Pattern Recognition Letters
journal homepage: www.elsevier.com

Ranking Cell Tracking Systems without Manual Validation

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

[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*

ARTICLE INFO

Article history:

Cell tracking
Tracking quality
Optimal assignment
Tracker selection
Bayes theorem

ABSTRACT

Automated cell segmentation and tracking can significantly increase the productivity of research in biology. In order to tune a tracking system for a particular video, researchers usually have to manually annotate a part of the video, and tune the algorithm with respect to this ground truth. However, large variability in cell video characteristics leads to different trackers and parameters being optimal for different videos. Therefore for any new video, manual annotation and tuning has to be performed again. Alternatively, suboptimal parameters have to be used which may result in a significant amount of manual post-correction being required. The challenge that we address in this paper is automated selection and tuning of cell tracking systems without the need for manual annotation. Given an estimate of the cell size only, our method is capable of ranking the trackers according to their performance on the given video *without the need for ground truth*. Our evaluation using real videos and real tracking systems indicates that our method is capable of selecting the best or nearly best tracker and its parameters in practical scenarios.

1. Introduction

Over the past decades single cell tracking results combined with mathematical modeling are having an increasing impact on cell biology (Bakstad et al., 2012; Hawkins et al., 2009). Automated cell segmentation and tracking can significantly increase the productivity of biological research. A major challenge in developing such a system is the large diversity of cell morphology and motility as well as variations in recording conditions (Figure 1). This diversity has resulted in a large number of proposed cell tracking systems, where each system has parameters that need to be specified (Meijering et al., 2012). By a cell tracking system (CTS) we mean a combination of algorithms that is capable of both locating cells in video frames (cell segmentation) and maintaining cell identities throughout the video (data association). These two tasks can be approached separately (Padfield et al., 2011) or within a single algorithm (Delgado-Gonzalo et al., 2011). Furthermore, by a CTS we mean a combination of such algorithms with their parameters fixed to specific values. For example, we treat the same software with different parameters as two different CTSs. Occasionally by "CTS for the given video" we also mean "results of the CTS given the video as input".

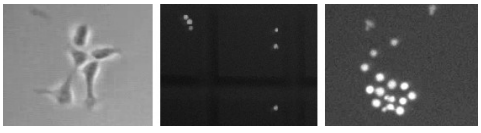

Fig. 1. Examples of a diversity in cell morphology and recording conditions. The images show neural progenitor cells (left, with permission from Al-Kofahi et al., 2006) and B-lymphocytes in micro-grids (middle, right).

Due to the variability in experimental conditions, optimal combinations of algorithms and parameters can vary for different videos, even for the same cell type (Kan et al., 2013). In order to find the best CTS for a given video, a practical solu-

**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), jmarkham@wehi.edu.au (J. Markham), Rajib.Chakravorty@nicta.com.au (R. Chakravorty), baileyj@unimelb.edu.au (J. Bailey), caleckie@csse.unimelb.edu.au (C. Leckie).

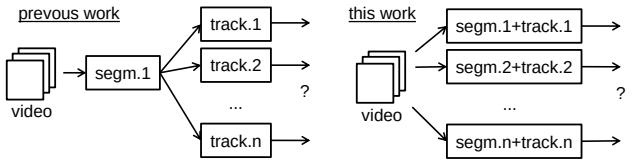

Fig. 2. Where segmentation and tracking steps are separated, the previous method is capable of choosing the best tracker, given a fixed segmentation (Kan et al., 2013). In this work, we address a more general problem of choosing among a number of CTS, where a CTS can have a combined segmentation and tracking steps. Both methods operate in the absence of the ground truth.

tion is to annotate a part of a video and use the resulting ground truth to evaluate the performance of different CTSs. However, if a video from another experiment needs to be processed, the previously best found CTS may not be the best anymore, and ideally, a part of the new video needs to be manually annotated. Even within a single long video, a CTS tuned on one part of it may not be the best for another part. Manual annotation for each new video or different parts of the same video can severely compromise the effectiveness of an automated CTS.

Consider the following real example from an Australian medical research institution. A lab recorded a set of novel cell videos. The analysis of results required cell tracking, and multiple software packages appeared suitable for this task (e.g., Chakravorty et al., 2014; de Chaumont et al., 2012). Initial ground truth could have been created totally manually, but it was found beneficial to use a CTS with imperfectly guessed parameter values and then correct the automated results. It then took a researcher-biologist a few hours to choose and guess parameters and a few more hours to manually correct results in order to produce the ground truth (cell outlines and identities over 200 frames) for only 3 cells. This is a large amount of manual time, given that more cells are required for a representative ground truth and that the lab usually produces a few novel videos each year.

Here we address this challenge with a system for ranking CTSs *without the need for ground truth*. To the best of our knowledge, this is the only system of its kind in cell tracking literature. Given a video (or a fragment) and a range of candidate CTSs, the user is only asked to provide an estimate of the cell size. Our system then ranks the CTSs according to their performance on the given video.

