A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video

Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen,
Jong Taek Lee, Saurajit Mukherjee, J. K. Aggarwal, Hyungtae Lee, Larry Davis, Eran Swears,
Xioyang Wang, Qiang Ji, Kishore Reddy, Mubarak Shah, Carl Vondrick, Hamed Pirsiavash,
Deva Ramanan, Jenny Yuen, Antonio Torralba, Bi Song, Anesco Fong, Amit Roy-Chowdhury, Mita Desai

{sangmin.oh, anthony.hoogs, amitha.perera, naresh.cuntoor}@kitware.com,
{ccchen, jongtaeklee,saurajit, aggarwaljk}@mail.utexas.edu,{htlee, lsd}@umiacs.umd.edu,
{sweare, wangx16,qji}@ecse.rpi.edu, {kkreddy,shah}@eecs.ucf.edu, {jenny,torralba}@csail.mit.edu,
{cvondric,hpirsiav,dramanan}@ics.uci.edu, {bsong,anesdo, amitrc}@ee.ucr.edu, mita.desai@darpa.mil

Abstract

We introduce a new large-scale video dataset designed to assess the performance of diverse visual event recognition algorithms with a focus on continuous visual event recognition (CVER) in outdoor areas with wide coverage. Previous datasets for action recognition are unrealistic for real-world surveillance because they consist of short clips showing one action by one individual [15, 8]. Datasets have been developed for movies [11] and sports [12], but, these actions and scene conditions do not apply effectively to surveillance videos. Our dataset consists of many outdoor scenes with actions occurring naturally by non-actors in continuously captured videos of the real world. The dataset includes large numbers of instances for 23 event types distributed throughout 29 hours of video. This data is accompanied by detailed annotations which include both moving object tracks and event examples, which will provide solid basis for large-scale evaluation. Additionally, we propose different types of evaluation modes for visual recognition tasks and evaluation metrics along with our preliminary experimental results. We believe that this dataset will stimulate diverse aspects of computer vision research and help us to advance the CVER tasks in the years ahead.

1. Introduction

Visual event recognition—the recognition of semantic spatio-temporal visual patterns such as "walking", "getting into vehicle", and "entering facility"—is a core computer vision problem, as evidenced by the plethora of publications in the academic literature [8, 11, 12, 14, 13, 15, 19, 20]. Much of the progress has been enabled by the availability of public datasets, such as the KTH [15] and Weizmann [8] datasets. However, the current state of the art has surpassed these existing datasets, i.e., performance on these datasets have been saturated, and there is a need to for a new, larger, and more complex dataset to stimulate progress.

In this paper, we introduce *VIRAT Video Dataset* which is a new large-scale surveillance video dataset designed to assess the performance of event recognition algorithms in realistic scenes[1]. The dataset includes videos collected from both stationary ground cameras and moving aerial vehicles. We expect the dataset to further research in "*continuous visual event recognition* (CVER)", where the goal is to both recognize an event and to localize the corresponding space-time volume from large continuous video . This is far more closely aligned with real-world video surveillance analytics needs than the current research which aims to classify a pre-clipped video segment of a single event. Accurate CVER would have immediate and far reaching impact in domains including surveillance, video-guided human behavior research, assistive technology, and video archive analysis.

Existing datasets [15, 8] for action recognition are unrealistic for real-world surveillance because they consist of short clips showing one action by one individual. Datasets have been developed for movies [11] and sports [12], but, these actions and scene conditions do not apply effectively to surveillance videos. Our dataset consists of 16 outdoor scenes with actions occurring naturally by non-actors in continuously captured videos of the real world. The dataset includes large numbers of instances for 23 event types distributed throughout 29 hours of video, which is two to three orders of magnitude larger than existing datasets such CAVIAR [7]. TRECVID 2008 airport dataset [16] contains 100 hours of video, but, it provides only frame-level annotations, which makes it a difficult benchmark for most learning-based computer vision approaches. More specifically, existing datasets [7, 11, 14, 15, 19, 13, 12, 8, 16] are often limited for CVER on surveillance videos for one of the following ways: (1) unnatural appearance because the events are acted in constrained scenes; (2) lack of support for CVER because examples are cropped in space and time; (3) limited spatial and temporal coverage, which limits the use of advanced methods to exploit spatio-temporal context; (4) lack of event type diversity (particularly for multi-object

[1] Available from: www.viratdata.org

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California, Irvine,Department of Computer Science ,Irvine,CA,92697-3435**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Computer Vision and Pattern Recognition (CVPR) Colorado Springs, Colorado, 20-23 June 2011. U.S. Government or Federal Rights License

14. ABSTRACT
We introduce a new large-scale video dataset designed to assess the performance of diverse visual event recognition algorithms with a focus on continuous visual event recognition (CVER) in outdoor areas with wide coverage. Previous datasets for action recognition are unrealistic for real-world surveillance because they consist of short clips showing one action by one individual [15, 8]. Datasets have been developed for movies [11] and sports [12], but, these actions and scene conditions do not apply effectively to surveillance videos. Our dataset consists of many outdoor scenes with actions occurring naturally by non-actors in continuously captured videos of the real world. The dataset includes large numbers of instances for 23 event types distributed throughout 29 hours of video. This data is accompanied by detailed annotations which include both moving object tracks and event examples, which will provide solid basis for large-scale evaluation. Additionally, we propose different types of evaluation modes for visual recognition tasks and evaluation metrics along with our preliminary experimental results. We believe that this dataset will stimulate diverse aspects of computer vision research and help us to advance the CVER tasks in the years ahead.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

Figure 1. An example end-to-end sequence: a person walks into a scene, loads the bag he was carrying into a vehicle, gets into the vehicle, and then leaves while driving. The corresponding person and the car are marked in red and light-green boxes respectively.

events); (5) lack of concurrent events; (6) lack of variability in viewpoints, subjects, and/or scenes; (7) lack of detailed annotations; and (8) lack of aerial datasets (with the exception of [1]) which hinders us from delving into the new application areas for intelligent aerial vehicles. All these issues make it difficult to assess real world performance of CVER algorithms for surveillance. For example, the results from recent competition [13] noted that none of the participating teams could produce any results for CVER (in time for the competition), although a few teams produced fairly accurate N-way classification results on cropped examples. Such results indicate that CVER is one of the next challenges that computer vision research will address.

Our major contributions are as follows: (1) we introduce a new public ground and aerial camera surveillance video dataset, which are two or three orders of magnitude larger than existing datasets in many dimensions. This dataset provides realistic and diverse event examples, with no visibly evident acting. Collective effort has been made by nine research groups to obtain natural examples from wide variety of sites under different weather conditions and for annotation (Sec.2); (2) this data is accompanied by detailed annotations which include both object tracks and localized events, which will provide solid basis for large-scale quantitative evaluation for computer vision research, e.g., CVER and tracking. In addition, we share our two-stage annotation methodologies used for our large-scale annotation efforts where both Mechanical Turks and domain experts are involved. (Sec.2.3); (3) we propose different evaluation settings which include independent and learning-enabled CVERs, our considerations for objective evaluation which include definitions of hits, and separate training/test/sequestered sets for each evaluation mode, and showcase exemplary experimental results (Sec.3).

2. VIRAT Video Dataset

Challenging datasets and support for objective evaluation inspire the innovations in computer vision. For example, The PASCAL VOC Challenge [6] provides large-scale multi-class object image datasets annually and provide venues to measure performance of diverse approaches from different aspects, leading to scientific advances in visual object recognition. In much the same way, our goal in introducing this dataset is to provide a better benchmark and to help identify the limitations of current CVER capabilities in real-world surveillance settings, and thus focus research effort on innovations to overcome these difficulties. In the following sections, we present both stationary ground (Sec. 2.1) and aerial (Sec. 2.2) video datasets, along with the developed annotation standard.

Compared to existing datasets, our new dataset has richer event diversity, and includes events that involve interactions between multiple actors, vehicles, and facilities. The following lists enlist some of the 23 event types in our dataset:

- **Single Person Events** (8): walking, running, standing, throwing, gesturing, carrying, loitering, picking up
- **Person and Vehicle Events** (7): getting into or getting out of vehicle, opening or closing trunk, loading, unloading, dropping off, bicycling
- **Person and Facility Events** (2): entering or exiting facility

It can be seen that our new dataset captures events beyond the usual single person actions provided in existing datasets such as KTH, Weizmann, and HOHA where the number of single person activities are below 10 classes. Our dataset contains a richer set of multiple-object actions, and may be the first large public dataset to include diverse type of vehicle-person interactions, annotated in detail with many examples per category. In particular, characteristics such as (1) existence of incidental objects and activities and (2) recording of end-to-end activities attribute our dataset to be more suitable for CVER in real world. By incidental objects and activities, we mean that there are multiple moving objects or activities occurring simultaneously at multiple space locations in the same scene. By end-to-end activities, event instances in our dataset capture the entire temporal context of events. For example, an example sequence in Fig. 1 shows a person with full temporal context where he appears into a scene carrying a bag, approaches a vehicle, loads the bag into the vehicle, gets into the vehicle, and finally leaves out of scene driving the vehicle. The comparison of our dataset and other existing datasets are shown in Table 1[2]. In particular, Table 1 shows that the amount of pixels occupied by moving objects (especially people) in this dataset constitute very small portion of captured images. For example, HOHA 1 dataset consists of movie clips which sometimes include only 20% of entire human figure (e.g., above shoulder), amounting to 500% human to video

[2]Some statistics are approximate, obtained from the CAVIAR 1st scene and TRECVID dry-run data only, due to limited public information.

	KTH	Weizmann	HOHA 1	TRECVID	This Work
# of Event Types	6	10	8	10	23
Avg. # of samples per class	100	9	~85	3~1670	10~1500
Max. Resolution (w x h)	160 x 120	180 x 144	~540 x 240	720 x 576	1920 x 1080
Human Height in Pixels	80~100	60~70	100~1200	20~200	20~180
Human to video height ratio	65~85%	42~50%	50~500%	4~36%	2~20%
# Scenes	N/A	N/A	Many	5	17
Viewpoint Type	Side	Side	Varying	5 / Varying	Varying
Natural Background Clutter	No	No	Yes	Yes	Yes
Incidental Objects/Activities	No	No	Yes, Varying	Yes	Yes
End-to-end Activities	No	No	Yes, Varying	Yes	Yes
Tight Bounding boxes	Cropped	Cropped	No	No	Yes
Multiple annotations on movers	No	No	No	No	Yes
Camera Motion	No	No	Varying	No	Varying

Table 1. Comparison of characteristics of datasets

height ratio. By contrast, in this dataset, a 20-pixel-tall person frequently appears in 1080p video, amounting to 2% of the video height ratio, which makes this dataset to be an excellent batch for CVER tasks. In terms of annotation, we provide multiple annotations, e.g., a person may be marked by simultaneous events of `walking` and `carrying`.

2.1. First Part: VIRAT Ground Video Dataset

The first portion of the video dataset consists of stationary ground camera data. We collected approximately 25 hours of stationary ground videos across 16 different scenes, amounting to approximate average of 1.6 hours of video per scene. The snapshots of these scenes are shown in Fig. 2, which include parking lots, construction sites, open outdoor spaces, and streets. These scenes were selected based on the observation that human and vehicle events occur frequently in these areas. Multiple models of HD video cameras recorded scenes at 1080p or 720p to ensure that we obtain appearance information from objects at distance, and frame rates range 25~30 Hz. The view angles of cameras towards dominant ground planes ranged between 20 and 50 degrees by stationing cameras mostly at the top of buildings to record large number of event instances across area while avoiding occlusion as much as possible. Heights of humans within videos range 25~200 pixels, constituting 2.3~20% of the heights of recorded videos with average being about 7%. In terms of scene diversity, only two pairs of scenes (total 4) among 16 scenes had FOV overlap, with substantial outdoor illumination changes captured over days. In addition, our dataset includes approximate homography estimates for all scenes, which can be useful for functions such as tracking which needs ground coordinate information.

Most importantly, most of this stationary ground video data captured *natural events* by *monitoring scenes over time*, rather than relying on recruited actors. Recruited multi-actor acting of both people and vehicles was involved in the limited subset of 4 scenes only: total acted scenes are approximately 4 hours in total and the remaining 21 hours of data was captured simply by watching real-world events. Originally, more than 100 hours of videos were recorded in monitoring mode during peak activity hours which include morning rush hour, lunch time, and afternoon rush hour, from which 25 hours of quality portions were manually selected based on the density of activities in the scenes.

For comparison purpose, six scenes in our dataset and snapshots from other datasets are shown in Fig.2 and Fig.3 respectively. It can be observed that our new dataset provides a new benchmark for CVER: it is more challenging in terms of pixel resolution on humans, wide spatial coverage of scenes, background clutters, and diversity in both collection sites and viewpoints. In comparison to surveillance datasets such as CAVIAR [7] and TRECVID [16] shown in Fig. 3 (d) & (e), our dataset provides diversified outdoor scenes and event categories. In addition, our stationary datasets include image-to-ground homographies for researchers interested in exploiting geometry.

Figure 4. An example person: 140, and 50, 20, and 10 pixels tall

Downsampled Versions. In the stationary dataset, we include downsampled versions of dataset obtained by downsampling the original HD videos to lower framerates and pixel resolution. For CVER, it is expected that different approaches will demonstrate varying performance based on the characteristics of videos. For example, STIP-based approaches may be more accurate in high-resolution imageries while approaches based on histogram of gradients (HOG) [3] may be superior when video framerates and pixel resolutions are low. This is a relatively unexplored area

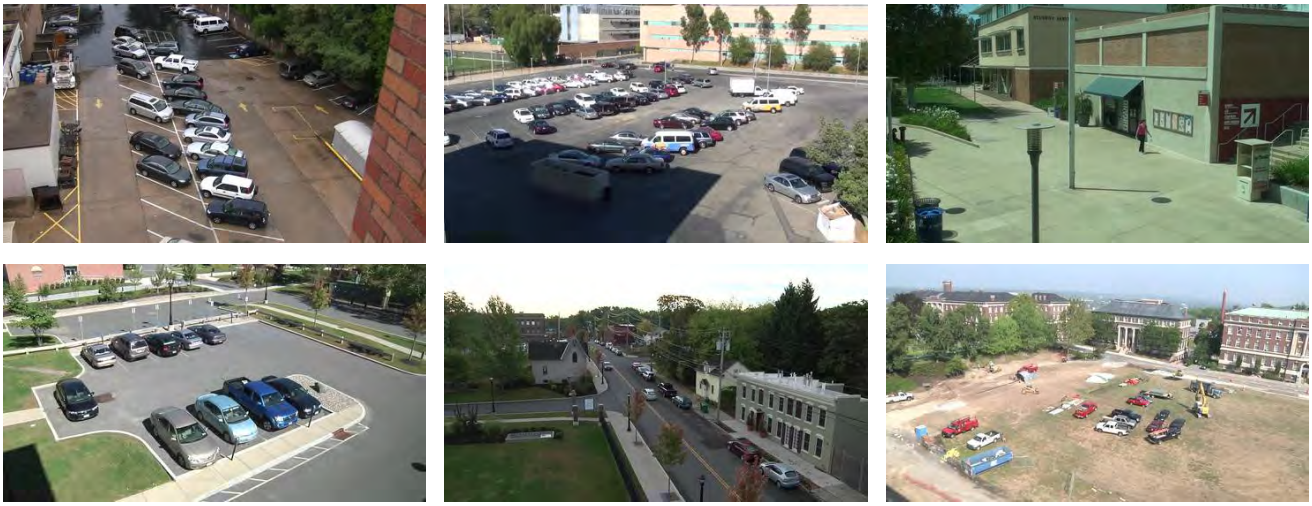

Figure 2. Six example scenes in VIRAT Video Dataset (total 16 scenes)

(a) KTH (b) Weizmann (c) HOHA (d) CAVIAR (e) TRECVID

Figure 3. Example images from existing video datasets: (a) KTH, (b) Weizmann, (c) HOHA, (d) CAVIAR, and (e) TRECVID

and it is important to understand how existing approaches will behave differently based on video characteristics. Accordingly, we provide several different versions of datasets downsampled both spatially and temporally. For temporal downsampling, we provide datasets sampled at three different framerates of 10, 5, 2 Hz. In fact, large number of existing surveillance cameras operate at 2 Hz or lower, and studies on downsampled data with lower framerates will provide important insights into this largely open area for research on surveillance. Spatial down-sampling needs more attention because vanilla spatial downsampling by fixed ratios will not result in the type of data that will be most useful. In fact, datasets with movers exhibiting similar amount of pixel appearance information, i.e., similar heights, will be more useful for performance evaluation. Accordingly, we measure the average pixel heights of people in each scene, and created downsampled versions in such a way that average downsampled people are at 3 consistent pixel heights: 50, 20, 10 pixels. A set of downsampled examples of a person in the dataset is shown in Fig. 4. After both temporal (3 cases) and spatial down-sampling (3 cases), 9 different downsampled versions of videos are additionally provided.

2.2. Second Part: VIRAT Aerial Video Dataset

The second part of our benchmark dataset includes aerial video datasets. Again, the goal of this aerial dataset is to provide a large number of instances of each visual event, collected from aerial vehicles. The CVER problem is more challenging for aerial videos due to the auxiliary variables such as changing viewpoints, illumination, and visibility. Due to privacy and cost constraints, the events in this dataset are acted by hired human actors and vehicles at a designated site. The overall scene of the site and example imageries across the collection site are shown in Fig. 5 (a). It can be seen that the site includes various types of facilities such as buildings and parking lots where people and vehicles are acting events of interest. In addition, examples of evolving image snapshots from a video sequence are shown in Fig. 5 (b) where imageries are roughly 10 seconds apart in time. Defining characteristics of aerial videos such as changing viewpoints and scales can be observed, which presents crucial challenges such as stabilization of videos and dealing with any imperfect pixel-to-pixel alignment during CVER.

The resolution of aerial videos are at 640x480 with 30Hz framerate and the camera is on a gimbal on a manned aircraft where the typical pixel height of people in collections are about 20 pixels tall. From a total of 25 hours of original videos recorded at this site, a subset of dataset which exhibit relatively smooth camera motion and good weather conditions (no severe cloud) were manually selected and included in this dataset with the total amount of 4 hours of videos. Downsampled versions were not considered because cap-

Overall site scene | Building facilities | Vehicle lots | Storage facilities

(a) Aerial dataset contains diverse scenes with varying scales and viewpoints.

time ——————————————————————————————————————>

(b) Sample image shots containing person/vehicle/facility events with viewpoints changing over time due to flight motion.

Figure 5. Aerial dataset: (a) Diversity of scenes (b) Sample images containing person activities over time (avg. 10 secs apart).

tured moving objects are already at fairly low resolution.

2.3. Annotations

Annotating a large video dataset presents a challenge on its own. Two major trade-off factors are: quality and cost. In particular, the annotation of the dataset includes two different types of ground-truths: tracks consisting of bounding boxes for moving objects and localized spatio-temporal events. In our work, a two-step procedural approach brings us a plausible solution: (1) creating tracks of bounding boxes for moving objects by Internet Mechanical Turks (MTs), and (2) event annotation by associating related objects with identified time intervals by experts.

2.3.1 Ground-truth for moving multi-object tracks

Moving objects in this dataset are marked and tracked by bounding boxes. Only the visible part of moving objects are labeled, and they are not extrapolated beyond occlusion by guessing. For example, if upper body of a person is the only visible part, then, only the upper body is labeled as 'person'. This consideration is important because it allows us to measure the performance of multiple moving object trackers more accurately. Bounding boxes around the objects are targeted to be 'whole' and 'tight'. By 'tight', we mean that bounding boxes should be as tight as possible and should not extrapolate beyond the objects being labeled. On the other hand, 'whole' means that all related parts are captured in the bounding boxes. For example, all the visible limbs of people should be in the bounding box, not just the person's torso. These two characteristics are enforced in our dataset to ensure that high-quality event examples with minimal irrelevant pixels are provided to learning modules. Occasionally, other important static parts of scenes such as entrances of buildings, and other related objects such as a bag being carried are labeled when possible.

The annotation of bounding boxes were mostly conducted by MTs, following the insights reported recently in VATIC [18]. More specifically, videos were broken up into segments of ten seconds each and placed on MT. We instructed workers to annotate periodically and used automatic interpolation to recover the annotations in-between key frames. In order to guarantee plausible quality, workers were required to pass an automated test designed to evaluate their annotation ability where only the high quality workers were allowed to continue. During the annotation sessions, we occasionally sampled annotations from each worker to detect whether a worker's quality had decreased since their initial evaluation. This approach eliminated most systematic low-quality jobs. Results were then vetted and edited manually as needed, in a minimalistic fashion. An example standard for ground-truth tracking annotation is shown in Fig. 6 where individual objects, both person and vehicle are tracked over time maintaining identities.

2.3.2 Visual event annotations

Once bounding boxes are labeled, spatio-temporal events are annotated by experts. One of the fundamental issues in event labeling is ambiguity. To avoid ambiguity in event labeling tasks, we define events very explicitly and enforce that events are always annotated based on visual pixel information and not on guesses. For example, even though a person goes behind a car and disappear, which is a strong indicator for an event `person_getting_into_a_car`, it will not be annotated so without an apparent pixel evidence because it can be an instance of `person_crawling_under_car` as well. In

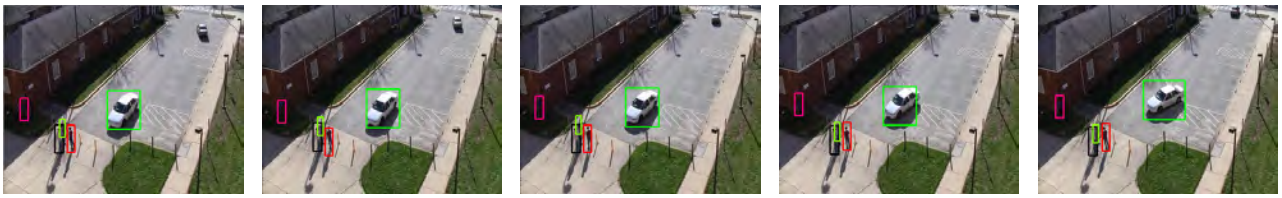

Figure 6. Example ground-truth tracking annotations on stationary dataset. Individual objects, either person or vehicle, are tracked over time at every fixed frame intervals (20 frames in this work), by human annotators who are mostly MTs. For unannotated frames, annotations are automatically generated by interpolation.

(a) (b) (c)

Figure 7. An example snapshot of event annotation system: Events are annotated from individual objects by identifying time interval of every event and all associated objects. (a) a red event box shows an instance of person (bright green) getting out of vehicle (blue). (b-c) additional example annotations with labels on static scene elements such as parking lots and sidewalk.

addition, we define the start and end moments of events very specifically. As an example, the event of `carrying` is defined in detail as follows:

- **Description**: A person carrying an object. The object may be carried in either hand, with both hands, or on one's back.
- **Start**: The event begins when the person who will carry the object, makes contact with the object. If someone is carrying an object that is initially occluded, the event begins when the object is visible.
- **End**: The event ends when the person is no longer supporting the object against gravity, and contact with the object is broken. In the event of an occlusion, it ends when the loss of contact is visible.

For event annotation, we have extended the approaches studied in Video LabelMe [21]. Example snapshots of annotation system is shown in Fig. 7 where events are annotated from individual objects by identifying time interval of every event and associating related objects. For example, in Fig. 7(a), it can be seen that a red event box shows an instance of person (bright green) getting out of vehicle (blue). List of annotated objects and events in videos are available as part of annotation system. The additional example annotations in Fig. 7(b-c) show the quality and density of the annotation efforts where stationary scene elements such as sidewalk and parking lots are annotated as well. It can be observed that our event annotations encode logical interaction information between objects and/or scene, which makes it suitable for the evaluation of algorithms involving probabilistic logic approaches such as [9].

3. Evaluation Approaches

Our new dataset can be used for evaluation of diverse computer vision algorithms. Examples include CVER, multi-object tracking, and functional scene recognition, among others. The available ground-truth annotations provide crucial grounds to conduct large-scale evaluations and obtain measures closer to real-world performance. In this section, we describe potential evaluation areas and modes for diverse computer vision problems, potential metrics, our considerations to support objective evaluations such as data splitting, along with sample experimental results.

3.1. Continuous Visual Event Recognition (CVER)

CVER involves localizing events of interest both spatially and temporally. In the following sections, we propose a useful definition of metrics and two different types of primary evaluation modes along with sample results.

3.1.1 Metrics

It is important to devise well-defined metrics for the evaluation of CVER algorithms[3]. We propose to use a standard definition for 'hits' where a recognition result is a 'hit' if both the ratio of spatial and temporal intersection divided by both ground truth and recognition are above designated threshold. However, we propose that 'hits' are counted from the ground truth point of view - every instance of ground truth example only contributes towards one correct detection even though there are many reasonable overlapping recognition results for it, but, avoiding penalizing multiple positive overlapping detections towards false alarm. This is to avoid the cases where algorithms which produce many similar overlapping recognition results for high-likelihood regions do not get advantage by such operations. Then, we

[3]Exact definitions of diverse metrics and scoring software tools are provided from the data webpage at: www.viratdata.org.

Event Categories	1	2	3	4	5	6
# Ground Truth	11	16	18	19	61	63
# Hit Upper-bound	6	8	8	9	18	14

Table 2. Statistics of the dataset subset used for evaluations.Total number of ground truth examples across three scenes and the upper-bounds for hits after tracking are shown.

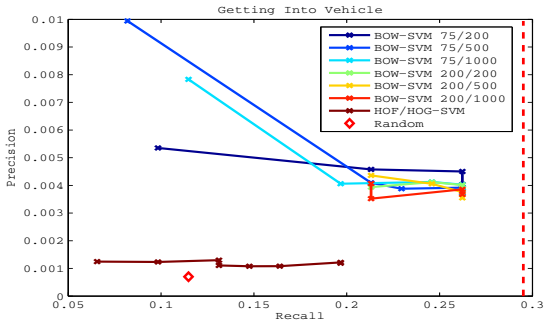

Figure 8. PR curves by 7 evaluations sets for the event 'getting_into_vehicle'. The red diamond at the bottom middle corresponds to random performance. The red vertical dotted line at the right represents the best potential recall after tracking.

propose to use standard metrics such as precision, recall, and false alarm rates to characterize CVER performance.

3.1.2 Independent Recognition Mode

The first mode for CVER is independent recognition - we mean that all the ground truth examples used to learn event models and all the test examples are considered to be independent from each other. More specifically, any prior probability distributions or patterns such as subject appearance or location prior regarding specific subjects and scenes are not to be learned and used during testing.

To support independent recognition mode carefully, it is critical to have different scenes for training and testing datasets. Frequently-used leave-one-out approach, e.g., [12], is challenging due to the large dataset size, but, also is not desirable for independent recognition mode because scene-based correlation across examples can lead to optimistic results. Accordingly, the video dataset is divided into multiple sets based on scenes and support users to easily obtain reliable independent recognition results via N-fold cross validation. Furthermore, a few scenes have been completely sequestered, included in neither training nor testing datasets. The purpose is to provide a completely overfitting-free environment to assess the performance of algorithms. Official venue or process will be provided to aid official evaluation on the sequestered data.

Experimental Results To understand the performance of existing approaches on this dataset, a preliminary evaluation was conducted on a subset of the ground video dataset[4], with focus on 6 event categories: loading(1), unloading(2), opening_trunk(3), closing_trunk(4), getting_into_vehicle(5), and getting_out_of_vehicle(6). This subset contained examples from three different scenes with approximately equal number of examples, and the total video duration is approximately three hours. The total number of ground truth examples per category are shown in Table 2. We used two learning-based approaches similar to [10, 2]. The training and testing were conducted via leave-one-scene-out evaluation, and results were averaged afterwards. For the first approach, separate one-vs-all event detectors were trained for every six event types, and used individually to detect events, making it possible that a single example is labeled with multiple event types. For the second approach, N-way classification was used. For training, available ground truth spatio-temporal volumes were used. For testing, spatio-temporal volumes were obtained by a frame-differencing multi-object tracker with automatic track initialization. In particular, tracking was used to reduce the huge video volume. Once tracks are obtained, tracks were divided into detection units of 3~4 second segments (with 2 seconds overlap) each, which resulted in the total of more than 20K detection units. Note that some of the ground truth examples exhibited durations longer than detection units, which we will miss in most cases. Each detection had confidence score generated by event detectors where a 'hit' is defined between a detection and a ground truth if both the spatial and temporal overlap are over 20%. Note that the imperfect tracking results will limit the final performance. While many tracks are correctly formed to capture spatio-temporal extent of events correctly, there are also many tracks formed on shadows or partial spatio-temporal volumes of events, e.g., only on one arm of a person. Accordingly, we measured the quality of tracking results and computed the best possible number of hits, which are shown both in Table 2 and Fig. 8.

In detail, our first approach (BOW-SVM) is similar to [10] where we use the spatio-temporal interest point detector and descriptor analogous to [5]. Six different sets of evaluations were conducted with different numbers of interest points per detection unit and codebooks with different sizes. For every ground truth spatio-temporal volume, we used adaptive thresholds to extract two different number of interest points, which are 75 and 200. Then, the collected vectors are clustered by K-means to form codebooks after PCA-based dimension reduction where we obtained three codebooks with different sizes (200/500/1000). Finally, a bag of words (BOW) representation is computed for every detection unit, which is classified by SVM with histogram intersection kernel. We also tried a nearest-neighbor approach, but, SVM results were superior. Our second method (HOG/HOF-SVM) uses an approach similar to [2] which uses both HOG and HOF [4] descriptors from spatio-temporal volumes. The final activity descriptor is formed by

[4]Available as Release 1.0 training dataset at the dataset webpage.

concatenating time series of dimension-reduced HOG and HOF features, which are classified by a linear SVM.

We obtained the best results for the event category 'getting_into_vehicle', shown in Fig. 8 along with the PR performance by a notional random classifier and best possible recall after tracking. For the other five categories, results were less accurate and omitted for brevity. It can be observed that tracking imposes fairly low upperbound. Accordingly, we will explore searching entire video volume and reducing false alarms in the future evaluation efforts[5].

3.1.3 Learning-Enabled Recognition Mode

In addition to the conventional independent recognition mode, another mode for CVER is learning-enabled recognition. By learning-enabled recognition, we mean that the learning of important patterns regarding current scene is allowed for future recognition tasks. For example, event or trajectory priors are allowed to be learned from scenes to aid future recognition. To support this type of research goals, we provide another data splitting where ground truth examples in every scene are divided into multiple sets per scene and event type, in time-linear way. For every testing set for a particular scene, examples from other subsets from the same scene and all the rest of out-of-scene examples can be used for training. In particular, training examples from the same scene can be used towards learning scene-dependent priors. Again, to support object evaluation of learning-enabled CVER, certain portions of collected data are sequestered from all available scenes, as a way to provide overfitting-free evaluation setting.

Functional Scene Recognition Potentially related research area is the problem of scene understanding based on activities. Recent work [17, 9] showed that functionalities of different parts of scenes can be automatically reasoned based on the observed activities. We expect that such contextual models can benefit CVER performance.

4. Conclusion

We have presented the VIRAT dataset, a large-scale real-world surveillance video dataset containing diverse examples of multiple types of complex visual events. We believe that this dataset will stimulate diverse aspects of computer vision research in the years ahead.

Acknowledgments

Thanks to Kamil Knuk, Stefano Soatto, Arslan Basharat, and many others for help on this work. This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-08-C-0135. The views expressed are those of the authors and do not reflect the position of the the U.S. Government.

[5]Results will be collected and updated on the dataset webpage.

References

[1] *UCF Aerial Video Dataset. http://server.cs.ucf.edu/ vision/aerial/index.html.*

[2] *C.-C. Chen and J. K. Aggarwal. Recognizing human action from a far field of view. In* IEEE Workshop on Motion and Video Computing (WMVC)*, Utah, USA, 2009.*

[3] *N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In* CVPR*, 2005.*

[4] *N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In* ECCV*, 2006.*

[5] *P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In* VS-PETS*, 2005.*

[6] *M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results.*

[7] *R. B. Fisher. The PETS04 Surveillance Ground-Truth Data Sets. 2004.*

[8] *L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes.* PAMI*, 29(12):2247–2253, Dec 2007.*

[9] *A. Kembhavi, T. Yeh, and L. S. Davis. Why Did the Person Cross the Road (There)? Scene Understanding Using Probabilistic Logic Models and Common Sense Reasoning. In* ECCV*, 2010.*

[10] *I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In* CVPR*, 2008.*

[11] *I. Laptev and P. Perez. Retrieving actions in movies. In* ICCV*, 2007.*

[12] *J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos ïn the wild: In* CVPR*, 2009.*

[13] *M.S.Ryoo, C.-C. Chen, J. Aggarwal, and A. K. Roy-Chowdhury. An Overview of Contest on Semantic Description of Human Activities. 2010.*

[14] *M. Rodriguez, J. Ahmed, and M. Shah. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In* ICPR*, 2004.*

[15] *C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In* ICPR*, 2004.*

[16] *A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In* MIR*, 2006.*

[17] *M. Turek, A. Hoogs, and R. Collins. Unsupervised Learning of Functional Categoreis in Video Scenes. In* ECCV*, 2010.*

[18] *C. Vondrick, D. Ramanan, and D. Patterson. Efficiently Scaling Up Video Annotation with Crowdsourced Marketplaces. In* ECCV*, 2010.*

[19] *D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes.* CVIU*, 104(2):249–257, 2006.*

[20] *Y.Ke, R. Sukthankar, and M. Hebert. Volumetric Features for Video Event Detection.* IJCV*, 88(1), 2010.*

[21] *J. Yuen, B. C. Russell, C. Liu, and A. Torralba. LabelMe video: Building a Video Database with Human Annotations. In* ICCV*, 2009.*