An Information Theoretic Approach for Tracker Performance Evaluation

Edward K. Kao, Matthew P. Daggett, Michael B. Hurley
Intelligence and Decision Technologies Group
MIT Lincoln Laboratory
244 Wood Street, Lexington, MA 02420-9185
{edward.kao, daggett, hurley}@ll.mit.edu

Abstract

Automated tracking of vehicles and people is essential for the effective utilization of imagery in wide area surveillance applications. In order to determine the best tracking algorithm and parameters for a given application, a comprehensive evaluation procedure is required. However, despite half a century of research in multi-target tracking, there is no consensus on how to score the overall performance of these trackers. Existing evaluation approaches assess tracker performance through measures of correspondence between ground truth tracks and system tracks using metrics such as track detection rate, track completeness, track fragmentation rate, and track ID change rate. However, each of these only provides a partial measure of performance and no good method exists to combine them into a holistic metric. Towards this end, this paper presents a pair of information theoretic metrics with similar behavior to the Receiver Operating Characteristic (ROC) curves of signal detection theory. Overall performance is evaluated with the percentage of truth information that a tracker captured and the total amount of false information that it reported. Information content is quantified through conditional entropy and mutual information computations using numerical estimates of the probability of association between the truth and the system tracks. This paper demonstrates how these information quality metrics provide a comprehensive evaluation of overall tracker performance and how they can be used to perform tracker comparisons and parameter tuning on wide-area surveillance imagery and other applications.[1]

1. Introduction

As video surveillance systems continue to cover larger areas at higher resolutions, the need for automated tracking systems becomes increasingly strong so that the full use of all collected imagery can be accomplished by narrowing human analysis to those frames where automated trackers have detected interesting features. An overall scoring function is needed to evaluate different trackers so that users can select the one that is most effective for their application. Such a scoring function must be mathematically sound and comprehensive in capturing the significant characteristics of tracker performance.

The same users usually have no difficulty articulating what they want in a tracking system and can often provide narratives of what they expect it to do. These descriptions suggest that a utility scoring function is what is desired, but the users often do not have sufficient knowledge of the costs and benefits of the tracking system for an analysis of utility to be conducted. The tracker behaviors that are articulated often relate to different track pathologies, such as missed tracks, false tracks, track fragmentations, track merges, and tracks that are too long or too short.

In order to capture the various track behaviors and pathologies, tracker performance metrics like detection rate, completeness, fragmentation rate, and ID change rate, etc. [1][2][3][4][5] have been proposed. However, each metric only addresses a subset of the track pathologies and a single pathology often affects multiple metrics. To obtain an overall score of tracker performance, these metrics are often combined through an ad-hoc weighted sum. The weights are used to adjust the score to account for users' unique needs. However, choosing a set of weights that is fair for algorithm comparison and reflects the requirements of a given application is difficult, especially when the metrics are in different units and are often highly correlated.

An overall scoring function based upon a theoretical foundation eliminates the problems of selecting weights and resolving conflicting results when some metrics improve and others degrade. This paper presents an information theoretic metric that provides an overall

[1] This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense or the United States Air Force.

2009 IEEE 12th International Conference on Computer Vision (ICCV)
978-1-4244-4419-9/09/$25.00 ©2009 IEEE

scoring function and captures the effects of common track pathologies. The approach is related to prior work in medical test performance evaluation [6], medical decision making [7] and inspection performance [8] but extends that work from a binary detection problem to a multi-assignment problem. Furthermore, an additional false information quantity is added to construct ROC-like performance plots and an overall scoring function is proposed over this two-dimensional metric space.

2. Existing Tracking Metrics

There has been an on-going effort in the design of tracker evaluation procedures in the last few years with much work presented in the Workshop on Performance Evaluation of Tracking and Surveillance (PETS). In general, the evaluation procedure involves running a tracking algorithm on a test set and comparing the resulting system tracks to the ground truth tracks [9]. The correspondence between the system tracks and truth tracks is established through an association algorithm. Different metrics have been proposed to capture the quality of the correspondence. Earlier works by Senior [1] and Ellis [2] assess correspondence between truth tracks and system tracks through metrics similar to those of a detection ROC curve. Their proposed metrics focus on the fraction of truth tracks observed by the system (analogous to the probability of detection), as well as the fraction of system tracks that correspond to the truth tracks (one minus this quantity is analogous to the false alarm rate). For the purpose of discussion, these metrics will now be referred to as truth completeness (C_T) and track completeness (C_S). Later works by Brown [3] and Bashir [4] continue to use the completeness metrics; however, they emphasized its limit in evaluating performance in the presence of track fragmentation and track merging. Towards this end, they proposed a new way to compute completeness using a many-to-many association between the system tracks and the truth tracks. Equation (1) and (2) define the completeness metrics in the many-to-many association case (i.e. C_T^M and C_S^M) and the one-to-one association case (i.e. C_T^O and C_S^O) respectively, where $\mathbf{T}$ is the set of all truth tracks, $\mathbf{S}$