The problem of automated performance estimation has been previously addressed for medical image segmentation and tracking for surveillance cameras (SanMiguel et al., 2010; Warfield et al., 2004). The previous methods rely on knowledge specific to their respective domains and are not directly applicable to CTSs. In the context of cell tracking, given a video and a fixed segmentation step, a previous method (Kan et al., 2013) is capable of comparing data association algorithms (Figure 2). A more general problem of choosing among a number of CTS, where both segmentation and tracking steps can differ, remains an open research challenge that we address in this work.

Here we propose a novel method for ranking CTSs without the need for ground truth and using minimum user input. We

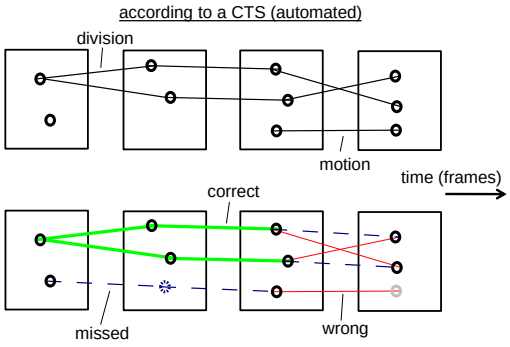

Fig. 3. A CTS produces a detection (set of circles) and a set of links, where each link can denote a tentative cell division or motion. After a manual validation, each link can be classified into correct (bold green, connects two correct measurements), wrong (thin red, wrong association or incorrect measurements), and missed (dashed blue, absent in the output of the CTS).

design a scheme of pairwise comparisons (Section 2.2), and employ a special case of an optimal assignment problem to match detections (Section 2.3). Finally, we develop a general face validity test for a CTS (Section 2.4). We find that together these components based on minimum prior information can be effective in practical scenarios (Section 3).

2. Methods

2.1. Cell Tracking Preliminaries

A *CTS* is a function that takes a video as input and produces a detection and a set of links as output. A *detection* is a set of measurements, and a *measurement* is a vector that comprises a numerical description of each located object (tentative cell). A CTS can produce different numbers of measurements for different frames. Each measurement can contain information such as the cell centroid location, mean brightness, size, etc. While different CTSs may differ in the type of information they produce, we assume that cell centroid location is always present, and in this paper, we usually treat measurements as cell centroid locations. Finally, a *link* is a pair $\{\vec{m}_{i,j}, \vec{m}_{i+1,k}\}$ of measurements from consecutive frames i and $i+1$ (j,k index measurements). Links represent tentative moves or division events (Figure 3).

Performance of a CTS is a measure of accuracy of detections and links. Given a video, a resulting detection and a set of links, performance is measured with respect to some manual annotation (*ground truth*). Based on the manual annotation, each measurement from the detection can be classified as either correct or spurious (Figure 3), where the *spurious measurement* is a measurement that does not originate from any cell. Additionally there can be *missed* measurements when a cell is not represented with any matching measurement. Furthermore, each link can be classified as either correct, wrong, or missing (Figure 3). The CTS performance is then computed as $F = N_{corr}/(N_{corr} + N_{wrong} + N_{miss})$. This equation defines an F-score, and it has been shown that such a performance measure adequately represents tracking accuracy (Kan et al., 2013). Importantly, the above performance measure is defined with respect to a fixed input video. For example, one CTS can be better for one video, and another for a different video.

Algorithm 1 A method for ranking CTSs

Input: $\mathbb{S} = \{CTS_1, \ldots, CTS_K\}$; par ={3 parameters, see text};
Output: $\mathbb{S}$, ordered according to the performance from high to low;

1: **for each** $CTS_i \in \mathbb{S}$
2: **if not** $FaceValidity(CTS_i, par)$
3: $\mathbb{S} = \mathbb{S} \backslash CTS_i$;
4: **end**
5: **end**
6: $\{\#wins_i\} = PairwiseComparisons(\mathbb{S}, par)$;
7: $\mathbb{S}$ =order according to decreasing $\{\#wins_i\}$;
8: **return** $\mathbb{S}$

Finally, CTSs can be implemented as two sequential steps: cell segmentation that produces a detection, and data association that creates links over the detection (Kan et al., 2011; Kanade et al., 2011). Here, a cell segmentation algorithm can produce a detection first, and then different tracking algorithms can produce different versions of the links for the same detection. Given a fixed detection, each version of the links can be characterized by a quantity called the *ED-score* (see the Appendix). It has been shown that the ED-score correlates with the F-score in practical situations (Kan et al., 2013). Note that ED-score does not require ground truth to be computed. In contrast, the computation of the F-score is based on manual annotation. Our CTS ranking system employs the ED-score in pairwise comparisons. Among other additions, we supplement the ED-score with a new method of detection matching presented below.

2.2. Method Overview

The overall aim of this study is a system capable of comparing relative the performance of CTSs while requiring minimal input from the user. While employing a previously proposed ED-score for this task, the three major challenges are (1) matching detections from different CTSs; (2) using minimum prior knowledge; and (3) turn pairwise comparisons into a ranking. We address the first challenge in Section 2.3. Furthermore, where possible we use rather general assumptions. Finally, the ranking challenge is addressed using cross-comparison based rankings with the exclusion of infeasible solutions.

