Description of a Tracking Metric Inspired by KL-divergence

T. M. Adams [*]

April 20, 2018

Abstract

A unified metric is given for the evaluation of tracking systems. The metric is inspired by KL-divergence or relative entropy, which is commonly used to evaluate clustering techniques. Since tracking problems are fundamentally different from clustering, the components of KL-divergence are recast to handle various types of tracking errors (i.e., false alarms, missed detections, merges, splits). Preliminary scoring results are given on a standard tracking dataset (Oxford Town Centre Dataset). In the final section, prospective advantages of the metric are listed, along with ideas for improving the metric. We end with a couple of open questions concerning tracking metrics.

1 Adapting KL-divergence as a Performance Metric

In 1951, Simon Kullback and Richard Liebler introduced a measure of divergence or discrimination between two distributions. It has since been referred to as Kullback-Liebler (KL) divergence, and been used in many areas of statistics, machine learning and applied neuroscience. Its formulation is based on conditional entropy and Shannon's entropy. It enjoys nice properties such as being additive for independent distributions.

By itself, KL-divergence is not a metric, but the symmetrized version is a premetric and induces a topology on the space of probability distributions. It has commonly been used as a metric for clustering algorithms. In this setting, it is often referred to as relative entropy. Given a single reference clustering Q, then the error of a computed clustering P can be measured as

$$H(P|Q) + H(Q|P).$$

This error will evaluate to zero, if $P = Q$, and equals $H(P) + H(Q)$, if P is independent of Q.

[*] Terry Adams is with IARPA.

For many problems in computer vision, or more broadly in statistical estimation, the goal is to track or detect the presence of an object or event. See the references [1, 2, 3, 4], as well as the citations contained therein for a background on tracking metrics. The tracking problem is fundamentally different from the clustering problem, in that, it can be an open universe problem. Often in clustering, there is a well-defined superset of objects or elements that are being clustered. From this closed set, a probability space may be established. This makes the problem of clustering tractable for probabilistic methods. However, the problem of tracking can take place in a very large universe of possible tracks. In particular, if the goal is to track a person in space and time from one or more video sequences, the set of possible tracks becomes vast. For high-def video at 30 frames per second, a spatio-temporal track may include billions of pixels. Still, we are able to extend the probabilistic notion of divergence to a unified metric for evaluating tracking problems, as well as event detection. This has the benefit that notions such as track fragmentation and track purity can be handled with a single metric, along with false alarms or partial missed detections. See [4] for a definition of track fragmentation and track purity.

In order to extend the notion of divergence to the tracking regime, we separate the standard formula for the symmetrized KL-divergence (and cross entropy) into multiple parts. The component parts from the KL formulas naturally map to various types of errors. In all, we identify three main types of errors, referred to as:

1. Inner divergence,

2. Outer divergence,

3. Track density divergence.

The inner divergence error represents error obtained from splitting tracks or merging tracks. The outer divergence error measures error produced from false alarms or missed detections. Also, included is a track density divergence which measures error introduced from duplicate system outputs, and more generally the difference between densities produced from system tracks and densities produced from ground-truth tracks. In a statistical sense, these three classes of error can be rolled up into a single metric for measuring tracking performance.

Prior to defining the complete metric, we give further heuristics for this KL-inspired tracking metric that associate common tracking error types with the corresponding components of the KL-divergence formulas. Suppose that $\mathcal{T} = \{\tau_1, \ldots, \tau_n\}$ is a set of ground-truth tracks, and $\mathcal{S} = \{\sigma_1, \ldots, \sigma_m\}$ is a set of system tracks. At the moment, assume that there is no intersection between ground-truth tracks, and likewise, there is no intersection between system tracks. Place a probability measure ν on a given set of tracks in the following manner. Assume ν is defined on $\mathcal{T}$ where each track is given equal weight, $1/n$. Each track τ_i may be thought of as a spatio-temporal volume. Define ν to give uniform mass on τ_i based on the number of pixels in τ_i. Thus, the formula for symmetrized KL-divergence, or more directly, for cross entropy, is

$$H(\mathcal{S}|\mathcal{T}) + H(\mathcal{T}|\mathcal{S})$$

where $H(*|*)$ represents standard conditional entropy using the measure ν. In our setting, this total is referred to as the total inner divergence. Splitting errors are associated with $H(\mathcal{S}|\mathcal{T})$, and merging errors are associated with $H(\mathcal{T}|\mathcal{S})$.

