SLOVENSKI STANDARD
oSIST prEN 18281:2026

01-maj-2026

Umetna inteligenca - Metode vrednotenja za natančne sisteme računalniškega vida

Artificial Intelligence - Evaluation methods for accurate computer vision systems

KI-Aufgaben und Bewertungsmethoden für Computer-Vision-Systeme

Ta slovenski standard je istoveten z: **prEN 18281**

ICS:

35.020	Informacijska tehnika in tehnologija na splošno	Information technology (IT) in general
35.240.01	Uporabniške rešitve informacijske tehnike in tehnologije na splošno	Application of information technology in general

oSIST prEN 18281:2026 **en,fr,de**

EUROPEAN STANDARD

NORME EUROPÉENNE

EUROPÄISCHE NORM

DRAFT
prEN 18281

March 2026

ICS 35.240.01

English version

Artificial Intelligence - Evaluation methods for accurate computer vision systems

KI-Aufgaben und Bewertungsmethoden für Computer-Vision-Systeme

This draft European Standard is submitted to CEN members for enquiry. It has been drawn up by the Technical Committee CEN/CLC/JTC 21.

If this draft becomes a European Standard, CEN and CENELEC members are bound to comply with the CEN/CENELEC Internal Regulations which stipulate the conditions for giving this European Standard the status of a national standard without any alteration.

This draft European Standard was established by CEN and CENELEC in three official versions (English, French, German). A version in any other language made by translation under the responsibility of a CEN and CENELEC member into its own language and notified to the CEN-CENELEC Management Centre has the same status as the official versions.

CEN and CENELEC members are the national standards bodies and national electrotechnical committees of Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Republic of North Macedonia, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Türkiye and United Kingdom.

Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

Warning : This document is not a European Standard. It is distributed for review and comments. It is subject to change without notice and shall not be referred to as a European Standard.

CEN-CENELEC Management Centre:
Rue de la Science 23, B-1040 Brussels

Ref. No. prEN 18281:2026 E

Contents

Page

European foreword 4

1 Scope 5

2 Normative references 5

3 Terms and definitions 5

4 Abbreviated terms 5

5 Overview 5
5.1 General 6
5.2 Purpose of evaluation 7
5.3 Types of evaluation methods 7
5.4 Basic evaluation concepts 7
5.4.1 General 7
5.4.2 Intrinsic vs. extrinsic evaluation 7
5.4.3 Component vs. system evaluation 8
5.4.4 Quantitative vs. qualitative methods 8

6 Definition of evaluation methods 8
6.1 General 8
6.2 Accuracy, precision, recall, F1 8
6.3 Absolute Relative Error (AbsRel or REL) 8
6.4 Absolute Trajectory Error (ATE) 9
6.5 Association Accuracy (AssA) 9
6.6 Average Precision (AP) 10
6.7 Chamfer Distance 10
6.8 Detection Accuracy (DetA) 10
6.9 Dice Similarity Coefficient 11
6.10 Expected Calibration Error (ECE) 11
6.11 Frechet Inception Distance (FID) 11
6.12 Hausdorff Distance (HD) 12
6.13 Higher Order Tracking Accuracy (HOTA) 12
6.14 Identification F1 (IDF1) 13
6.15 Inception Score (IS) 14
6.16 Intersection over Union (IoU) 14
6.17 Learned Perceptual Image Patch Similarity (LPIPS) 16
6.18 Mean Average Precision (mAP) 16
6.19 Mean Intersection over Union (mIoU) 16
6.20 Mean Per Joint Position Error (MPJPE) 17
6.21 Multiple Object Tracking Accuracy (MOTA) 17
6.22 Mutual Information and Normalized Mutual Information 18
6.23 Panoptic Quality (PQ) 20
6.24 Peak Signal-to-Noise Ratio (PSNR) 20
6.25 Percentage of Correct Keypoints (PCK) 21
6.26 Percentage of Correct Parts (PCP) 21
6.27 Pixel Accuracy (PA) 21
6.28 Relative Pose Error (RPE) 22
6.29 Root Mean Square Error (RMSE) 22
6.30 Scale-Invariant Logarithmic Loss (SILog) 23
6.31 Structural Similarity Index Measure (SSIM) 23
6.32 Surface Coverage (SC) 24
6.33 Target Registration Error (TRE) 24
6.34 Temporal Intersection over Union (tIoU) 25
6.35 Hausdorff Distance 95 Percentile 26
6.36 Average Symmetric Surface Distance 26
6.37 Mean Average Surface Distance 26