A high level overview of our solution is presented in Algorithm 1. Given a set of results from candidate CTSs for the given video sequence, we first eliminate infeasible CTSs using the *FaceValidity* test (Section 2.4). We then perform pairwise comparisons of the remaining CTSs, and for each CTS compute the total number of wins across all comparisons. At the core of our method is a single comparison of two CTSs, in which one of the the CTS results is estimated to be better than another ("win"). We start a detailed presentation of our method with a description of a single pairwise comparison in the next section.

Given a video sequence, and the results of two CTSs for this sequence, the aim of the comparison is to estimate which set of results is better without the need for ground truth. Recall that the results of a CTS can be represented as a detection and a set of links constructed over this detection. In practice, different CTSs tend to produce different detections for the same video. For example, by making different mistakes with spurious objects. Therefore, the main idea behind our comparison is to align the two detections, and produce a synthetic detection as a new basis for the two sets of links (Figure 4). We then use the synthetic detection and the links for computing ED-scores, and choose the CTS with the lowest ED-score as the winner. It has been shown that in this setup a lower ED-score indicates more accurate tracking (Kan et al., 2013).

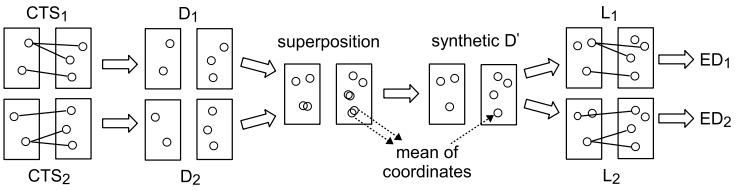

Fig. 4. An outline of the pairwise comparison procedure. Cells are represented with circles for convenience only. Cells need not to be round in videos.

Note that in any given detection, some measurements correspond to real cells and some measurements are spurious. In two different detections D_1 and D_2, some measurements may correspond to the same cell, despite having different values. We therefore start with an assignment procedure that determines which of the measurements from D_1 and D_2 are likely to originate from same cells. Such measurements are then merged (averaged) in the synthetic detection. The ED scores and the decision are ultimately based on a synthetic detection and links that are adjusted accordingly. However, as we show below, the synthetic detection is constructed in a way that does not severely distort original detections. Therefore, the decision based on the ED scores reflects the original results from CTSs (Section 3).

2.3. Detection Matching Scheme

The synthetic detection is generated via matching detections from two CTSs. Our matching is based on a linear assignment formulation. Linear assignment has been previously explored in the context of cell tracking (Jaqaman et al., 2008; Kirubarajan et al., 2001). However, one important difference is that previous methods assign measurements across frames, where the measurements are produced by a single CTS, whereas we assign measurements from two different CTSs produced for the same frame. This means that in our case, a measurement from one frame cannot be assigned to a measurement from another frame. In general, the probability of assignment within one frame depends on the assignment in the previous frame via tracking links. However, a priori we do not know how good the tracking is (which motivates the present work), and therefore we perform the assignment independently for different frames.

For a given frame, let $U = \{\vec{u}_i, i = 1, \ldots, n\}$ and $V = \{\vec{v}_j, j = 1, \ldots, m\}$ be two sets of measurements produced by the two CTSs ($n = |U|$, $m = |V|$ and in general $n \neq m$). A measurement from U can have a corresponding measurement from V which would mean that the two measurements correspond to

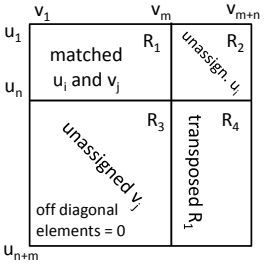

Fig. 5. Assignment and scoring matrices are divided into 4 regions, where R_1 corresponds to matched measurements, R_2 and R_3 correspond to unassigned measurements, and R_4 is a technical region. Our convention is that in an assignment, off diagonal elements in R_2 and R_3 are zero and R_4 is a transpose of R_1.

the same cell. Alternatively, any measurement from U or V can remain unassigned. The latter implies that the measurement is either spurious or valid, but with a lost (not detected) matching measurement from the other CTS.

An *assignment* is a square binary matrix where each row and each column has exactly one non-zero element. In application to our case an assignment is a matrix A with $n + m$ rows and columns. The first n rows (respectively, m columns) correspond to measurements from U (respectively, V). The last m rows and n columns artificially added to allow unassigned measurements. Matrix A can be divided into four rectangles R_1 to R_4 (Figure 5). Here $A(i, j) = 1$ in R_1 denotes a matching between $\vec{u}_i$ and $\vec{v}_j$: these measurements are assumed to originate from the same cell. Furthermore, $A(i, j) = 1$ in R_2 or R_3 denotes a non-matched $\vec{u}_i$ or $\vec{v}_j$. Finally, R_4 does not have a problem-specific interpretation and is only maintained to ensure that each row and column in the assignment has exactly one non-zero element.

We now define a *scoring matrix* W of the same size as an assignment matrix. For a given assignment A, elements of W reflect the probability of the assignment. The *total score* is then defined as $P_{tot}(A) = \prod_{A(i,j)=1} W(i, j)$, and the *optimal assignment* is the one that maximizes the total score $A^* = \arg\max_A P_{tot}(A)$. Note that our definition of the assignment is compatible with the general combinatorial optimization problem, and the optimal assignment can be found using the Hungarian method (Munkres, 1957). The problem that we address next is how to define W.