The outer divergence is defined in a similar framework as the inner divergence, but accounts for false alarms and missed detections. Also, in the following section, we show how the track density divergence relates to KL-divergence.

2 Components of Track Divergence

In this section, we define the individual components that make up the total divergence between two spatio-temporal tracks. Given a video frame, a subframe S is a collection of pixels from the frame. This collection can be thought of as a 2-dimensional region with area $a(S)$. A video track τ is a sequence of subframes $S_1, S_2, \ldots, S_k$ (also known as a tuboid). It has a volume $v(\tau)$ defined as $v(\tau) = \sum_{i=1}^{k} a(S_i)$. To view the formulas in a probilistic framework, the measure is normalized to assign equal weight to each track. In particular, if τ is one of n ground-truth tracks, then

$$v(\tau) = \sum_{i=1}^{k} a(S_i) = \frac{1}{n}.$$

2.1 Inner Divergence

The inner divergence of track τ from track σ is computed in the following manner:

$$D_{id}(\tau|\sigma) = -\frac{v(\tau \cap \sigma)}{v(\sigma)} \log \frac{v(\tau \cap \sigma)}{v(\sigma)}$$

Given a set of tracks $\mathcal{T}$ and single track σ,

$$D_{id}(\mathcal{T}|\sigma) = \sum_{i=1}^{n} D_{id}(\tau_i|\sigma). \tag{1}$$

Provided two sets $\mathcal{T} = \{\tau_1, \ldots, \tau_n\}$ and $\mathcal{S} = \{\sigma_1, \ldots, \sigma_m\}$ of tracks, the inner divergence of $\mathcal{T}$ from $\mathcal{S}$ is

$$D_{id}(\mathcal{T}||\mathcal{S}) = \frac{1}{m}\sum_{j=1}^{m} D_{id}(\mathcal{T}|\sigma_j).$$

2.2 Outer Divergence

Given a set of tracks $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_n\}$ and single track σ, let

$$\alpha = \frac{v\big(\sigma \cap (\bigcup_{i=1}^{n} \tau_i)\big)}{v(\sigma)}.$$

Define the outer divergence of $\mathcal{T}$ from track σ as:

$$D_{od}(\mathcal{T}|\sigma) = \log\Big(\frac{1+n}{1+\alpha n}\Big). \tag{2}$$

The outer divergence of track set $\mathcal{T}$ from track set $\mathcal{S}$ is defined as,

$$D_{od}(\mathcal{T}||\mathcal{S}) = \frac{1}{m}\sum_{j=1}^{m} D_{od}(\mathcal{T}|\sigma_j).$$

2.3 Track Density Divergence

If a system outputs the same ground-truth track twice, while the track occurs only once in the ground-truth, then we expect an error to be generated. The distance between these sets of tracks should not be zero. Observe that neither the inner divergence nor the outer divergence, as described previously, penalize a system for outputting the same ground-truth track multiple times. The track density divergence described here assigns a penalty or error for generating spurious duplicative tracks.

Kullback-Leibler divergence is a common method for measuring the difference between two distributions. When comparing two discrete distributions, $P = \{p_1, p_2, \ldots, p_n\}$ and $Q = \{q_1, q_2, \ldots, q_n\}$,

$$D_{KL}(P||Q) = \sum_{i=1}^{n} p_i \log\big(\frac{p_i}{q_i}\big).$$

For this formula, it is assumed that P is absolutely continuous with respect to Q. I.e., $p_i = 0$ whenever $q_i = 0$. This assumption fits well with our needs, since the outer divergence measures errors when the two distribuitons do not overlap. Here we will compute the KL-divergence relative to each track, and then average this across all tracks. To make a symmetrized version, this is done relative to ground-truth tracks, and system tracks separately.

