NISTIR 7972

Performance Metrics for Evaluating Object and Human Detection and Tracking Systems

Afzal Godil
Roger Bostelman
Will Shackleford
Tsai Hong
Michael Shneier

http://dx.doi.org/10.6028/NIST.IR.7972

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 7972

Performance Metrics for Evaluating Object and Human Detection and Tracking Systems

Afzal Godil
Roger Bostelman
Will Shackleford
Tsai Hong
Michael Shneier
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.7972

July 2014

U.S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Willie May, Acting Under Secretary of Commerce for Standards and Technology and Acting Director

This publication is available free of charge from http://dx.doi.org/10.6028/NIST.IR.7972

DISCLAIMER

Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

1. Abstract:

In this report, we provide an overview of various performance evaluation metrics for object detection and tracking for robot safety applications in smart manufacturing. We present three different types of performance evaluation metrics based on detection, tracking, and perimeter intrusion. The basis for comparing the strengths and weaknesses of different object detection and tracking algorithms is to evaluate their results on a set of tasks with known ground-truth data using the same performance metrics. The tasks, the ground-truth data, and performance evaluation metrics and test procedures can help vendors justify claims about the performance of their systems and assist users and manufacturers to compare systems for their particular automation tasks. They will also allow researchers to fully understand the strengths and limitations of different approaches. This is an essential step towards establishing the credibility of object detection and tracking for real time manufacturing and robotic applications. The performance metrics and evaluation methods are an essential first step towards providing scientific foundations for developing robot safety standards that enable the use of perception systems in manufacturing applications and particularly in providing confidence in systems to be used for safety-critical applications.

2. Introduction

Next generation robotic systems are expected to perform highly complex tasks in dynamic manufacturing environments. Collaboration between humans and robots can take advantage of their complementary strengths. Humans can perform complex and precise tasks that require intelligence, while robots can do dangerous and repetitive tasks well. However, human-robot interaction in a manufacturing environment can be dangerous for workers because of possible collisions with robots or other objects. To mitigate the risks, the robot system must know the location of people and objects at all times. This situational awareness requires the use of perception systems that can recognize, localize, and track objects in their environment. This is a challenging task because a manufacturing environment can be cluttered, there may be objects occluding each other, and there could be illumination and viewpoint variations. While prototypes of such perception algorithms are being developed, a science-based methodology for their performance evaluation does not exist. We are currently developing the necessary metrics and methods, with an initial focus on the ability to detect people and objects as they move about the workspace. We will build test-beds and conduct experiments to assess the methodology. The results can be used to develop new standards that will help enable the use of perception systems in manufacturing applications. Our previous efforts are reported in [24][25].

In our methodology, the strengths and weaknesses of different object detection and tracking algorithms are compared using known ground-truth data on a common set of tasks using a common set of performance metrics. The tasks, the known ground-truth data, and the performance metrics and test procedures will help vendors justify claims about the performance of their systems and assist users and manufacturers to evaluate the systems for their specific automation tasks. They will also allow researchers to fully understand the strengths and limitations of different approaches for different tasks. This

is a first step towards establishing the credibility of perception systems used for object detection, recognition, and tracking for manufacturing and robotic applications. The results are planned to lead to the development of safety standards.

Object detection and localization are important for many other practical applications such as manufacturing automation, navigation, part inspection, and computer aided design/computer aided manufacturing (CAD/CAM). Our main interest is evaluating algorithms used to detect and track objects for robot safety applications and for smart manufacturing applications in a dynamic indoor factory environment. We emphasize the detection of position and orientation, motion, and classification of objects. The following scenarios are examples of those for which we will evaluate perception systems.