2.3.1. Scoring Matrix

Recall that we have introduced dummy rows and columns in order to allow disconnected $\vec{u}_i$ and $\vec{v}_j$, while maintaining A as an assignment. By our convention, a non-zero diagonal element in R_2 ($A(i, j) = 1$, $i \le n$, $j = m + i$) is a unique way to denote disconnected $\vec{u}_i$, which implies a restriction that $A(i, j) = 0$ for $i \le n$, $j \ne m + i$ (similarly for R_3). Moreover, having all $\vec{u}_i$ connected implies all zeroes inside R_2 while an assignment requires having non-zero elements in leftmost columns. Therefore we use compensating region R_4: missing non-zero elements can be added here. In order to have a unique way of compensation, we restrict R_4 to be a transposed image of R_1: $A(n+j, m+i) = A(i, j)$ for $(i, j) \in R_1$. Importantly, we introduce the above restrictions by means of scoring. This way, we can use a general purpose optimal assignment method without modifications.

Lemma: Let scores $W(i, j) > 0$ and $W(n + j, m + i) = W(i, j)$ for $(i, j) \in R_1$ and let off diagonal scores in R_2 and R_3 be zero and diagonal scores in these regions be positive. There exists an optimal assignment that satisfies the restrictions from the above paragraph.

Proof: There exists a trivial assignment with all diagonal elements of R_2 and R_3 set to ones and zeros elsewhere, and a positive total score. On the other hand, any assignment that involves non-zero off diagonal elements in R_2 and R_3 has the total score of 0. Therefore, any optimal assignment does not have non-zero off diagonal elements in R_2 and R_3. Furthermore, consider an assignment with 0 off-diagonal elements and some non-zero diagonal elements in R_2 and R_3. These non-zero elements restrict the use of certain rows and columns in R_1 and corresponding columns and rows in R_4, which means that it is possible to have R_4 as a transposition of R_1. Finally, consider an assignment where R_1 is not a transposition of R_4. If products of scores from R_1 and scores from R_4 differ, we can choose the region with the highest score and make the other region as a transposition. This will produce a valid assignment with a higher total score. □

After adopting the scoring scheme described in the lemma, we still need to choose weights for elements in R_1 and diagonal elements in R_2 and R_3. The problem with choosing such scores is the lack of detailed information about CTSs and the video (e.g., rate of spurious measurements). Therefore, we observe that $A(i, j) = 1$, $(i, j) \in R_1$ implies $A(i, m+i) = 0$, $A(n+j, j) = 0$ and vice versa. Here $A(i, j) = 1$ is the hypothesis $H_1 = \{\vec{u}_i$ and $\vec{v}_j$ originate from the same cell$\}$, and $A(i, j) = 0$ is the hypothesis $H_0 = \{\vec{u}_i$ and $\vec{v}_j$ originate from different cells OR both are spurious measurements OR one originates from a cell and another one is spurious$\}$. Therefore, finding an optimal assignment is related to a binary classification problem: given $\vec{u}_i$ and $\vec{v}_j$ we need to decide which hypothesis is more plausible. We can use the Euclidean distance $d_{ij} = \|\vec{u}_i - \vec{v}_j\|$ as a classification feature, and then we have $P(H_s|d_{ij}) = P(H_s) \cdot P(d_{ij}|H_s)/P(d_{ij})$, where $s = 0, 1$. In order to choose the hypothesis we only need to compare numerators.

Let d_{99} be the 99^{th} percentile of the cell diameter (provided by a user). We have that if $d_{ij} > d_{99}$, then $P(d_{ij}|H_1)$ is small. In this case, it is relatively safe to choose H_0, since H_1 is highly unlikely regardless of the prior $P(H_1)$. Now consider $P(d_{ij}|H_0)$. We assume that the locations of both real cells and spurious measurements follow a uniform distribution within a circle with diameter L (a dimension of a video frame). The cumulative distribution of distances between two random points within the circle (Hammersley, 1950) is then

$$F_2(x) = (2x/L)^2 \, I_{1-\left(\frac{x}{L}\right)^2}(1.5, 0.5) + I_{\left(\frac{x}{L}\right)^2}(1.5, 1.5), \qquad (1)$$

where $I_a(b, c)$ is the incomplete beta function. Note that L characterizes dimensions of a video frame and it is reasonable to assume $d_{99} \ll L$. For example, if $d_{ij} \le d_{99}$ and $d_{99} \le 0.1 \cdot L$, then $F_2(d_{ij}) \le 0.04$. We conclude that if $d_{ij} \le d_{99}$ then $P(d_{ij}|H_0)$ becomes unlikely, and as an approximation in this case we can choose H_1.

