1 score) and defined as $F_1 = 2\frac{prec \cdot recl}{prec + recl}$.

Given detection $\mathbb{D}$, and a set of feasible trackers, we aim to select a tracker that maximizes the F-score, without using a GT.

There have been other proposed tracking performance measures, such as the proportion of swap errors [6], and the proportion of lost tracks [4]. We note that F-score and the other measures are related[2], and it is not surprising that in our evaluation we find correlation between different measures (Section 5). The derivation of our method is based on the definitions of precision and recall, and therefore is closely related to F-score. For convenience, in this paper, we mainly use F-score as the measure of tracking performance, although we also report some of the results with respect to other measures.

[2] An extended discussion is omitted due to space limitations. The discussion is available in the extended version of the paper at http://people.eng.unimelb.edu.au/akan/perfrank.html

4. Measuring Performance without Ground Truth

Our approach for solving the problem stated in the previous section is to develop a measure that estimates the tracking performance from the information available as a result of tracking. For example, one can know the number of links made by the tracker in every frame. Furthermore, one can know the lengths of the links made by the tracker. Our baseline measure presented in the next section uses the number of links a tracker makes between consecutive frames.

4.1. Number of Links

Intuitively, a "good" tracker is one that is consistent in making links across frames. Consider the output $\mathbb{L}_{trk}$ of a tracker. Let $L_i \subseteq \mathbb{L}_{trk}$ be the subset of links that end in frame i, that is, $L_i = \{l = \{t, \vec{x}, \vec{y}\} : l \in \mathbb{L}_{trk}; t + 1 = i\}$. Now $n_i = |L_i|$ is the number of links that end in frame i according to the cell tracker. Our naive measure of tracking performance is called *VN-score*. It is defined as a sample variance $VN = Var\left(\{n_2, \ldots, n_k\}\right)$, where k is the number of frames in the video. Note that VN-score does not require a GT.

Surprisingly, such a simple measure is well correlated with F-score in our real video experiments, but it does not perform well in the synthetic video experiments where tracking conditions are harsher. In fact, VN-score has intrinsic limitations. For example, swapping tracks is a common tracking error in practice. However, after a swap, the number of cells or links in a frame as perceived by the tracker does not change compared to the real number of links. Therefore a swap error is not captured by the VN-score.

Moreover, if a cell enters or leaves the field of view, the number of true links changes. However this legitimate change will be reflected in the VN-score as an error. Therefore, we proceed with developing a more reliable measure.

4.2. Lengths of Links

Another kind of information that is available without a GT is the lengths of links made by the tracker. For a given link $l = \{t, \vec{x}, \vec{y}\}$, where $\vec{x}, \vec{y} \in \mathbb{R}^k$ the *length of the link* is a function $R(l) : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$. For example, if $\vec{x}$ and $\vec{y}$ are the cell centroids in two consecutive frames, the length of the link can be the Euclidean distance between $\vec{x}$ and $\vec{y}$. Essentially, the length of a link is a distance between two locations. However, in our method, we treat a pair of locations (i.e., a link) as a single object. From this perspective it is more convenient to use term "the length of the link" instead of "the distance between the end points of the link".

In this paper, we define the link length as the Euclidean distance between centroid locations $R(\{t, \vec{x}, \vec{y}\}) = ||\vec{x} - \vec{y}||$. Note that, in general, the definition of the "link length" is not limited to the use of centroid locations only. In principle, the length might be calculated using other available information, such as cell fluorescence and area. Our evaluation shows that satisfactory results can be achieved using the centroids alone, but it would be an interesting direction for future work to consider other types of information.

Note that if the lengths of true and false links come from different distributions, then given a set of links, we can look at their lengths and make some conclusions about the proportion of true links in the set. We next develop in three steps a measure that correlates with F-score. We first introduce a measure that correlates with precision (Section 4.3). We then introduce another measure that correlates with recall (Section 4.4). Finally, we combine these two measures into one (Section 4.5).

4.3. Mirrored Precision

Consider a cell tracker $\mathbb{CT}$, with input $\mathbb{D}$ and output $\mathbb{L}_{trk}$. Let $E_f(\mathbb{L}_{trk})$ denote the event "a link l that is randomly chosen from a set $\mathbb{L}_{trk}$ is a false

link". For brevity we denote $E_f(\mathbb{L}_{trk})$ as E_f. Our main observation is that the posterior probability $P\{E_f|\mathbb{L}_{trk}\}$ is correlated with the proportion of false links in $\mathbb{L}_{trk}$ and hence has negative correlation with the precision $prec(\mathbb{L}_{trk})$.[3]

Based on this observation, we define our first measure called the *mirrored precision* (also MP-measure or MP-score) as

$$MP(\mathbb{L}_{trk}) \equiv \frac{P\{E_f|\mathbb{L}_{trk}\}}{P\{false\}} = \frac{1}{N}\sum_{i=1}^{N}\frac{P_f(R_i)}{P_{all}(R_i)}. \quad (1)$$

Here the middle and right parts of the equation are related via the Bayes' theorem. $P_{all}(R_i)$ (respectively, $P_f(R_i) \equiv P_{all}\{R_i|false\}$) are the probabilities of observing a link (respectively, false link) with a particular length R_i. That is, P_{all} (respectively P_f) is the PDF of the lengths of all (respectively, false) links. We assume that $P_f \not\equiv P_{all}$. $P\{false\}$ is the a priori probability that a link is a false link. For a given input $\mathbb{D}$, $P\{false\}$ is an (unknown) proportion of false links implied by $\mathbb{D}$, that is, a fixed value.

A useful property of the mirrored precision is that it correlates with precision as formalized in the proposition below (a proof sketch is given in Appendix A). This property is essential for our work.

Proposition 1. *Let $X = MP(\mathbb{L}_{trk})$ and $Y = prec(\mathbb{L}_{trk})$ be real-valued random variables (on the sample space Φ where an outcome is a tracker). Then if $P_f \not\equiv P_{all}$, $Cov(X,Y) < 0$, where $Cov(\ldots)$ denotes the covariance.*

4.3.1. Estimation of PDFs of Link Lengths

We estimate the PDFs of the lengths of all links and false links from the input detection. We estimate P_{all} from a sample of all links that one can make

[3]We define the event E_f as "... a link ... being a false link", not a true link, because it is more convenient for us to recover the distribution of lengths of false links, as we show in Section 4.3.1. To reflect the fact the the correlation is negative we call our measure *mirrored* precision.

in the given detection. That is, we connect each location in a frame with each location in the subsequent frame. We estimate the PDF from a sample using the kernel density estimation method from Botev et al. [21].

We also need to estimate P_f. However, without a GT, it is not known which links are false links. Therefore, we build upon a previous observation that the distribution of false links can be estimated from the the distances between locations *within a frame* [6]. Let P_{within} be the PDF of a random variable that denotes the Euclidean distance between a random pair of locations within a single frame in the video. We can estimate P_{within} from a set of pairwise distances between locations within each frame. We then use the estimated P_{within} instead of P_f based on the following assumption.