For each pixel location x, define

$$\mathcal{T}(x) = \sum_{i=1}^{n} I_{\tau_i}(x)$$

where $I_{\tau_i}(x)$ is the indicator function for τ_i. For $x \in \tau_i$, let

$$\mathcal{S}_i(x) = \sum_{j=1}^{m} I_{\sigma_j}(x) I_{\tau_i}(x).$$

Define

$$N(\tau_i) = \sum_{x \in \tau_i} \mathcal{S}_i(x)$$

and

$$\tau_i^* = \{x \in \tau_i : \mathcal{S}_i(x) > \mathcal{T}(x)\}.$$

Let the track density divergence of $\mathcal{S}$ given track τ_i be

$$D_{td}(\mathcal{S}|\tau_i) = \frac{1}{N(\tau_i)} \sum_{x \in \tau_i^*} S_i(x) \log\big(\frac{\mathcal{S}_i(x)}{\mathcal{T}(x)}\big)$$

and the total track density divergence given the set of tracks $\mathcal{T}$:

$$D_{td}(\mathcal{S}||\mathcal{T}) = \frac{1}{n} \sum_{i=1}^{n} D_{td}(S|\tau_i) = \frac{1}{n} \sum_{i=1}^{n} \Bigg(\frac{1}{N(\tau_i)} \sum_{x \in \tau_i^*} S_i(x) \log\big(\frac{\mathcal{S}_i(x)}{\mathcal{T}(x)}\big) \Bigg).$$

3 Total Track Divergence

The total track divergence is obtained as a sum of the inner divergence, outer divergence and track density divergence. We include the symmetrized version for each of these; thus, the error measurement becomes the addition of six terms:

$$D_{TD}(\mathcal{T}, \mathcal{S}) = D_{id}(\mathcal{T}||\mathcal{S}) + D_{id}(\mathcal{S}||\mathcal{T}) + D_{od}(\mathcal{T}||\mathcal{S}) + D_{od}(\mathcal{S}||\mathcal{T}) + D_{td}(\mathcal{T}||\mathcal{S}) + D_{td}(\mathcal{S}||\mathcal{T}) \tag{3}$$

This is used only when individual tracks within $\mathcal{T}$ are disjoint, and individual tracks within $\mathcal{S}$ are disjoint. In the following section, it is shown how to handle track sets containing overlapping tracks.

3.1 Purified Track Divergence

Notice that if the reference tracks $\mathcal{T}$ have overlaps within individual tracks, then $D_{TD}(\mathcal{T}, \mathcal{T}) > 0$. In particular, $D_{id}(\mathcal{T}, \mathcal{T}) > 0$. To produce a divergence that assigns 0 error, when a system outputs reference tracks, define purified inner divergence as,

$$D_{pid}(\mathcal{T}||\mathcal{S}) = \begin{cases} 0, & \text{if } D_{id}(\mathcal{T}||\mathcal{T}) > D_{id}(\mathcal{T}||\mathcal{S}) \ , \\ D_{id}(\mathcal{T}||\mathcal{S}) - D_{id}(\mathcal{T}||\mathcal{T}), & \text{otherwise.} \end{cases} \tag{4}$$

Thus, for any track set $\mathcal{T}$,

$$D_{pid}(\mathcal{T}||\mathcal{T}) = 0.$$

3.2 General Formula for the KL-inspired Tracking Metric

Given any two sets of tracks $\mathcal{T}$ and $\mathcal{S}$, define

$$D_{TD}(\mathcal{T},\mathcal{S}) = D_{pid}(\mathcal{T}||\mathcal{S}) + D_{pid}(\mathcal{S}||\mathcal{T}) + D_{od}(\mathcal{T}||\mathcal{S}) + D_{od}(\mathcal{S}||\mathcal{T}) + D_{td}(\mathcal{T}||\mathcal{S}) + D_{td}(\mathcal{S}||\mathcal{T}) \tag{5}$$

4 Results on the Oxford Town Centre Tracking Data

The Active Vision Laboratory at Oxford University has collected a rich video dataset for evaluating pedestrian tracking algorithms. It is publicly available:

`http://www.robots.ox.ac.uk/ActiveVision/Research/Projects/`
`2009bbenfold_headpose/project.html#datasets` .

The dataset contains reference tracks which give highly accurate bounding boxes for all movers (people, strollers). Also, provided are outputs from two systems, one described in a 2009 BMVC paper, and another described in a 2011 CVPR paper.