Returning to the assignment scoring problem, we define an *allowed set* $R_1^* \subseteq R_1$ as a set of index positions $(i, j) \in R_1$, where $d_{ij} \leq d_{99}$. For elements from $R_1 \backslash R_1^*$ (not allowed), we set $W(i, j) = 0$, which essentially prohibits matching if the assignment hypothesis H_1 is unlikely. On the other hand, we set diagonal elements in R_2 and R_3 to some ε such that $\min_{(i,j)\in R_1^*} W(i, j) > \varepsilon > 0$. This penalizes leaving the measurements unassigned when hypothesis H_0 is unlikely. We do not completely prohibit unassigned measurements where $d_{ij} > d_{99}$, because if $n \neq m$ then it is impossible to assign all measurements to each other.

Furthermore, scores $W(i, j)$, $(i, j) \in R_1^*$ reflect conditional probabilities $P(\vec{u}_i, \vec{v}_j | \vec{u}_i, \vec{v}_j$ originated from $\vec{x})$, where $\vec{x}$ is an unknown true measurement. Here we make another assumption, that given cell $\vec{x}$, then $\vec{u}_i$ and $\vec{v}_j$ are independent and identically distributed according to a multivariate normal distribution with mean $\vec{x}$ and some unknown positive definite covariance matrix S. Without the loss of generality, components of $\vec{u}_i$ can be rescaled (change of units), and it is feasible to assume independence of variations, so that S is diagonal and $S(i, i) = \sigma$ (consider for example, a distribution of measured 2-dimensional cell centroid locations: it is reasonable to expect a symmetric distribution). A maximum likelihood estimate of $\vec{x}$ is then $(\vec{u}_i + \vec{v}_j)/2$, and we have

$$P(\vec{u}_i | \vec{u}_i, \vec{v}_j \text{ from } \vec{x}) = (2\pi)^{-\frac{k}{2}} \cdot |S|^{-\frac{k}{2}} \times \\ \times \exp\left\{ -\left(\vec{u}_i - \vec{v}_j/2\right)^T S^{-1} \left(\vec{u}_i - \vec{v}_j/2\right)/4 \right\} \cdot \quad (2)$$

Note that with our scoring matrix the set R_1^* is fixed (from the data), and the choice of optimal assignment depends only on the relative scores within R_1^*. We therefore can ignore terms related to S, and note that $\log P(\vec{u}_i, \vec{v}_j | \vec{u}_i, \vec{v}_j \text{ from } \vec{x}) \propto -\|\vec{u}_i - \vec{v}_j\|^2$.

Finally, we apply a log transformation to the total score (equation 2.3), negate, and define the *total cost function* that we seek to minimize: $C = \sum_{i,j \leq n+m} A(i, j) \cdot W(i, j)$. We use a special value MAX_1 for cases that involve $\log 0$, and then the scores for allowed assignments $(i, j) \in R_1^*$ are $W(i, j) = \|\vec{u}_i - \vec{v}_j\|^2$; for prohibited assignments $(i, j) \in R_1 \backslash R_1^*$ are $W(i, j) = MAX_1$; for region R_4 (transposed scores from R_1) are $W(n+j, m+i) = W(i, j)$, $(i, j) \in R$; the scores for off-diagonal elements in R_2 and R_3 are set to MAX_1; and the scores for diagonal elements in R_2 and R_3 are set to MAX_2. Here $MAX_1 > MAX_2 > \max_{(i,j)\in R_1^*} W(i, j)$.

Using the above scoring scheme, we prohibit making assignments for measurements that are too far apart ($d_{ij} > d_{99}$). The remaining measurements are assigned in a way that minimizes the sum of Euclidean distances between the assigned measurements. After the optimal assignment is found, matched measurements are replaced with their means in the synthetic detection D'. Unmatched measurements from either U or V are copied to synthetic detection without modifications. As a result, we have two sets of links $\mathbb{L}_1$ and $\mathbb{L}_2$ that are defined on the common detection D' (Figure 4). We can now calculate the ED-score for each set, and the CTS that results in a smaller ED-score wins.

2.4. Face Validity Test

The previous section focused on matching detections from different CTSs. In this final part of our methodology we focus on results from a single CTS. The aim here is to perform an initial screening and make a decision whether a particular CTS appears to be feasible ($FaceValidity(CTS_i)$ in algorithm 1). Recall, that for a given detection a link is a pair of measurements. Let *a length of the link* be $r = \|m_{i,j} - m_{k,l}\|$, where i, k are frame numbers and j, l index measurements within a frame. (Note that the two measurements are now taken from the same detection, not from different detections as in the previous section.) We can then define three probability density functions (PDFs). First, P_{all} is the distribution of lengths of all possible links in a detection. For example, if frame 1 has two measurements and frame 2 has three, then there are 6 possible links. Second, P_{wrong} is the PDF of lengths of wrong links as determined by the ground truth. Third, P_{within} is the PDF of links that connect all possible measurement pairs within frames. For example, if frame i has three measurements, there are three possible within-links. Note that P_{all}, P_{wrong} and P_{within} are defined over the entire input video sequence.

Importantly, P_{all} and P_{within} can be computed without the ground truth. Furthermore, it was shown theoretically and experimentally that in videos that can be tracked it is safe to assume $P_{within} \approx P_{wrong}$ (Kan et al., 2013). Therefore, we compare P_{within} and P_{all} using the Kolmogorov-Smirnov test (KST). Failing to reject the null hypothesis that $P_{within} \approx P_{wrong} = P_{all}$ at significance level α, indicates that there are no correct links that can be made in D_i (note that P_{all} is a convex combination of the PDFs of correct and wrong links). In this case, CTS_i fails our face validity check.