Assumption 1. *We assume that* $P_{within} \approx P_f$.

It is important to note that the assumption was empirically validated on a range of real and synthetic cell videos (Section 5). Moreover, below we explain why we expect the assumption to hold in practical tracking scenarios.

Consider two locations chosen at random within a single frame, let one of the locations be randomly chosen as the origin, and let dx_{within} be a random variable denoting the difference between the x coordinates of the locations (the same argument applies to the y coordinate). For a pair of consecutive frames, let dx_{true} and dx_{false} be random variables denoting the differences between the x coordinates of the locations in the two frames. Here dx_{true} corresponds to a true link and dx_{false} corresponds to a false link. We can express the differences in coordinates for a false link as $dx_{false} = dx_{true} + dx_{within}$ (Figure 2).

By construction, the mean of dx_{within} is 0. We also assume zero means for dx_{true} (otherwise there is a common offset that can be subtracted). Note that if $Var(dx_{true}) \ll Var(dx_{within})$ then $dx_{false} \approx dx_{within}$ and Assumption 1 is satisfied. For example, in our real videos, typical variances satisfy $Var(dx_{true}) <$

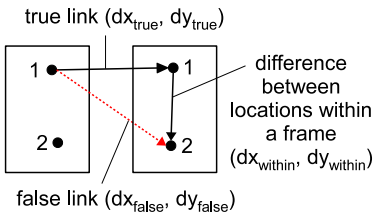

Figure 2: The length of a false link can be expressed with the length of the corresponding true link and the distance between locations within a frame.

36 pixels and $Var(dx_{within}) > 5100$ pixels.

Furthermore, as $Var(dx_{true})$ approaches $Var(dx_{within})$, cells tend to travel larger distances between consecutive frames compared to the distances between different cells. This usually results in a degraded tracking performance. In turn, if the maximum achievable tracking performance is low, an inaccurate MP-score is not a critical problem. More generally, in order for cells to be identifiable, one can expect the existence of a property (or combined property) such that the variation in the property for the same cell in two consecutive frames is small compared to the variation across different cells in the same frame. In this case, the assumption holds with respect to this property.

In summary, in order to estimate the distribution of false links, we collect the distances between locations within individual frames. We then use kernel density estimation to obtain the PDF of these within-frame distances, and use this PDF as an approximation of P_f. From our evaluation, we conclude that our estimation method is reliable in practical cell tracking scenarios.

4.4. Mirrored Recall

We now have a measure to estimate the precision of tracking, and we need a measure to estimate the recall. We note that under certain circumstances, recall is closely related to precision. For example, when the outputs of different trackers all have the same number of links N, then the precision of such trackers depends only on the number of true links included in their output (true posi-

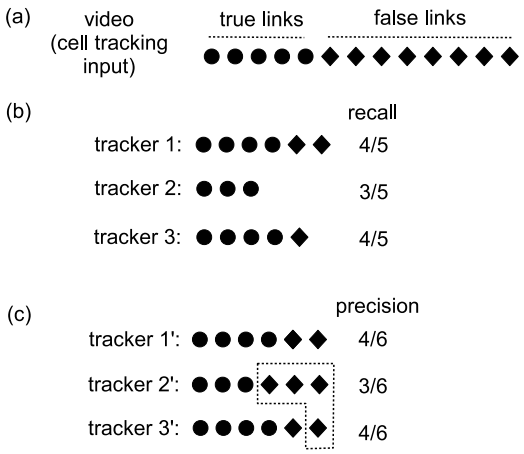

Figure 3: Equalizing the sizes of tracking results. (a) An output of a cell detection algorithm implies sets of true and false links. (b) Different trackers can produce different results, and different results can have different recall values. (c) Each tracker's results are appended with false links randomly drawn from the set of false links in the video. The precision of the padded results is correlated to the recall of the original results.

tives). At the same time, recall also depends on the number of true positives. Therefore, if different trackers produce outputs of the same size then precision is proportional to recall.

In practice, different trackers can produce outputs of different sizes. Therefore we equalize the sizes of different outputs, in order to express tracking recall via precision. We append shorter outputs to the size of the largest output N_{max} with dummy links with lengths independently randomly drawn from the PDF of lengths of false links P_f (Figure 3). In the rest of this section, we summarize the calculation of mirrored recall using the equalized output sizes. The proposed dummy link generation process is justified in Appendix B.

We denote the padded output as $\mathbb{L}'_{trk} = \mathbb{L}_{trk} \cup \mathbb{L}_{pad}$. Let $prec'(\mathbb{CT}(\mathbb{D})) = prec(\mathbb{L}'_{trk}) = \frac{T}{N_{max}}$, that is, $prec'$ is the precision of the padded output. Here $T = |\mathbb{L}_{true} \cap \mathbb{L}_{trk}|$ is the number of true links in $\mathbb{L}'_{trk}$, which is the same as the number of true links in $\mathbb{L}_{trk}$. Note that $recl(\mathbb{CT}(\mathbb{D})) = \frac{T}{T^*}$, where $T^* = |\mathbb{L}_{true}|$. For a given input $\mathbb{D}$ and set of trackers Φ, both T^* and N_{max} are fixed. Hence

we have that $prec'(\mathbb{CT}(\mathbb{D})) \propto recl(\mathbb{CT}(\mathbb{D}))$.

We now define the *mirrored recall* (also MR-measure or MR-score) as $MR = MP(\mathbb{L}'_{trk})$, where MP is calculated using Equation 1. From Proposition 1 it follows that $MR = MP(\mathbb{L}'_{trk})$ is correlated with $prec'(\mathbb{L}_{trk})$, and hence MR-score is negatively correlated with the recall $recl(\mathbb{L}_{trk})$.

4.5. Combined Performance Measures

In order to produce a ranking of trackers, we need to combine the MP and MR-scores into a single measure. Our approach is based on representing each cell tracker as a point in a two-dimensional Euclidean space. Note that smaller values of MP and MR-scores correspond to higher precision and recall respectively. Therefore, an ideal tracker is expected to have both MP and MR-scores equal to zero. We define our second combined measure *ED-score* as the Euclidean distance to this ideal point $ED = \sqrt{MP^2 + MR^2}$.[4]

An alternative approach to combine the scores is to use a dimensionality reduction approach. We apply principal component analysis (PCA) to the MP and MR-scores, and use the first principal component as a combined measure. We denote this measure as *PC-score*. Note that PCA does not guarantee optimality in terms of correlation with F-score. For example, MP-score can have a perfect correlation with F-score and have a small variance, while MR-score can be uncorrelated with F-score and have a large variance. In this case, PC-score will be essentially equal to MR-score, which is not the best possible combination.

In our evaluation, we also tested the combination of VN, MP and MR-scores (i.e., three measures). We denote the respective combined scores as *ED3* and *PC3-scores*. ED3-score is the quadratic mean of VN, MP, and MR-scores with

[4]We find that using the Euclidean distance is a more reliable method than using the harmonic mean of MP and MR-scores. An extended discussion is omitted here due to space limitations, but it is available online at http://people.eng.unimelb.edu.au/akan/perfrank.html

Table 1: Summary of tracking performance measures that we introduce in this paper. None of these measures requires GT. Details of our evaluation are given in Section 5.

