An End-to-End System for Content-Based Video Retrieval using Behavior, Actions, and Appearance with Interactive Query Refinement

A. Hoogs[1], A. G. A. Perera[2], R. Collins[1], A. Basharat[1], K. Fieldhouse[1], C. Atkins[1], L. Sherrill[1], B. Boeckel[1], R. Blue[1], M. Woehlke[1], C. Greco[2], Z. Sun[1], E. Swears[1], N. Cuntoor[2], J. Luck[3], B. Drew[4], D. Hanson[3], D. Rowley[3], J. Kopaz[3],T. Rude[3], D. Keefe[3], A. Srivastava[5], S. Khanwalkar[5], A. Kumar[5], C. C. Chen[6], J. K. Aggarwal[6], L. Davis[7], Y. Yacoob[7],A. Jain[8], D. Liu[9], S.-F. Chang[9], B. Song[10], A. Roy-Chowdhury[10], K. Sullivan[11], J. Tešić[11], S. Chandrasekaran[11], B. S. Manjunath[11], X. Wang[12], Q. Ji[12], K. Reddy[13], J. Liu[13], M. Shah[13], K. Chang[15], T. Chen[14], M. Desai[16]

Kitware[1], work performed at Kitware[2], Lockheed Martin[3], work performed at Lockheed Martin[4], Raytheon BBN Technologies[5]
U. Texas[6], U. Maryland[7], work performed at U. Maryland[8], Columbia University[9], U. California at Riverside[10], Mayachitra[11]
Rensselaer Polytechnic Institute[12], U. Central Florida[13], Cornell University[14],work performed at Cornell[15],work performed at DARPA[16]
Corresponding author: roddy.collins@kitware.com; full author contact information at [1]

Abstract

We describe a system for content-based retrieval from large surveillance video archives, using behavior, action and appearance of objects. Objects are detected, tracked, and classified into broad categories. Their behavior and appearance are characterized by action detectors and descriptors, which are indexed in an archive. Queries can be posed as video exemplars, and the results can be refined through relevance feedback. The contributions of our system include the fusion of behavior and action detectors with appearance for matching; the improvement of query results through interactive query refinement (IQR), which learns a discriminative classifier online based on user feedback; and reasonable performance on low resolution, poor quality video. The system operates on video from ground cameras and aerial platforms, both RGB and IR. Performance is evaluated on publicly-available surveillance datasets, showing that subtle actions can be detected under difficult conditions, with reasonable improvement from IQR.

1. Introduction

The increasing volume of surveillance video collected by ground cameras and aerial platforms vastly exceeds the processing capacity of human analysts. Research into automated exploitation of such data has mostly focused on object detection and tracking, person re-identification, image retrieval and matching, anomaly detection and face recognition. Far-field surveillance, where image resolution is a primary challenge, receives less attention, but typically accounts for most of the activities and objects of interest in aerial surveillance and ground camera scenes with significant depth of field.

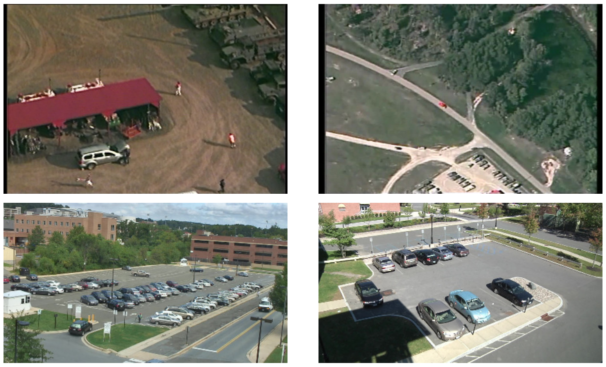


Figure 1. Sample frames from the VIRAT Video Dataset. Top row: aerial video; bottom: ground cameras.

To address this challenge, as part of the DARPA VIRAT program[6], we have developed an end-to-end video analytics system that ingests video from a stationary or moving platform; stabilizes and enhances the video; detects, tracks, and classifies all movers; characterizes their motion and behaviors globally and locally with action-independent descriptors, which are indexed into a compact archive; enables user querying of the archive via image/video exemplars or pre-defined action types; and enables interactive query refinement to improve results and learn new event or action types. A GUI enables analysis and result browsing.

Much of the extensive research in content-based retrieval for general images and video [11] and CBR systems (e.g. en.wikipedia.org/wiki/List_of_CBIR_engines) does not translate effectively to surveillance video, whose

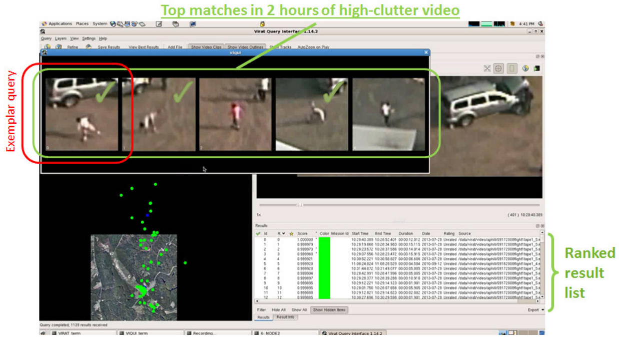


Figure 2. A query video of a "cartwheel" returns two other instances in the top four results (excluding query video), from a two hour archive containing more than 1600 computed tracks. Query results are displayed on a map and a ranked list (lower left, right).

visual content has lower diversity due to continuous scene coverage at lower resolutions [16]. Compared to recent commercial and prototype video analytics systems [8], our approach offers significant contributions: **Activity-based retrieval**, enabling queries based on not just appearance but also kinematics, actions, and events; **relevance feedback** via Interactive Query Refinement (IQR), dynamically improving the query model's precision via user feedback; and performance under **far-field and poor video quality** conditions, including far-field resolution where persons are as small as 10 pixels in height.

Our algorithms are designed to handle poor image quality conditions such as low contrast, high levels of sensor noise, compression artifacts, and graphics burned into the video frames (issues particularly common in aerial video.) In ground-camera video, operating at lower resolutions significantly increases the effective field-of-view. The system processes both RGB and infra-red video; queries in one modality may yield results in either, see [1] §5.

Figure 1 shows samples of full-motion video (FMV) from the VIRAT Video Dataset [16]. Consider the frame in the upper left; at this resolution adults are about 30 pixels high, and the imagery is blurred from sensor motion. Multiple events are occurring simultaneously, as is typical in busy scenes. With our system, an analyst may specify an interval of an object track showing a particular action, and find instances of the same action across the video archive, without any prior training or knowledge of the action type.

In the top right of Figure 2, the user selected the person performing a cartwheel as the query exemplar. The archive contains two hours of aerial video with over 1600 tracks, mostly on moving people. The results adjacent to the query show that two other instances of gymnastics were found in the top four results, out of a total of less than 10 similar actions in the archive, performed by three individuals. For common actions such as "person running", "vehicle stopping", etc., detectors trained offline are included, which may be run as queries without video exemplars.

The system has been evaluated on aerial and ground camera video datasets, particularly the VIRAT Video Dataset [16]. In our experiments, we measure baseline retrieval accuracy on challenging actions such as carrying, and improvement through IQR iterations across different descriptor types. We also compare the accuracy of popular action descriptors included in the system such as space-time histogram of gradients [13]. Although accuracy is lower than non-surveillance action recognition datasets such as UCF 50 [17], it is comparable to reported results on the most similar dataset and evaluation, the Surveillance Event Detection task of the TRECVID evaluation [15].

2. Architecture

The system is organized into functional modules (Figure 3, and scales to large data volumes by parallelization using standard IPC mechanisms. A 720x480@30fps video stream can be processed in real-time on a 16-core workstation.

Video enters the stabilization process, which computes frame-to-frame homographies. If available, sensor position metadata is used for geo-location and scene scale estimation. Tracking uses the ingested video, homographies, and metadata to detect moving objects, and to initialize and update tracks. Tracks are then classified as either "Person", "Vehicle", or "Other" (hereafter, PVO) by a classification module. Using tracks and the PVO scores, descriptors are computed and indexed in an archive. Descriptors include characterizations of kinematic (track-level) motion, behavior as articulated motion, events, and object appearance.

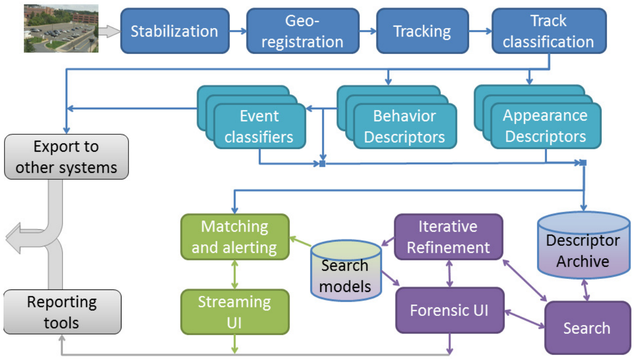


Figure 3. The system architecture.

To search the archive, an analyst (via the GUI) forms a query from either pre-trained action types (e.g. "person running") or using a single video exemplar. If given, the exmplar is processed; the system then searches the archive for similar content. A single exemplar often leads to noisy initial results, but using IQR the user can provide relevance feedback, guiding the system to find more relevant results.

The major system components are described in the following sections.

3. Multi-Object Tracking

Detection and tracking of moving objects reduces the volume of video drastically, segmenting the video into spatio-temporal trajectories used in all subsequent processing. Challenges in low-quality video include: a wide range of scales; very small objects, often < 20 pixels high; EO and IR video; abrupt camera motion with significant zoom changes; on-screen metadata burn-in; missing or unreliable sensor metadata; and corrupt images due to data transmission errors. We address these and related challenges in the proposed system. The techniques in [7] are used to automatically detect and mask pixels with on-screen burn-in. The video is stabilized using frame-to-frame homographies estimated via KLT feature points [26] and RANSAM [12, 10]. When available, sensor metadata is used to estimate the ground-sample-distance (GSD) and image-to-ground transformation to rectify distortions. Empirical evaluation shows that fusing stabilization with metadata significantly reduces geo-localization error when the metadata is inaccurate.

Tracks are initiated on moving object detections from either Gaussian Mixture Models (GMM) [25] or three-frame differencing [29]; typically the latter is used for aerial video due to memory and processing requirements of GMMs. Following [4], multi-frame analysis of motion detections in a stabilized reference plane is used for track generation. Tracks are updated via adaptive integration of appearance models and motion detections; when the motion detections are unreliable or unavailable, the appearance or foreground tracker are used as in [2]. The system detects and tracks person and vehicles simultaneously; a track reconciliation process merges and suppresses duplicates. PVO classification (Section 4) further reduces false alarms.

4. Descriptors

Object appearance and activity are captured by descriptors, the critical part of the proposed system. During development, we investigated a wide variety of descriptor algorithms [18, 13, 30, 5, 3, 14, 19, 27, 31, 22], focusing first on algorithms suitable for low-resolution, low-quality video, and further down-selecting based on descriptor quality, runtime, and software reliability; see [1], §4 for more details.

The system uses two types of descriptors, *classifier* and *raw*. Classifiers are computed by an action or event detector, and encode the probability that a track interval is an instance of the event. Raw descriptors capture low-level information such as gradients or kinematics variance within a spatiotemporal volume on a track, and are typically used for exemplar matching and IQR (Section 6).

Descriptors can be grouped into three broad categories based on their focus of attention: *trajectory*, *articulation*, or *interaction*. The first group contains kinematic features extracted from the stabilized trajectories of tracked objects [31], and classifiers built upon them to detect kinematic events such as vehicle start, stop, turn, u-turn, etc. Articulated descriptors encode part-based motion and shape deformations over time, in order to represent actions such as opening a door, closing trunk, digging, etc. A number of these descriptors rely on extensions [3, 18, 13] of histogram of gradients (HoG) [5] and histogram of optical flow (HoF) features. The *UTECE HOG* descriptor [3] models a series of human poses as a time series of HOG and HoF along with Supervised Principal Component Analysis for action classification. [18] creates a 3D representation of the HoG feature. Other descriptors involve action template matching [19], flow categorization [27], visual bag-of-words, and Partial Least Squares (PLS) [22]. Interaction descriptors represent multiple object events. Dynamic Bayesian Networks similar to [30] are used to model relational activities such as person entering vehicle/facility, while [14] uses temporal logic for modeling human-vehicle interaction with dynamic programming for optimal search.

Appearance descriptors are also used to represent object shape, color and distinctive parts. These are treated similarly to action descriptors for indexing and matching. Stationary scene elements can also be represented and matched, although this capability has not been evaluated.

The PVO (person / vehicle / other) estimates are computed using appearance and behavior models trained offline. Features used include HOG [5], trajectory, object size, and scale priors, all fused via a tree of single-class SVM classifiers, each trained on a single feature type. This approach performs well with low resolution and large variation in the target appearance. PVO is vital in suppressing false alarms and associating the relevant event type.

5. Indexing and Retrieval

In the system, indexing and retrieval speeds searching through large volumes of video data. Tree based structures can be efficient for low-dimensional data, but scale poorly in in terms of storage and access time for high-dimensional feature spaces like those used in visual descriptors. Uniform space partitioning [28], while highly scalable with large dimensions and number of data points, does not capture the high-dimensional space well, resulting in undetermined upper boundary on the selected set candidates. Locality Sensitive Hashing (LSH) [9] assigns similar data (for a given distance metric) to the same bucket while relying on the probabilistic boundaries on approximate search. Our indexing engine is specifically tuned for high-dimensional descriptors, using a data driven indexing approach. Prototypes of data clusters in the high dimensional space become search indices; these are matched at query time to the query vector, and nearest neighbors to the best matching indices are retrieved. These nearest neighbors are then sorted by similarity to the query vector to generate the ranked retrieval list. [1] §6 has more discussion.

6. IQR

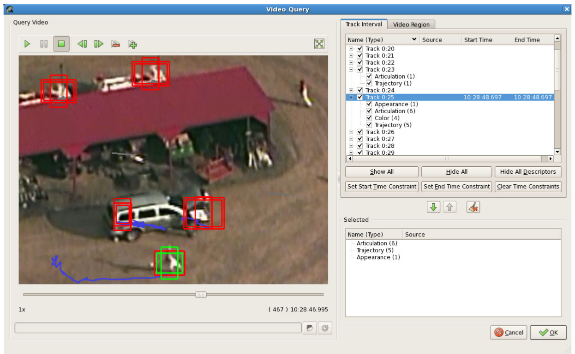
The system provides a "query-by-example" capability via the GUI (see Section 7), allowing the user to indicate what they are looking for without having to learn a complex query specification language or to understand the abstract representations embodied by the system descriptors (Section 4). However, naive use of a video exemplar as the basis for a query is likely to return only a few relevent results among many irrelevant ones. Furthermore, the result set from querying a large archive likely cannot be practically reviewed in its entirety. IQR focuses the search on characteristics of interest that may not have been emphasized in the initial query, re-ranking the results to prioritize relevant results so they may be found more quickly.

Our IQR system uses Relevance Feedback (RF), in which a user attempts to improve the quality of the result set based on both positive feedback, for results that match or nearly match the desired result characteristics, and negative feedback, for results that do not match [21]. RF has been shown to be quite useful in text applications [20], and is also seeing application in content-based information retrieval systems [23, 24].

Starting with the initial result set based on the exemplar, the user selects and provides feedback on a subset, submits it to the system for re-ranking, and then iterates as many times as desired constructing a customized descriptor-based model of the activity of interest. The IQR algorithm typically converges within several iterations, such that further iterations are of minimal value. The actual re-ranking is very fast, typically on the order of a few seconds for 1500 results; the full user-in-the-loop process may require several minutes. Fully developed IQR models may be saved as a new system query, allowing analysts to leverage previous refinement efforts. Saved queries may be copied, branched, and further refined as necessary.

7. GUI and Workflow

The GUI is the interface for executing queries and browsing results. The basic workflow is to *(a)* initiate a query, *(b)* review the results, and then *(c)* optionally refine the query to focus the results for further review, repeating *(b)* and *(c)* as necessary. Query options include simple classifiers (Section 4 selected from a predetermined list, such as "Walking" or "Running", and video exemplars, which the user constructs from descriptors extracted from an exemplar video clip (Figure 4). Exemplars may be any returned result clip, or a novel clip supplied by the user. A typical workflow starts with a classifier query such as "Person Moving", then switches to an exemplar query based on an interesting classifier result. Queries can be constrained both in time and space. The GUI also enables IQR (Section 6) by allowing the user to indicate whether results are relevant or not.



Name (Type)	Source	Start Time	End Time
☑ Track 0:20			
☑ Track 0:21			
☑ Track 0:22			
☑ Track 0:23			
☑ Articulation (1)			
☑ Trajectory (1)			
☑ Track 0:24			
☑ Track 0:25		10:28:48.697	10:28:48.697
☑ Appearance (1)			
☑ Articulation (6)			
☑ Color (4)			
☑ Trajectory (5)			
☑ Track 0:26			
☑ Track 0:27			
☑ Track 0:28			
☑ Track 0:29			

Name (Type)	Source
Articulation (6)	
Trajectory (5)	
Appearance (1)	

Figure 4. Analyst initiates an "exemplar" query by selecting the activity of interest in the video (green box, left); analyst may select all or a subset of the corresponding descriptors to form the query.

8. Experiments

The system was evaluated using the VIRAT Video Dataset [16], containing both aerial and ground camera data (Figure 1). The aerial archive contains roughly two hours of data collected over three days, on which the system computed 1651 tracks. The ground camera archive contains three hours of surveillance video collected at three different sites, and contains 5600 tracks. We present results on event retrieval, combinations of raw descriptors, and IQR trials. An extensive end-user evaluation was also conducted by an independent third party (MITRE Corp.) as part of the DARPA VIRAT program; more details are at [1] §7.

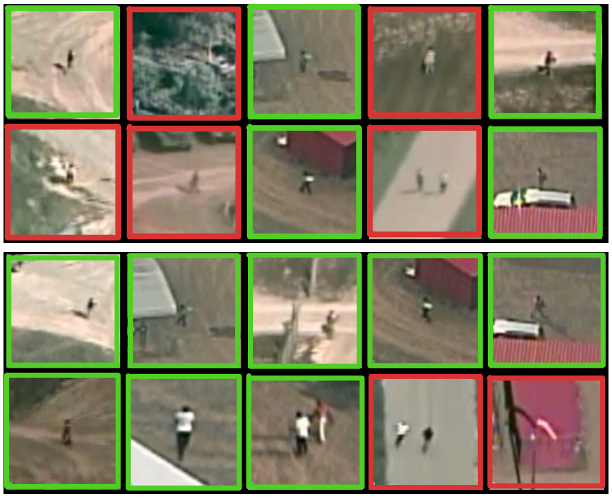


Figure 5. Top ten results from the initial query (top) and after two rounds of IQR (bottom) for the carrying example. Observed matches are in green, misses in red.

The first result is a "Person Carrying" query, based on an exemplar track chosen from the archive. [1] §3 discusses exemplar choice sensitivity. This clip, selected by the user in the GUI is associated with 19 descriptors: 2 appearance, 14 articulation (5 different types), and 3 PVO. Color and trajectory descriptors are not selected, assuming they are less useful for detecting carrying; those remaining form the search query basis. Note that no semantic hint is supplied that we are looking for “carrying”; we are merely looking for clips in the archive whose descriptors match our search query. The first 10 results of this initial query are shown in the top row of Figure 5; first is the query, which matched itself; 4 of the remaining 10 are hits. The aerial dataset ground truth has 1505 events; 119 (7.9%) were labelled “Person Carrying”. The ROC and P/R curves labelled *initial* in Figure 6 quantify this initial retrieval performance, which is encouraging given the challenging problem of finding such events data at this low resolution. [1] §1 includes videos.

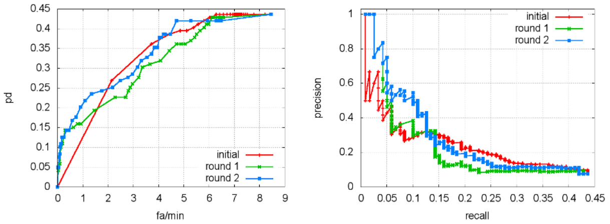


Figure 6. ROC and P/R curve for the carrying example; IQR has doubled the precision for the first set of matches (left side of curves.) See also [1] §2.

Next, the user performs IQR on these initial results, giving feedback for the top 20 returns as to whether or not the clip seemed to match “person carrying”. The system also nominated seven clips whose rank ranged from 19 to 765 as “feedback requests”. The system incorporates the feedback into the model and re-ranks the results. After two rounds of this process, 8 of the top 10 are hits; IQR has doubled the top-10 precision of the initial query as shown Figure 6. This emphasizes that IQR favors improving response for higher-ranked results, discussed further in [1] §2.

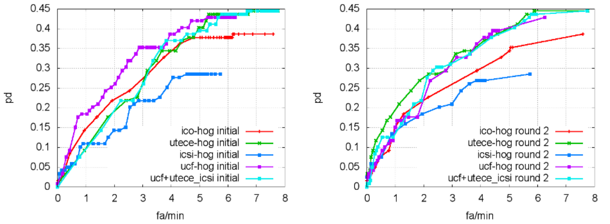


Figure 7. ROC curves for the initial query (left) and after two rounds of IQR (right), for individual and combinations of articulated descriptors.

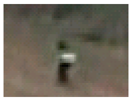
name	backpack	shovel	vehicle
sample frame			
dataset	groundcam	aerial	aerial
est. prior	6% (1.2/20)	7%(1.4/20)	1% (0.2/20)
initial	3/20	3/20	3/20
feedback	+5, -20	+4, -4	+2, -25
round 1	9/20	7/20	5/20
feedback	+5, -15	+2, -0	+2, -14
round 2	17/20	7/20	6/20
gain vs first	5.6x	2.3x	2.0x
gain vs prior	**14x**	**5x**	**30x**

Table 1. Ad hoc IQR trials: observed results and feedback schedules. Initial query precision is typically at least doubled by IQR.

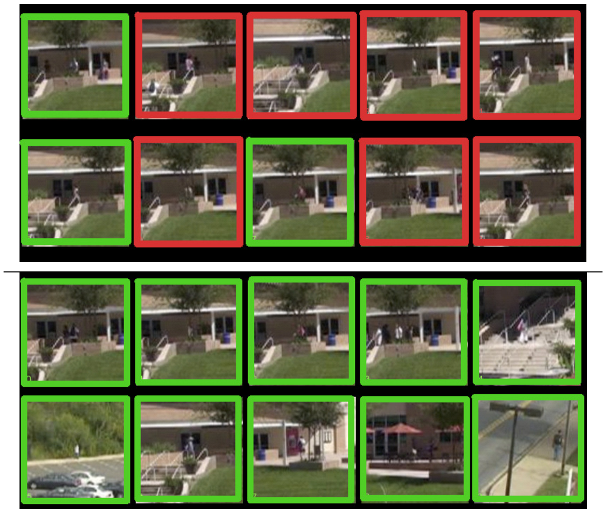


Figure 8. Initial (top) and final (bottom) top 10 results for the IQR “backpack” experiment. Initial results had 3 hits (in green, including query) in the top 20, or 1.8x random; final results included 17 hits in the top 20, or 14x random. Negative feedback is in red.

We experimented with a variety of descriptor combinations; the contribution of individual descriptor types can be analyzed by re-running the experiment with specialized sets of descriptors. Figure 7 shows the ROC curves from replacing the full suite of 14 articulation descriptors with (in experimental order) 2 instances of the UCF_BoW descriptor, 3 of UTECE HOG, 5 of icosahedron HOG, 3 of ICSI HOG, and finally, a suite made up of UTECE, UCF, and ICSI descriptors. The figure shows that descriptors vary both in absolute performance and in how they respond to IQR.

We now show results for three *ad hoc* queries without *a priori* ground truth, as might dynamically arise during analysis session. The first (“Wearing a backpack”) was computed against the ground camera archive; the other two (“Carrying a shovel”, “Interacting with a vehicle.”) from the aerial data archive. Using an IQR protocol similar to that for the carrying example above, two rounds of feedback were given after the initial query; results are summarized in Table 1. The precision of the initial query is typically 2x better than chance, it is further doubled by IQR. The estimated priors were computed from the top 100 “person moving” results and projected to top 20 for consistency. IQR results are shown in Figure 8 (backpack query, appearance and articulation); results for the shovel and vehicle experiments are available in [1] §8. For the vehicle query, Table 1 shows that the initial query performed 15x better than random, and large amounts of feedback (4 positive, 39 negative) allowed two rounds of IQR to again double the initial precision.

9. Conclusion

Our state-of-the-art system for surveillance video analytics is based upon the fusion of behavior and action descriptors with appearance. Our focus is on low-resolution, low-quality video conditions where typical methods for tracking and action recognition can fail. We have developed methods that account for these conditions, and can operate on video from an airborne moving sensor as well as ground cameras. Performance is demonstrated on the VIRAT Video Dataset, showing improvement from IQR.

Acknowledgement

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-08-C-0135. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

[1] AVSS 2015 supplemental material. http://www.kitware.com/virat_avss_2015.

[2] A. Basharat, A. Hoogs, et al. Multi-target tracking in video with adaptive integration of appearance and motion models. In *AIPR*. IEEE, 2014.

[3] 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, 2010.

[4] N. Cuntoor, A. Basharat, A. Perera, and A. Hoogs. Track initialization in low frame rate and low resolution videos. In *ICPR*. IEEE, 2010.

[5] N. Dalal. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[6] DARPA. BAA 08-20 Video and Image Retrieval and Analysis Tool (VIRAT). FedBizOpps.gov, Mar. 2008.

[7] M. Dawkins, A., A. Perera, and A. Hoogs. Real-time heads-up display detection in video. In *AVSS*. IEEE, 2014.

[8] J.-Y. Dufour. *Intelligent Video Surveillance Systems*. John Wiley and Sons, 2012.

[9] A. Gionis, P. Indyk, and R. Motwani. Similarity Search in High Dimensions via Hashing. In *Proc. 25th International Conf. on Very Large Data Bases*, 1999.

[10] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. 2000.

[11] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank. A survey on visual content-based video indexing and retrieval. *IEEE Trans. on Systems, Man, and Cybernetics*, Nov 2011.

[12] R. Iser, D. Kubus, and F. M. Wahl. An efficient parallel approach to random sample matching (pRANSAM). In *Proc IEEE Intl Conf on Robotics and Automation*, 2009.

[13] A. Klser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[14] J. T. Lee, C.-C. Chen, and J. K. Aggarwal. Recognizing human-vehicle interactions from aerial video without training. In *IEEE CVPR Workshops*, 2011.

[15] M. Michel, J. Fiscus, and P. Over. TRECVID 2011 Video Surveillance Event Detection Task. In *NIST TRECVID Workshop*, 2011.

[16] S. Oh et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, 2011.

[17] K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5), 2013.

[18] K. K. Reddy, N. Cuntoor, A. Perera, and A. Hoogs. Human action recognition in large-scale datasets using histogram of spatiotemporal gradients. In *AVSS*. IEEE, 2012.

[19] M. Rodriguez, J. Ahmed, and M. Shah. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[20] Y. Rui, T. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 8(5):644 –655, sep 1998.

[21] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41:288–297, 1990.

[22] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis. Human detection using partial least squares analysis. In *International Conference on Computer Vision*, 2009.

[23] D. M. Squire, W. Mller, H. Mller, and J. Raki. Content-based query of image databases, inspirations from text retrieval: Inverted files, frequency-based weights and relevance feedback. In *PATTERN RECOGNITION LETTERS*, pages 143–149, 1999.

[24] D. M. Squire and T. Pun. A comparison of human and machine assessments of image similarity for the organization of image databases. In *10th Scandinavian Conf. on Image Analysis*, 1997.

[25] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*. IEEE, 1999.

[26] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report, CMU, 1991.

[27] S. Wu, B. Moore, and M. Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *CVPR*, 2010.

[28] G. Wyvill, C. McPheeters, and B. Wyvill. Data structure for soft objects. *Visual Computer*, 2, August 1986.

[29] J. Xiao, H. Cheng, H. Sawhney, and F. Han. Vehicle detection and tracking in wide field-of-view aerial video. In *CVPR*. IEEE, 2010.

[30] Z. Zeng and Q. Ji. Knowledge based activity recognition with dynamic bayesian network. In *ECCV 2010*.

[31] Y. Zhu, N. Nayak, and A. Roy-Chowdhury. Context-aware modeling and recognition of activities in video. In *CVPR*, 2013.