Moreover, we note that the results of certain CTSs are poor due to a large number of spurious locations. In this case, the density of such locations can be abnormally high. Therefore, in each frame, for each location we find its nearest neighbor and compare the distance to the nearest neighbor d_{NN} with our estimate of the cell size d_{99}. Relation ($d_{NN} < d_{99}$) indicates that the CTS reports two distinct cell locations that are suspiciously close to each other. We then look at the proportion of such suspicious locations among all locations in D_i. If the proportion β is high, then CTS_i fails our face validity check.

In total, our method has 3 parameters (d_{99}, α, β). Here, d_{99} is an estimate of the largest cell size in the video, α is the significance level for the KST, and β is the proportion of the suspicious locations that we can tolerate. Note that α and β are independent of the video. We set $\alpha = 0.1$ and $\beta = 0.5$. In our evaluation, we find that our results are not very sensitive to variations of α and β.

3. Evaluation Results

We validate the utility of our method by considering the following practical situation. A number of experiments resulted in 5 cell videos. One of the videos (*ak*) shows the devel-

Table 1. Eleven CTSs used in our evaluation. The CTSs are constructed from previously reported algorithms: Sobel edge detection, Hough transform, Otsu thresholding, Gaussian filtering, cell tracker nenia (Kan et al., 2011), particle tracker u-track (Jaqaman et al., 2008).

CTS name	Description
hough − 15 − *nenia* − 20, *hough* − 20 − *nenia* − 50	Sobel edge detection; Hough transform with the accumulator array threshold set to 15 and 20; nenia tracker with the gating distance set to 20 and 50
otsu − *nenia* − 20, *otsu* − *nenia* − 50	Otsu thresholding; nenia tracker with the gating distance set to 20 and 50
gauss − 0.22 − *nenia* − 20, *gauss* − 0.22 − *nenia* − 50, *gauss* − 0.25 − *nenia* − 20, *gauss* − 0.25 − *nenia* − 50	Gaussian filtering; locating local maxima with the fluorescence threshold set to 0.22 and 0.25; nenia tracker with the gating distance set to 20 and 50
gauss − 0.5 − *utrack* − 2, *gauss* − 0.5 − *utrack* − 10	Gaussian filtering; locating local maxima with the fluorescence threshold set to 0.5; u-track tracker with default parameters, except for the standard deviation multiplication factor which is set to 2 and 10
rand − 30 − *nenia* − 200	30 random uniformly distributed locations; nenia with the gating distance of 200

opment of neural progenitor cells[1] plated into poly-L-lysine coated Terasaki plate micro-wells and imaged using an inverted Olympus microscope (Al-Kofahi et al., 2006). The other videos show proliferation of naive B cells stimulated with CpG DNA through Toll-Like Receptor 9 as described previously (Hawkins et al., 2009). The stimulated cells were placed in either $250\,\mu m$ hexagonal wells (*hex.*$*$) or $125\,\mu m$ micro-grids (*square*) and imaged every 2 minutes using an Axiovert 200*m* microscope (Figure 1). On the other hand, we have a number of CTSs constructed from previously reported algorithms (Table 1). We set most of the algorithm parameters to their default values, set some parameters arbitrarily, and some parameters based on our knowledge of video frame sizes. We construct 10 fully parametrized CTSs and in addition we use a random CTS that produces results irrelevant to the input video (Table 1).

We note that different CTSs can have different performance on different videos (Kan et al., 2013). Therefore, the task is to find the best CTS for each of the given videos. The task can be approached by manually annotating a part from each video and maximize F-scores with respect to these annotations (Section 2.1). However, such an approach tends to be tedious and subjective, and instead we ask whether it is possible to achieve a similar result but without the manual annotation. Furthermore, different parts of the same video can present different cell dynamics (e.g., changing from a low to a high cell density). Therefore, we elaborate our question: whether it is possible to find the best CTS for different short fragments of each video without manual annotation. To this end, we define several video sequences (Table 2) such that the sequences include a variety of conditions (e.g., cell divisions, leaving the field of view, high cell density, and abrupt movements). Cell sizes (parameter d_{99}) are estimated manually by considering about ten cells in each sequence. We then use our system (algorithm 1) to rank CTSs for each sequence, and verify the correlation between our results to the results obtained using manual annotation. See Table 2 for results, and note that this table has been produced using 11 fully parameterized CTSs mentioned above and summarized in Table 1.

With each video sequence as input, we run our 11 CTSs, and rank the CTSs using our method. For the purposes of evaluation, we also record the F-scores for each CTS using a ground truth. It is perhaps not surprising that for every input there is a large variation in the F-score, which means that choosing a good CTS (e.g., proper parameter settings) is necessary for effective tracking. Importantly, in every case one of our top 3 results achieves the maximum or near maximum F-score, which indicates that our method can eliminate the need for ground truth collection in practical cases.