Meas.	Definition	Evaluation
MP	Equation 1	Strong negative correlation with *prec*
MR	Equation 1 (padded set)	Strong negative correlation with *recl*
ED	$\sqrt{MP^2 + MR^2}$	Strong negative correlation with F-score
PC	PCA of $[MP, MR]$	Weak negative correlation with F-score
VN	Var. of the number of links	Unreliable (sometimes performs well)
PC3	PCA of $[MP, MR, VN]$	Does not perform well
ED3	$\sqrt{MP^2 + MR^2 + VN^2}$	Does not perform well

equal weights, and PC3-score is the first principal component after applying PCA to VN, MP and MR-scores. We summarize all our measures in Table 1.

In the next section, we show how our combined measures can be used in practical scenarios, and compare the effects of using different measures.

5. Evaluation

The aims of our evaluation have been to (i) validate the proposed measures in practical scenarios, and (ii) study the reliability of the measures under conditions that are expected to hamper our method. In order to achieve these aims, we use different cell trackers previously proposed in the literature, and vary the parameter values used in these trackers. Furthermore, we use a range of real and synthetic videos as inputs.[5]

Our experiments are grouped in three classes. The first class comprises experiments on real videos (Section 5.1). These experiments are used to evaluate our measures in practical tracking scenarios. Furthermore, we have two groups of experiments on synthetic videos (Section 5.2). These experiments are designed to test our measures under more challenging tracking conditions.

In all groups, each experiment has the same structure. We run the trackers

[5]Selected real videos and cell trackers are available online at http://people.eng.unimelb.edu.au/akan/perfrank.html

on the same input. For each tracker, we compute the F-score using the GT, and we compute the ED-score without the GT. We then select an optimal tracker based on the computed ED-scores. As our baselines we use the performance of the tracker selected using the VN-score. Another baseline is the F-score of a randomly selected tracker. Finally, in each experiment, we also compute Spearman's correlation between the F-score and the proposed measures.

5.1. Real Cell Videos

In our evaluation, we use five real videos. These videos have been previously published [6], and we only briefly summarize them here (Table 2). One of the videos (named ak[6]) shows neural progenitor cells, and the remaining four videos show B lymphocyte cells. The videos show cells at different densities and speeds. The difficulty of tracking cells in a video may depend on a variety of factors, such as density and relative sizes of cells. However, to our knowledge, there is no standard way to formalize the "difficulty". Here, just to give an additional insight into our videos, we adopt the concept of a normalized cell density β_n [22]. It describes how fast cells move compared to the distance between cells $\beta_n = N \cdot \sigma^2 \cdot \pi / L^2$. Here N is the number of cells in a frame, σ^2 is the variance of cell offsets between consecutive frames, and L is the side of a square frame. Higher values of β_n correspond to faster motions and more difficult videos. Our videos contain a few cell divisions and deaths, so the number of cells changes across frames. We use the average number of cells per frame to estimate β_n. Furthermore, in some of our videos, cells tend to group in clusters in the middle of the frame, so we estimate L as the side of the smallest square that outlines all cells in a frame. This square can be smaller than the actual frame.

[6]This video is an excerpt from supplementary movie 1 for the work of [9]. The supplementary file is available on-line from the Cell Cycle journal website at http://www.landesbioscience.com/journals/cc/supplement/alkofahi.zip

Table 2: Real videos cover a variety of tracking conditions. #cells is the number of distinct cells. For example, if a mother cell divides into two daughter cells, then there are three distinct cells. β_n is the normalized object density defined in the main text. To obtain values of β_n, divide the numbers in the table by 10^2.

video	**#frames**	**#cells**	$\beta_n \times 10^2$
ak	300	18	0.25
hex.6	100	6	5.94
hex.16	100	16	3.45
hex.22	100	22	3.18
square	50	35	0.14

For each video, we manually produce the GT for cell locations. We then add random noise to the GT, in order to simulate different levels of detection quality. We generated detection levels 0, 1, 3, 5, 10 and 20, where level k means that there are k% of spurious locations *(false positives)* and k% of missing true locations *(false negatives)*. Therefore, in total we have 5 real videos each at 6 detection levels, and hence 30 inputs. We denote an input as *movie*@k. For example, *hex*.16@3 denotes the detection for the *hex*.16 video that contains 3% of false positives and 3% of false negatives.

5.1.1. Cell Trackers

We use three previously proposed cell tracking methods with different parameter settings (Table 3). We asked the authors of the corresponding methods which parameters, in their opinion, are most influential for tracking. Based on their responses we select up to three parameters. The authors of the algorithms did not suggest any specific values for the parameters, but, where appropriate, pointed out a range of reasonable values for each parameter. We then set arbitrary values for the selected parameters such that they fall into the appropriate ranges. For example, the probability should be in the range $[0, 1]$, and the maximum allowed spatial distance should be comparable with the frame dimensions. In total, different combinations of the tracking methods and selected parameter values give us $5 + 27 + 27 = 59$ different trackers (see Table 3).

Table 3: Real trackers and their parameters. The right column shows the number of trackers.

Method, authors	**Parameters**	**trackers**
NENIA (version A), Kan et al. [6]	gating distance = $\{5, 10, 20, 50, 100\}$	5
LJIPDA, Musicki and Evans [23]; Chakravorty [unpublished]	process covariance = $\{0.1, 1, 10\}$ detection probability = $\{0.7, 0.8, 0.9\}$ termination probability = $\{0.1, 0.05, 0.01\}$	$3\times3\times3 =$ 27
u-track, Jaqaman et al. [10]	maximum search radius = $\{5, 20, 100\}$ time window= $\{1, 5, 10\}$ st. dev. multiplication= $\{2, 3, 4\}$	$3\times3\times3 =$ 27

The NENIA (version A) tracker has only one parameter. LJIPDA (hybrid) and u-track[7] have more than three parameters. We set the remaining parameters to their default values as provided by the authors of the corresponding methods.

LJIPDA requires a supplementary module to resolve cell divisions. We did not implement this module in our evaluation. As a result, when division occurs LJIPDA can only follow up to one of the daughter cells, and hence LJIPDA always misses the link to the second daughter. Consequently LJIPDA tends to perform slightly worse than the other algorithms. This is not important for our evaluation, since our aim is to rank different trackers.

Recall that we aim to find the best tracker for a given video. Two key components of this optimization problem are the choice of measure to optimize and the optimization strategy. In this paper, we focus on the choice of the performance measure. As we argue in the Introduction, this is not a trivial task without the ground truth. Consequently, we adhere to a naive grid search based strategy: (1) collect a pool of available trackers; (2) identify a reasonable range and steps for the parameters of the given trackers (e.g., from documentation); and (3) try all combinations of trackers and parameters. In the case where there

[7] $u-track$ is available online at http://lccb.hms.harvard.edu/software.html. Strictly speaking, this is a multiple particle tracker. However, it is often used and cited in the cell tracking domain. For convenience of narration we refer to this tracker as a cell tracker in our paper.

are too many trackers available at step (1), the researcher might like to select a smaller subset based on practical considerations (e.g., popularity or ease of use). For some trackers, there might be a more efficient strategy for searching the parameter space, but such a method is beyond the scope of this paper.