To demonstrate our metrics, we implemented them in a C++ library. The library contains a few different classes including

- GanitaMetrics
- GanitaMetricsTrackSet
- GanitaMetricsTrack
- GanitaMetricsTopDetection
- GanitaMetricsVisualize

Also, it includes a driver program gmetrics for calling a few different functions. Below is output from running on the two different system outputs. Note, the BMVC results generate a better score than the CVPR 2011 output.

```
time ./gmetrics ../data/duke_tracker_output/TownCentre/TownCentre-groundtruth.top
../data/duke_tracker_output/TownCentre/TownCentre-output-BenfoldReidBMVC2009.top
File ../data/duke_tracker_output/TownCentre/TownCentre-groundtruth.top size = 5553300
File ../data/duke_tracker_output/TownCentre/TownCentre-output-BenfoldReidBMVC2009.top
size = 4390814
Total number of tracks = 230
Total number of tracks = 244
***************************************************
*                    Summary                      *
***************************************************
* Inner div relative to reference      0.286446   *
* Inner div relative to system         0.445220   *
* Total inner div error               +0.731667   *
* Missed detection error              +0.969750   *
* Missed detection proportion          0.438332   *
* Density score rel to reference      +0.012585   *
* False alarm error                   +0.244259   *
* False alarm proportion               0.198179   *
* Density score rel to system         +0.047035   *
*-------------------------------------------------*
* Total KL-track error                =2.005296   *
***************************************************

real    1m1.863s
user    0m58.641s
sys    0m3.216s


time ./gmetrics ../data/duke_tracker_output/TownCentre/TownCentre-groundtruth.top
../data/duke_tracker_output/TownCentre/TownCentre-output-BenfoldReidCVPR2011.top
File ../data/duke_tracker_output/TownCentre/TownCentre-groundtruth.top size = 5553300
File ../data/duke_tracker_output/TownCentre/TownCentre-output-BenfoldReidCVPR2011.top
 size = 5728155
Total number of tracks = 230
Total number of tracks = 455
***************************************************
*                    Summary                      *
***************************************************
* Inner div relative to reference      0.757959   *
* Inner div relative to system         0.192347   *
* Total inner div error               +0.950307   *
* Missed detection error              +0.719617   *
* Missed detection proportion          0.322062   *
* Density score rel to reference      +0.035313   *
* False alarm error                   +0.529964   *
* False alarm proportion               0.252416   *
* Density score rel to system         +0.031307   *
*-------------------------------------------------*
* Total KL-track error                =2.266509   *
***************************************************

real    1m13.410s
user    1m9.821s
sys     0m3.583s
```

The CVPR 2011 system outputted 455 tracks whereas ground-truth contains 230 tracks. This leads to a higher false alarm rate as noticed by 0.529964, and a higher "inner div relative to reference": 0.757959 compared to 0.286446.

5 Properties of the KL-inspired Tracking Metric

1. No need to introduce thresholds (i.e., IoU);
2. Error measurements are continuous with respect to system tracks;
3. No need to enforce a one-to-one track alignment;
4. Split tracks produce lower error rates than false alarms;
5. Error levels do not necessarily tend to infinity with the number of reference tracks.

Typically, when thresholds are introduced into tracking metrics, this causes discontinuities in the scoring. Many times this can impact the rankings of the systems that are being measured. Also, thresholds when applied to IoU[1] tend to incentivize focus at a specific parameter level, while leading to weaker performance toward the overall goal of tracking. Since our metric does not introduce arbitrary thresholds, the metric is continuous with respect to system tracks (as well as ground-truth tracks).

A key property with KL-divergence is the use of the function $f(x) = x\log(x)$ for computing entropy or counting discrepancies between discrete distributions. This function is horseshoe shaped and maps both $x = 0$ and $x = 1$ to zero. It's role in the relative entropy formula has the effect of assigning the largest error when a reference track is split evenly into two halves by two system tracks. Instead, if two system tracks were to cover a reference track, but one system track covers 9/10 while the other system track covers 1/10, then the inner divergence error will be lower. This reduction in error occurs without the need to program specific thresholds for temporal overlap.

Other characteristics of this metric is that the overall score does not necessarily depend strongly on the number of reference tracks. In particular, certain scenarios produce a similar overall error, regardless of the number of tracks. Suppose that there are n disjoint reference tracks of equal length. Suppose a system generates n disjoint system tracks, and for each reference track, there is a single system track that overlaps it by half. In this case, our metrics generate error in two places: inner divergence relative to reference, and outer divergence relative to reference. For the inner divergence error, we get $0.5 \times \log(2)$ error for each reference track. Then a uniform mean is computed to give a final inner divergence error of $0.5 \times \log(2)$. The outer divergence relative to a reference track is

$$\log\left(\frac{1+n}{1+\frac{1}{2}n}\right) < \log\left(\frac{1+n}{\frac{1}{2}+\frac{1}{2}n}\right) = \log(2).$$

To get the overall outer divergence relative to the reference track set, these errors are averaged uniformly over all reference tracks. In this case, the total error will approach $1.5\log(2)$, as $n \to \infty$.