$$C_T^M = \frac{\sum_{i\in\mathbf{T}}\sum_{j\in\mathbf{S}} \ell_{ij}}{\sum_{i\in\mathbf{T}} \ell_i} \qquad C_S^M = \frac{\sum_{j\in\mathbf{S}}\sum_{i\in\mathbf{T}} \ell_{ij}}{\sum_{j\in\mathbf{S}} \ell_j} \tag{1}$$

$$C_T^O = \frac{\sum_{i\in\mathbf{T}} \max_{j\in\mathbf{S}}(\ell_{ij})}{\sum_{i\in\mathbf{T}} \ell_i} \qquad C_S^O = \frac{\sum_{j\in\mathbf{S}} \max_{i\in\mathbf{T}}(\ell_{ij})}{\sum_{j\in\mathbf{S}} \ell_j} \tag{2}$$

the set of all system tracks, ℓ_{ij} the length of association between truth track i and system track j, ℓ_i the length of truth track i, and ℓ_j the length of system track j. Figure 1 demonstrates the computation of completeness using a one-to-one association and a many-to-many association through an example where multiple track pathologies are present. In this example,

$$C_T^M = (7+6+4)/(10+12) = 77\%,$$
$$C_S^M = (7+4+6)/(11+6+8) = 68\%,$$
$$C_T^O = (7+6)/(10+12) = 59\%,$$
$$C_S^O = (7+6)/(11+6+8) = 52\%.$$

The differences between the two completeness computation methods are made evident here by the track fragmentation and merge.

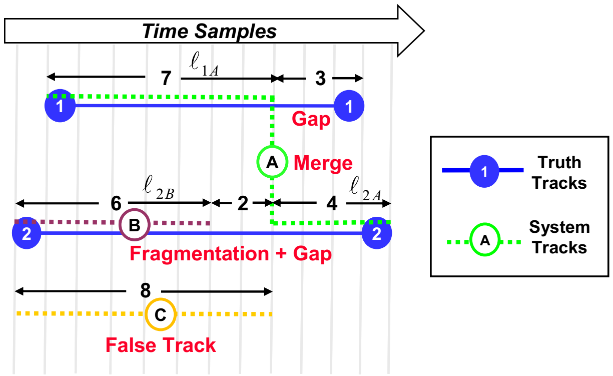


Figure 1. Two truth tracks and three system tracks with multiple track pathologies.

However, neither of the two completeness computation methods gives a comprehensive overall correspondence between the truth tracks and the system tracks. Figure 2 demonstrates six scenarios when tracking two vehicles that pass each other at a traffic intersection. Five cases are presented with common track pathologies. A desirable

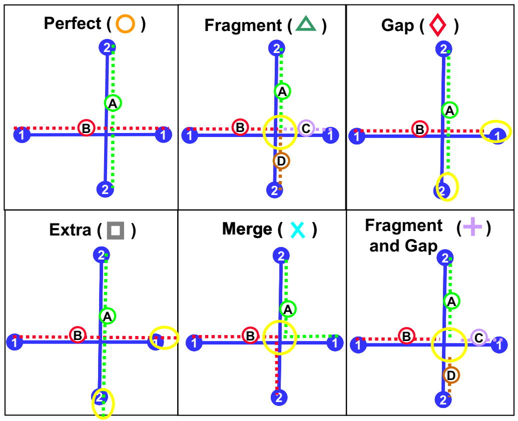


Figure 2. Six scenarios at a traffic intersection with common track pathologies.

