# PETS vs. VACE Evaluation Programs: A Comparative Study

Vasant Manohar, Matthew Boonstra, Valentina Korzhova,
Padmanabhan Soundararajan, Dmitry Goldgof and Rangachar Kasturi
Computer Science & Engineering, University of South Florida, Tampa, FL
{**vmanohar, boonstra, korzhova, psoundar, goldgof, r1k**}**@cse.usf.edu**

Shubha Prasad and Harish Raju
Video Mining Inc., State College, PA
{**sprasad, hraju**}**@videomining.com**

Rachel Bowers and John Garofolo
National Institute of Standards and Technology, Gaithersburg, MD
{**rachel.bowers, john.garofolo**}**@nist.gov**

## Abstract

*There are numerous distributed efforts on performance evaluation of computer vision algorithms. Information exchange between programs can expedite achievement of the common goals of such projects. As a step towards this, this paper presents a qualitative comparison of the VACE and the PETS programs, the synergy of which, we believe, will have a significant impact on the progress of research in the vision community. The comparison is based on the vital aspects of an evaluation project – the framework, the tasks, performance metrics and ground truth data.*

## 1. Introduction

Performance evaluation of computer vision algorithms has received significant attention in the past few years with increasing demand and awareness to establish a global platform for benchmarking various algorithms against a common dataset, metrics, and evaluation methodology. This is substantiated by the need to identify specific aspects of a task that has further scope for improvement, quantify research progress and most important of all acquire diverse data that reflect the inherent problems to be confronted if the system were to be deployed in a real world scenario.

Object detection and tracking is an important and challenging topic in computer vision which addresses the problem of spatially locating an object of interest in the video. This is an important step in object identification since it is necessary to localize an object before recognition can be accomplished. Similarly, for a system to detect the occurence of an event, it has to first detect and further track the relevant objects involved in the event.

There are several current efforts towards the evaluation of object detection and tracking in video. The Video Analysis and Content Extraction (VACE) program, supported by Advanced Research and Development Activity (ARDA) is one such endeavor, established with the objective of developing novel algorithms and implementations for automatic video content extraction, multi-modal fusion and event understanding. The program has several phases and is currently nearing the end of Phase II. During VACE Phase I and Phase II, the program achieved significant progress in video content analysis; specifically in techniques for the automated detection and tracking of scene objects such as faces, hands, and bodies of humans, vehicles, and text in four primary video domains: broadcast news, meetings, surveillance, and Unmanned Aerial Vehicle (UAV) motion imagery. Initial results have also been obtained on automatic analysis of human activities and understanding of video sequences. The performance evaluation initiative in VACE is carried out by the University of South Florida (USF) under the guidance of National Institute of Standards and Technology (NIST).

The Performance Evaluation of Tracking and Surveillance (PETS) program was launched with the goal of evaluating visual tracking and surveillance algorithms. The first PETS workshop was held in March, 2000. Since then, there have been seven such workshops exploring a wide range of surveillance data. While the theme of the initial workshops was target detection and tracking, in the past few workshops the program has matured to address event level tasks. PETS Metrics[1] is a related, but newer initiative to provide an online service for automatically evaluating surveillance

[1] http://petsmetrics.net

results. Currently, the website supports the motion segmentation metrics but in due course is expected to extend to tracking and event detection.

Video Event Recognition Algorithm Assessment Evaluation (VERAAE), Computers in the Human Interaction Loop (CHIL)[2], Evaluation du Traitement et de l'Interprétation de Séquences Vidéo (ETISEO)[3], and Augmented Multiparty Interaction (AMI)[4] are a few programs that share the central idea of developing new algorithms for video understanding and evaluating them for tracking research progress.

Since the motivation for all the above programs is essentially the same, technology transfer among individual programs will result in faster research growth. The CLassification of Events, Activities and Relationships (CLEAR)[5] Evaluation Workshop was the first international evaluation effort that brought together two programs – VACE and CHIL. The benefits of collaboration are obvious – availability of more data for the research community for algorithm development and evolution of widely accepted performance metrics that provide an effective and informative assessment of system performance, are some of the more important ones.

The inspiration for this paper is akin to the CLEAR evaluation attempt. To that extent, we present a qualitative comparison between the PETS and VACE programs in the following viewpoints – evaluation framework (Section 2), evaluation tasks (Section 3), ground truth data (Section 4) and performance metrics (Section 5). We conclude with a discussion on the factors that influence the decision on collaboration between any two programs in Section 6.

## 2. Evaluation Framework

In any evaluation task it is important that each step is clearly defined and executed accordingly. Following this ideology, both the PETS and the VACE programs have a well defined evaluation framework that shares a common perspective.

In VACE the evaluation of a task begins with the task definition. For this to be finalized, the process follows a sequence of steps on a set schedule.