Potential downsides of this metric are that the metric does not appear to be transparent in the manner in which each KL-based error is generated. This could be improved with tools for visualizing the error contribution from individual system tracks. Also, the algorithm, that computes errors, scales at $m * n$ or essentially n^2 where n is the number of reference

[1] IoU stands for Intersection over Union. Often, detectors or trackers apply a threshold for correct detections such as $\{\ \text{IoU} \geq 0.5\}$.

(or system tracks). Nevertheless, we found computation times reasonable as seen with the Town Centre example. (The metric was implemented in C++, and the software is likely to benefit from code optimization.)

Although, no thresholds are used to generate errors, design choices were made to formulate the KL-divergence in this setting. We found that errors were not generated in an obviously counter-intuitive manner. For example, systems that split a reference track produced less error than another system that moved the split piece to be a false alarm. There are cases where errors can tend to infinity with the number of tracks (reference or system). If there are n disjoint reference tracks, and a system generates n more disjoint tracks that do not intersect any other track, the error will be encapsulated in the outer divergence as

$$2\log(1+n).$$

This might be considered the worst case scenario for n reference tracks and n system tracks. There is a scenario which produces a higher error rate. Suppose there are n disjoint reference tracks. Suppose that each system track intersects each reference track at exactly $1/n$ proportion of the reference track. Also, suppose each of the n system tracks are identical. Let's compute the error produced from this scenario. Inner divergence of a system track relative to a reference track is $(1/n)\log(n)$. This is summed over all system tracks to produce $\log(n)$. Then it is uniformly averaged over all reference tracks to give the same $\log(n)$. Inner divergence relative to the system gives the same $\log(n)$ error. The total inner divergence is $2\log(n)$. Outer divergence will produce an error:

$$\log\left(\frac{1+n}{1+(\frac{1}{n})n}\right) = \log\left(\frac{1+n}{2}\right).$$

Also, error is produced by the track density divergence, when identical system tracks are outputted. This produces another $\log(n)$ term. The total error in this scenario is $3\log(n) + \log\left(\frac{1+n}{2}\right)$. This is a very pathological case, and we are not aware of anything like this appearing from a video tracking system. Possibly, this output should have the highest error rate, since the system generated tracks that were independent of the reference tracks, and then continued outputting the same independent track over and over again.

Another potential way to compute the inner divergence would be to take a reference track and compute the relative entropy of the partition induced by the system tracks on this reference track. Then average these relative entropies uniformly across all reference tracks. Finally, flip the role of reference and system tracks and compute the symmetrized version. This would reduce the inner divergence error in the previous example. However, it appears to be much more computationally intensive and does not clearly lead to an improved metric.

We would find it interesting if theoretical results could give credence to one tracking metric, as opposed to other metrics. It would also be interesting, if one could show that any tracking metric satisfying conditions (1) - (5) above, also has the property that error levels can tend to infinity with the number of tracks, for certain pathological cases, otherwise the error rates may tend to zero for systems that are deemed imperfect.

Acknowledgements: We would like to thank Reuven Meth for valuable feedback on the development of these metrics.

References

[1] Godil, A., Bostelman, R., Shackleford, W., Hong, T. and Shneier, M., Performance Metrics for Evalluating Object and Human Detection and Tracking Systems, *NIST*, http://dx.doi.org/10.6028/NIST.IR.7972 (2014).

[2] Bernardin, K., Elbs, A. and Stiefelhagen, R., Multiple Object Tracking Performance Metrics and Evaluation in a Smart Room Environment (2006)

[3] Milan, A., Leal-Taixe, L., Reid, I., Roth, S. and Schindler, K., MOT16: A Benchmark for Multi-Object Tracking, https://arxiv.org/pdf/1603.00831v2.pdf (2016)

[4] Perera, A., Hoogs, A., Srinivas, C., Brooksby, G. and Hu, W., Evaluation of Algorithms for Tracking Multiple Objects in Video, AIPR (2006).