- Human and object detection and tracking
- Articulated human motion tracking
- Tracking of robots, automated guided vehicles (AGVs), and industrial parts
- Human-robot Interactions

In most previous work, the ground-truth was created in the sensor coordinate system by manually annotating the objects, for example by drawing a bounding box around each object in a sequence of images. In our case the 3D ground-truth is captured in world coordinates with the help of a tracking sensor (for more details, see [24][25]). In the annotation-based case, matches are evaluated using the area of intersection of bounding boxes in the ground truth data and data from the system being evaluated. For our case (3D ground-truth), the matches are evaluated based on the Euclidean distance between the 3D ground-truth data and the data from the system being evaluated. Other performance evaluation metrics for our study are similar to performance evaluation metrics used for video surveillance systems. Hence, a review of that work is presented.

During the last decade, several performance evaluation projects for video surveillance systems have been developed **Error! Reference source not found.**[2][3][4][5][6][9], each with different emphasis and motivation. The PETS workshops [8] focused on algorithm development and performance evaluation of tasks such as multiple object detection, event detection, and recognition. Nascimento and Marques[22] proposed a novel way to evaluate the performance of object detection systems by comparing algorithm results to ground-truth data and calculating performance metrics such as correct detections, false alarms, detection failure, and splitting and merging errors. CLEAR [3]provides performance evaluation of people, faces, cars, and object tracking and ETISEO [7]was a video understanding and evaluation project for tracking systems that used an event detection algorithm. The i-LIDS [4] is a United Kingdom government initiative that conducts performance evaluations of vision-based detection systems to ensure that they meet Government requirements. Other papers specific to tracking-based metrics are Brown et al [12] who suggest a motion tracking evaluation framework that estimates the number of True Positive, False Positive and False Negative, Merged, and Split trajectories. Yin et al. [11] proposed a large set of metrics to assess different aspects of the performance of motion tracking and to help identify shortcomings of motion trackers under specific conditions. Lazarevic-McManus et al [13] developed a tracking metric to enable evaluation of motion detection based on Receiver Operating Characteristic (ROC)-like curves and the F-measure. Bashir and Porikli [10] presented

metrics based on the spatial intersection of ground-truth and system generated bounding boxes and then calculated a number performance metrics, which they then averaged for all the sampled frames. Black et al. [21] used synthetic video to evaluate tracking performance. They varied the scene complexity of the tracking task by adding occlusions and clutter and increasing the number of objects and people in the scene and presented results based on a number of metrics. Several other performance evaluation metrics were developed and discussed in [16][17][18][19][20][23].

The National Institute of Standards and Technology (NIST) has a long history in this field, having helped to develop performance metrics for object and human detection in a number of different applications, ranging from videoconferences through surveillance to counting and tracking people in stores and commercial establishments. NIST has worked with the United States (US) Department of Homeland Security, with the British Home Office, and with the European CHIL [14] program and the CLEAR [3] evaluations. NIST has also worked with the US Army Collaborative Technology Alliance (CTA) on Robotics to evaluate systems that locate and track human pedestrians from a moving vehicle [15].

In this report we describe performance evaluation metrics that can be used for evaluating the performance of a number of tasks, including object detection, tracking, and perimeter intrusion detection, and also mention some of the factors that affect performance. The performance metrics allow us to quantitatively compare different systems and measure performance improvements over time. With the ROC curve and Precision Recall curve, tradeoffs between performance and other parameters can be determined.

3. Performance Evaluation Metrics

The performance evaluation should be quantitative. It should report how many objects were detected correctly and how many false positives (false alarms) were produced. It should support one-to-one matches, one to many matches, and many to one matches, and the evaluation should scale up to larger test areas or multiple 3D scenes without losing its tracking capability (Figure 1 and Figure 2). There are three main types of performance metrics in our system: detection-based metrics; tracking-based metrics; and perimeter intrusion detection metrics. The detection-based metrics are used to evaluate the performance of a System Under Test (SUT) on individual frames from video sensor data. They do not monitor the identities of objects over the life of the test. All the objects are individually tested to see if there is a match between the SUT and the Ground-truth (GT) system for each video frame. The performance on each individual frame is then averaged over all the frames in the experiment to develop a performance score. The tracking-based metrics use the identity and the complete trajectory of each object separately over the test sequence and compare the GT tracks with the SUT tracks based on best correspondence. Then, based on the best matches, various error rates and performance metrics, described below, are computed. Finally, the perimeter intrusion detection measure is based on detecting any object when it enters a specified area.

