Neovision2 Performance Evaluation Protocol

Version 3.0

4/16/2012 – Public Release

Prepared by

Rajmadhan Ekambaram
rajmadhan@mail.usf.edu
Dmitry Goldgof, Ph.D.
goldgof@cse.usf.edu
Rangachar Kasturi, Ph.D.
r1k@cse.usf.edu

Department of Computer Science and Engineering
University of South Florida

and

Yang Ran, Ph.D.
yran@setcorp.com
Qinfen Zheng, Ph.D.
qzheng@setcorp.com

SET Corporation

1 Introduction

The DARPA/DSO's Neovision2 program aims to develop neuromorphic vision systems, that is, artificial vision systems based on the design principles employed by mammalian vision systems. These systems' performance is measured using a set of annotated video clips. This document describes the scope of the performance evaluation including the metrics and methodologies.

2 Scope

The evaluation dataset is derived from data sources (domains) which are labeled as Tower and Helicopter. Since the video source characteristics and perhaps the corresponding Neovision2 systems are different for these domains, the performance of the systems are evaluated and reported separately for each of these domains. There are ten object classes of interest in this program. The methodology in this effort is based on the VACE (Video Analysis for Content Extraction) performance evaluation methodology described in [1]. For a description of the object classes, their annotations, and ground truth file format refer to the annotation guidelines [2]. The performance of the system in detecting each of these object classes is evaluated independent of other classes using Normalized Multiple Object Thresholded Detection Accuracy (NMOTDA, defined in Section 3), missed detects and false positives; i.e., the evaluations will be run one at a time for each object class.

2.1 Datasets

This section describes the dataset used to support the NeoVision2 evaluations. The dataset is divided into three sets: Micro-corpus, Training, and Evaluation (Test) as shown in Table 1. The micro-corpus is used as seed for initial experiments and to provide participants with a concrete sampling of the datasets and the tasks. The numbers of sequences and times shown are upper bounds and the actual annotated numbers will be dependent on resource availability and annotation complexity. The data is distributed to participants as individual frame images in PNG format.

DATA	NUMBER OF SEQUENCES	TOTAL MINUTES	AVERAGE SECONDS PER SEQUENCE
MICRO-CORPUS	6	2	15-20
TRAINING	125	45	15-20
EVALUATION	125	45	15-20

Table 1 NeoVision2 corpus partitioning for each task.

The presence and average sizes for all ten object classes are shown in Table 2. Note that not all object classes are present in each domain.

TASK	DOMAIN	
	TOWER	HELICOPTER
	Object Present?	Object Present?
Car	✓	✓
Truck	✓	✓
Tractor-Trailer		✓
Bus	✓	✓
Container		✓
Boat		✓
Plane		✓
Helicopter		✓
Person	✓	✓
Cyclist	✓	✓

Table 2: Object Classes and their presence in the imagery domains.

2.2 Permitted Site Information

The following additional information will be supplied to the Neovision2 teams:

1. Object categories in each domain (as described in Table 2).
2. A range of Ground Sampling Distances, measured in the center of the frame, for the datasets as a whole (as described in Table 3)

	DOMAIN	
	TOWER	HELICOPTER
GSD Range (Pixels/meter)	30-40	25-40

Table 3: Range of Ground Sampling Distance in the imagery domains.

3 Performance Measures

This section describes the performance measures where the object definitions are for an object with a bounding box. A NeoVision2 object is denoted by a class ID and a bounding box (4 corners in (x, y) pixels in a frame, where (0,0) is the top left corner of the frame.). The following are the notations used in the remainder of the document:

- $G_i^{(t)}$ denotes the ith ground-truth object in frame t.

- $D_i^{(t)}$ denotes the ith detected object in frame t.
- $N_G^{(t)}$ and $N_D^{(t)}$ denote the number of ground-truth objects and the number of detected objects in frame t, respectively.
- N_{frames} is the number of frames in the video sequence.

3.1 Performance Measure per Object Pair

We use the spatial overlap between a pair of ground-truth object and a system output object for detection evaluation. For a pair of ($G_i^{(t)}$, $D_i^{(t)}$), the overlap ratio is calculated as:

$$\text{Overlap\_Ratio} = \frac{\left| G_i^{(t)} \bigcap D_i^{(t)} \right|}{\left| G_i^{(t)} \bigcup D_i^{(t)} \right|} \tag{1}$$

In the scope of NeoVision2, an object instance is considered as detected by its matched pair if the Overlap_Ratio of the spatial intersection between the ground truth and system output boxes to their union (in area) as defined in Eq. (1) is equal to or greater than a preset value of 0.2. In the scope of NeoVision2, an oriented box from annotation will be converted to a corresponding vertical box (minimal vertical rectangular envelope) before compared to system output using Equation 1.