6.38 Mean Squared Error (MSE) 26
6.39 Mean Absolute Error (MAE) 27
6.40 Pearson Correlation Coefficient (PCC) 27
7 Evaluation methods per task 27
7.1 General 27
7.2 Image analysis 28
7.2.1 General 28
7.2.2 Object localization 29
7.2.3 Object classification 29
7.2.4 Object detection 30
7.2.5 Semantic segmentation 32
7.2.6 Instance segmentation 33
7.2.7 Panoptic Segmentation 34
7.2.8 Image registration 35
7.2.9 Image similarity 39
7.3 Spatial analysis 40
7.3.1 General 40
7.3.2 Depth estimation 40
7.3.3 Object / 3D reconstruction 40
7.3.4 Object pose estimation 43
7.4 Temporal analysis 44
7.4.1 General 44
7.4.2 Multiple object tracking 44
7.4.3 Action recognition 48
7.4.4 Event detection 51
7.5 Tasks producing image as an output 52
7.5.1 General 52
7.5.2 Image generation 52
7.6 Tasks producing video as an output 53
8 Requirements on resources 53
Annex A (informative) Regularity Measures for Image Transformations 54
Bibliography 56

European foreword

This document (prEN 18281:2026) has been prepared by Technical Committee CEN/CENELEC JTC 21"Artificial Intelligence", the secretariat of which is held by DS.

This document is currently submitted to the CEN Enquiry.

This document has been prepared under a standardization request addressed to CEN by the European Commission. The Standing Committee of the EFTA States subsequently approves these requests for its Member States.

1 Scope

This document specifies the evaluation of computer vision systems, in the sense of measuring the quality of a system's results to assess its functional suitability. It provides a definition of evaluation methods for those systems, together with guidance on how to select, implement and interpret those evaluation methods. This document covers quantitative metrics as well as other evaluation methods. It includes requirements on the implementation of the described metrics, and further requirements on the technical resources involved in the evaluation process.

The accuracy and functional correctness metrics defined in this document provide point-in-time measurements under defined test conditions. These metrics can also be used as inputs to broader assessments of system robustness, as defined in prEN 18229-2. However, these metrics do not, on their own, guarantee stable performance across all deployment contexts or operating conditions. The selection, weighting, and interpretation of metrics are therefore also informed by the intended deployment context and the potential consequences of error.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC TS 4213:2022, *Information technology — Artificial intelligence — Assessment of machine learning classification performance*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— IEC Electropedia: available at http://www.electropedia.org/

— ISO Online browsing platform: available at http://www.iso.org/obp

3.1
computer vision
artificial vision capability of a functional unit to acquire, process, and interpret visual data

Note 1 to entry: Computer vision involves the use of visual sensors to create an electronic or digital image of a visual scene.

Note 2 to entry: Not to be confused with machine vision.

Note 3 to entry: Computer vision; artificial vision: terms and definition standardized by ISO/IEC [ISO/IEC 2382-28:1995].

[SOURCE: ISO/IEC 2382:2015 [1]]

4 Abbreviated terms

AI **Artificial Intelligence**

GT **Ground Truth**

5 Overview

5.1 General

Compared to evaluation practices for other AI systems, computer vision evaluation has its own specificities. While the main principles and caveats remain, the concrete practices, and in particular the evaluation methods themselves, can be substantially affected in computer vision, both due to the inherent constraints and variability associated with processing visual content (images, videos…) and to existing practices by computer vision practitioners.