metric would show performance degradation reflecting the presence and severity of each pathology. As shown in Figure 3, both existing completeness metrics are limited in their capacity towards this end. In order to show the analogy to a detection ROC curve, the x-axis represents track “incompleteness” which is simply one minus track completeness. Note that both types of completeness metrics fail to distinguish each of the five pathological scenarios. Completeness computed with one-to-one association only considers the single track with the longest association. It disregards all other tracks and fails to distinguish a single pathology from a combination of them, as demonstrated here with the single fragment case (△) versus the fragment and gap case (+). It also tends to overly penalize track fragmentation (△) and track merging (×). Correspondingly, completeness computed from many-to-many association does not penalize track fragmentation (△) and track merge (×). As a result, these completeness metrics are inadequate in evaluating a complete multi-target tracking system where track fragmentation and track merging are typical pathologies. Works by Bashir [4] and Yin [5] included the number of track fragmentations and merges as separate metrics. While all these proposed metrics together respond to all track pathologies, each of them only provides a partial performance measure and no good method exists to combine them into a holistic metric. In the next section, a single metric based on information theory will be described which captures all the aforementioned track pathologies and provides a comprehensive overall performance measure. This work is not to be confused with Loutas’ appearance-based information theoretic tracker evaluation approach in [10] where comparison is made between the object reference chip and the track chip. This paper takes a different approach by measuring directly the similarity between the truth and the observed track states.

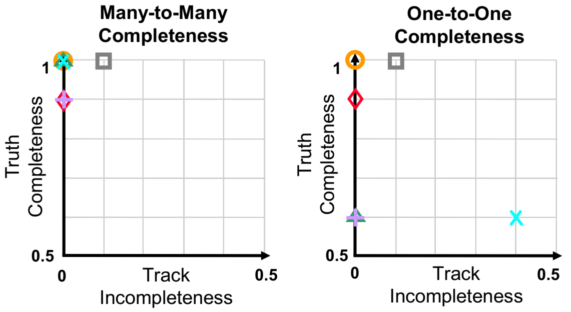


Figure 3. A comparison on the two completeness computation methods in six different scenarios of two crossing tracks at a traffic intersection.

3. Information Theoretic Metrics

Information theoretic metrics can be used to evaluate tracker performance by interpreting tracker data as messages received over a communications channel where the truth data are the original messages. If the truth dataset is known, the information theoretic metrics can be easily calculated. For this paper, it is assumed that the truth dataset is a complete description of the objects of interest. The typical joint entropy diagram [11] in Figure 4 displays a graphical representation of the five information measures that are relevant to tracker performance evaluation: truth track entropy $H(T)$ representing the total amount of the truth tracks’ information, system track entropy $H(S)$ representing the total amount of the system tracks’ information, mutual information $I(T;S)$ representing the amount of matching information between the truth tracks and system tracks, truth conditional entropy $H(T\,|\,S)$ representing the amount of truth tracks information missed by the system tracks, and tracker conditional entropy $H(S\,|\,T)$ representing the amount of false information introduced by the system tracks.

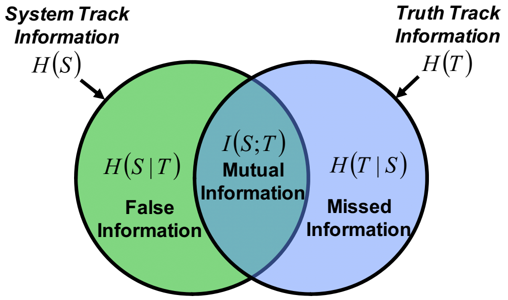


Figure 4. Relationship between truth tracks and system tracks in information space.

A single score based upon information alone can be calculated by summing the two conditional entropies. This score measures the amount of true information that a tracker missed plus the amount of false information that it generated. An optimal tracker from an information theoretic perspective minimizes this score:

$$S_I = H(T\,|\,S) + H(S\,|\,T). \tag{3}$$

Considering the quantities in $H(T) = I(T;S) + H(T\,|\,S)$ and $H(S) = I(T;S) + H(S\,|\,T)$, the five unique terms are interrelated and can be collapsed into three terms. Truth entropy, mutual information, and tracker conditional entropy have been selected as the three terms to use for more in-depth evaluation.