Furthermore, we used two alternative methods for selecting the best CTS without manual annotation. The first is the random choice, and the second is based on the variance in the number of links. The intuition here is that a good CTS is expected to exhibit consistency in the linking, and one can select the CTS with the minimum variance in the number of links (which can be computed without manual annotation). Our method clearly outperforms both baselines (Table 3). Moreover, we note that the poor performance of the variance based method is often due to the inability of this approach to identify the random CTS. We therefore tested a hybrid method that first uses our face validity test to exclude CTSs that appear to be inappropriate, and then selected the CTS with the minimum variance. Our selection method outperforms this hybrid method as well (Table 3). We conclude that our results validate the utility of our method.

4. Discussion and Conclusion

In this work, we address the problem of ranking CTSs for a given video. It is assumed that we are given a set of pre-selected CTSs, where each CTS includes some fixed parameter setting. The pre-selection itself is outside of the scope of this study. Importantly, this is not a disadvantage specific to our method, because even with manual annotation, a researcher still needs to identify a set of CTSs to try. A practical approach can be to choose a small set of well-established methods for which a reliable implementation is available, set parameters to recommended default values where possible, use common sense for other parameters, and, finally, for the remaining parameters try

[1] Available online from the Cell Cycle journal website at http://www.landesbioscience.com/journals/cc/supplement/alkofahi.zip

Table 2. We ran each of the 11 CTSs on each of the 10 sequences (extracted from 5 real cell videos). For each sequence, we then ranked the 11 CTSs according to the tracking quality as perceived by our method (but without ground truth). Subscript of the sequence indicates the number of frames. "Valid" is the number of CTSs that pass our validity test (out of 11). "Mean±SD" and "Max" show the mean, standard deviation, and maximum of the F-scores among 11 CTSs for each sequence. The columns on the right show the F-scores of the top 3 CTSs selected using our method. The maximum score is underlined (if achieved).

Sequence	Cells/Frame	Valid	Mean±SD	Max	1st	2nd	3rd
ak_{10}	7 ∼ 9	3	0.16 ± 0.20	0.76	0.76	0.04	0.23
ak_{15}	5 ∼ 7	2	0.18 ± 0.24	0.83	0.83	0.40	0.06
$hex.6_{10}$	2 ∼ 3	10	0.80 ± 0.32	1.00	0.89	1.00	1.00
$hex.6_{15}$	3 ∼ 4	9	0.69 ± 0.35	0.96	0.96	0.95	0.95
$hex.16_{10}$	16	9	0.68 ± 0.33	0.94	0.94	0.94	0.94
$hex.16_{15}$	16	9	0.69 ± 0.34	0.98	0.97	0.98	0.96
$hex.22_{10}$	18 ∼ 20	10	0.62 ± 0.33	0.94	0.92	0.92	0.94
$hex.22_{15}$	20 ∼ 21	9	0.66 ± 0.34	0.98	0.97	0.97	0.98
$square_{10}$	32 ∼ 33	8	0.51 ± 0.45	0.97	0.97	0.97	0.97
$square_{15}$	32 ∼ 33	9	0.53 ± 0.46	0.98	0.98	0.98	0.98

Table 3. Performance comparison between our method, random selection, variance based selection (VB), and the hybrid method. For each sequence, the best CTS is selected without ground truth, and the resulting F-score is noted. In addition, we report the mean of each column.

Sequence	Our method	Rand.	VB	Hybrid
ak_{10}	0.76	0.06	0.02	0.23
ak_{15}	0.83	0.08	0.02	0.40
$hex.6_{10}$	0.89	0.89	0	0.89
$hex.6_{15}$	0.96	0.95	0	0.95
$hex.16_{10}$	0.94	0.31	0	0.94
$hex.16_{15}$	0.98	0.98	0.01	0.98
$hex.22_{10}$	0.92	0.92	0.01	0.59
$hex.22_{15}$	0.97	0.71	0.01	0.98
$square_{10}$	0.97	0.97	0.01	0.01
$square_{15}$	0.98	0.79	0.98	0.98
mean	0.92	0.67	0.11	0.69

different random values within reasonable bounds. Of note, our method can be trivially parallelized, e.g., by running different comparisons on different cores.

Recall that we focus on the F-score (Section 2.1) as a measure of tracking performance. In fact, there are three aspects of the CTS performance: accuracy of outlining cells in frames, detection errors (missed cells, spurious detections), and tracking errors (swapping tracks, losing tracks). As was explained earlier (Kan et al., 2013) the F-score definition of performance, while targeting the links, indirectly addresses the accuracy of outlines and detections. Indeed, a missing detection implies a missing link accounted for in the F-score. Moreover, the accuracy of cell boundaries is categorized by a user into "accepted" leading to a correct detection and "not accepted" leading to a missing detection. Finally, the F-score correlates with some other common performance measures (Kan et al., 2013).

In summary, we propose a novel method for ranking cell tracking systems with minimum input (estimate of the cell size) required for each new video. Our method identifies CTSs that are likely to perform well, and runs pairwise comparisons of such CTSs. A pairwise comparison is implemented using an optimal assignment-based augmenting and a previous approach for ranking data association algorithms. Our results indicate that the new method can effectively assist in CTS selection and tuning in practical scenarios. Furthermore, our method is sufficiently general for potential application to tracking systems, e.g., for multiple particle tracking.