This document approaches the specificities of computer vision evaluation from the angle of tasks. The task is a suitable way to formalize the processing done by the AI system, at a level of abstraction that enables to capture the methodological specificities of certain types of processing, while not delving into the specific considerations of the use case and the exact data that is being processed.

EXAMPLE AI systems performing a classification task can be evaluated through metrics like precision, recall or F1, as specified in ISO/IEC 4213. This is, however, not applicable to image inpainting or image super-resolution and different evaluation methods are used for such systems. Object detection cases fall somewhere in between, as metrics like precision, recall, and F1 score are still relevant but are applied differently compared to classification. It is important to clarify this by outlining the process for matching predicted bounding boxes with ground truth boxes before calculating the F1 score.

This document is therefore to be read and used in conjunction with prEN 18288. The description of the various tasks addressed in this document can be found in prEN 18288, and this document focuses on which evaluation methods can be applied to each task and how.

This document is complementary to ISO/IEC 4213 which addresses evaluation methods for classification, regression, clustering and recommendation, and to EN ISO/IEC 23282 which cover the NLP evaluation methods in a similar way. Evaluation methods from ISO/IEC 4213 are built upon by some evaluation methods in this document, and evaluation methods from EN ISO/IEC 23282 are typically combined with the present ones, either in the context of multi-modal processing (for which evaluating the processing on both aspects is warranted) or due to the joint presence of computer vision components and NLP components within the same AI system.

Clause 6 describes a set of evaluation methods used in computer vision, including their mathematical definition and an indication of any aspect of the metric computation for which reporting is needed to ensure comparability and reproducibility across different implementations of the same metric. Only automated evaluation methods are handled in this document.

Clause 7 then relates each computer vision task with these evaluation methods, providing task-specific considerations for the methodology to apply these to the particular task, complemented with further specifications (both in terms of application and practical implementation) meant to ensure the reproducibility of the evaluation in the context of that task, as well as additional informative material to calibrate the typical ranges of values obtained by this evaluation method and also indicating the aspects of the task evaluated by that method in order to guide the interpretation of results.

Depending on the task and the evaluation method, the provided specifications differ:

— For tasks and evaluation methods that are a direct application of practices not specific to computer vision (e.g. usage of F1 for object classification), this document serves as an entry point to identify the relevant applicable standards for evaluating these tasks.

— For tasks and evaluation methods that are a computer vision-specific application of practices that exist in another form out of computer vision (e.g. usage of F1 for object detection), this document builds on existing standards to complement their specifications with the additional ones that apply in the case of computer vision (e.g. procedure for matching bounding boxes).

— For tasks that have their own computer vision-specific evaluation methods, this document provides a full specification of the evaluation methods as well as their usage for these tasks.

Finally, Clause 8 complements the specifications on the evaluation methods themselves and their application to tasks, with further specifications on the technical resources needed for applying

evaluation methods to computer vision systems. This includes, for instance, the consideration of the test data size or diversity.

5.2 Purpose of evaluation

The primary objective of computer vision model evaluation is to quantify their performance reproducibly, and to ensure that the systems operate as intended in real-world scenarios.

There are no <<one size fits all>> metrics to evaluate AI computer vision algorithm and the right metric depends on the task's purpose and what constitutes "success". For example,

— For detection, success means finding *what* and *where* things are.

— For generation, success means producing *realistic* and *diverse* visual content.

Using a single universal metric would therefore give misleading or even meaningless results. Therefore having a clear understanding of a computer vision accuracy metrics and its pitfalls is key to select the appropriate metrics to measure the performance of a model.

The reliability of evaluation results depends on the characteristics of the evaluation dataset. Evaluation datasets should be statistically representative of real-world operating conditions, exhibit appropriate class and condition balance, and employ consistent and validated labelling and annotation techniques. Further information on datasets can be found in prEN 18284.

5.3 Types of evaluation methods

The evaluation of computer vision models relies on a diverse set of methods and metrics, the choice of which is intrinsically linked to the nature of the task to be performed (classification, detection, segmentation, etc.) and to the specific characteristics of the problem (class imbalance, boundary importance, etc.).