Because comparisons between trackers or algorithm parameters are best made with the same truth dataset, the entropy of the truth dataset can be used to normalize the mutual information and tracker conditional entropy and reduce the number of variables directly used in the evaluation to two:

$$f(T;S) = I(T;S)/H(T), \tag{4}$$

$$r(S\,|\,T) = H(S\,|\,T)/H(T). \tag{5}$$

The score S_I can be normalized as well when comparing different trackers and parameters against a common truth set. The truth information completeness $f(T;S)$ is a measure of the fraction of truth information collected by a tracker, which has been previously called the 'fraction of uncertainty removed' by Raz [8]. The false information ratio $r(S\,|\,T)$ is the ratio of false information to truth information. Figure 5 shows the information coverage plot composed of these two quantities. The plot has many similarities to the ROC curve: perfect tracker performance is at the point (0, 1) (zero false information generated and all truth information captured). Also shown is the notional impact of the aforementioned track pathologies on the performance in this metric space. Curves can be traced through the two-dimensional space by the variation of a single tracker parameter; overall better trackers will generate curves closer to the upper-left corner in the graph than overall poorer trackers.

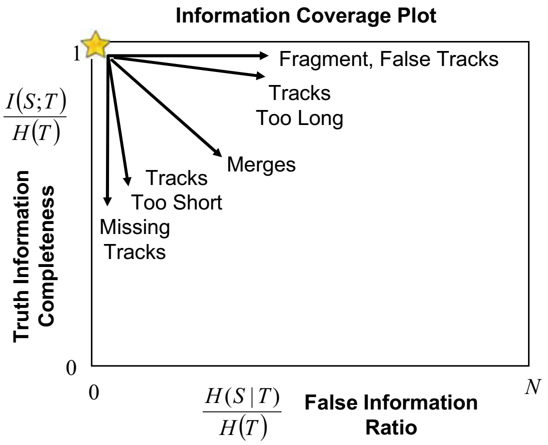
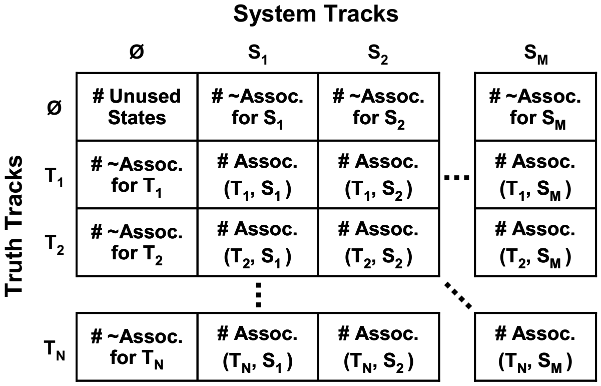


Figure 5. Different track pathologies drive the performance metrics in different directions in the proposed information theoretic metric space.

The principal function needed to calculate the information metrics is a joint probability density function between the truth and system tracks. A simple technique for generating a joint probability density function is to select a series of times to compare tracker data with truth data. The best association between the system tracks and the truth tracks is determined for each time sample. The numbers of times that tracks in the two sets associate and do not associate are counted. As shown in Figure 6, the counts are stored in a two-dimensional matrix and used to generate the joint probability density function. The first row (column) of the matrix is used to record the counts of the number of times that truth (system) tracks did not associate with any system (truth) tracks. Each additional row (column) in the matrix is for a specific truth (system) track.



System Tracks

Truth Tracks	Ø	S_1	S_2		S_M
Ø	# Unused States	# ~Assoc. for S_1	# ~Assoc. for S_2		# ~Assoc. for S_M
T_1	# ~Assoc. for T_1	# Assoc. (T_1, S_1)	# Assoc. (T_1, S_2)	⋯	# Assoc. (T_1, S_M)
T_2	# ~Assoc. for T_2	# Assoc. (T_2, S_1)	# Assoc. (T_2, S_2)		# Assoc. (T_2, S_M)
		⋮		⋱	
T_N	# ~Assoc. for T_N	# Assoc. (T_N, S_1)	# Assoc. (T_N, S_2)		# Assoc. (T_N, S_M)

Figure 6. Track association table used to generate the joint probability.

An association algorithm is used to determine the assignment of system tracks to truth tracks. Both sets of tracks are defined on a common state space and represented as Gaussian probability density functions. Association costs between truth and system tracks are generated by integrating pairs of Gaussian functions. A threshold association cost must be selected to set the threshold between assignments and non-assignments. This is an evaluation parameter that should be analyzed to determine the sensitivity of the information scores to different thresholds. The results of this paper were generated through the use of normalized Mahalanobis distances and a linear assignment algorithm. Additional details on track association can be found in [12] and [13].