Finally, we validate our strategy in a practical setup. The most time consuming part of our evaluation comprises individual tracker runs. The time for a single run depends on the tracker, but we only use a short portion of the video in a run. For example, on our five real videos at detection level 3, single run times in seconds are $[2.8; 3.7]$, $[0.6; 0.8]$, $[2.5; 2.9]$, $[3.3; 4.0]$, $[2.1; 2.5]$ (95% bootstrap confidence intervals for the mean of a sample of 59 runs) on a modern PC. In other words, a grid search for the best tracker among 59 combinations can be completed within 5 minutes in a sequential execution. This can be considerably faster than creating a ground truth manually. Moreover, single runs can be executed in parallel.

5.2. Synthetic Videos

The goal of our synthetic experiments is to cover tracking conditions that were not covered in our real videos. Such conditions include (i) large numbers of cells, (ii) abrupt movements and frequent divisions, and (iii) changing density.

We implement two cell density changing models in two groups of synthetic videos. In the first group, the number of cells increases due to divisions, but all cells are constrained within fixed frame boundaries. In the second group, cells are not constrained and the number of cells does not change, however, cells have a directed motion outwards from the middle of the frame. This kind of motion can be observed in wound healing assays [2]. In a wound healing assay cells usually move towards the wound (inwards rather than outwards), but this does not make a difference for our evaluation. More details on our synthetic videos are given in the next section.

We do not test with videos with high numbers of frames, because in general a long video can be divided into parts. Some tracking algorithms are global in the sense that they can achieve higher performance given the whole video at once. Such trackers can be given the complete video as input. We then can select a part or a few parts of the video and estimate the performance of the tracker on each part.

For each synthetic video, we added random noise to the detection, such that each detection contains 3% of false positives and 3% of false negatives. In total, we have 20 synthetic inputs: 10 synthetic videos in each of two groups, and one detection level for all videos.

5.2.1. Group One Synthetic (G1)

We generate two groups of synthetic videos. Each group comprises 10 synthetic videos. Every video has 10 frames, and starts with 100 cells uniformly placed in the first frame.

In the first group (G1), we use a Brownian motion model with the same variance of the offset σ^2 in both spatial directions. The variance takes different values in different videos in such a way that β_n takes values from 0.1 to 1 with a step of 0.1 (in our real videos, the maximum estimated β_n is around 0.06, i.e., the motion is much slower). By definition $\beta_n = N \cdot \sigma^2 \cdot \pi / L^2$, where we use the number of cells $N = 100$, and the side of the square frame in pixels $L = 100$.

In our G1 videos, the motion of cells is constrained within the frame boundaries. If a generated inter-frame offset for a cell leads outside the frame, we move the cell towards the direction of the offset, but leave it within the frame at some random distance from the boundary. A visual inspection of the produced videos shows that this motion amendment does not lead to dramatic skews in the spatial distribution: the placement of cells remains approximately uniform.

Every odd frame, 20% of the cells divide, and so the second frame has 120

cells, the fourth frame has 144 cells, and so on. The last 10-th frame contains 200 cells. We do not recalculate the offset variance σ^2. This implies that β_n increases further (i.e., tracking becomes harder) throughout the video.

5.2.2. Group Two Synthetic (G2)

In the second group (G2) of synthetic videos, there are no divisions and the number of cells is 100 in each frame. After placing the cells in the first frame, we draw an imaginary vertical line that divides the frame into halves. We then move cells to the left of the line to the left with some constant offset plus random noise, while we move cells to the right of the line to the right with the same constant offset and noise.

We characterize the constant offset in relation to the expected distance to the nearest cell denoted as D. It is not convenient to measure the offset in relation to the frame side L because such a measurement does not take into account the initial density of cells. Different videos in the second group have different offsets αD, where α takes values from 0.1 to 0.55 with a step of 0.05. We estimate D as $D = L/\sqrt{\pi \cdot N}$. The intuition is that if D is the distance to the nearest cell then there is only one cell expected to be in a circle with radius D. On the other hand, there are N cells in a square frame with side L, and hence $D = L/\sqrt{\pi \cdot N}$. This is only a coarse estimation of D, but it suffices for the purposes of our evaluation.

Note that L denotes the side of the first frame where we place cells uniformly. In the subsequent frames, cells move apart and the effective frame size grows (the field of view is not restricted). In our simulations, we need a control over D and hence it is slightly more convenient to move cells outwards.

5.3. Summary of Results

In what follows we report correlations using Spearman's rank correlation coefficient r, because (a) we are interested in ranking trackers, and (b) this

coefficient operates on ranks rather than absolute values and therefore is more reliable to outliers than Pearson's coefficient.

Apart from F-score, in our evaluation we also use proportion of swap errors $\frac{\#\,swaps}{\#\,locations}$, and proportion of lost tracks $\frac{\#\,lost\,tracks}{\#\,all\,tracks}$. All these measures require GT, and all of them are correlated. On average, over all our real and G1 experiments, correlation between the F-score and the proportion of swap errors (respectively lost tracks) is -0.89 (respectively -0.62). Furthermore, we note that F-score is the harmonic mean of precision and recall, whereas our ED-score is the quadratic mean for mirrored precision and recall. We therefore also look at the correlation between the harmonic and quadratic means for precision and recall. In every real and synthetic experiment we find a strong correlation of approximately 0.99 between the F-score and the quadratic mean of precision and recall. Therefore, in what follows we report results with respect to F-score.[8]

5.3.1. Preliminary Experiments

We first verify the validity of Assumption 1 on real videos by comparing the empirical distributions of false links (obtained using the ground truth) and the distances measured within the frames. We find that in all cases the two distributions are similar (representative cases are shown in Figure 4). These results justify Assumption 1 on a range of practical scenarios tested.