Three main categories of evaluation methods can be distinguished:

a) **Ground-Truth–Based Evaluation.** This category applies to tasks where an objective, well-defined reference exists and where the system output can be directly compared to a labelled dataset. Typical examples include object detection, image segmentation or pose estimation. Ground-Truth–Based evaluations require precise alignment of spatial coordinates and label consistency.

b) **Generative Evaluation.** This category applies to tasks that do not have a single correct answer, such as image generation, style transfer, super-resolution, or inpainting. Since the output space is inherently multimodal, no unique ground truth can be defined. Generative Evaluation-Based metrics estimate the alignment between the distribution of generated images and that of real-world samples rather than pixel-wise correctness.

c) **Semantic Evaluation.** This category applies to tasks that involve high-level reasoning or multimodal understanding, such as image captioning, visual question answering (VQA), or scene graph generation. Evaluation depends on the semantic equivalence between the predicted description and the reference. Multiple acceptable outputs can exist, and evaluation needs to account for linguistic variability, contextual relevance, and factual consistency.

5.4 Basic evaluation concepts

5.4.1 General

The evaluation of computer vision tasks is a rich field that relies on several fundamental concepts enabling the objective assessment of a model's quality and performance for a given application.

5.4.2 Intrinsic vs. extrinsic evaluation

Intrinsic evaluation measures a computer vision model's performance using predefined metrics (e.g. accuracy) on labelled data, focusing on how well it performs the task itself.

Extrinsic evaluation assesses the model's impact on an application (e.g. improved safety in autonomous driving), focusing on real-world usefulness rather than internal performance.

5.4.3 Component vs. system evaluation

Component evaluation measures the performance of an individual computer vision module (e.g. object detection accuracy or segmentation quality).

System evaluation assesses the end-to-end performance of the complete application that integrates multiple components, focusing on overall functionality, reliability, and real-world outcomes.

5.4.4 Quantitative vs. qualitative methods

Quantitative methods use objective, numerical metrics (e.g. precision, recall) to measure a computer vision model's performance.

Qualitative methods rely on human judgment or visual inspection (e.g. perceptual realism, interpretability) to assess how convincing or meaningful the output of the computer vision algorithm is.

6 Definition of evaluation methods

6.1 General

Evaluation methods in computer vision encompass both quantitative and qualitative approaches, applied at the component or system level, and assessed through intrinsic or extrinsic evaluations.

They ensure that models are measured not only by numerical accuracy on benchmark data, but also by their overall performance.

Certain metrics are sensitive to class or instance imbalance in evaluation datasets. Where imbalance is present, users shall implement macro-averaging, Balanced Accuracy, weighted variants or other established mitigation approaches.

Implementations of metrics shall guard against numerical conditions that produce undefined or misleading results. These conditions include, but are not limited to, division by zero, logarithm of zero or negative values and computation over an empty valid sample set.

The computation of every quantitative metric shall be fully reproducible.

NOTE The evaluation methods in clause 6 are organized in alphabetical order to facilitate the reading. It is planned to reorganize the evaluation methods in a more meaningful manner in a further version of the document.

6.2 Accuracy, precision, recall, F1

For accuracy, ISO/IEC 4213:—, 6.2.3 applies.

For precision, ISO/IEC 4213:—, 6.2.4 applies.

For recall, ISO/IEC 4213:—, 6.2.4 applies.

For F1, ISO/IEC 4213:—, 6.2.5 applies.

For Jaccard Index, ISO/IEC 4213:—, 6.5.4 applies.

6.3 Absolute Relative Error (AbsRel or REL)

Absolute Relative Error normalizes the error for each pixel compared to the ground truth. In contrast to RMSE (6.29) , this metric is less sensitive to large absolute errors at long range. Absolute Relative Error score shall be computed according to Formula (1):

$$AbsRel = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \quad (1)$$

where $y_i$is the ground truth value, $\hat{y}_i$is the predicted value at pixel i. Lower values of $AbsRel$rerpesent more accurate predictions. This metric is appropriate where the tolerance for variations is higher at far ranges than at close ranges.