An important subtlety with the conversion of association counts to probabilities is that an estimate of the true negatives is needed, i.e. the number of times that a tracker did not report a track when there was no real track present. This value populates the (Ø, Ø) entry in the accumulation matrix. The approach adopted here is to estimate the total number of states that can reasonably exist in the tracker state space and then subtract the total number of counts in all the other entries in the matrix. For video data, the number of frames, times the total size of the state space, divided by the resolution limit in the state space is a reasonable estimate for the total number of states. The maximum number of position states per frame is no more than the number of pixels. The velocity limits and resolutions can usually be determined from the dynamics of the objects of interest and tracker performance requirements.

The accumulation matrix is then normalized by the total state estimate to generate a joint probability density function $P(s,t)$. The system and truth track probability functions $P(s)$ and $P(t)$, can be calculated by the normalized sums along rows or columns in the association table. The conditional probability density functions $P(t\,|\,s)$ and $P(s\,|\,t)$ are simply $P(s,t)/P(s)$ and $P(s,t)/P(t)$. The information metrics can be easily computed from these probability terms using the following equations [11]:

$$H(T) = -\sum_{t} P(t)\log(P(t)), \tag{6}$$

$$H(S) = -\sum_{s} P(s)\log(P(s)), \tag{7}$$

$$H(S\,|\,T) = -\sum_{s,t} P(s,t)\log(P(s\,|\,t)), \tag{8}$$

$$H(T\,|\,S) = -\sum_{s,t} P(s,t)\log(P(t\,|\,s)), \tag{9}$$

$$I(T;S) = \sum_{s,t} P(s,t)\log\left(\frac{P(s,t)}{P(s)P(t)}\right). \tag{10}$$

Figure 7 shows the common track pathology cases presented earlier (see Figure 2). Unlike the existing completeness metrics, the information theoretic metrics are not only able to distinguish each case but provide results that are proportional to the intuitive severity of the track pathologies. The ordering of each case here is according to the overall score function described in Equation (3). It is worth noting the metric's preference of fragments (△) over merges (×). Such a preference is intuitive because truth tracks can be perfectly reconstructed from connecting segments of system tracks while truth tracks cannot be reconstructed after a merge. Additional information on the precise location of the merge is needed to reconstruct the truth tracks. In other words, merges result in an informational loss on the truth track's coverage. Another nice feature is the information metrics' ability to handle combinations of track pathologies. This can be observed by the similarity in the deterioration of performance from the perfect case (○) to the gap case (◇) and from the fragment case (△) to the fragment-plus-gap case (+).

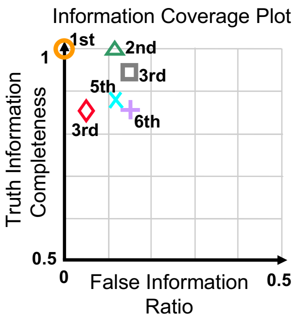


Figure 7. The information coverage plot effectively evaluates common track pathologies.

4. Applications

Wide area surveillance for activity recognition and event reconstruction requires information on the movement of vehicles and people. The large area of regard and hours upon hours of video imagery make the exploitation of such data very daunting for human analysts. Automated tracking can significantly reduce analyst workloads, enabling analysts to quickly find all the objects that had entered a certain area at a certain time, as well as all the places an object of interest has visited. Moreover, post-processing of tracks in an automated event detection algorithm can cue analysts to more thoroughly inspect specific data for anomalous behavior.

However, automated tracking in video imagery remains a difficult problem, especially in an urban area where traffic volume can be large and occlusion is prevalent in the scene. Figure 8 shows an example of such operating environment. The wide area coverage was accomplished by stitching images collected by an array of six cameras from an airborne platform flying over Ohio State University. As shown by the example system tracks, track fragmentation and track merge are common pathologies at a traffic intersection as crowds of vehicles come to stop next to each other and become difficult for the tracker to distinguish. Also shown are false tracks caused by spurious detections from imperfect image registration, as well as gaps in the system tracks caused by occlusions.