We then validate our MR-score, which is based on dummy links generation (Section 4.4 and Appendix B). In the real and G1 experiments MR-score is well correlated with recall. The weakest (respectively average) correlation coefficient for the real experiments is -0.77 (-0.95), and for the G1 experiments it is -0.65 (-0.83). MP-score and precision are also well correlated, except for a few experiments. This correlation is weak only when the variance of precision

[8]Results with respect to some other measures can be found at http://people.eng.unimelb.edu.au/akan/perfrank.html

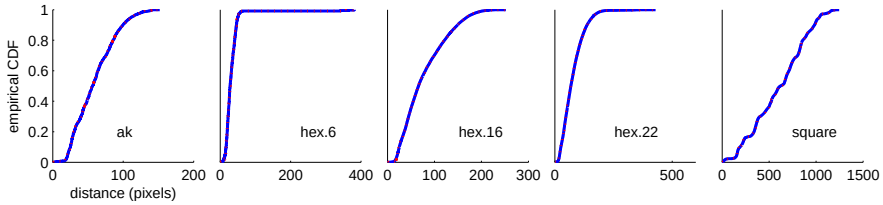

Figure 4: Empirical CDF of the distances measured within frames (red dashed line) closely resemble the CDF of the length of false links (blue solid line). The figure shows real datasets at detection level 3. The p-value of the Kolmogorov-Smirnov test for each of the plots is > 0.4.

is small. In all experiments, where the correlation coefficient between mirrored precision (MP-score) and precision is weaker than -0.3, the standard deviation of precision is below 0.01. In such experiments, precision does not affect the variation in tracking performance, and therefore the overall correlation between ED and F-scores is strong, as we quantitatively confirm in the next section.

Finally, we study the effect of different parameters on the tracking performance. Figure 5 shows the F-score against ED-score in a representative experiment on real videos. This and other experiments (not shown) confirm that different parameter settings can lead to significantly different performance. Furthermore, on different inputs the best performance can be achieved with different trackers. There is not necessarily a "winning" cell tracker. The relation between ED-score and F-score is explored in the next section.

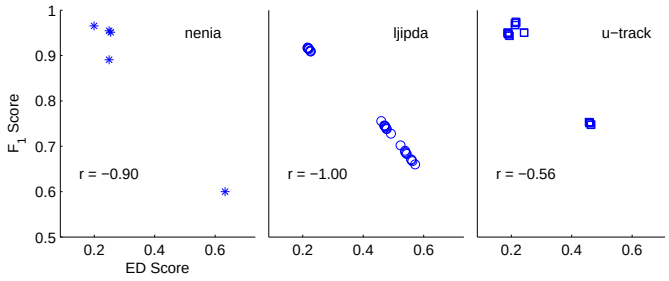

Figure 5: Experiment with the real video *hex*.16 at detection level 3. There are in total 59 points in three plots. Each point is the result of a run of a tracker with its parameters fixed to certain values.

5.3.2. Real and G1 Experiments

We present the results of our evaluation in Figure 6. The top row shows observed correlation r between F-score and ED-score for each input (video @ detection level). This plot also shows bootstrapped confidence intervals for r. Bootstrapping is performed by generating 300 samples drawn with replacement from the set of candidate trackers on a fixed input.

Overall, ED-score is strongly correlated with F-score ($|r| > 0.5$, note that the correlation is negative). The correlation is weaker on inputs "ak@10" and "ak@20", because in these cases all candidate trackers perform similarly and there is little variation in measures (F-score remains around 0.95). In our G1 videos, ED-score is also strongly correlated with F-score. The weakest correlation observed across 10 G1 experiments is -0.62, and the mean value for the correlation coefficient is -0.79 (data not shown).

The middle row in Figure 6 shows the tracking performance achieved using random selection, VN-score ED-score, and the ground truth. ED-score outperforms random selection (for each real or G1 input $p < 0.01$, Wilcoxon signed-rank test applied to a bootstrapped sample as described above). The bottom row compares the performance achieved using random selection and ED-score where candidate trackers are grouped by software package. In this evaluation, for a given input we tested different parameter settings for NENIA, LJIPDA or u-track. In each group ED-score outperforms random selection ($p < 0.01$).

On real videos, VN-score performs similarly to ED-score. However, in each of our G1 videos ED score outperforms VN-score, PC-score, as well as, ED3 and PC3 scores (Table 4). We note that in our real videos trackers tend to have high precision, and hence the F-score is effectively determined by recall only. In this scenario, the VN-score performs well. In the G1 videos, both precision and recall vary, and the VN-score does not estimate the performance

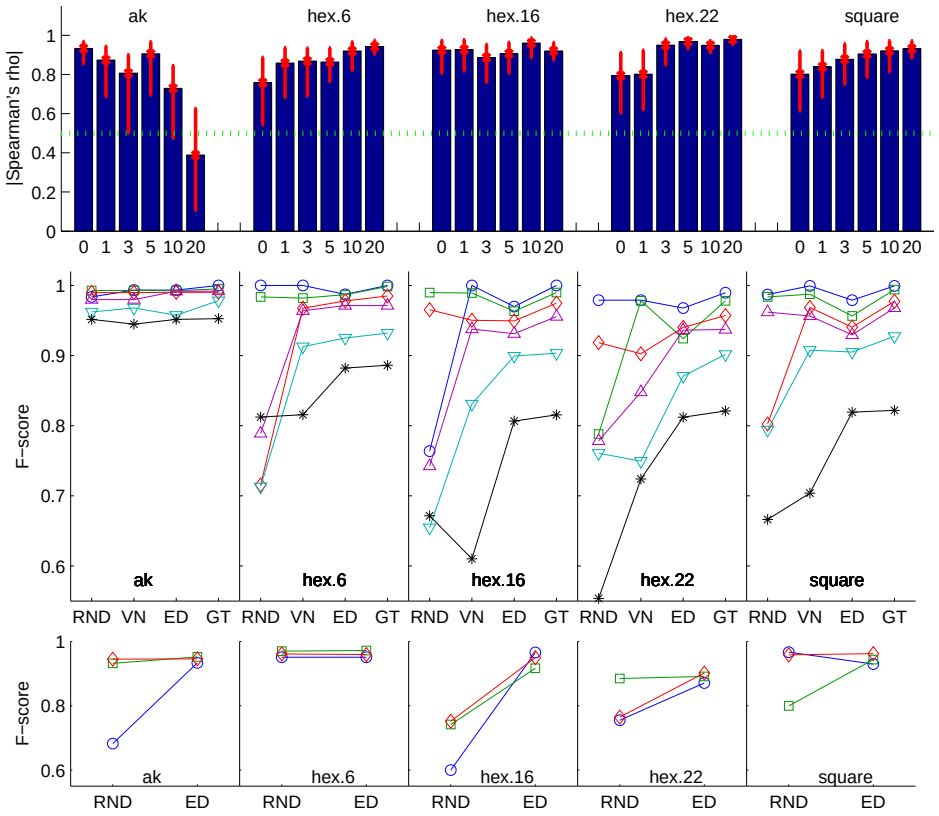

Figure 6: ED-score is strongly correlated with F-score, and selecting a tracker using ED-score is better than random selection. Top row: Spearman's correlation r and bootstrapped confidence intervals for r between ED-score and F-score for different inputs. Bars correspond to detection levels. Middle row: performance achieved using random selection (RND), VN-score, ED-score, or ground truth (GT) on different videos and detection levels (0 - circle, 1 - square, 3 - diamond, 5 - arrow up, 10 - arrow down, 20 - asterisk). Bottom row: performance achieved using random selection or ED-score for trackers grouped by tracking software (NENIA - circle, LJIPDA - square, u-track - diamonds, inputs: randomly chosen detection levels for each video).