This publication is available free of charge from http://dx.doi.org/10.6028/NIST.IR.7972

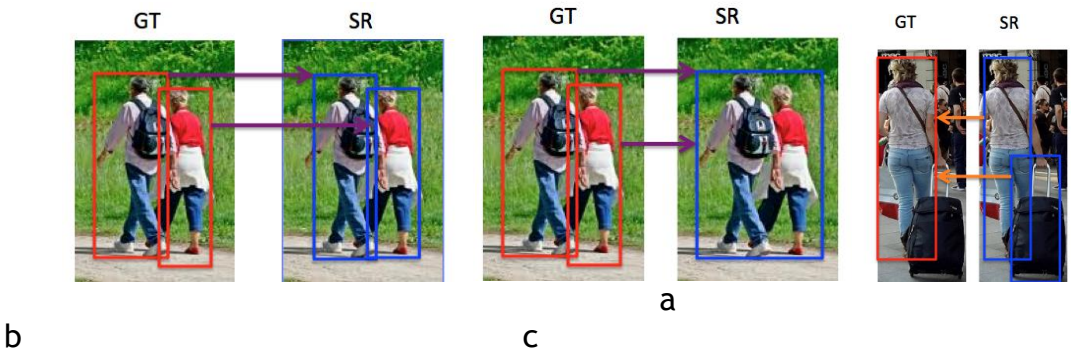

Figure 1. (a) One-to-one matching, (b) many-to-one matching and (c) one-to-many matching.

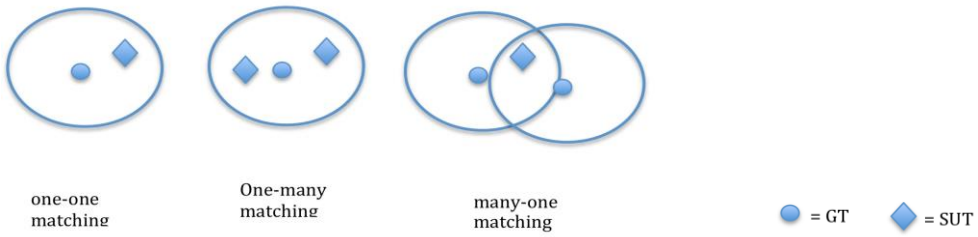

Figure 2. One-to-one matching, many-to-one matching, and one-to-many matching.

3.1. Object Correspondence

The methods used for determining object correspondences between the SUT's data and the GT data significantly affect the values of the performance measures. The object matching and correspondence methods that we use are discussed in the following sections. In the case of annotation-based GT, the system results and the GT data are compared in the sensor coordinate (or sensor image). The main criteria are:

1) Object matching based on the object area intersection criterion is measured by calculating the overlapping area of the SUT-reported bounding box with the GT bounding box at each frame [22], with a threshold selected for a successful match.
2) Object matching using object centroids is based on measuring the Euclidean distance between the object's centroid as reported by the SUT and the GT data at each frame, with a threshold selected for a successful match. Normalization based on the size of the bounding box is often also used [15].

In the case of 3D GT data in world coordinates, object matching is based on a centroid criterion. The threshold value used for matching may increase with the distance from the sensor, since the accuracy of the depth reported by some sensors goes down with distance. Sometimes there is also an allowance for a time threshold for object matching if there is a time lag between the GT and SUT data. We only test for correspondence when

the objects are in the field of view of the SUT sensor. It is also possible to project the 3D GT data back to the 2D sensor image coordinates and use object matching based on object area intersection as described above for annotation-based GT systems.