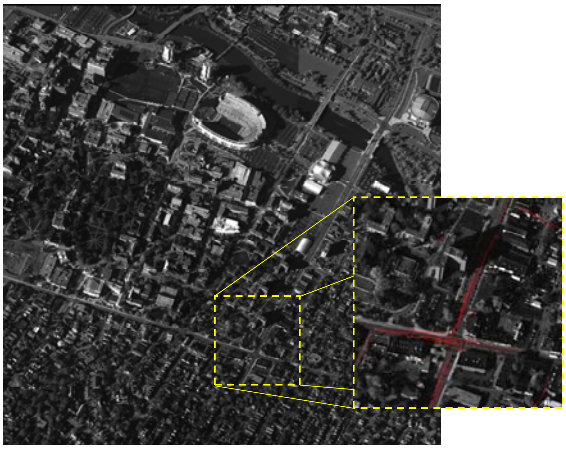


Figure 8. Wide area surveillance in urban environment with some example system tracks. Image from the AFRL CLIF dataset [14].

In order to optimize tracking performance under such challenging environment, the proposed information theoretic metric is used to determine the tracking algorithm and parameter value that produces system tracks that capture the most information on the truth tracks while minimizing the amount of false information introduced.

Figure 9 shows the performance of two different tracking algorithms as the detection threshold parameter is varied. The dotted lines represent the points in the metric space with equal amount of erroneous information (recall Equation (3)). Lines towards the upper left hand corner represent better overall performance. In applications where the amount of missed information is not equally important to the amount of false information, one may add weights to the terms in Equation (3). This simply causes the slope of the dotted lines to change. In this example, tracking algorithm 1 dominates algorithm 2 in the detection threshold parameter space. However, in cases where the performance curves cross, the choice of the winning algorithm will depend on the operating region (i.e. the relative importance between missed and false information).

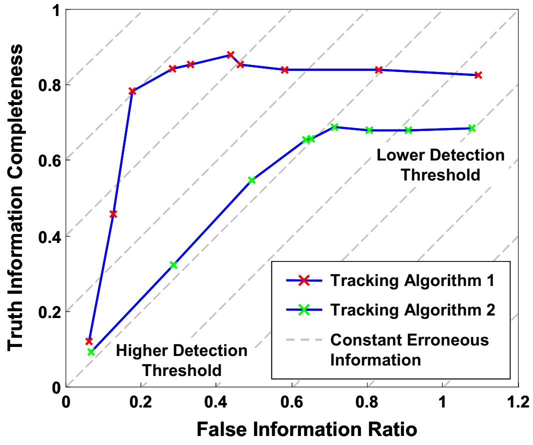


Figure 9. Algorithm comparison using the information coverage plot.

Figure 10 shows the performance curves of tracking algorithm 1 when two different parameters are varied. Varying the detection threshold adjusts the detection sensitivity so it is not surprising to see the corresponding curve take a shape of a typical detection ROC curve. On the other hand, the tracker process noise curve displays a very different shape. The optimal parameter value can be simply determined as the point closest to the most upper-left dashed line. In theory, the tracker process noise parameter is orthogonal to the detection threshold parameter in its impact on the performance. Therefore, the two parameters can be tuned independently. However, in practice, an incomplete tracking model can cause the two parameters to be correlated. In this case, iterative joint parameter tuning can be performed, with each iteration focusing on a smaller parameter space.

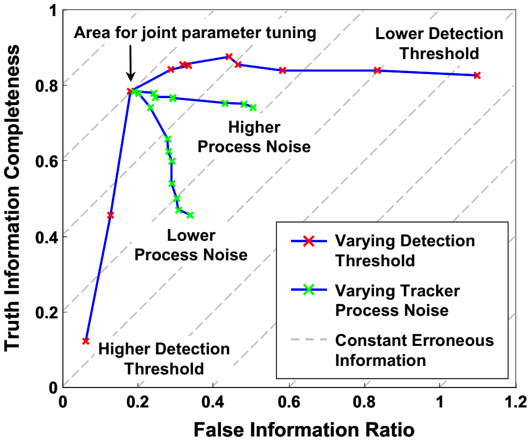


Figure 10. Parameter tuning using the information coverage plot.

5. Conclusion

An information theoretic metric has been derived to evaluate the overall performance of multi-target trackers. Results on a set of common track pathologies demonstrate the new metric's effectiveness over the existing completeness metrics by capturing the impact of common track pathologies on overall performance. The proposed metric has been applied to a set of wide-area-surveillance video imagery for tracker algorithm comparison and parameter tuning. The ROC-like information coverage plot provides an effective means by which to compare and visualize the overall performance between different trackers. It also allows performance curves to be generated as tracker parameters are adjusted for optimal performance.

The information theoretic metrics can be used to evaluate the effects from changes to any part of the wide-area video surveillance processing chain, from sensors, to data registration and reduction, movement detection, and target tracking. The metrics are general so that they can be easily applied for performance evaluation in other application areas where the system observation and the ground truth can be expressed in an N-to-M mapping, such as for classification and recognition algorithms.

Additional work remains in order to fully evaluate the utility of this information theoretic evaluation technique. First, the sensitivity of the information theoretic metrics to parameters within its own algorithm remains to be fully analyzed. Although not presented, preliminary evaluation on the sensitivity of the information theoretic metrics to the estimation of the size of the overall state space and the association threshold parameter has shown that, while the magnitudes of the resulting metrics value change, the relative positions on the information coverage plot for different trackers and parameter sets remain consistent with respect to one another. Further evaluation is required to more accurately quantify these sensitivities.

Furthermore, the current association algorithm is a linear assignment algorithm that makes a one-to-one assignment between truth and system tracks at each time sample. Further study is planned to examine the utility of using a linear programming algorithm to make many-to-many assignments of truth and system tracks at each time sample via fractional weights and estimate the information theoretic metrics based upon these weights.

Lastly, instances where the truth dataset is not fully complete need to be investigated. Specifically, the affect of incomplete truth on the False Information Ratio needs to be characterized.

Acknowledgement

The authors would like to thank several Lincoln Laboratory colleagues: Gary Condon, Peter Jones, and Luke Skelly for technical discussions and feedback; Luke Skelly, Robert Garnier, and Andrew Wang for infrastructure support.

References

[1] Senior, A. et al., *Appearance Models for Occlusion Handling*, IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Kauai, HI, December, 2001.

[2] T. Ellis, *Performance Metrics and Methods for Tracking in Surveillance*, IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Copenhagen, Denmark, June 2002.

[3] L. M. Brown, A. W. Senior, Ying-li Tian, Jonathan Connell, Arun Hampapur, Chiao-Fe Shu, Hans Merkl, Max Lu, *Performance Evaluation of Surveillance Systems Under Varying Conditions*, IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Colorado, USA, Jan 2005.

[4] F. Bashir, F. Porikli. *Performance Evaluation of Object Detection and Tracking Systems*, IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, New York, USA, June 2006.

[5] Fei Yin, Dimitrios Makris, Sergio Velastin, *Performance Evaluation of Object Tracking Algorithms*, IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Rio De Janeiro, Brazil, October 2007.

[6] Metz, Charles E., David J. Goodenough, Kurt Rossman, *Evaluation of Receiver Operating Characteristic Curve Data in Terms of Information Theory, with Applications to Radiography*, Radiology, vol. 109, no. 2, pp 297-303, November 1973.

[7] Eugene Somoza and Douglas Mossman, *Comparing and Optimizing Diagnostic Tests; an Information-theoretical Approach*, Medical Decision Making, vol.12, no. 3, pp. 179-188, 1992.

[8] Tzvi Raz, *Information theoretic measures of inspection performance*, Int. J. Prod. Res., vol. 29, no. 5, pp 913-926, 1991.

[9] Robert Collins, Xuhui Zhou, Seng Keat Teh, *An Open Source Tracking Testbed and Evaluation Web Site*, The Joint IEEE International Workshop on Visual Surveillance, Beijing, China, October 2005.

[10] E. Loutas, N. Nikolaidis, I. Pitas, *Evaluation of Tracking Reliability Metrics based on Information Theory and Normalized Correlation*, 17th International Conference on Pattern Recognition (ICPR'04), vol. 4, pp 653-656, 2004.

[11] Thomas M. Cover and Joy A. Thomas, Elements of Information Theory, Wiley-Interscience, 1991.

[12] Samuel S. Blackman, Multiple Target Tracking with Radar Applications, Dedham, MA, Artech House, Inc., 1986.

[13] Michael B. Hurley, *Track Association with Bayesian Probability Theory*, Technical report ADA417987, DTIC, 10 October 2003.

[14] AFRL CLIF 2007 dataset over Ohio State University https://www.sdms.afrl.af.mil/datasets/clif2007/