6.4 Absolute Trajectory Error (ATE)

The metric reflects the correctness of an estimated trajectory relative to a reference trajectory. The metric is calculated by first aligning the estimated trajectory to the reference trajectory thereby minimizing the quadratic position error between corresponding trajectory points. The ATE is then defined as the quadratic mean of the Euclidean distances between the adjusted estimated positions and the reference positions in dependence of time and/or space.

Absolute Trajectory Error score shall be computed according to Formula (2):

$$ATE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}||(p_i - (R\hat{p}_i + t))||^2} \quad (2)$$

Where:

— N is the total number of trajectory points (e.g. camera poses, 2D or 3D positions).

— p_i is the Ground Truth position of the ith camera or object, expressed as a 2D or 3D vector.

— $\hat{p}_i$ i is the predicted position of the ith camera or object, before alignment.

— $R \in SO(n)$ is the rotation matrix, $SO(n)$ is the n-D rotation group, and $t \in \mathbb{R}^n$ is the n-D translation vector, where n ∈ [2,3], both obtained from rigid-body alignment (e.g. via Horn's method or Umeyama algorithm). These align the predicted trajectory to the ground truth to remove global drift.

6.5 Association Accuracy (AssA)

Association Accuracy (AssA) focuses on assessing how well an object tracking system correctly associates the same identities, given the known associations and identities of a sequence. This is done by comparing the alignment between the detected objects and their associated track to the matching ground-truth tracks.

AssA score shall be computed according to the following Formula (3):

$$AssA = \frac{1}{|TP|}\sum_{c \in TP}\frac{|TPA(c)|}{|TPA(c)| + |FNA(c)| + |FPA(c)|} \quad (3)$$

Where TP the set of matched ground-truth and predicted trajectories (denoted as c) , TPA denotes objects correctly associated with their respective ground-truth trajectory, FNA denotes ground-truth tracks that exist but fail to maintain the correct identity, and FPA denotes trajectories where a predicted trajectory maintains the wrong ID.

AssA focuses on evaluating the accuracy of object-to-track associations across the entire sequence. It measures how well a tracker preserves identity links over time, accounting for both switches and fragmentation. AssA' sensitivity to track fragmentation makes it particularly applicable in scenarios where maintaining consistent trajectories is crucial. By emphasizing long-term association quality rather than frame-level detection, AssA complements detection-focused metrics and provides a more

comprehensive evaluation of a system's ability to maintain consistent object identities across time, especially in scenarios involving occlusions.

6.6 Average Precision (AP)

The Average Precision (AP) is the weighted sum of precision values at multiple thresholds, with the weight for the n^{th} precision value being the increase in recall from the $(n-1)^{\text{th}}$ threshold to the n^{th} threshold. AP summarizes the Precision- Recall curve but is distinct from calculating the area under the Precision-Recall curve using e.g. the trapezoid rule.

Average precision is calculated as:

$$\text{AP} = \sum_{n} (R_n \quad - \quad R_{n-1}) \quad P_n \tag{4}$$

where R_n and P_n are the recall and precision at the n^{th} threshold respectively.

NOTE This metric could potentially be included in the new version of ISO/IEC 4213.

6.7 Chamfer Distance

Chamfer distance measures the average squared distance between the points in one point cloud to their nearest neighbours in the other point cloud, and vice versa. It is calculated using the following Formula (5):

$$CD\left(A, B\right) = \frac{1}{|A|} \sum_{a \in A} \min_{b \in B} \| a - b \|^2 + \frac{1}{|B|} \sum_{b \in B} \min_{a \in A} \| a - b \|^2 \tag{5}$$

where A and B are two sets of points with their cardinalities $|A|$ and $|B|$, respectively.

Parameters:

— A: Set of points from the reconstructed model.

— B: Set of points from the reference model.

— $|A|$, $|B|$: Number of points in A and B, respectively.

Calculation:

— For each point in A, find the closest point in B and compute .

— For each point in B, find the closest point in A and compute .

— Chamfer Distance shall be computed according to the formula given in Formula (5).