Table 4: The results of the G1 experiments (in rows). Columns show the F-score achieved using random selection (RND), VN, ED, ED3, PC, and PC3-measures, and a ground truth (MAX). ED-score outperforms random selection ($p < 0.01$). However, the performance achieved by VN, ED3, PC, and PC3-scores is not statistically different from random performance.

β_n	**RND**	**VN**	**ED**	**ED3**	**PC**	**PC3**	**MAX**
0.1	0.605	0.689	0.715	0.689	0.689	0.689	0.721
0.2	0.651	0.596	0.619	0.596	0.600	0.596	0.654
0.3	0.447	0.498	0.520	0.498	0.492	0.498	0.545
0.4	0.419	0.406	0.434	0.406	0.405	0.406	0.481
0.5	0.382	0.371	0.393	0.371	0.371	0.371	0.438
0.6	0.317	0.338	0.379	0.338	0.320	0.338	0.412
0.7	0.276	0.306	0.368	0.306	0.353	0.306	0.373
0.8	0.298	0.290	0.355	0.290	0.328	0.290	0.355
0.9	0.277	0.280	0.302	0.280	0.302	0.280	0.322
1.0	0.247	0.240	0.313	0.240	0.273	0.240	0.313
min	0.247	0.240	0.302	0.240	0.273	0.240	0.313
max	0.651	0.689	0.715	0.689	0.689	0.689	0.721
mean	0.392	0.401	0.440	0.401	0.413	0.401	0.462
st.dev.	0.141	0.148	0.136	0.148	0.138	0.148	0.140

adequately. The poor performance of the VN-score in the G1 videos contributes to the poor performance of the combined ED3 and PC3-scores on these videos. Furthermore, the G1 experiments illustrate that the first principal component is not necessarily strongly correlated with F-score.

5.3.3. Stress Test with G2 Experiments

The ED-score relies on the assumption that the distribution of false links can be estimated from the distances between locations in individual frames (Assumption 1). Our evaluation on G1 videos reveals that when cells are uniformly distributed within a frame, the assumption holds even when cells move fast, and when intensive division process is present.

The G2 videos allow us to perform a stress test for another scenario when groups of cells move apart. The results of the G2 experiments are presented in Table 5. An increased cell motion affects Assumption 1. As a result, the estimation of the length distribution of false links becomes poorer (as measured

Table 5: The results of our evaluation on the G2 videos. "alpha" characterizes the relative magnitudes of the directed motion components of cells, "r" is the correlation between ED and F-scores, "p-val" is the p-value of the Kolmogorov-Smirnov test for the real and estimated distributions of false links, and "Max" is the maximum F-score that we observe among our real trackers.

alpha	p-val	r	Max
0.10	0.990	-0.820	0.915
0.15	0.999	-0.757	0.905
0.20	0.973	-0.719	0.899
0.25	0.946	-0.700	0.900
0.30	0.949	-0.789	0.894
0.35	0.786	-0.772	0.892
0.40	0.536	-0.731	0.883
0.45	0.264	-0.568	0.874
0.50	0.119	-0.422	0.872
0.55	0.033	0.123	0.858

with the p-value of the Kolmogorov-Smirnov test for the real and estimated distributions). Accordingly, ED-score becomes less correlated with F-score. At the same time, as the motion becomes faster, the maximum F-score achieved among our real trackers decreases. In other words, on videos with such fast motion one is unlikely to achieve reasonable tracking anyway.

6. Discussion

The advent of high content assays in screening applications makes it impracticable to tune a cell tracker for every new video. Consider a number of videos produced in a course of a biological experiment. In the experiment, cells of the same type are recorded under different treatment conditions, and the treatment is expected to affect cell motility or lifetimes. A visual inspection of a few frames from one of the videos can guide the choice of a cell detection algorithm and its parameters. The same detection algorithm can then be used for all the videos. At the same time, as the motility of the cells may vary across videos, one might like to choose a different tracking algorithm (or different tracking parameters) for different videos. For a given video, one can run all candidate trackers and

select the tracker that achieves the lowest ED-score.

ED-score performs reasonably well even in the presence of abrupt movements and intensive divisions. Furthermore, ED-score performs well in our wound healing type synthetic videos, when the relative directed speed of cells satisfies $\alpha < 0.4$. We obtained three previously reported real videos of wound healing assays [2]. We do not have the complete GT for these videos, but we estimated the relative magnitude of the directed speed component. We find that, in these real videos, $\alpha \approx 0.3$ which indicates that ED-score can be applicable in real wound healing assays.

Our evaluation shows that ED-score can be applicable to videos where objects exhibit different motion styles. Recall that we have evaluated ED-score in the situation where there is increasing cell density, and in the situation where two groups of cells move in the opposite directions. Also note that in our real videos (first introduced in [24]) some cells occasionally move notably faster than others within the same video. Essentially, our real videos show a mixture of different motions. However given practical frame rates, these can be considered as a single population of moving objects. Note that our method does not make any direct assumptions on cell motion style. As we discuss in Section 4.3, it is not the motion style per se that affects the validity of Assumption 1, but instead the amount of variance in inter-frame displacements compared to the distance between the objects. For example, if different groups of objects move apart this will increase the variance. However, an increased frame rate would decrease the variance such that Assumption 1 holds again.

Finally, we comment on the applicability of ED-score in other tracking domains. In this work, we focus on cell tracking and test our measures in the scenarios specific to this domain (e.g., the in presence of divisions and deaths of cells). However, our definitions of the *location* and the *link* are general enough

to be applied to a wide range of domains including tracking people, cells or particles. According to our definition, the location is merely a point in some space of features that describe an object to be tracked. For a given tracking problem, one can identify the appropriate features and define the link length accordingly. The only requirement is that tracked objects satisfy Assumption 1. As we explain in Section 4.3.1, if the objects are well identifiable by the features the assumption is likely to be satisfied.

7. Conclusions

In this paper, we address the problem of ranking cell trackers according to their performance. By cell tracker we mean a point association algorithm, and by performance we mean the accuracy of track reconstruction. We have developed several novel measures for estimating the tracking performance without the need for the ground truth. Our measures can be used to automatically select the most appropriate tracker and its parameters for a given set of cell detection results. One of our measures (VN-score) uses the variance in the estimated number of links across frames. Other measures (ED and PC-scores) use the lengths of links in the tracker's output. In our evaluation, we find that ED-score correlates with previously proposed measures for tracking performance, and is the most reliable proxy for the performance among the proposed measures.

Acknowledgments

We would like to thank Dr Khuloud Jaqaman (Harvard Medical School) for the advice regarding u-track; Dr Zhaozheng Yin (Carnegie Mellon University) for sharing the videos of wound healing assays; and Dr Daniel Day (Swinburne University of Technology) for supplying the 125 μm microgrids. This work is partially supported by National ICT Australia (NICTA). NICTA is funded

by the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.