Acknowledgments

This work is partially supported by National ICT Australia. National ICT Australia is funded by the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.

Appendix

Full details of ED-score computation can be found in (Kan et al., 2013), and here we only briefly outline the computation steps. Consider a tracker (data association algorithm) that takes a detection as input and produces a set of links as output. Within the detection, let S_{all} be the set of all possible inter-frame links (pairwise connections across subsequent frames), and S_{wn} be the set of all within-frame links (pairwise connections within each frame). Then PDFs of link lengths in these sets can be empirically estimated and denoted accordingly P_{all} and P_{wn}. Let the output of the tracker comprise N links with lengths R_i. Now the *mirrored precision* is defined as $MP = (1/N)\sum_{i=1}^{N} P_{wn}(R_i)/P_{all}(R_i)$.

Furthermore, recall that ED-score is used to estimate relative performance within a group of trackers, where each tracker produces a set of links. Let N^* be the size of the largest set, whereas the tracker for which ED-score is currently computed produces N links. The *mirrored recall* is then defined as

$$MP = \frac{1}{N^*}\left[\sum_{i=1}^{N}\frac{P_{wn}(R_i)}{P_{all}(R_i)} + \sum_{j=1}^{\delta}\frac{P_{wn}(R_j^*)}{P_{all}(R_j^*)}\right], \qquad (.1)$$

where R_j^* is $\delta = (N^* - N)$ random dummy lengths sampled from distribution P_w. Finally, we define $ED = \sqrt{MP^2 + MR^2}$.

References

Al-Kofahi, O., Radke, R.J., Goderie, S.K., Shen, Q., Temple, S., Roysam, B., 2006. Automated Cell Lineage Construction : A Rapid Method to Analyze Clonal Development Established with Murine Neural Progenitor Cells. Cell Cycle 5, 327–335.

Bakstad, D., Adamson, A., Spiller, D.G., White, M.R.H., 2012. Quantitative measurement of single cell dynamics. Current opinion in biotechnology 23, 103–9.

Chakravorty, R., Rawlinson, D., Zhang, A., Markham, J., Dowling, M.R., Wellard, C., Zhou, J.H.S., Hodgkin, P.D., 2014. Labour-Efficient In Vitro Lymphocyte Population Tracking and Fate Prediction Using Automation and Manual Review. PLoS ONE 9, e83251.

de Chaumont, F., Dallongeville, S., Chenouard, N., Hervé, N., Pop, S., Provoost, T., Meas-Yedid, V., Pankajakshan, P., Lecomte, T., Le Montagner, Y., et al., 2012. Icy: an open bioimage informatics platform for extended reproducible research. Nature methods 9, 690–696.

Delgado-Gonzalo, R., Chenouard, N., Unser, M., 2011. A New Hybrid Bayesian-Variational Particle Filter With Application To Mitotic Cell Tracking, in: 8th IEEE International Symposium on Biomedical Imaging, Chicago, Illinois, USA. pp. 1917–1920.

Hammersley, J., 1950. The distribution of distance in a hypersphere. The Annals of Mathematical Statistics 21, 447–452.

Hawkins, E.D., Markham, J.F., McGuinness, L.P., Hodgkin, P.D., 2009. A single-cell pedigree analysis of alternative stochastic lymphocyte fates. Proceedings of the National Academy of Sciences of the United States of America 106, 13457–13462.

Jaqaman, K., Loerke, D., Mettlen, M., Kuwata, H., Grinstein, S., Schmid, S.L., Danuser, G., 2008. Robust single-particle tracking in live-cell time-lapse sequences. Nature Methods 5, 695–702.

Kan, A., Bailey, J., Leckie, C., Markham, J., Dowling, M.R., Chakravorty, R., 2011. Automated and semi-automated cell tracking: Addressing portability challenges. Journal of Microscopy 244-2, 194–213.

Kan, A., Leckie, C., Bailey, J., Markham, J., Chakravorty, R., 2013. Measures for Ranking Cell Trackers without Manual Validation. Pattern Recognition 46, 2849–2859.

Kanade, T., Yin, Z., Bise, R., Huh, S., Eom, S., Sandbothe, M., Chen, M., 2011. Cell image analysis: Algorithms, system and applications, in: IEEE Workshop on Applications of Computer Vision, Kona, Hawaii, USA. pp. 374–381.

Kirubarajan, T., Bar-Shalom, Y., Pattipati, K., 2001. Multiassignment for tracking a large number of overlapping objects. IEEE Transactions on Aerospace and Electronic Systems 37, 2–21.

Meijering, E., Dzyubachyk, O., Smal, I., 2012. Methods for cell and particle tracking. Methods in enzymology 504, 183–200.

Munkres, J., 1957. Algorithms for the assignment and transportation problems. Journal of the Society for Industrial and Applied Mathematics 5, 32–38.

Padfield, D., Rittscher, J., Roysam, B., 2011. Coupled minimum-cost flow cell tracking for high-throughput quantitative analysis. Medical Image Analysis 15, 650–668.

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

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging 23, 903–921.