3.2. Detection Metrics

The purpose of a detection-based metric is to get meaningful measures of the system's ability to perform object detection tasks. Metrics include the number of correctly detected objects, falsely detected objects, or misdetected objects. Other widely used detection measures are detection rate/precision and sensitivity. The detection-based metrics (also called frame-based metrics) are used to evaluate the performance of a SUT on individual frames from video sensor data. They do not take into account the identities of objects over the lifespan of the test. All the objects are individually validated to see if there is a corresponding match between SUT and GT systems for each frame during the test. To compute the performance of the SUT compared to the GT data, we mainly followed the work of Bashir and Porikli [10] and Nascimento and Marques [22]. When associating GT data with SUT-detected objects, six cases can occur [10][22]: zero-to-one, one-to-zero, one-to-one, many-to-one, one-to-many, and many-to-many associations. According to [10] and [22], these associations correspond to false alarms (the detected object has no correspondence), misdetection (the GT data has no correspondence), correct detection (the detected object matches one and only one object), merge error (the detected object is associated with several GT objects), split error, and split-merge. The performances for each individual frame are then averaged over all the frames in the experiment to provide a performance evaluation measure.

The notation used for evaluation is as follows:

- SUT—System Under Test
- FP—False positive, an object present in the SUT, but not in the GT (also called a False Alarm)
- FN—False negative, an object present in the GT, but not in the SUT (also called a Detection Failure)
- TP—True positive, an object present in the GT and the SUT (also called Correct Detection or one-to-one match)
- TN—True negative, an element present in neither the GT nor the SUT
- CGT—Complete Ground Truth is the total number of GT objects.

The following metrics are calculated:

3.2.1. False Positive Rate (FPR)

FPR = FP/(FP + TN), the number of false positives relative to the sum of the number of false positives and true negatives. It is a measure of how well the system correctly rejects false positives.

3.2.2. False Alarm Rate (FAR)

FAR = FP/(TP+FP), the number of false positives relative to the sum of the number of true positives and the false positives. It provides a measure of the likelihood that a detected target is correctly reported.

3.2.3. Detection Rate (DR)

DR = TP/(TP+ FN), the number of true positives relative to the sum of the true positives and the false negatives. It is a measure of the percentage of true targets that is detected.

3.2.4. False Negative Rate

False Negative Rate = FN/(TP+FN), the number of false negatives relative to the sum of the true positives and the false negatives. It is a measure of the likelihood that a target will be missed given the total number of actual targets.

3.2.5. True Negative Rate (TNR)

TNR = TN/ (TN + FP), the true false detections relative to the sum of the true false detections and the false positive. This provides a measure of the likelihood of a negative response given the total number of actual negative detections.

3.2.6. Accuracy

Accuracy = (TP+TN)/CGT, the sum of the true positives and the true negatives relative to the total number of GT objects. This is a measure of the actual performance of the system with regard to both correctly detecting and correctly rejecting targets.

3.2.7. Precision

Precision = TP/ (TP + FP), the number of true positives relative to the sum of the true positives and the false positives. That is, precision is the fraction of detected items that are correct.

3.2.8. Recall

Recall = TP/ (TP + FN), the number of true positives relative to the sum of the true positives and the false negatives. Recall is the fraction of items that were correctly detected among all the items that should have been detected.

3.2.9. F-Measure

F-Measure = $(1 + b^2)$ x (Precision * Recall) / (b^2 x Precision + Recall)

Where b^2 is a non-negative real valued weighting factor [23]. The F-measure gives an estimate of the accuracy of the system under test.

3.2.10. Receiver Operating Characteristic (ROC) Curve

Detection rate vs. False Positive Rate (many other ROC-like curves are possible)

A single performance number is inadequate to measure system performance. Since the system performance has many critical measurement points, it is best represented by a performance curve. In the next paragraph, we discuss two curves widely used for this purpose.

Most of the SUTs will detect different objects and report their locations and detection confidence values. By varying the detection confidence value, the ROC curve can be calculated. It shows the Detection Rate vs. False Positive Rate curve for different factors at different value levels as shown in Figure 3. An ideal ROC curve will show a very steep rise followed by a flat response.

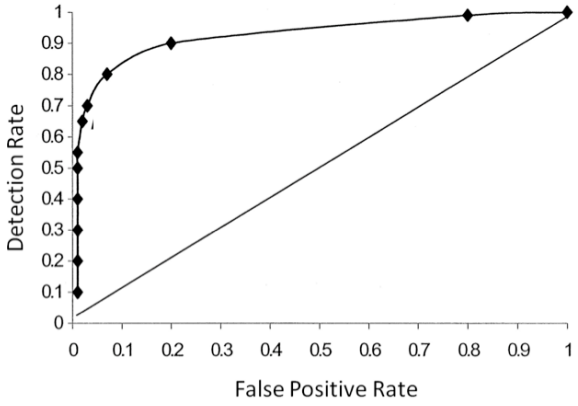

Figure 3. An example ROC curve (Detection Rate vs. False Positive Rate)

3.2.11. Detection Error Trade-off Curve (DET Curve)

A DET Curve is a graph of Miss Rate (or False Negative Rate) vs. False Positive Rate. The DET curve is a plot of the error rate for a binary classification system.

3.2.12. Precision-Recall Curve (PR Curve)

By varying the confidence value, it is also possible to create the PR curve. In pattern recognition and information retrieval, precision (also called positive predictive value) is the fraction of labeled or retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of labeled or relevant instances that are retrieved. Both precision and recall are therefore measures of relevance.

3.2.13. Time Detection Lag

The time detection lag is the delay that the SUT has compared to the GT system. This value should be accounted for in calculating all of the other metrics. Its effects should be evaluated rather than trying to calibrate the lag and remove it from the computations.

3.2.14. Object Localization Metrics

The 2D/3D localization metrics measure the distances between the centers of SUT-detected objects and the corresponding GT centers of gravity. This metric determines the detection precision.

3.3. Tracking Metrics

The tracking based metrics measure the ability of a SUT to track objects over time. The tracking-based metrics (also called object-based metrics) take the identity and the complete trajectory of each object separately over the test sequence and compare the GT tracks with the SUT tracks based on best correspondence. Then, based on these correspondences, various error rate and performance metrics are computed.

Since the GT track(s) could correspond to more than one SUT track, a correspondence mapping has to be established first. Based on this mapping between the object tracks, the track-based metrics are computed. The correct match requires both spatial and temporal

overlap between GT tracks and SUT tracks as shown in Figure 4. Some of the measures that we have selected are based on [2][10][11][12][13]. Requirements for these metrics include:

1. Finding a mapping between the objects or people indicated by the GT and the hypotheses of the tracker (correspondence problem).
2. For each individual mapping, determining the precision with which the object's or person's position was estimated.
3. Counting all GT persons as misses for which no SUT tracker hypothesis was output.
4. Counting all SUT hypotheses for which no GT exists as false positives.
5. Making sure that the objects and people were tracked correctly over time. This includes checking that objects and people were not substituted for each other, for example when they passed close to each other, and checking that a track was correctly recovered after it was lost, for example when an object or person was occluded.

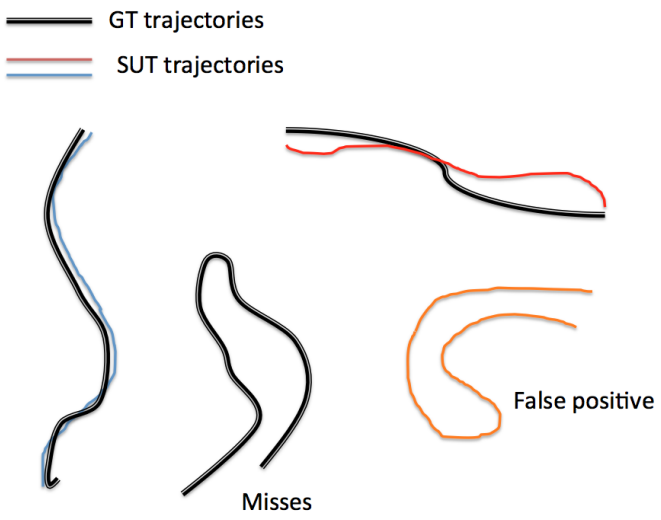

Figure 4. Shows the GT and SUT trajectories and a miss and false positive

Two measures are used to express the performance of the tracker. The first is the tracking precision, which expresses how well the tracker estimates the exact positions of objects or people. The second is the tracking accuracy, which measures how well the system keeps track of people or objects and how many mistakes are made in terms of misses, false positives, mismatches, failures to recover tracks, etc.

3.3.1. Object Tracking Time delay

This is the estimated delay between the SUT algorithm's detection of an object or person and that of the GT [11]. It could be positive or negative.

3.3.2. Tracker Detection rate (TRDR)

This is the precision. TRDR= Total True Positives/ Total Number of GT tracks[11].

3.3.3. Identifier Change (IDC)

The metric IDC is the number of times the identifier changes for each SUT track, while the GT identifier is unchanged. This is a very basic metric for a tracking test [11].

3.3.4. Track Matching Error (TME)

This TME metric is the positional error between the SUT trajectory and the GT trajectory and measures the average distance error between the GT and SUT track. The smaller the TME number, the better the tracking accuracy [11].

3.3.5. Track Completeness (TC)

TC is defined as the time for which the SUT track overlapped with the GT track divided by the total duration of the GT track [11].

3.3.6. Latency of the SUT track (LT)

Latency (LT) is the time delay of the SUT track start compared to the GT track start. The optimal latency is zero or less than zero if the GT sensor has latency [11].

3.3.7. Occlusion success rate (OSR)

Occlusion success rate is not easy to calculate in our case. OSR = Number of successful dynamic occlusions/ Total number of dynamic occlusions [11]. A successful occlusion occurs when the track and object identity are not lost during the occlusion or are correctly recovered immediately following the occlusion.

3.3.8. ROC Curve

It is possible to calculate ROC curves based on tracking by varying the different parameters (threshold, confidence value, etc.)

3.3.9. Precision Recall Curve

Precision and recall measures can also be used as metrics. The tracking methods can be evaluated on the basis of whether or not they generate correct trajectories. In the context of tracking, precision and recall measures can be defined as in [13]:

Precision = TP/ STR, the percentage of the selected trajectory that is correct,
Recall = GT / STR, the percentage of the GT trajectory that overlaps with the selected trajectory,

where STR is one of the selected trajectories out of all the trajectories reported by the SUT. The precision and recall curve identifies the track that maximizes recall for a given precision and determines the value of parameters (threshold, confidence value, etc.)

3.3.10. Multiple Object Tracking Precision (MOTP)

MOTP is the precision of the tracker in determining the exact position of a tracked person or object. MOTP is calculated as follows:

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} ,$$

where d_t^i is the Euclidian distance error between the matched GT location and the matched SUT location and c_t is the total number of matches made. The MOTP is a Euclidian distance error for matched GT-SUT pairs over all frames, averaged by the total number of matches made. It shows how well positions of persons or objects are estimated.

3.3.11. Multiple Object Tracking Accuracy (MOTA)

MOTA is the accuracy of the tracker in keeping correct correspondences over time, estimating the number of people or objects, recovering tracks, etc.

$$MOTA = 1 - \frac{\sum_t m_t + fp_t + mme_t}{\sum_t g_t} ,$$

Where m_t, fp_t, mme_t and g_t are the number of misses, of false positives, of mismatches and the number of GT objects respectively for time t. It is the sum of all errors made by the tracker over all frames, averaged by the total number of GT objects and people. MOTA is similar to accuracy metrics widely used in other domains and gives a very intuitive measure of the tracker's performance independent of its ability to determine the exact person locations.

3.4. Perimeter Intrusion Detection Metric

In this section, we focus on detection performance metrics for safety systems and flexible automation. In order for more advanced flexible automation to operate in an environment that may contain humans, the probability of the robot or automation system injuring any person must be acceptably low. The human detection system is only part of the entire automation safety system and thus by itself cannot guarantee a safe operating environment. Human detection systems provide varying amounts of information about the humans they detect. The more information provided and the more accurate that information, the more options are available to the automation system designer as to how to use that information.

3.4.1. Mean Time to False Detection

The true and false detection rates are measured separately for each type of moving or movable object by deliberately moving that object into the protected area of the human detection system. In addition, the system must be tested to determine how frequently the perimeter intrusion is triggered with no stimulus present. The system's false detection rate will equal the sum of the false detection rate for each class of object multiplied by the frequency associated with that class of object plus the false detection rate without a stimulus. The mean time to false detection is the reciprocal of the false detection rate. It

is important to vary the speeds and positions where the intrusion occurs. The rates should also be reported separately for each object type if the results are to be extrapolated to environments with different frequencies.

3.4.2. Mean Time to Missed Detection for Perimeter Intrusion Accuracy

Separate missed detection rates must be measured to capture the contribution of each activity, demographic group, and type of clothing likely to occur in the environment. The planned activity will dictate a range of positions and velocities around the perimeter that need to be tested. It is important to report the rates separately if the results are to be extrapolated to environments with different frequencies. Activities that cannot be performed while crossing the perimeter need not be considered.

3.4.3. Presence Accuracy for Perimeter Intrusion Accuracy

Presence accuracy includes the same two metrics as perimeter intrusion, but adds two additional metrics which measure the system's ability to report when all people have left the protected area.

3.4.3.1. Mean Time to False Clear for Human Presence

The average time while the human presence is being reported until the system falsely indicates the area is clear.

3.4.3.2. Mean Time to Missed Clear for Presence Accuracy

The average time while presence is being reported until the system fails to report that the area has become clear.

3.4.3.3. Falsely Clear Regions

A falsely clear region is an area or volume for which the GT reports a person but the SUT does not.

3.4.3.4. Falsely Occupied Regions

A falsely occupied region is an area or volume that the SUT reports as containing a person but the GT reports as being clear.

3.4.3.5. Mean Time To Tracking Failure (MTTF).

This measure gives the confidence level of the expected time until one of the following tracking failure events occurs.

- Losing track of a person
- Swapping identifiers for two different people
- Creating a new unnecessary identifier for a person already being tracked
- Treating a group of two or more people as a single person with a single identifier
- Treating a single person as two or more people

3.4.4. Human Identity Tracking

The SUT and GT systems provide an identifier for each person that should remain constant and unique to that person even if he/she leaves the environment for extended periods and returns.

3.4.5. Probability/Confidence

The system may provide a probability or confidence value that each tracked person is in fact a human. The system may also provide a polygon, polyhedrons, or grids or matrices

to describe 2D or 3D areas within the protected area where sensors will be unable to detect people due to either temporary or permanent occlusions.

4. Conclusion

In this report we have presented performance evaluation metrics for object detection and tracking in manufacturing and safety applications. In particular we have discussed three types of performance evaluation metrics based on detection, tracking, and perimeter intrusion.

Currently, we have implemented some of these performance metrics and test procedures for the evaluation of human detection and tracking for a robot safety evaluation. The results will provide scientific foundations for development of new standards that enable the use of perception systems in manufacturing applications.

5. Bibliography

[1] Ogale, N. A., "A survey of techniques for human detection from video," University of Maryland Technical Report, 2006.

[2] Sage, K. H., Nilski, A.J., and Sillett, I. M. "Latest Developments in the iLids Performance Standard: from Multiple Standard Camera Views to New Imaging Modalities," SPIE Vol. 74860F, 2009.

[3] CLEAR, Classification of events, activities and relationships—evaluation campaign and workshop. http://www.clear-evaluation.org/, Feb. 2013.

[4] i-LIDS, "Image library for intelligent detection systems," http://scienceandresearch.homeoffice.gov.uk/hosdb2/ physical-security/detection-systems/i-lids/.

[5] CREDS, "Call for real-time event detection solutions (CREDS) for enhanced security and safety in public transportation," http://www.visiowave.com/pdf/ISAProgram/CREDS.pdf.