3.2 Performance Measure for Sequence

Performance measure for a sequence is obtained by matching the ground truth and performer's output at each frame from that sequence. In each frame there can be multiple instances of the same object class, for example N cars represented by N ground truth boxes. A Neovision2 car detection system may output M boxes for this image. Ideally M is equal to N but in general it may be less or more than N. We must first assign the system output boxes to the corresponding ground truth boxes in a one-to-one fashion. When N=M there are N! possible combination of such assignments. Our assignment algorithm finds the combination that maximizes the detection performance measure described below.

To find the best match for each class and each frame, we will use the Hungarian algorithm [3], which is a numerical search algorithm that guarantees arriving at one optimal solution. The basic algorithm has a series of steps, which are followed iteratively, and has a polynomial time complexity; specifically, some implementations are $O(N^3)$. Faster implementations have been known to exist and, to the best of our knowledge, the current best bound is $O(N^2 \log N + NM)$ [3]. There are many variations of the basic strategy, most of which exploit constraints from the specific problem domains they consider. In our case, since the spatial overlap matrix between the ground truth objects and the system objects is expected to be sparse, we use a hash function for mapping subinputs from the whole set of inputs. This algorithm has been used by NIST for similar evaluations.

For a pair of ground truth box and system output box (G_i, D_j), we assign a score of 1 (denotes object detection) when the Overlap_Ratio is greater than or equal to 0.2; if it is less than the threshold we have a Missed Detect AND a False Positive. In addition we may have more system output boxes than ground truth objects (when M>N) or more ground truth objects than system output boxes (when M<N) resulting is unmatched boxes. Thus each ground truth box that has no matching pair at all or a match with a spatial overlap of less than 20% is a Missed Detect. Each system output box which has no matching pair at all or a match with a spatial overlap of less than 20% is a False Positive.

Aggregating these over all images in the entire sequence, we get the Normalized Multiple Object Thresholded Detection Accuracy, NMOTDA given by

$$NMOTDA = 1 - \frac{\sum_{t=1}^{N_{frames}} (c_m m^{(t)} + c_f f^{(t)})}{\sum_{t=1}^{N_{frames}} N_G^{(t)}} \quad (2)$$

where c_m and c_f are the cost functions for missed detects and false positives, $m^{(t)}$, $f^{(t)}$ and $N_G^{(t)}$ are the number of missed detects, false positives and ground truth objects in the t^{th} frame in the sequence. The summations are carried out over all evaluated frames. In Neovision2 evaluations, the cost functions c_m and c_f are set equal , that is, c_m=1 and c_f =1. This is a sequence-based measure which penalizes false detections, missed detections and spatial fragmentation. Note that maximizing NMOTDA for the sequence is the same as finding the optimal assignment of ground truth boxes to system output boxes at each frame image. NMOTDA can take a value from (-infinity) to (+1).

3.3 Performance Measure for Domain

In each data domain (Tower or Helicopter), multiple sequences will be used for evaluation. The NMOTDA score will be calculated using Equation 2 at the domain level by summing up all the individual frames from all sequences. Scores will be computed for each class present in that domain. Therefore, the evaluation software will generate a NMOTDA score for each data domain, each object class and each performer.

A "Detection Only" score will be generated using Equation 2 by ignoring the classification labels. All detections will be reassigned the same label for calculation. The evaluation software will generate a "Detection Only" NMOTDA score for each data domain and each performer. Before scoring, identical boxes will be merged into one. Overlapped boxes (if Overlap_Ratio is over 20%) will be merged into one and the union of them will be used instead.

To summarize, the result is 25 NMOTDA scores for each team – 22 for classification (using Table 2) plus 3 for detection.

3.4 Performance Measure for All Sequences and All Object Classes

DARPA has not made a final decision about how to compute a total score from the previously calculated NMOTDA scores. The calculation may use the scores for all object categories in one domain, one object category in all domains, or all object categories in all domains.

Candidate methods include the mean, median and weighted mean. For example, a weighted mean would be calculated as:

$$Score = \frac{\sum_{t=1}^{N} N^{(t)} * NMOTDA^{(t)}}{\sum_{t=1}^{N} N^{(t)}}$$

Where $N^{(t)}$ is the total number of groundtruth objects when calculating $NMOTDA^{(t)}$

3.5 Generating ROC Curves

The procedure described in the Section 3.2 generates a single operating point in the Receiver Operating Characteristic (ROC) curve for each object class and each domain. To generate additional points the system developers are asked to include a degree of confidence (a float number from 0.0 to 1.0, where 1.0 means almost sure) for each of their system output boxes. The evaluation software then thresholds these values at ten levels (95%, 85%, …, 5%) and retains only those boxes which are at or above the threshold; those boxes which are below the threshold are ignored. The performance measure (# of detections and # of false positives) is then computed for each set. Note that lower thresholds correspond to higher detection score with a corresponding increase in false positives and vice versa. We will use detection rate (# of total detections divided by # of total targets, in percentage) and average # of false positives per frame (in #/frame) as the x and y axis when rendering the ROC curves.

3.6 Scores Based on ROC Curves

The previous section described the process of generating a ROC curve for a given object class. This section describes how to create a score based on the generated ROC curves. Let Az represent the area under the ROC curve. A larger value of Az represents better performance and will be used. As in Section 3.4, DARPA will explore various combinations of these Az values, searching for results that are both intuitively meaningful and statistically sound.

3.7 Handling Limitations in the Ground Truth

Sometimes we want to exclude certain frames from evaluation because they contain frame-level events which place them outside of the scope of the task. To address this issue, **Don't Care Frames (DCFs)** will be established prior to scoring the test results using information in the reference annotations. Reasons for DCF include but not limited to blurred frames, corrupted frames etc. Frames which are designated as DCFs by annotators will be automatically ignored by the scoring procedure.

Likewise, sometimes we want to exclude certain annotated objects from the target object class because they contain attributes which place them outside the scope of the task. To address this issue, **Don't Care Objects (DCOs)** will be established prior to scoring the test results using information in the reference annotations. Objects in these DCOs will be effectively treated as not existing in both the reference and system output.

Where **DCOs** are used to annotate individual objects which can be spatially annotated but which can't be reliably identified, there may be groups of objects which are too crowded or too difficult to localize and cannot be bounded. To address this problem, **Don't Care Regions (DCRs)** are used to mark areas in frames which are to be eliminated entirely from the mapping and scoring process. Detected objects which fall inside a **DCR** or whose area is contained primarily within a **DCR** will be eliminated prior to the mapping/scoring process and will thus not generate false alarm errors. System detection boxes that have more than 20% (same number is used in calculating Overlap_Ratio) area inside a DCR will be eliminated from evaluation.

4 Output File Formats

The CSV (comma separated value) format is used for both the system output (same as reference annotations). Both the input and output files may contain appropriate tags required for evaluation.

Results will be requested to output in a series of csv (comma separated value) files. The csv files will be stored in a top level directory indicating the data source e.g., TowerF. Within this directory, there will be one csv file for each sequence, named after the input image sequence ID, e.g. 001.csv for sequence 001. Within each csv file there will be a single row for each detected object following the format specified in **Table 4**. At the beginning of each csv file, the first line shows the format info as a reference.

An example annotation file can be found in **Table 4**. Instructions for reading and generating CSV files will be provided along with the scoring software. Both oriented box and vertical box are supported in the evaluation software.

Frame,BoundingBox_X1,BoundingBox_Y1,BoundingBox_X2,BoundingBox_Y2,BoundingBox_X3,BoundingBox_Y3,BoundingBox_X4,BoundingBox_Y4,ObjectType,Occlusion,Ambiguous,Confidence,SiteInfo,Version
0,701,489,870,489,870,560,701,560,Car,FALSE,FALSE,1.0,,1.0
4,861,768,1051,768,1051,839,861,839,Car,TRUE,FALSE,1.0,,1.0
4,452,492,621,492,621,563,452,563,Car,FALSE,FALSE,0.5,,1.0
4,452,492,621,492,621,563,452,563,Truck,FALSE,FALSE,0.7,,1.0

Table 4: Example of system output file in csv format.

For multiple detections in the same location (one box with multiple labels), performers must record them into the result file using multiple entries (lines, or rows). The last two lines of **Table 4** shows an example: they have identical boxes but different labels and confidence values. Please note that for a single box with multiple labels, if one label is correct, that result will be awarded; if another label is incorrect, that result will be penalized.

5 Score Reporting

SET Corporation will provide scoring software with documentation. Support and bug tracking will also be provided via a dedicated page on the NeoVision2 project web site. In the Formative evaluation, both NTT and NeoVision2 performers will use the same software to score the output of the systems and compare the scores to make sure they are consistent. In the Summative evaluation, SET will run independent tests with the output of the systems and provide the scores to DARPA.

References

[1] Rangachar Kasturi, Dmitry Goldgof, Padmanabhan Soundararajan, Vasant Manohar, John Garofolo, Rachel Bowers, Matthew Boonstra, Valentina Korzhova, and Jing Zhang. Framework for Performance Evaluation of Face, Text, and Vehicle Detection and Tracking in Video: Data, Metrics, and Protocol," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.31, Feb. 2009.

[2] Matthew Parent, Rajeev Sharma, Yang Ran, Qinfen Zheng, *Neovision2 Annotation Guidelines,* 2011

[3] J.R. Munkres, “Algorithms for the Assignment and Transportation Problems,” J. SIAM, vol. 5, pp. 32-38, 1957.