6.8 Detection Accuracy (DetA)

Detection Accuracy (DetA) focuses on evaluating how well the tracking system localizes objects, regardless of their identity assignment. This is done by comparing the detected objects to all of the known ground-truth objects.

Detection Accuracy score shall be computed according to the Formula (6):

$$DetA_{\alpha} = \frac{|TP|}{|TP| + |FN| + |FP|} \tag{6}$$

Where α denotes the IoU (6.16) matching threshold, TP denotes the number of detections that match the ground truth, FN denotes the number objects that are present in the ground-truth but were missed, and FP denotes the number of objects that were detected but do not match any ground-truth object.

The IoU threshold α determines whether a predicted bounding box has sufficient overlap with a given ground truth bounding box to be considered a match. Therefore, the selected threshold (typically $\alpha = 0.5$), determines how accurately the predicted boundingbox fits to the object.

DetA is particularly useful for scenarios where reliable object localization is crucial (e.g. detecting all pedestrians in autonomous driving). By isolating detection accuracy from association performance, DetA provides a complementary perspective on tracker performance.

6.9 Dice Similarity Coefficient

The Dice Similarity Coefficient (DSC), also known as the Dice-Sørensen Coefficient or SørensenDice Index is an overlap metric similar to the Jaccard Index. Given two sets A and B, DSC is defined as:

$$DSC = \frac{2 |A \cap B|}{|A| + |B|} \tag{7}$$

6.10 Expected Calibration Error (ECE)

Expected Calibration Error (ECE) is a quantitative metric that measures the difference between a model's confidence and its accuracy. It is calculated by dividing the model's predictions into a number of equally-spaced bins based on confidence and then computing a weighted average of the absolute difference between the accuracy and confidence for each bin.

The **ECE** score shall be computed according to Formula (8):

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{8}$$

Where:

— M is the number of bins.

— $|B_m|$ is the number of predictions in bin m.

— n is the total number of data points.

— $acc(B_m)$ is the accuracy in bin m.

— $conf(B_m)$ is the average confidence in bin m.

NOTE The number of bins M used to calculate this metric shall be stated when using this metric.

6.11 Frechet Inception Distance (FID)

Frechet Inception Distance (FID) is a metric used to assess the difference between two distributions. It is commonly used for several tasks such as image generation quality assessment and domain gap evaluation. It is a combination of Frechet distance and Inception V3 model [2]

The FID shall be computed according to Formula (9):

$$FID = \| \mu_1 - \mu_2 \| + Tr\left(\Sigma_1 + \Sigma_2 - (\Sigma_1 \Sigma_2)^{1/2}\right) \tag{9}$$

Where μ_1 and Σ_1 are the mean and covariance matrix of the first feature multivariate normal distribution and μ_2 and Σ_2 are the mean vector and the covariance matrix of the second feature multivariate normal distribution.

The distance is divided into two steps:

a) Feature extractor based on pre-trained models, often Inception V3 is used, where for each image a vector of features is extracted form its layers. This extraction occurs generally the 3rd pool layer.

b) Fréchet Distance computes the metric based on extracted features distribution in (a) following Formula (9).

NOTE 1 FID metric is usually used in case of image input for image generation model evaluation. Other variant of audio and video inputs for video and audio generation models are also developed under FVD (Fréchet Video Distance) and FAD (Fréchet Audio Distance) names, respectively. This is an adaption of FID by modifying only the model extractor for feature given in (a).

FID is usually useful for image generation evaluation, during training of e.g. models with GAN architectures and monitoring such as domain gap assessment.

The interpretation of FID score between two distributions:

— A lower FID score indicates a higher degree of similarity between the two distibutions. In case of FID=0, then the two distributions are identical.

— As the discrepancy between the two distributions distributions increases the FID score also increases.

NOTE 2 In case of generative models, such as GAN, the two distributions are extracted from real and generated images, respectively.