- Micro-corpus: Task definitions are formalized.
  - Two annotators annotate the same clip using a carefully defined set of attributes. (This helps fine tune the annotation guidelines)
  - These annotations are released to the research community.
  - Modifications are made according to the researchers and evaluators need. These include object definitions and attributes.
- Dry Run: Task definitions are frozen. Annotation of training data is initiated and first implementation of the defined metrics in the scoring tool is completed. During this phase, participants familiarize themselves with the output file formatting by using the scoring software.
  - A sub-sample (usually 5 annotated clips) of the training data is released to the participants along with the *index* file on which the participants are asked to submit their results on.
  - Evaluators score and send scores to the participants for verification.
  - Participants use the scoring tool to score their output with the annotation reference and finally verify.
  - Any issues are resolved iteratively with consultations between evaluators and participants.
- Formal Evaluation
  - Release ground truth for the training data (50 clips of 2.5 mins length approx.).
  - Release test data (*index*) file (another 50 clips of 2.5 mins length approx.).
  - Participants submit the results (similar to the Dry Run process).
  - Evaluators score and release scores along with annotations for verification.

PETS also follows a similar process except that the researchers (participants) can submit their results on the web and get their scores generated through PETS Metrics, the online evaluation service. To reach this level of maturity in VACE, the tasks and the framework needs to be frozen beforehand. The schedule so far prevented the evaluators in having a similar setup. Possible helpful scripts can be written so that the participants can submit the results and automated error messages including filenaming conventions, formatting, etc can be generated on the web. This can potentially alleviate generic problems that arise due to the participants not fully implementing the required submission formats.

## 3. Evaluation Tasks

Object detection and tracking is the primary task evaluated in VACE-II. Broadly speaking, the goal of the detection task is to identify the location of the object of interest in each

[2] http://chil.server.de
[3] http://www.silogic.fr/etiseo
[4] http://www.amiproject.org
[5] http://www.clear-evaluation.org

| TASK \ DOMAIN | MEETINGS | BROADCAST NEWS | SURVEILLANCE | UAV |
|---|---|---|---|---|
| Text detection & tracking | – | VACE | – | – |
| Face detection & tracking | VACE | VACE | – | – |
| Hand detection & tracking | VACE | VACE | – | – |
| Person detection & tracking | VACE | – | PETS & VACE | VACE |
| Vehicle detection & tracking | – | – | PETS & VACE | VACE |

Table 1: Evaluation Tasks Supported in VACE-II (– indicates tasks not supported in either PETS or VACE).

frame. On the other hand, the goal of the tracking task is to locate and track the object by a unique identifier in the video sequence.

Evaluation support for a particular task-domain pair depends on the richness of the target objects being evaluated in that domain. After careful deliberation, 10 task-domain pairs were identified for evaluations (Table 1).

The primary focus of PETS is on surveillance technologies. Target detection & tracking and event recognition are the predominant problems addressed in the workshops. A major difference between the definitions of the detection task between the two programs is the fact that PETS definition additionally includes object segmentation which requires systems to output a binary bitmap providing the shape of the object. The tracking task is analogous to the VACE definition with the only difference that the first instance of tracking failure terminates evaluation in PETS. In VACE, among multiple output tracks for a single ground truth object in the video, the single longest track in time is picked as the corresponding hypothesis, while in all other frames the ground truth is treated as a miss.

Since object detection and tracking is the main problem evaluated in VACE-II, the scope of this paper is restricted to the comparison of these tasks. Metrics, annotations and definitions for event level tasks are not discussed. Table 1 summarizes the evaluations supported in VACE-II and PETS for different task-domain pairs.

## 4. Ground Truth Data

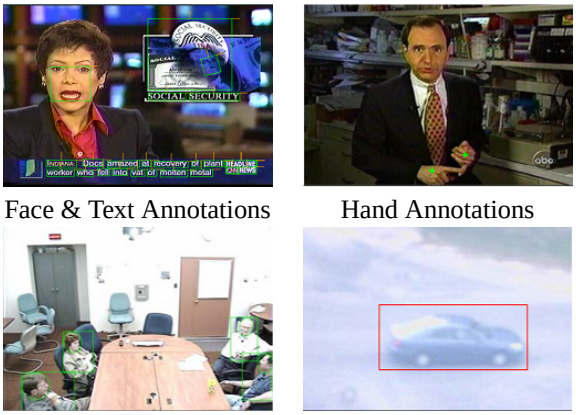
Ground truthing is the process of manually marking what an algorithm is expected to output. Creating a reliable and consistent reference annotation against which algorithm outputs are scored is an extremely involved procedure. Establishing a valid reference is mandatory to carry out a systematic and authentic evaluation.

The ground truth annotations in VACE-II are done by *Video Mining Inc.* using ViPER[6] (Video Performance Evaluation Resource), a video truthing tool developed by the University of Maryland. The source video is in MPEG-2 standard in NTSC format encoded at 29.97 frames per second at 720 x 480 resolution. Currently, every I-frame (every 12 or 15 frames) in the video is ground truthed.

The annotation approach for marking objects varies depending on the task definition and the properties of the video source. Domain specific characteristics such as spatial resolution of objects, time duration of objects' presence, object movement in the sequence and shot boundaries are few such video attributes that influence this decision. For these reasons, the annotation method is classified into two broad categories in VACE [6]:

1. Object bounding annotations – limits of the box edges are based on features of the target object
   - Simple object box (e.g. oriented bounding box for *vehicle* tasks in UAV, *face* and *text* tasks in broadcast news)
   - Composite object box (e.g. head & body box for *person* tasks in meetings)
2. Point annotations – location is based on the features of the target object
   - Single point (e.g. *hand* tasks in meetings)

Figure 1 shows examples of annotations for few representative VACE tasks.




Face & Text Annotations | Hand Annotations

Person Annotations | Vehicle Annotations

Figure 1: Sample Annotations for VACE Tasks.

[6] http://viper-toolkit.sourceforge.net

In addition to the bounding box location, extent, and orientation each object is associated with a set of descriptive attributes, characterizing the region with meta data that are useful during the analysis of evaluation results. For instance, a face object is annotated with the following set of attributes –

- Visible (boolean type)
  - 'TRUE' if 1 eye, nose and part of the mouth is seen
  - 'FALSE' otherwise
- SYNTHETIC (boolean type)
  - 'TRUE' if it is an artificial face
  - 'FALSE' otherwise
- HEADGEAR (boolean type)
  - 'TRUE' if the person is wearing goggles/caps
  - 'FALSE' otherwise
- AMBIGUITY FACTOR (integer type)
  - '0' = Conclusive evidence – when eyes, nose and mouth are clearly seen
  - '1' = Partially conclusive evidence – when two out of three features are seen
  - '2' = Completely inconclusive – when only one or none of the three features can be seen

The ground truthing process in PETS is more involved when compared to that of VACE. This is reasonable considering the point that producing the reference annotation for object segmentation requires demarcation of the foreground which is significantly intricate than just selecting bounding boxes. The University of Reading ground truth annotation tool from the AVITRACK[7] project was adapted to PETS Metrics to generate the reference data. Currently, every 10th frame in the video is annotated. Figure 2 shows an example of the PETS ground truth for the segmentation task[8].

Objects are labeled with different tags (*unoccluded object boundaries*, *partial occlusions* and *complete occlusions*) based on the extent of the visible region. Cases where the target is split by an occluding object are marked with a flag. While the binary bitmap is used to measure the precision of object segmentation, a bounding box is marked to define the extent of an object. The bounding box coordinates are used in the evaluation of tracking. Figure 3 shows an example of the PETS reference annotation for the tracking task[9].

[7] http://www.avitrack.net

[8] The raw image in this example is from [1]

[9] This example is from the PETS On-Line Evaluation Service website – http://petsmetrics.net

Sample Image from Video.

Corresponding Ground Truth.

Figure 2: Sample Annotation for PETS Motion Segmentation Task.

Figure 3: Sample Annotation for PETS Tracking Task.

From the above discussions, it can be noted that among other factors, the definition of the task being evaluated plays a major role in deciding how the ground truth is generated.

## 5. Performance Metrics

In VACE, we have developed a suite of diagnostic measures that capture different aspects of an algorithm's performance [3] and a comprehensive measure which captures many aspects of the task in a single score [4, 5]. The diagnostic measures are useful to the researchers in their failure analysis while the comprehensive measures will be used to provide a summative analysis of overall system performance. All metrics are performance measures normalized between 0 and 1, with 0 being the worst and 1 being the best.

Depending on the annotation approach for a specific task-domain pair, the evaluations can be classified into two categories:

1. **Area-based metrics for object bounding annotations**

   An area-based metric, that is based on spatial overlap between ground truth objects and system output objects to generate the score, is used in the case of an object bounding annotation. For the detection task, the metric (Sequence Frame Detection Accuracy, *SFDA*) captures both the detection accuracy (misses and false alarms) and the detection precision (spatial alignment). Similarly, for the tracking task, both the tracking accuracy (number of correct trackers) and the

| METRIC | DESCRIPTION |
|---|---|
| Negative Rate (NR) | measures the pixel-wise mismatches between the ground truth and the system output in each frame |
| Misclassification Penalty (MP) | measures the segmentation performance on an object basis. Penalty for misclassified pixels is based on distance from an object's boundary |
| Rate of Misclassifications (RM) | measures the average misclassified pixel's distance to an object boundary |
| Weighted Quality Measure (WQM) | measures the spatial alignment error between the ground truth and the system output as a weighted average of false positive and false negative pixels |

Table 2: Motion Segmentation Metrics Currently Implemented in PETS Metrics.

tracking precision (spatial and temporal accuracy) are measured in a single score (Average Tracking Accuracy, *ATA*).

2. **Distance based metrics for point annotations**
   The distance-based metrics (Sequence Frame Detection Accuracy – Distance-based & Average Tracking Accuracy – Distance-based) are parallel to the area-based metrics. The only difference being the fact that in computations for spatial proximity, the spatial overlap between bounding boxes is replaced by a distance measure between corresponding points.

In PETS, the motion segmentation metrics are formulated at the pixel level. Unlike VACE, all metrics are coined as error measures, meaning lower the score, better the performance. Table 2 briefs the four metrics that are currently implemented in PETS Metrics for motion segmentation evaluation [2, 7].

The following five criteria are used in case of the tracking task: [1]

1. Percentage of dataset tracked – is the ratio of the number of frames in which the algorithm tracked the object to the total number of frames. Evaluation is terminated when a track is lost once.

2. Average overlap between bounding boxes – is the percentage overlap between ground truth and system output bounding boxes over the percentage of dataset tracked.

3. Average overlap between bitmaps – measures the bitmap generated by the algorithm to the ground truth object bitmap.

4. Average *chamfer* distance using the ground truth object bitmap to compute the distance transform against the algorithm generated bitmap.

5. Average *chamfer* distance using the algorithm generated bitmap to compute the distance transform against the ground truth object bitmap.

A debate on using multiple performance metrics vs. a comprehensive metric is difficult to arbitrate. The task is challenging primarily because of the conflicting requirements between algorithm developers and end users. From the research community point of view, multiple metrics aid them in debugging their algorithm by identifying failure components. However, from the perspective of an end user who is presented with numerous systems to choose among, comprehensive metrics assist them in their initial analysis through which entirely naïve systems can be instantly discarded. A deeper analysis of selected algorithms can later be done using diagnostic measures to choose the right system for a given application.

# 6. Discussion

From the above sections, it is clear that a common ground needs to be established before the initiation of a collective effort. In this section, we present some of the factors to be considered during such a decision of collaboration between two programs.

1. *Data domain* – The target domain being addressed in each of the programs should be of sufficient interest to both the projects.

2. *Task definition* – The task definitions should be comparable, i.e. the goal of the program in the given domain should be compatible.

3. *Data exchange format* – It is mandatory to agree on a common data format for both ground truth and algorithm output. This is required for a smooth transition in using tools developed in each program.

4. *Tools* – It is optional to agree on the usage of common tools (ground truth authoring tool, scoring tool) as long as the data format is the same. By means of technology transfers, improvements can be made to tools developed in each program.

5. *Ground truth data* – It is necessary that the ground truth annotations are done in a similar manner. Since

ground truth is intended to be what an algorithm is expected to generate, differences in the ground truthing approaches will prohibit cross evaluations on different datasets.

6. *Metrics* – Metrics from each program can be retained as long as the quantities they measure are equivalent, which essentially depends on the task definitions. Through continuing collaboration, the right set of metrics that is most representative of the system performance can be identified.

# References


[1] J. Aguilera, H. Wildenauer, M. Kampel, M. Borg, D. Thirde, and J. Ferryman. Evaluation of Motion Segmentation Quality for Aircraft Activity Surveillance. In *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 293–300, October 2005.

[2] R. Collins, X. Zhou, and S. K. Teh. An Open Source Tracking Testbed and Evaluation Web Site. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (WAMOP-PETS)*, pages 17–24, January 2005.

[3] R. Kasturi, D. Goldgof, P. Soundararajan, and V. Manohar. (Supplement document) Performance Evaluation Protocol for Text and Face Detection & Tracking in Video Analysis and Content Extraction (VACE-II). Technical report, University of South Florida, 2004.

[4] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, M. Boonstra, and V. Korzhova. Performance Evaluation Protocol for Text, Face, Hands, Person and Vehicle Detection & Tracking in Video Analysis and Content Extraction (VACE-II). Technical report, University of South Florida, 2005.

[5] V. Manohar, P. Soundararajan, H. Raju, D. Goldgof, R. Kasturi, and J. Garofolo. Performance Evaluation of Object Detection and Tracking in Video. In *Proceedings of the Seventh Asian Conference on Computer Vision*, volume 2, pages 151–161, January 2006.

[6] H. Raju, S. Prasad, and R. Sharma. Annotation Guidelines for Video Analysis and Content Extraction (VACE-II). Technical report, Video Mining Inc., 2006.

[7] D. Young and J. Ferryman. PETS Metrics: On-Line Performance Evaluation Service. In *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 317–324, October 2005.