[6] CAVIAR, "Context aware vision using image-based active recognition," http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[7] ETISEO, "Video understanding evaluation," http://www-sop.inria.fr/orion/ETISEO/.

[8] IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), http://pets2007.net/.

[9] VACE, "Video analysis and content extraction," http://www.perceptual-vision.com/vt4ns/vace brochure.pdf.

[10] Bashir, F. and Porikli, F., "Performance evaluation of object detection and tracking systems," in Proceedings of the 9th IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS 06), New York, NY, USA, June 2006.

[11] Yin, F., Makris, D., and Velastin.S. A., "Performance evaluation of object tracking algorithms." 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS2007), Rio de Janeiro, Brazil. 2007.

[12] Brown, L. M., Senior, A. W., Tian,Y., Connell, J., Hampapur, A., Shu, C., Merkl, H. and Lu, M., "Performance Evaluation of Surveillance Systems Under Varying Conditions", IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance, Colorado, Jan 2005.

[13] Lazarevic-McManus, N., Renno, J.R., Makris, D. and Jones, G.A., "An Object-based Comparative Methodology for Motion Detection based on the F-Measure", in 'Computer Vision and Image Understanding', Special Issue on Intelligent Visual Surveillance, 2007.

[14] Stiefelhagen, R., Bernardin, K., Ekenel, H.K., and Voit, M., "Tracking Identities and Attention in Smart Environments - Contributions and Progress in the CHIL Project." Eighth IEEE Int'l Conference on Face and Gesture, Amsterdam, 2008.

[15] Bodt, B., Camden, R., Scott, H., Jacoff, A., Hong, T., Chang, T., Norcross, R., Downs, A., Virts, A., "Performance Measurements for Evaluating Static and Dynamic Multiple Human Detection and Tracking Systems in Unstructured Environments," PerMIS09, September 21-23, 2009, Gaithersburg, MD, USA. ACM 978-1-60558-747-9/09/09

[16] Davis, J. and Goadrich, M. "The relationship between Precision-Recall and ROC curves", Proceeding ICML '06 Proceedings of the 23rd international conference on Machine learning, Pages 233-240, 2006.

[17] Kalal , Z., Matas, J., and Mikolajczyk, K., "Online learning of robust object detectors during unstable tracking", The 3rd On-line Learning for Computer Vision Workshop Kyoto, Japan 2009.

[18] Popoola, J. and Amer, A., "Performance Evaluation for Tracking Algorithms Using Object Labels", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2008.

[19] Bernardin, K., Elbs, A., and Stiefelhagen, R., "Multiple Object Tracking Performance Metrics and Evaluation in a Smart Room Environment", The Sixth IEEE International Workshop on Visual Surveillance, VS 2006, Graz, Austria, 01. May 2006.

[20] Moeslund, T. B. and Granum, E., "A Survey of Computer Vision-Based Human Motion Capture", Computer Vision and Image Understanding, March, 2001.

[21] Black, J., Ellis, T. and Rosin, P., "A Novel Method for Video Tracking Performance Evaluation", The Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, October, Nice, France, pp. 125-132. (2003).

[22] Nascimento, J.C., and Marques, J.S., "Performance evaluation of object detection algorithms for video surveillance", Multimedia, IEEE Transactions on 8.4 (2006): 761-774.

[23] Axel, B., Marco, B., Julia, E., Matthias, K., Hartmut S, L., Marcel, M., and Jie, Y. (2008). A review and comparison of measures for automatic video surveillance systems. *EURASIP Journal on Image and Video Processing*, *2008*.

[24] Godil A., Bostelman R., Saidi K., Shackleford W., Cheok G., Shneier M., and Hong T., 3D Ground-Truth Systems for Object/Human Recognition and Tracking, CVPR'13 workshop, June 2013

[25] Godil A., Eastman R., Hong T., Ground Truth Systems for Object Recognition and Tracking, NISTIR 7923, April 2013.