NOTE 3 In case of domain gap assessment, the two distributions are extracted from source (usually training dataset) and target dataset (usually operational dataset), respectively

6.12 Hausdorff Distance (HD)

The Hausdorff Distance (HD) $d_H(X, Y)$ is defined over two non-empty subsets X and Y of a metric space (M, d), where M is a set and d is a metric on M. HD is defined as:

$$d_H(X,Y) \quad = \quad \max\left(\sup_{x \in X} d(x,Y),\ \sup_{y \in Y} d(y,X)\right) \tag{10}$$

where sup is the supremum operator (also known as the least upper bound), and the distance $d(x, Y)$ between element $x \in X$ and set Y is defined as :

$$d(x,Y) \quad = \quad \inf_{y \in Y} \ d(x,y) \tag{11}$$

where inf is the infimum operator (also known as greatest lower bound). Distance $d(y, X)$ is defined similarly to $d(x, Y)$.

NOTE 1 HD is symmetric, i.e $d_H(X,Y) \quad = \quad d_H(Y,X)$

NOTE 2 When applied to the points, pixels, or voxels making up the boundaries of regions, HD is also known as Maximum Symmetric Surface Distance (MSSD)

If the HD is a given value then every point in each set is within that distance of some point of the other set. Thus HD is sensitive to outliers, but is useful for giving a worst-case measure of how far apart two sets are.

6.13 Higher Order Tracking Accuracy (HOTA)

The Higher Order Tracking Accuracy (HOTA) quantifies the overall performance of an object tracking system by assessing the accurate position of many objects among successive video frames. It provides a quantitative measure that combines X key types of errors into a single figure:

a) Detection Errors: ratio of objects present in the ground truth that were not detected (recall) and the rate of detection reflecting the objects in the ground truth (precision).

b) Association Errors: ratio of correctly associated objects throughout the entire input sequence compared to total associations of the entire input sequence.

HOTA score shall be computed according to Formula (12):

$$HOTA_\alpha = \sqrt{\frac{\sum_{c \in \{TP\}} A(c)}{|TP| + |FN| + |FP|}} \tag{12}$$

Where:

— α denotes the localization threshold for accepting detection as a true positive,

— $|TP|$, $|FN|$ and $|FP|$ denote the sum of objects that are correctly detected, incorrectly detected and missed respectively.

— $A(c) = \frac{|TPA(c)|}{|TPA(c)| + |FNA(c)| + |FPA(c)|}$

$$A(c) = \frac{|TPA(c)|}{|TPA(c)| + |FNA(c)| + |FPA(c)|} \tag{13}$$

Where:

— c denotes a given object id,

— $|TPA(c|$, $|FNA(c)|$ and $|FPA(c)|$ denote correctly associated, incorrectly associated and missed associations for all instances of a given object.

NOTE This is computed for the entire input sequence.

HOTA leverages the Hungarian algorithm to perform matching which requires a manually selected threshold for the maximum allowed cost to accept a given association. The value of this threshold shall be stated when reporting this metric.

6.14 Identification F1 (IDF1)

Identification F1 (IDF1) focuses on assessing how long a given tracker is correctly identifying an object. Rather than just accumulating frame-by-frame error it calculates a joint metric for all frames in one go. Thus it heavily focuses on consistent identity association throughout. Thus, this metric suffers less from the number of objects in a given video. This is done by employing Identity Precision (IDP) and Identity Recall (IDR), where IDP computes how many of the detections are associated correctly, and IDR computes how many detections are correctly associated concerning all relevant objects.

As it is identity-focused the two types of errors in the tracking systems that IDF1 measures are:

a) Improper associations: Objects that are detected and associated in a manner that does not match the ground truth.

b) Missed trajectories: Trajectories that have not been tracked but are present in the ground truth.

IDF1 score shall be computed according to Formula (14):

$$IDF1 = \frac{|IDTP|}{|IDTP| + 0.5|IDFN| + 0.5|IDFP|} \tag{14}$$

where,

— $IDTP$ denotes objects that are correctly detected and associated,

— $|IDFP|$ denotes objects that have been associated but are not present in the ground truth,

— $|IDFN|$ IDFN denotes objects that are present in the ground truth but have not been detected and associated.

Formulas of IDP and IDR are given below: