Evaluation Test Bed for Object Detection/Tracking and Text Recognition in the VACE Program

Vasant Manohar, Jonathan Fiscus, Martial Michel, Dmitry Goldgof, *Fellow, IEEE*, Rangachar Kasturi, *Fellow, IEEE*.

Abstract—The Video Analysis and Content Extraction (VACE) program supported the development of several key component technologies which could be integrated into systems for automatic analysis of human activities and understanding of video sequences. In order to measure system performance and gauge research progress, a series of evaluations focusing on 15 evaluation tasks were administered in this program during Phase II. Our earlier paper titled "Framework for Performance Evaluation of Face, Text, and Vehicle Detection and Tracking in Video: Data, Metrics, and Protocol" described how the individual aspects of evaluation were accomplished through a consistent employment of a formal infrastructure that consists of a set of data, metrics, tools, and processes required for evaluation. Six evaluation tasks were described. In this follow up article, we detail the remaining nine evaluation tasks. We show that, with minor adaptation, the generic framework proposed earlier can be extended to other target objects in various data domains. Additionally, we report the latest scoring tools we have developed in the context of detection and tracking evaluation. It is our intention that this paper, together with our earlier publication, will serve as a complete report of the various evaluations that were conducted as part of the VACE program.

Index Terms—VACE, CLEAR, performance evaluation, data annotation, face, hand, person, detection and tracking, text recognition.

1 Introduction

THE Video Analysis and Content Extraction (VACE) [1] program was established to develop novel algorithms for automatic video content extraction, multimodal fusion, and event understanding. During VACE Phases I and II, the program made significant progress in the automated detection and tracking of moving objects including faces, hands, people, vehicles, and text in three primary video domains: broadcast news, meetings, and street surveillance. In VACE Phase III, participants developed systems for automatic analysis of human activities and understanding of video sequences. The program supported several evaluation conditions. Each evaluation condition was a combination of a *technology class*, *target object type*, and a *data domain*.

Two classes of technologies were supported – *detection* technologies and *tracking* technologies. The program defined detection as the process of identifying the spatial location of a target object-type (face, person, etc.) within a video frame by outputting an oriented bounding box or a Cartesian point. Tracking was defined as the process of detecting the target object types and then assigning a unique ID for each specific object, thus keeping *track* of the object throughout the video sequence. The data was annotated at the tracking level in that subject bounding boxes were linked together to provide a full track of the subject. Since detection scoring was performed disregarding the cross-frame information, no further annotation was needed to score a detection system. As an initial study, the performance of optical character recognition (OCR) on overlay *videotext* was assessed. There were six target object types supported, namely – *face*, *hand*, *person*, *moving vehicle*, *text line*, and *text word*. Three data domains were supported, namely – broadcast news (*BNews*), meeting room (*MRoom*), and street surveillance (*Surveillance*).

Three performance evaluations were conducted under the auspices of VACE between 2004 and 2007. The evaluation which occurred in 2005 was administered by the University of South Florida (USF) in collaboration with the National Institute of Standards and Technology (NIST) and guided by an advisory forum including the evaluation participants. The reference annotations were generated by Video Mining Inc. In 2006 and then again in 2007, the VACE program and the European Union's Computers in the Human Interaction Loop (CHIL) [2] collaborated to hold the CLassification of Events, Activities, and Relationships (CLEAR) Evaluation Workshop [3] series, which was the first international computer vision evaluation effort that brought together two programs both in terms of data sharing, metrics, and evaluation task definition. During the CLEAR evaluation series, the

- *V. Manohar was the Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620. He is currently with the Speech, Language, and Multimedia Group, Raytheon BBN Technologies, 10 Moulton St., Cambridge, MA 02138. E-mail: vmanohar@bbn.com.*
- *J. Fiscus and M. Michel are with the Multimodal Group, National Institute of Standards and Technology, 100 Bureau Dr., Stop 8940, Gaithersburg, MD 20899. Email: { jonathan.fiscus, martial.michel}@nist.gov*
- *D. Goldgof and R. Kasturi are with the Department of Computer Science and Engineering, University of South Florida, 4202 E. Fowler Ave., ENB 118, Tampa, FL 33620. E-mail: {r1k, goldgof}@cse.usf.edu.*

VACE program contributed the evaluation infrastructure (data, scoring tools, etc.) for a specific set of evaluation tasks and the CHIL consortium, coordinated by the University of Karlsruhe, contributed a separate set of evaluation infrastructure. To the extent possible, the VACE and CHIL programs harmonized their evaluation protocols and metrics.

Table 1 shows the matrix of the 15 supported evaluation tasks in VACE. Kasturi et al. [4] detailed the evaluation framework including the data, metrics, and protocol for 6 evaluation tasks – face detection and tracking in BNews, text line detection and tracking in BNews, and moving vehicle detection and tracking in Surveillance. In this paper we describe the remaining 9 evaluation tasks – face detection and tracking in MRoom, hand detection and tracking in MRoom, person detection and tracking in MRoom, person detection and tracking in Surveillance, and text word recognition in BNews. Additionally, we report the latest tools we have developed in the context of detection and tracking evaluation. This paper, together with [4], provides a complete account of the various evaluations that were conducted as part of the VACE program.

TABLE 1
EVALUATION TASKS SUPPORTED IN VACE
(✓ - TASKS DESCRIBED IN THIS PAPER)

Target Objects	Technology Class	Data Domain		
		BNews	*MRoom*	*Surveillance*
Face	Detection	Kasturi et al. [4]	✓	
	Tracking			
Hand	Detection		✓	
	Tracking			
Person	Detection		✓	✓
	Tracking			
Vehicle	Detection			Kasturi et al. [4]
	Tracking			
Text Line	Detection	Kasturi et al. [4]		
	Tracking			
Text Word	Recognition	✓		

For each detection and tracking task and supported *object-domain* combination, we developed a 50-clip training set and a 50-clip test set. As part of CLEAR, the face detection and tracking task in MRoom had an additional 50 clips collected at multiple sites (see Section 2.1). For the text word recognition task in BNews, we developed a 5-clip training set and a 25-clip test set. Each data clip was approximately 2.5 minutes long resulting in around 32.5 hours of video data that was completely spatially/ temporally annotated at the I-frame level.

2 SOURCE VIDEO DEFINITIONS

2.1 Meetings

The Meeting Room corpus can be categorized into two sets – the ones collected at the National Institute of Standards and Technology and a set of multi-site meeting room recordings collected at 5 different participating organizations.

2.1.1 NIST Meeting Room

NIST constructed a Meeting Data Collection Laboratory to collect corpora to support research, development, and evaluation in meeting recognition technologies. This data is captured in a NIST-internal file format. The video data is extracted from this format and is then encoded using the MPEG-2 standard in NTSC format (720 x 480 at 29.97 fps).

2.1.2 Multi-Site Meeting Data

The Multi-Site Meetings test set consists of datasets from the following 5 sites: Carnegie Mellon University (CMU), University of Edinburgh (EDI), National Institute of Standards and Technology (NIST), Netherlands Organization for Applied Scientific Research (TNO), and Virginia Tech (VT). Each site had their own independent camera setup, different illuminations, viewpoints, people, and topics in the meetings. Most of these datasets also included High-Definition (HD) recordings but were subsequently formatted to MPEG-2 for evaluation purposes. The video data provided for the multi-site data meeting data was transcoded to a near-common format (CMU, NIST, and VT – 720 x 480 resolution at 29.97 fps; EDI and TNO – 720 x 480 at 25 fps).

2.2 Surveillance

The surveillance data set was from the Imagery Library for Intelligent Detection Systems (i-LIDS) corpus which was developed by the UK Home Office Scientific Development Branch (HOSDB) and the Centre for the Protection of National Infrastructure (CPNI). It consists of real CCTV footage based on scenarios such as abandoned baggage detection, parked vehicle detection, etc. The original videos were in QuickTime MJPEG format but were subsequently formatted to MPEG-2 for evaluation purposes (720 x 480 at 25 fps).

3 REFERENCE ANNOTATIONS

The reference annotations for VACE Phase II evaluations were generated by Video Mining Inc. The Video Performance Evaluation Resource (ViPER) [5] was used as the tool for ground truthing video sequences. Objects were marked by geometric shapes such as oriented and non-oriented bounding boxes, ellipses, and points. The truth data was formatted in XML following the ViPER schema[1].

During VACE, a clear and exhaustive set of annotation guidelines was established for each object-domain pair. Each object block was associated with a set of attributes that characterized the object both from an evaluation and informational point of view. In order to empirically evaluate the consistency of the annotation, 10% of the entire corpus was doubly annotated with two independent annotators and were used to compute inter-annotator disagreements using the developed performance metrics. The representative disagreements were logged and checked

[1] http:/ / viper-toolkit.sourceforge.net/ owl/ viper/ datatypes/ .

for on the remaining 90% of the singly annotated ground truth through a quality control process that employed both manual and automatic testing. For more details on the annotation quality control procedure, we refer the reader to our earlier publication [4]. The remainder of this section explains the set of guidelines and additional flags used in the face, hand, person, and text word annotation tasks.

3.1 Face Annotations

Each face object was annotated with the following set of attributes – [LOCATION], [VISIBLE], [AMBIGUITY], [SYNTHETIC], and [HEADGEAR].

The [LOCATION] attribute was defined by the *oriented box* data type (X, Y, WIDTH, HEIGHT, ORIENTATION). Facial features were used as cues to mark the bounds of this box. The [VISIBLE] attribute was set to *true* as long as one eye, part of the nose, and part of the mouth were seen.

The [AMBIGUITY] attribute described how clear the face was in the frame. It had three possible values ("0" when all of the three features were visible, "1" when two of the three features were visible and "2" when none of the three features was visible). The [SYNTHETIC] attribute was set to *true* for cases when the face was not human (cartoons, sketches, etc.). The [HEADGEAR] attribute was used in cases when the person was wearing hats, sunglasses, etc. Fig. 1 shows sample face annotations from the multi-site meeting room data.

Fig. 1. Sample face annotations in multi-site meeting room data (face marked in *yellow* has [OCCLUSION] = *true*).

3.2 Hand Annotations

The five fingers and the area between the base of the fingers and the wrist defined a hand. Each hand object was annotated with the following set of attributes – [LOCATION], [LARGE HAND], [VISIBLE], [OCCLUSION], [SYNTHETIC], and [HANDGEAR].

The [LOCATION] attribute was annotated using the *point* data type (X, Y). When a hand appeared, a point was placed at the center of the palm region of the hand. If a hand occupied more than 25% of the frame, then the hand was additionally bound with an oriented box denoted as [LARGE HAND].

The [VISIBLE] attribute was set to *true* if at least 50% of the hand region was present in the scene. If the fingers alone were seen without the palm region shown, [VISIBLE] was set to *false*.

There were a number of scenarios when a hand was marked with [OCCLUSION] = *true*. This was based on a general observation that the hand was clearly seen in the beginning and then one of the six features was blocked by an object in the scene. The six features were the five fingers and the complete palm region.

The [SYNTHETIC] attribute of the hand was marked *true* when the hand was not an in-scene human hand such as statues, pictures, etc. The [HANDGEAR] attribute was set to *true* when a hand was covered by something such as gloves. Fig. 2 shows sample hand annotations from the NIST meeting room data.

Fig. 2. Sample hand annotations in NIST meeting room data (hands marked in *green* have [OCCLUSION] = *true*).

3.3 Person Annotations

3.3.1 Meeting Room Videos

A person was defined as a human entity composed of a head and a torso. Each person object in the meeting room domain was annotated with the following set of attributes – [HEAD LOCATION], [BODY LOCATION], [VISIBLE], [OCCLUSION], [AMBIGUOUS], [SYNTHETIC], and [HEADGEAR].

The spatial location of a person was decomposed in two parts – [HEAD LOCATION] represented an elliptical region marking the head region of the person and [BODY LOCATION] defined a bounding box enclosing the head and the upper torso regions of the person. The orientation of the [BODY LOCATION] was parallel to the shoulder plane of the person irrespective of the body posture. This was done to provide consistency and repeatability across multiple annotators.

A person was annotated (the [VISIBLE] attribute was set to *true*) as soon as any portion of the head or shoulders was seen in the scene. The [OCCLUSION] attribute was set when some features were hidden or not seen in the scene. Further, an [AMBIGUOUS] attribute was included to handle extreme occlusion and complex situations. For instance, in the meeting room domain, some people were cut off by the camera view or were almost completely hidden behind a file cabinet. In these cases, [AMBIGUOUS] was set to *true*. The disambiguation between the usage of [OCCLUSION] and the [AMBIGUOUS] attributes was quantified with the following rules:

- If only a portion of the head and the shoulders were occluded, then [OCCLUSION] was set to *true*.
- If the entire shoulder region or a major part of the

head and the shoulder regions were occluded, then [AMBIGUOUS] was set to *true*.

The [SYNTHETIC] attribute of the person was marked *true* when the person was not an in-scene human. The [HEADGEAR] attribute was set to *true* when a person's head was covered by a hat, etc. Fig. 3 shows sample person annotations from the NIST meeting room data.

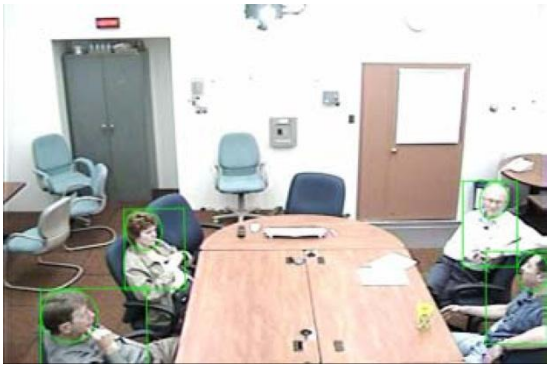

Fig. 3. Sample person annotations in NIST meeting room data.

3.3.2 Surveillance Videos

A person was defined as a human entity composed of a head, a torso, and legs. Each person object in the surveillance domain was annotated with the following set of attributes – [LOCATION], [PRESENT], [OCCLUSION], [AMBIGUOUS], [MOBILITY], and [SYNTHETIC].

The [LOCATION] attribute was defined by a non-oriented bounding box (X, Y, WIDTH, HEIGHT) drawn to cover the head, torso, and leg regions of the person. A person was annotated (the [PRESENT] attribute was set to *true*) when at least 25% of the person's body was seen. The [OCCLUSION] attribute was set to *true* if more than 50% of the person body frame was hidden by another object. As in the case of meeting room domain, the [AMBIGUOUS] attribute was used in situations when defining an accurate spatial bound was challenging. Further, in order to describe the motion of a person in the frame, the [MOBILITY] attribute was set to *mobile* or *stationary*. Finally, the [SYNTHETIC] attribute was used to denote artificial cases. Fig. 4 shows sample person annotations from the i-LIDs surveillance data.

Fig. 4. Sample person annotations in i-LIDs surveillance data.

3.4 Text Word Annotations

In [1], we described the reference annotations for the text line detection task in broadcast news. The truth data for the task had the following attributes – [LOCATION], [TYPE], [READABILITY], [OCCLUSION], [LOGO], and [CONTENT]. As noted in [1], text lines were transcribed and stored in the [CONTENT] attribute when [TYPE] = *Overlay*, [READABILITY] = *2*, [OCCLUSION] = *false*, and [LOGO] = *false*. This set of transcriptions along with the word offsets marked in the ground truth for these lines was used for the text recognition task. Fig. 5 shows sample word-level annotations in the broadcast news data.

Fig. 5. Sample text word annotations in broadcast news (CNN marked as [LOGO] = *true*).

4 Evaluation Metrics

The VACE Program and CLEAR evaluation series collaborated to define a set of metrics to analyze detection and tracking system performance. We created a set of summative metrics that addresses the important types of errors (misses, false alarms, spatial mismatches, and tracking ID switches) in a holistic way. The metrics are tied to an error minimization algorithm [6] that maps the system output objects to the ground-truth objects. These metrics (Sequence Frame Detection Accuracy (*SFDA*), Multiple Object Detection Accuracy (*MODA*), and Multiple Object Detection Precision (*MODP*) for detection; Average Tracking Accuracy (*ATA*), Multiple Object Tracking Accuracy (*MOTA*), and Multiple Object Tracking Precision (*MOTP*) for tracking) were described in detail in [4]. The generalized area-based metric definition that uses the spatial overlap as a kernel function to map the ground truth and the system output objects, facilitated consistence of scoring of variety of object detection and tracking tasks where a bounding-box approach to annotations could be adopted.

4.1 Distance-Based Metrics

For tasks that require point-based annotations such as hand detection and tracking in MRoom, we defined metrics that were analogous to the area-based metrics. The only difference being, when computing the spatial proximity, the spatial overlap between bounding boxes was replaced by a distance measure between corresponding points.

Detection: For a given frame, the Frame Detection Accuracy – Distance (*FDA-D*) measure calculates the spatial proximity between the ground truth and the system output objects as a ratio of the Euclidean distance between

the two objects and a pre-defined constant that denotes the maximum allowable distance between an object pair. Object pairs farther than this distance were unmappable. The sum of all of the distances was normalized over the average of the number of ground truth and detected objects. For a single frame *t*, where there are $N_G^{(t)}$ ground-truth objects and $N_D^{(t)}$ detected objects, we defined *FDA-D(t)* as:

$$FDA\text{-}D(t) = \frac{\sum_{i=1}^{N_{mapped}^{(t)}} (1 - d_i^{'})}{\left[\frac{N_G^{(t)} + N_D^{(t)}}{2}\right]}$$

Here, $N_{mapped}^{(t)}$ is the number of mapped object pairs in frame *t*, where the correspondence is established between objects that have the best spatial proximity in the given frame *t*, using the Hungarian Solution [6], [7] to the Bipartite Graph matching problem by modeling object instances as nodes in the bipartite graph. d_i' is the normalized distance and can take values between 0 and 1.

The Sequence Frame Detection Accuracy – Distance-based (*SFDA-D*), which is essentially the average *FDA-D* over all the relevant frames, was defined as:

$$SFDA = \frac{\sum_{t=1}^{N_{frames}} FDA\text{-}D\ (t)}{\sum_{t=1}^{N_{frames}} \exists(N_G^{(t)}\ \text{OR}\ N_D^{(t)})} \quad (1)$$

Tracking: We extended the notion of spatiotemporal overlap in the case of area-based metrics in order to define the Sequence Track Detection Accuracy – Distance-based as:

$$STDA\text{–}D = \sum_{i=1}^{N_{mapped}} \frac{\sum_{t=1}^{N_{frames}^{i}} (1 - d_t^{'})}{N_{(G_i \cup D_i \neq \phi)}} \quad (2)$$

We defined the Average Tracking Accuracy – Distance-based (*ATA-D*), which can be termed as the *STDA-D* per object, as:

$$ATA\text{–}D = \frac{STDA\text{–}D}{\left(\frac{N_G + N_D}{2}\right)} \quad (3)$$

where, N_G and N_D are the number of unique objects in the video sequence (uniqueness defined by object ID).

4.2 Text Recognition Metric

The performance measure for the text recognition task was based on insertion (*I*), deletion (*D*), and substitution (*S*) errors at the word level. We computed the standard Word Error Rate (*WER*) metric used in speech recognition evaluations.

$$WER = \frac{(I + D + S)}{(\text{Total \# Word Instances in Ref})} \quad (4)$$

Fig. 6 illustrates the text recognition metric scoring for a synthetic output.

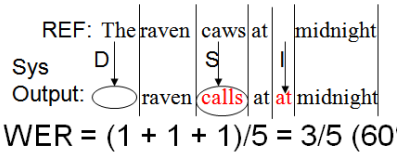

Fig. 6. WER computation for different system generated errors.

In the above example, the normalization was based on the number of word instances in the reference. We also computed an auxiliary metric where the normalization was based on the number of evaluation frames.

5 Evaluation Tools, Procedure, and Results

This section describes the F4DE (Framework for Detection Evaluations) scoring tool used for computing the performance scores by comparing the algorithm output with the annotated ground truth data. We present the scoring tool settings used for the various evaluation tasks. Finally, we report the VACE and CLEAR results for the 9 evaluation tasks described in this paper.

5.1 F4DE Scoring Tool

The Framework for Detection Evaluation (F4DE) toolkit[2] contains a set of evaluation tools for detection evaluations and was used in specific NIST-coordinated evaluations, such as the CLassification of Events, Activities, and Relationships (CLEAR) Evaluation Workshop, the TREC Video Retrieval Evaluation (TRECVID)'s Surveillance Event Detection (SED) [8] and Multimedia Event Detection (MED) [9], and the Advanced Video and Signal Based Surveillance (AVSS) Challenge [10]. The toolkit provides components for validating, scoring, and helping analysis of system files for such evaluations.

The CLEAR branch of F4DE was developed in Perl. Both the VACE (SFDA, ATA) and the CLEAR (MODA, MODP, MOTA, MOTP) detection and tracking metrics are supported by the tool. For the text recognition task, it computes the average WER (normalized both by the number of words and by the number of frames in the video). In addition to the metrics, the tool computes a set of auxiliary information such as the IDs of the ground truth objects that were missed, the system output objects that were treated as false positives, and the spatial overlap between mapped object pairs on a per frame basis. These features aided the developers in investigating the failure instances of their algorithm for debugging purposes.

5.2 Evaluation Settings

The reference was richly annotated with a variety of information, as described in Section 3. The additional set of attributes was used in deciding whether a particular object should be evaluated or not. The specific settings used for the face, hand, person, and text word tasks are shown in Table 2. All other annotated regions were treated as "Don't Care" where, the system output was neither penalized for missing nor given credit for detecting the unscored region. It has to be noted that each of these

[2] http:/ / www.nist.gov/ itl/ iad/ mig/ tools.cfm

attributes can be selectively specified to be included in evaluation through the scoring tool F4DE.

TABLE 2
EVALUATION SETTINGS FOR FACE, HAND, PERSON, AND TEXT WORD TASKS IN VACE AND CLEAR

TASK	EVALUATION SETTINGS
Face detect and track in MRoom	[VISIBLE] = *true* [AMBIGUITY] = *0* [SYNTHETIC] = *false* [HEADGEAR] = *false*
Hand detect and track in MRoom	[VISIBLE] = *true* [OCCLUSION] = *false* [SYNTHETIC] = *false* [HANDGEAR] = *false*
Person detect and track in MRoom	[VISIBLE] = *true* [OCCLUSION] = *false* [AMBIGUOUS] = *false* [SYNTHETIC] = *false* [HEADGEAR] = *false*
Person detect and track in Surveillance	[PRESENT] = *true* [OCCLUSION] = *false* [AMBIGUOUS] = *false* [MOBILITY] = *mobile/stationary* [SYNTHETIC] = *false*
Text word recognition in BNews	[TYPE] = *overlay* [READABILITY] = *2* [OCCLUSION] = *false* [LOGO] = *false*

5.3 Evaluation Results

In this section, we present the metric scores of the best performing system in each of the technology tracks administered during the three evaluation cycles of VACE. For Cycle I (2005), we report the results using the VACE metrics and for Cycle II (2006) [11] and Cycle III (2007) [12], we use the CLEAR metrics.

Cycle I. During this period, we evaluated the performance of text detection and tracking in *BNews* [4], face detection and tracking in *BNews* [4] and *NIST MRoom*, hand detection and tracking in *NIST MRoom*, person detection and tracking in *NIST MRoom*, and text word recognition in *BNews*. Fig. 7 shows the average SFDA and ATA for each evaluation condition in Cycle I.

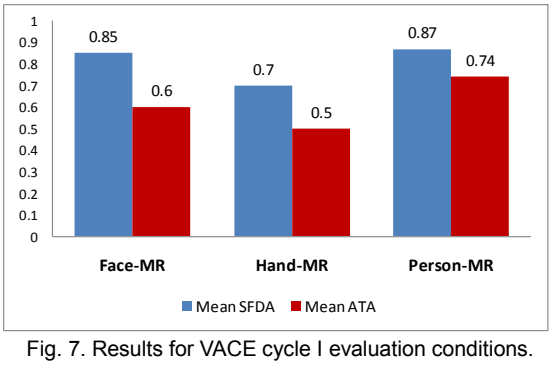

Fig. 7. Results for VACE cycle I evaluation conditions.

For the text recognition task, the WER of the best performing system was 0.53 when normalized by the number of words and 0.47 when normalized by the number of frames in the video.

Cycle II. During this period, we evaluated face detection and tracking in *Multi-Site MRoom*, person detection and tracking in *Surveillance*, and moving vehicle detection and tracking in *Surveillance* [4]. Fig. 8 shows the average MODA and MOTA for each evaluation condition in Cycle II.

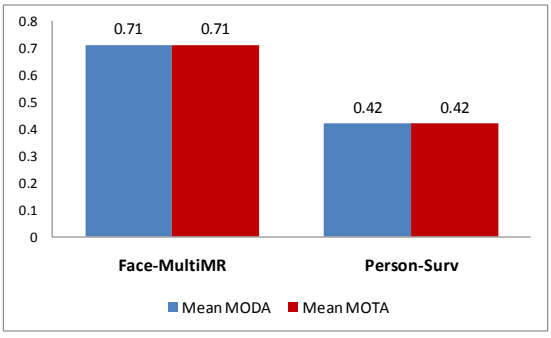

Fig. 8. Results for VACE cycle II evaluation conditions.

Cycle III. During this period, the focus was to measure the progress of the systems on the set of tasks undertaken in Cycle II using the same data. We evaluated face tracking in *Multi-Site MRoom*, person tracking in *Surveillance*, and moving vehicle tracking in *Surveillance*. Fig. 9 shows the average MOTA for each evaluation condition in Cycle III.

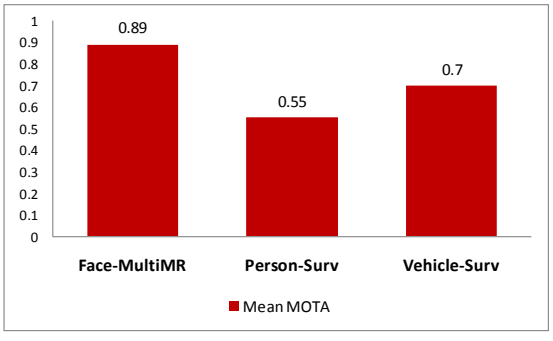

Fig. 9. Results for VACE cycle III evaluation conditions.

6 DATA PROCUREMENT PROCEDURE

The distribution includes the full set of data resources developed by the VACE program. It is accompanied by a detailed documentation that includes the following components: evaluation task definitions, source video data description, evaluation metrics, reference annotations and annotation specifications, and evaluation tools. The video corpora and reference annotations have been divided up into three individual sub corpora by data domain to simplify distribution. The following is the list of individual sub corpora –

- VACE Evaluation Resources for Text and Face Detection and Tracking and Text Recognition in Broadcast News: LDC[3] *(catalog number pending final publication by LDC).*
- VACE Evaluation Resources for Face, Hand, and Person Detection and Tracking in Meeting Data: LDC *(catalog number pending final publication by LDC).*

[3] http:/ / www.ldc.upenn.edu/ , ldc@ldc.upenn.edu, 3600 Market Street, Suite 810, Philadelphia, PA, 19104-2653, USA.

- VACE Evaluation Resources for Person and Vehicle Detection and Tracking in Street Surveillance Data: UK HOSDB[4] (request for "Parked Vehicle" corpus of i-LIDS).

7 SUMMARY

This paper summarized the VACE evaluation, which started early in 2004 and concluded with the CLEAR 2007 workshop. This article, together with our earlier publication, described the evaluation tasks performed in VACE, including description of metrics developed, databases used, ground truth annotations generated along with the specifications used, and also gave an overview of the individual results achieved by the best performing system in each track. The goal was to systematically address the challenges of object detection and tracking through a common evaluation framework that permits a meaningful objective comparison of techniques. To that extent, we have established sufficient resources that can support the necessary quantities of data for automatic modeling techniques, as well as a statistically significant comparison of the performance of algorithms.

ACKNOWLEDGMENT

The work presented here was funded by the US Government VACE program. The authors are grateful to John Garofolo (NIST) for coordinating the evaluation track in the VACE and the CLEAR programs. The authors would like to thank Padmanabhan Soundararajan, Valentina Korzhova, Matthew Boonstra, Jing Zhang, Matthew Shreve, Sridhar Godavarthy, and Gerry Hernandez (University of South Florida) for their contribution in the development of the tools and resources under VACE. They thank Rajeev Sharma, Harish Raju, and Shubha Prasad (Video Mining Inc.) for heading the annotation team that generated the ground truth for the VACE evaluations. They are indebted to Rainer Stiefelhagen and Keni Bernardin (University of Karlsruhe) for their efforts and cooperation in the development of the CLEAR metrics.

DISCLAIMER: *Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST nor is it intended to imply that the equipment, instruments, software, or materials are necessarily the best available ones for the purpose.*

REFERENCES

[1] *Video Analysis and Content Extraction (VACE),* https:/ / control.nist.gov/ dto/ twiki/ bin/ view/ Main/ WebHome.
[2] *Computers in the Human Interaction Loop (CHIL),* http:/ / chil.server.de.
[3] *CLassification of Events, Activities and Relationships (CLEAR),* http:/ / www.clear-evaluation.org.
[4] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 319-336, 2009.
[5] D. Doermann and D. Mihalcik, "Tools and Techniques for Video Performance Evaluation," *Proc. Int'l Conf. Pattern Recognition*, vol. 4, pp. 167-170, 2000.
[6] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistic Quarterly*, vol. 2, pp. 83-97, 1955.
[7] C.H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: Algorithms and complexity*, Prentice Hall, 1982.
[8] *TRECVid 2010 Evaluation for Surveillance Event Detection,* http:/ / www.itl.nist.gov/ iad/ mig/ tests/ trecvid/ 2010/ .
[9] *TRECVid 2010 Evaluation for Multimedia Event Detection,* http:/ / nist.gov/ itl/ iad/ mig/ med10.cfm.
[10] *AVSS 2010 Multi-Camera Tracking Challenge,* http:/ / www.itl.nist.gov/ iad/ mig/ tests/ avss/ 2010/ .
[11] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The CLEAR 2006 Evaluation," *Multimodal Technologies for Perception of Humans*, pp. 1-44, 2007.
[12] R. Stiefelhagen, K. Bernardin, R. Bowers, T. Rose, M. Michel, and J. Garofolo, "The CLEAR 2007 Evaluation," *Multimodal Technologies for Perception of Humans*, pp. 3-34, 2008.

[4] http:/ / www.homeoffice.gov.uk/ science-research/ hosdb/ , i-LIDS@homeoffice.gsi.gov.uk.