Appendix A. Proof Sketch for Proposition 1

Let $\mathcal{S}_i$ be the set of all possible trackers that yield outputs with N_i links and that all have T_i true links in their output. We choose arbitrarily two such sets and without loss of generality index them with $i = 1, 2$. Now let $\mathcal{EP}_i = E\left[MP(\mathbb{L}_i)\right]$ be the expected value for the MP-score, where the expectation is taken over all possible outputs $\mathbb{L}_i$ from trackers in $\mathcal{S}_i$. From equation 1 (main text) we have that

$$\mathcal{EP}_i = \frac{1}{N_i}\left(\sum_{t\in\mathcal{T}_i} E[\zeta(R_t)] + \sum_{f\in\mathcal{F}_i} E[\zeta(R_f)]\right). \tag{A.1}$$

Here $\zeta(R_j) = \frac{P_f(R_j)}{P_{all}(R_j)}$ is a summation term from equation 1; t and f enumerate individual links in $\mathbb{L}_i$, and $\mathcal{T}_i$ and $\mathcal{F}_i$ are the sets of indices for true and false links respectively. The expectations are taken over all possible outputs $\mathbb{L}_i$ from trackers in $\mathcal{S}_i$. For example, $E[\zeta(R_1)]$ is the expected value of $\zeta(R_1)$ where R_1 is the random variable denoting the length of the first link in the outputs from trackers in $\mathcal{S}_i$.

Given a tracker $s \in \mathcal{S}_i$ there exists another tracker $s^* \in \mathcal{S}_i$ that takes the result of s and permutes it. Consequently, for any two indices $k, l \in \mathcal{T}_i$, we have that $E[\zeta(R_k)]$=$E[\zeta(R_l)]$. The same reasoning applies to indices from $\mathcal{F}_i$. We can now let $\mathcal{ET}_i = E[\zeta(R_t)]$, $t \in \mathcal{T}_i$ and $\mathcal{EF}_i = E[\zeta(R_f)]$, $f \in \mathcal{F}_i$, and simplify equation A.1 to

$$\mathcal{EP}_i = \left(T_i \cdot \mathcal{ET}_i + (N_i - T_i) \cdot \mathcal{EF}_i\right)/N_i \tag{A.2}$$

Note that $\mathcal{ET}_i = \int_0^\infty P(R_t = r)\zeta(r)\,dr$, $t \in \mathcal{T}_i$. Further, assuming that the trackers in $\mathcal{S}_i$ are equally likely (according to the principle of indifference), we

have that $P(R_t = r) = P_t$. We can then put $\mathcal{ET} = \mathcal{ET}_1 = \mathcal{ET}_2$. The same reasoning apply to false links, and so we can put $\mathcal{EF} = \mathcal{EF}_1 = \mathcal{EF}_2$.

Proposition 1 states a negative correlation between the MP-score and the precision T_i/N_i. Alternatively this can be formulated as a condition $\mathcal{EP}_1 > \mathcal{EP}_2 \iff T_1/N_1 < T_2/N_2$.

From equation A.2, we have the difference

$$\mathcal{EP}_1 - \mathcal{EP}_2 = (\mathcal{ET} - \mathcal{EF}) \cdot (T_1 N_2 - T_2 N_1)/(N_1 N_2). \tag{A.3}$$

For Proposition 1 to hold, we require $\mathcal{EF} > \mathcal{ET}$, which can be expanded as $\mathcal{EF} - \mathcal{ET} = \int_0^\infty \zeta(r) \cdot (P_f(r) - P_t(r))\, dr > 0$.

The left part of the inequality can be written as $\int_0^\infty \frac{P_f(R_j)\cdot(P_f(r)-P_t(r))}{\alpha\cdot P_f(r)+(1-\alpha)\cdot P_t(r)} \cdot dr$, $1 > \alpha > 0$. Here we assume $P_t(r) \not\equiv P_f(r)$. We first define $S_0 \equiv \int_0^\infty \frac{P_f(R_j)\cdot(P_f(r)-P_t(r))}{1\cdot P_f(r)+0\cdot P_t(r)} \cdot dr = \int_0^\infty (P_f(r) - P_t(r))\, dr = 0$. We now express α as $\alpha = 1-\gamma$, $1 > \gamma > 0$, and rewrite the inequality as $S \equiv \int_0^\infty \frac{P_f(R_j)\cdot(P_f(r)-P_t(r))}{1\cdot P_f(r)+0\cdot P_t(r)+\gamma\cdot(P_t(r)-P_f(r))}$. In S, the integrand is positive when $(P_f(r) > P_t(r))$, which corresponds to $\gamma \cdot (P_t(r) - P_f(r)) \leq 0$. Therefore, on the intervals where the integrand of S is positive, it is greater than or equal to the value of the integrand of S_0 on the corresponding points. Similarly it can be shown that, on the intervals where the integrand of S is negative, its absolute value is smaller than or equal to the value of the integrand of S_0 on the corresponding points. Given that $P_t(r) \not\equiv P_f(r)$, we have that $S > S_0 = 0$ and hence $\mathcal{EF} - \mathcal{ET} > 0$. Therefore $\mathcal{EP}_1 > \mathcal{EP}_2 \iff T_1/N_1 < T_2/N_2$ is satisfied and Proposition 1 holds. □

Appendix B. Lengths of Dummy Links

In this section, we justify the proposed selection of lengths for dummy (padding) links introduced in Section 4.3. Let $\mathcal{S}_i$, $i = 1, 2$, be the set of all

possible trackers that yield output with N_i links and that all have T_i true links in their output. Padding involves adding δN_i links to the outputs of trackers from $\mathcal{S}_i$. Note that $(N_1 + \delta N_1) = (N_2 + \delta N_2)$.

Now let $\mathcal{ER}_i = E\left[MR(\mathbb{L}_i)\right]$ be the expected value for the MR-score, where the expectation is taken over all possible outputs $\mathbb{L}_i$ from trackers in $\mathcal{S}_i$. From the definition of MR-score (equation 1 on padded set), we have that

$$\mathcal{ER}_i = \left(\sum_{t \in \mathcal{T}_i} E[\zeta(R_t)] + \sum_{f \in \mathcal{F}_i} E[\zeta(R_f)] + \sum_{d \in \mathcal{D}_i} E[\zeta(R_d)] \right) / (N_i + \delta N_i). \quad \text{(B.1)}$$

Here $\zeta(R_j) = \frac{P_f(R_j)}{P_{all}(R_j)}$ is a summation term from equation 1; t, f, and d enumerate individual links in $\mathbb{L}_i$, and $\mathcal{T}_i$, $\mathcal{F}_i$, and $\mathcal{D}_i$ are the sets of indices for true, false and dummy links respectively. The expectation is taken over all possible outputs $\mathbb{L}_i$ from trackers in $\mathcal{S}_i$. For example, $E[\zeta(R_1)]$ is the expected value of $\zeta(R_1)$ where R_1 is the random variable denoting the length of the first link in the outputs from the trackers in $\mathcal{S}_i$.

Following the reasoning in Appendix A, we can have $\mathcal{ET} = \mathcal{ET}_1 = E[\zeta(R_k)] \approx \mathcal{ET}_2 = E[\zeta(R_l)]$, $k \in \mathcal{T}_1$,$l \in \mathcal{T}_2$. The same applies to $\mathcal{EF}$, and we simplify

$$\mathcal{ER}_i = \left(T_i \cdot \mathcal{ET} + (N_i - T_i) \cdot \mathcal{EF} + \delta N_i \cdot \mathcal{ED}_i\right) / (N_i + \delta N_i) \quad \text{(B.2)}$$

We want the MR-score to have a negative correlation with the number of true links (denoted as T_i) in the trackers' outputs, thus we want to satisfy $\mathcal{ER}_1 > \mathcal{ER}_2 \iff T_1 < T_2$. We now need to select the lengths of dummy links such that with the corresponding $\mathcal{ED}_i$ in equation B.2, the condition is satisfied.

Consider a strategy when the dummy links are drawn from the same distribution for $i = 1, 2$. In this case, $\mathcal{ED} = \mathcal{ED}_1 = \mathcal{ED}_2$. Let $(N_1 + \delta N_1) \cdot (\mathcal{ER}_1 - \mathcal{ER}_2) = K$. Note that if we set $\mathcal{ED} = \mathcal{EF}$ then we have $K = (T_1 - T_2) \cdot (\mathcal{ET} - \mathcal{EF})$. From Appendix A, we have that $\mathcal{ET} - \mathcal{EF} < 0$, and hence having $\mathcal{ED} = \mathcal{EF}$ satisfies

$\mathcal{ER}_1 > \mathcal{ER}_2 \iff T_1 < T_2$. We therefore suggest to generate the dummy links by drawing their lengths from the PDF of false links P_f, so that the expected value of $\zeta(R)$ for dummy links satisfies $\mathcal{ED} = \mathcal{EF}$.

References

[1] E. Hawkins, J. Markham, L. McGuinness, P. Hodgkin, A single-cell pedigree analysis of alternative stochastic lymphocyte fates, Proc. Natl. Acad. Sci. U.S.A. 106 (32) (2009) 13457–13462.

[2] T. Kanade, Z. Yin, R. Bise, S. Huh, S. E. Eom, M. Sandbothe, M. Chen, Cell image analysis: Algorithms, system and applications, in: IEEE Workshop on Applications of Computer Vision (WACV), 2011, pp. 374–381.

[3] D. Padfield, J. Rittscher, B. Roysam, Coupled minimum-cost flow cell tracking for high-throughput quantitative analysis, Med. Image Anal. 15 (2010) 650–668.

[4] K. Li, E. Miller, M. Chen, T. Kanade, L. Weiss, P. Campbell, Cell population tracking and lineage construction with spatiotemporal context, Med. Image Anal. 12 (5) (2008) 546–566.

[5] J. Degerman, T. Thorlin, J. Faijerson, K. Althoff, P. Eriksson, R. Put, T. Gustavsson, An automatic system for in vitro cell migration studies, J. Microsc. 233 (1) (2009) 178–191.

[6] A. Kan, J. Bailey, C. Leckie, J. Markham, M. R. Dowling, R. Chakravorty, Automated and semi-automated cell tracking: Addressing portability challenges, J. Microsc. 244 (2) (2011) 94–213.

[7] E. Meijering, O. Dzyubachyk, I. Smal, W. van Cappellen, Tracking in cell and developmental biology, Semin. Cell Dev. Biol. 20 (8) (2009) 894–902.

[8] J. Rittscher, Characterization of biological processes through automated image analysis, Annu. Rev. of Biomed. Eng. 12 (2010) 315–344.

[9] O. Al-Kofahi, R. Radke, S. Goderie, Q. Shen, S. Temple, B. Roysam, Automated cell lineage construction: A rapid method to analyze clonal development established with murine neural progenitor cells, Cell Cycle 5 (3) (2006) 327–335.

[10] K. Jaqaman, D. Loerke, M. Mettlen, H. Kuwata, S. Grinstein, S. Schmid, G. Danuser, Robust single particle tracking in live cell time-lapse sequences, Nat. Methods 5 (8) (2008) 695–702.

[11] J. Dobrucki, D. Feret, A. Noatynska, Scattering of exciting light by live cells in fluorescence confocal imaging: Phototoxic effects and relevance for FRAP studies, Biophys. J. 93 (5) (2007) 1778–1786.

[12] T. Bai, Y. Li, Robust visual tracking with structured sparse representation appearance model, Pattern Recognit. 45 (6) (2012) 2390–2404.

[13] M. Lázaro-Gredilla, S. Van Vaerenbergh, N. Lawrence, Overlapping mixtures of gaussian processes for the data association problem, Pattern Recognit. 45 (4) (2011) 1386–1395.

[14] S. Kang, S. Lee, Real-time tracking of multiple objects in space-variant vision based on magnocellular visual pathway, Pattern Recognit. 35 (10) (2002) 2031–2040.

[15] S. Wu, L. Hong, Hand tracking in a natural conversational environment by the interacting multiple model and probabilistic data association (immpda) algorithm, Pattern Recognit. 38 (11) (2005) 2143–2158.

[16] P. Assheton, A. Hunter, A shape-based voting algorithm for pedestrian detection and tracking, Pattern Recognit. 44 (5) (2011) 1106–1120.

[17] J. C. SanMiguel, A. Cavallaro, J. M. Martinez, Evaluation Of On-Line Quality Estimators For Object Tracking, in: Proceedings of 2010 IEEE 17th International Conference on Image Processing, 2010, pp. 825–828.

[18] M. Abdul-Karim, B. Roysam, N. Dowell-Mesfin, A. Jeromin, M. Yuksel, S. Kalyanaraman, Automatic selection of parameters for vessel/neurite segmentation algorithms, IEEE Trans. Image Process. 14 (9) (2005) 1338–1350.

[19] S. K. Warfield, K. H. Zou, W. M. Wells, Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation, IEEE Trans. Med. Imaging 23 (7) (2004) 903–21.

[20] Z. Yin, K. Li, T. Kanade, M. Chen, Understanding the optics to aid microscopy image segmentation, in: 13th MICCAI Conference, 2010, pp. 209–217.

[21] Z. Botev, J. Grotowski, D. Kroese, Kernel density estimation via diffusion, The Ann. of Stat. 38 (5) (2010) 2916–2957.

[22] S. Mori, K. Chang, C. Chong, Performance analysis of optimal data association with applications to multiple target tracking, in: Y. Bar-Shalom (Ed.), Multitarget-Multisensor Tracking: Applications and Advances, Vol. 2, Artech House, Boston, 1992, pp. 183–235.

[23] D. Musicki, R. Evans, Linear joint integrated probabilistic data association-LJIPDA, in: 41st IEEE Conference on Decision and Control, Vol. 3, 2002, pp. 2415–2420.

[24] A. Kan, J. Bailey, C. Leckie, J. Markham, M. R. Dowling, R. Chakravorty, Automated and Semi-automated Cell Tracking : Addressing Portability Challenges, J. Microsc. 244-2 (2011) 194–213.